REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE.
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE RECHERCHE SIENTIFIQUE.

UNIVERSITE MOULOUD MAMMERI, TIZI-OUZOU.
FACULTE : DES SCIENCES
DEPARTEMENT : MATHEMATIQUES

MEMOIRE DE MASTER
en
Mathématiques appliquées
Option: Processus aléatoires et statistique de la décision

Thème

Sur les méthodes variationnelles Bayésiennes et applications

Présenté par

GRAICHE Kahina

Devant le jury d'examen composé de :

M^r BOUDIBA Mohand Arezki	Maître de conférence A	U.M.M.T.O	Président
M^r FELLAG Hocine	Professeur	U.M.M.T.O	Rapporteur
M^{elle} ATIL Lynda	Maître de conférence B	U.M.M.T.O	Examinatrice

Soutenu le? /? / 2013

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude et ma sincére reconnaissance ***Monsieur FELLAG Hocine*** *qui m'a encadré pendant cette thèse. Je le remercie pour sa grande disponibilité, sa patience, ses précieux conseils et son optimisme contagieux. Je le remercie aussi pour les sujets passionnants vers lesquels il m'a dirigé et toute l'aide qu'il m'a fourni et grâce à laquelle j'ai pu mener à terme ce travail. Je lui en suis infiniment reconnaissante.*

j'adresse mes remerciements les plus respectueux à M^r ***HAMADOUCHE Djamel*** *pour l'honneur qu'il me fait en présidant le jury de ce mémoire.*

Merci également aux membres du jury qui ont accépté d'examiner ce travail: M^r ***BOUDIBA Mohand Arezki*** *et* M^{elle} ***ATIL Lynda***

Enfin, je ne trouve pas de mots pour exprimer ma gratitude et ma reconnaissance à mes parents, mes frères et mes soeurs. Merci de m'avoir soutenue durant ces années, d'avoir été là spécialement pendant les moments difficiles pour m'aider a réaliser ce rêve. Merci aussi à ma grand mère, mes deux oncles et leurs familles.

Vient le tour de mes amies qui m'ont accompagnées durant ce parcours.

Mes derniers mots vont à ma belle famille . Je voudrais remercier du fond du coeur mon fiancé ***Malik*** *pour sa présence à mes côtés, son soutien infini et sa positivité au quotidien.*

Je remercie tous ceux qui ont contribué de près ou de loin à la réalisation de ce modeste mémoire.

Table des matières

Introduction Générale

Le mot statistique désigne à la fois un ensemble de données d'observation et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation.
Les méthodes statistiques sont aujourd'hui utilisées dans presque tous les secteurs de l'activité humaine et font partie des connaissances de base de l'ingénieur, du gestionnaire, de l'économiste, du chercheur, etc. Parmi les nombreuses applications, on peut citer dans l'industrie : la fiabilité de matériel, le contrôle de qualité, la prévision ; et dans l'économie et les sciences humaines : les modèles économétriques, les sondages, les enquêtes d'opinion, etc.
La statistique Bayesienne est une théorie complémentaire à la statistique dite classique en ce sens que chacune d'elles propose vis-à-vis d'un même problème une approche est une résolution complètement différentes.
Les méthodes Bayésiennes sont un ensemble de techniques statistiques utilisées pour modéliser des problèmes, extraire de l'information de données brutes et prendre des décisions de façon cohérente et rationnelle. Son cadre d'application est général, mais ses avantages sont déterminants lorsque l'information disponible est incertaine ou incomplète. Bien que les premiers travaux d'inspiration Bayésienne datent du $XVII^{ème}$ siècle, cette méthode connaît un regain de popularité depuis quelques décennies. Ce renouveau est sensible dans des domaines très variés, en partie grâce à la disponibilité de calculateurs puissants, mais aussi à cause d'une évolution de la pensée statistique et des problèmes abordés.
Les méthodes variationnelles Bayésiennes réussissent à résoudre de divers problèmes où les autres méthodes ne résistent pas forcément, pour cela les méthodes variationnelles sont devenues un outil incontournable dans l'analyse bayésienne ainsi que dans l'analyse de données expérimentales et la présentation de résultats scientifiques.

Ce mémoire contient trois chapitres, afin d'aider le lecteur intéressé à avoir quelques idées sur la démarche Bayésienne.

Dans le premier chapitre, Nous présenterons superficiellement les notions et les outils sur lesquels se fonde une analyse Bayésienne, et dont nous aurons besoin pour établir les

prochains chapitres de ce mémoire. Dans un premier temps et après avoir donné le théorème de Bayse , nous allons parler de la loi a posteriori, sur laquelle l'approche Bayésienne se fond, et bien sur la loi a priori qui est le moteur de cette approche et au même temps la source de sa difficulté. Et enfin la dernière partie de chapitre présente des méthodes de calcule Bayésienne qui sont globalement numériques et qui ont rendu l'approche Bayésienne plus régide et performante.
Dans le deuxième chapitre, nous abordons les méthodes Bayesiennes variationnelles: définitions et principe.
Le troisième et dernier chapitre est consacré à un exemple d'application des méthodes bayésiennes variationnelle en neuroimagerie.

Chapitre 1

Analyse statistique Bayésienne

1.1 Introduction

Il subsiste, chez beaucoup, une représentation de la statistique Bayésienne fondée sur l'arbitraire de la loi a priori, point de référence qui ne saurait être remis en cause, tout en déterminant l'inférence résultante.
Un de nous objectifs est de guider le lecteur à se familiariser un peu dans la découverte de l'inférence Bayesienne. Quatre idées doivent motiver cette découverte :

- L'inférence Bayésienne n'est pas récente ;
- Elle apparait supérieure sur le plan théorique
- Elle est une inférence naturelle et flexible ;
- Elle va devenir de plus en plus facilement et largement utilisable.

1.2 Introduction à la théorie de la décision par l'approche Bayesienne

1.2.1 Le choix Bayésien

La statistique est un art interdisciplinaire de la quantification sous incertitudes utilisé par les physiciens, les économistes, les ingénieurs, les géographes, les biologistes, les assureurs, les psychologues,les météorologues etc. bref tous les praticiens soucieux de bâtir, sur des fondations solides, un pont entre théorie et données expérimentales. Depuis un siècle, la statistique s'est considérablement développée, initiant une révolution dans les modes de pensée, car elle porte un langage de representation du monde et des ses incertitudes.

C'est aujourd'hui une science mathématique dont l'objectif est de décrire ce qui s'est produit et de faire des projections quant à ce qu'il peut advenir dans le futur. Parfois, la situation peut être simplement décrite par quelques représentations graphiques d'analyse

élémentaire des données. Bien souvent, le problème est beaucoup plus compliqué car des multiples facteurs d'influences doivent être pris en compte.
Schématiquement, on construit deux ensembles avce ces facteurs. Un premier paquet contient les facteurs dits explicatifs, bien identifiés, ceux dont on souhaite étudier l'influence en détail. En ce qui concerne le second paquet de facteurs, on ne sait, ou ne veut pas, représenter leur effet pertubateur au cas par cas et, plus grossière par ses caractéristiques statistiques générales. Dans tous les cas, l'étude de la variabilité est au centre des débats: il s'agit d'abord de caractériser l'influence des facteurs identifiés et ensuite de représenter et d'évaluer le bruit résiduel dû à ces autres facteurs non pris en compte dans l'analyse de façon explicite. Dans une telle situation, le statisticien classique utilise à la fois un raisonnement déterministe par l'absurde, afin de proposer des valeurs acceptables pour les paramètres décrivant les effets des facteurs explicatifs et un raisonnement probabiliste, pour traduire la variabilté des résultats observés due au bruit. Ce mode de pensée s'appuie sur l'hypothèse de la réalité objective des paramètres ainsi que sur l'interpretation de la probabilité comme limite des fréquences des resultats observés. Par contre, le statisticien Bayésien utilise le même cadre de pensée pour traiter par le pari probabiliste l'interaction de ces deux niveaux d'incertitudes: ignorance quant aux valeurs possibles des paramètres et aléa des bruits entachant les résultats expérimentaux.

Choisir la piste Bayésienne paraîtra à certains inutilement trop sophistiqué si on se limite aux modèles élémentaires (binomial, normal, etc.), pour ces cas d'écoles simples, l'approche fréquentiste est facile (nombreux logiciels), et offre au praticien des résultats souvent trés proches de ceux que donnerait une analyse Bayésienne avec une distribution a priori peu informative. Mais pour peu que l'analyste souhaite prendre à bras le corps des problèmes plus proche de son réel quotidien, apparaissent variables multiples, données manquantes, effets aléatoires, grandeurs latentes..., la structure des modèles de la vie scientifique moderne se présente sous une forme où des couches successives de conditionnement s'êmboitent, et pour lesquels l'approche Bayésienne affirme sa véritable pertinence.

1.2.2 Notions de base

Un système est caractérisé par un certain nombre de variables définissant son état. Les valeurs de ces variables, les mesurandes, sont obtenues par le biais d'un système de mesure. Les résultats de la mesure, que nous appelerons aussi observations, ne permettent qu'une estimation de l'état, car elles interviennent dans le processus des phénomènes de nature aléatoire non maîtrisés par l'observateur. On obtient une correspondance entre l'état et l'observation qui est de nature statistique, associant à un état fixé une répartition des observations (gaussienne, poissonnienne ou autre). Le but de la mesure est alors

d'inverser cette relation, en ce sens que l'observateur ayant obtenu un résultat, doit en inférer l'état.

La relation donnant la repartition des observations pour état donné se nomme le modèle. Elle est objective, extérieure à l'observateur, car elle n'est dépendante que de l'objet mesuré et de l'instrument utilisé. En répétant l'experience en face du même état, l'observateur verra ses résultats se distribuer selon le modèle. On trouve une probabilité au sens fréquentiel.

Comme un état peut donner des résultats de mesure differents, une observation peut trés bien correspondance à plusieurs états. Quel est le bon? Ou mieux, quelle répartition peut-on imaginer sur les états ayant en main cette observation? L'incertitude passe du domaine des résultats dans celui des états. Cette incertitude est subjective, en ce sens qu'elle est propre à l'observateur.

En statistique classique, le modèle est donné par une fonction dite de vraisemblance $\ell(\theta|x)$, où θ désigne l'état du système et x l'observation. Cette fonction est normalisée et considérée comme densité de probabilité.

1.2.3 Principe général de l'inférence des paramètres en statistique Bayésienne

Au contraire de la statistique classique, en statistique Bayésienne les paramètres sont considérés comme des variables aléatoires aux quelles on affécte une densité de probabilitée, l'inférence des paramètres consiste à déterminer la densité de probabilité conjointe des grandeurs inconnues (paramètres, etc) à partir de toute l'information disponible sur les paramètres apportée par les données.

Cette remarque est connue sous le nom de principe de vraisemblance qui stipule que l'information apportée par une observation x sur le paramètre θ est entièrement contenue dans la fonction de vraisemblance $\ell(\theta|x)$.

L'inférence repose donc sur les notions de distribution conjointe et de distributions conditionnelles qui sont l'essentiel des méthodes Bayésiennes .

1.2.4 Théorème de Bayes

Soient A et B deux événements aléatoires tels que $P[A] \neq 0$ La probabilité de B, conditionnellement à la réalisation de A, est donner par la relation suivante:

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

$P[B,A]$ est la probabilité que les deux événements A et B aient lieu simultanément.
Puisque $P[B,A] = P[A,B]$ si dans l'expression de $P[A|B]$:

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

Donc

$$P[B,A] = P[B|A].P[A]$$

On remplace $P[A,B]$ par $P[B,A]$, en déduit la relation entre les deux probabilités conditionnelles $P[A|B]$ et $P[B|A]$:

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

Cette équation est une conséquence triviale de la définition de la probabilité conditionnelle, est appelée Formule de Bayes (ou aussi Théorème de Bayes) en l'honneur du Révérend.

Ce théorème (ou formule de Bayes) est l'un des plus célèbres de la statistique, est aussi un principe d'actualisation. Bayes (1763) donne en réalité une version continue de ces résultats, à savoir, pour deux variables aléatoires x et y, de distributions conditionnelle $f(x|y)$ et marginale $g(y)$, la distribution conditionnelle de y sachant x est:

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x)g(y)dy}$$

Bien que ce théorème d'inversion soit naturel d'un point de vue probabiliste, Bayes et Laplace sont allés plus loin et ont considéré que l'incertitude sur le paramètre θ d'un modèle peut être décrite par une distribution de probabilité π sur Θ. L'inférence est alors fondée sur la distribution de θ conditionnelle à x, cette probabilité conditionnelle s'écrit:

$$P(\theta|x) = \frac{\pi(\theta)P(x|\theta)}{M(x)} = \frac{P(x|\theta)\pi(\theta)}{\int_\theta P(x|\theta)\pi(\theta)}$$

Où Les différents termes de cette formule sont:
$M(x)$ est la distribution marginale des données qui a ici le rôle d'une constante de normalisation pour que la densité *a posteriori* s'intLgre bien à 1.

$\pi(\theta)$ est la distribution de probabilité *a priori* du paramètre inconnu θ.

$P(x|\theta)$ est la probabilité des observations conditionnellement à la valeur θ du paramètre du modèle statistique qu'on utilise pour leur description. Il s'agit de la vraisemblance des données, sous le modèle paramètrée par θ.

$P(\theta|x)$ est la distribution de probabilité *a posteriori* du paramètre du modèle, sur la base de la connaissance *a priori* et de l'information apportée par les données. L'appellation *a posteriori* vient du fait que, logiquement, elle suit l'observation des données.

Le dénominateur, indépendant de θ, est uniquement une constante de normalisation.

Alors que la statistique classique repose sur la loi des observations, la statistique Bayésienne repose sur la loi *a posteriori* qui peut s'interpréter comme un résumé (en un sens probabiliste) de l'information disponible sur θ, une fois x observé. L'approche Bayésienne réalise en quelque sorte l'actualisation de l'information *a priori* par l'observation x, au travers de $\pi(\theta|x)$.

Le shéma ci-dessous résume la démarche Bayésiene dans le cadre de la statistique paramétrique inférentielle. Il fait également apparaître, la modélisation stochastique des x_i comme étant des réalisations de variables aléatoires X_i (cette modélisation est caractéristique de la statistique inférentielle), ainsi que la modélisation stochastique de l'information a priori disponible sur le paramétre θ, au travers de la loi *a priori*.
Le passage de la distribution *a priori* à la distribution *a posteriori* des paramètres du modèle statistique, exprimé par la formule Bayes, peut être alors interprété comme une mise à jour de la connaissance, sur la base des observations (figure 1.1).

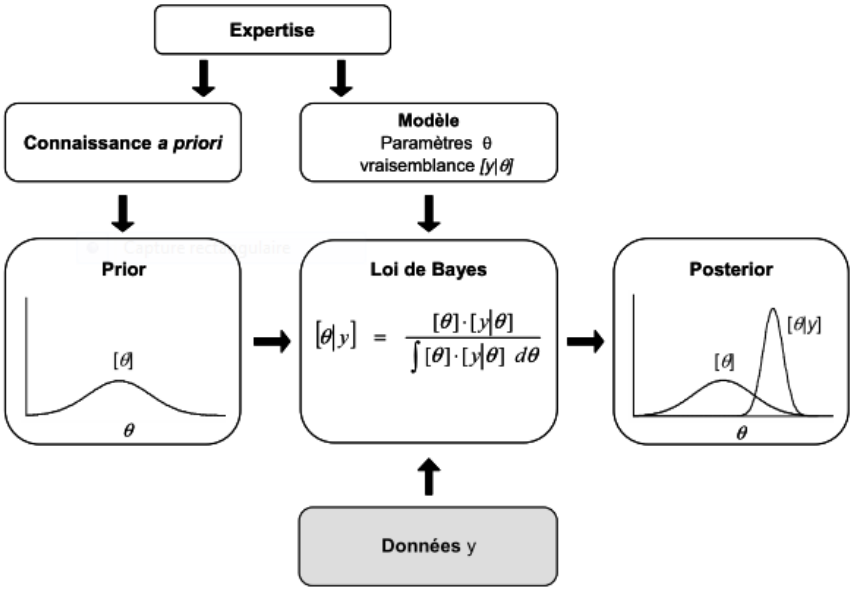

FIG. 1.1 – L'approche Bayesienne .

Exemple 1.1.

Soit un échantillon de variable aléatoire qui suit une loi Bernoulli de paramétre P $(B(P))$ $\theta = P$ est une variable aléatoire On prend la loi *a priori* $\pi(\theta) = 1$ Calculons alors la densité *a posteriori* $f(\theta|x)$:

$$f(\theta|x) \propto \ell(\theta,x)\pi(\theta)$$

donc:

$$f(\theta|x) \propto \theta^s(1-\theta)^{n-s}$$

Avec $s = \Sigma_{i=1}^{n} x_i$

On remarque que:

$$f(\theta|x) \propto \theta^{(s+1)-1}(1-\theta)^{(n-s+1)-1}$$

ce qui a implique:

$\theta|x$ suit une loi $Beta(s+1,n-s+1)$

Exemple 1.2. *(Jean-Michel Marin, Christian P.Robert)*

Dans le cadre d'une observation normale de moyenne inconnue θ, $x \sim \mathcal{N}_1(\theta,1)$[1], la loi *a posteriori* associée à la loi *a priori* $\theta \sim \mathcal{N}_1(0,10)$ est

$$\pi(\theta|x) \propto \exp -\frac{1}{2}\{10^2 + (\theta - x^2)\}$$

ce qui équivaut à la loi

$$\theta|x \sim \mathcal{N}(\frac{10x}{11},\frac{10}{11})$$

L'espérance *a posteriori* de θ est donc $\frac{10x}{11}$.

Remarquons que l'approche bayesienne respecte la vraisemblance.

1.3 Théorie de la décision statistique

1.3.1 Estimateur de Bayes

Comme nous l'avons déjà fait remarquer, la prise d'une décision, ici le choix d'un estimateur, va engendrer un coût que l'on va quantifier à l'aide de la fonction de perte. En pratique, on cherche une décision qui minimise en moyenne la fonction de coût.

Soit une fonction de coût $\ell(\theta,\delta)$, et une loi de probabilté *a priori* (ou une loi impropre) π, pour trouver l'estimateur de Bayes $\delta^\pi(x)$, on applique la règle suivante:

$$\delta^\pi(x) = \min_\delta \mathbb{E}^\pi[\ell(\theta,\delta)/x]$$

L'estimateur $\delta^\pi(x)$ sera déterminé analytiquement ou numériquement ceci dépendre de la fonction perte, sa nature et complexité.

Généralement, les solutions associées à des coûts classiques sont formellement connues et correspondents aux caractéristiques usuelles d'une distribution (moyenne, mediane, fractiles etc...).
Par exemple, l'estimateur de Bayes associé au coût quadratique est la moyenne *a posteriori*. Cette construction formelle des estimateurs de Bayes classiques n'évite pas toujours

1. Le vecteur aléatoire X à valeurs dans $\mathbb{R}^p$ distribué suivant une loi normale d'espérance μ et de structure de covariance Σ,$\mathcal{N}_p(\mu,\Sigma)$, admet comme densité de probabilté:

$$f_X(X|\mu,\Sigma) \propto \exp[-0.5(x-\mu)^T\Sigma^{-1}(x-\mu)]$$

, *où* A^T désigne la transposée de la matrice A.

le recours à une approximation numérique, particuliérement dans des cas multidimensionnels.

Lemme 1.3.1. *(Christian P. Robert 2006)*

Soit $f(x|\theta)$ une distribution de probabilité appartenant à une famille exponentielle. Pour toute loi a *priori* π la moyenne *a posteriori* de θ est donnée par:

$$\delta^{\pi}(x) = \nabla \log m_{\pi}(x).\nabla \log h(x)$$

où ∇ est l'opérateur gradient et m_{π} est la loi marginale associée à π.

Preuve: *L'espérance a posteriori est donnée par*

$$\begin{aligned} {}^{\pi}[\theta_i|x] &= \frac{\int_{\Theta} \theta_i h(x) e^{\theta.x-\psi(\theta)} \pi(\theta) d\theta}{m_{\pi(x)}} \\ &= \frac{\partial}{\partial x_i} \int_{\Theta} h(x) e^{\theta.x-\psi(\theta)} \pi(\theta) d\theta \frac{1}{m_{\pi}} - (\frac{\partial}{\partial x_i}) \\ &= \frac{\partial}{\partial x_i} [\log m_{\pi}(x) - \log h(x)]. \end{aligned}$$

Corollaire 1. *(Christian P.Robert 2006)*

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

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

est la moyenne *a posteriori*, $\delta^{\pi} = \mathbb{E}^{\pi}[\theta|x]$, pour toute matrice $Q(p \times p)$ symétrique défine positive.

Proposition 1. *(Christian P. Robert 2006)*

L'estimateur de Bayes associé à la loi *a priori* π et à la fonction de coût linéaire par morceaux est le fractile $(k_1/(k_1 + k_2))$ de $\pi(\theta|x)$.

En particulier, si $k_1 = k_2$, dans le coût absolu, l'estimateur de Bayes est la médiane *a posteriori*, qui est l'estimateur obtenu par Laplace.

Proposition 2. *(Christian P. Robert 2006)*

L'estimateur de Bayes associé à π et au coût 0-1 est

$$\delta^{\pi}(x) = \begin{cases} 1 & \text{si } P(\theta \in \Theta_0|x) > P(\theta \nsubseteq \Theta_0|x) \\ 0 & \text{sinon} \end{cases}$$

donc $\delta^{\pi}(x)$ vaut 1 si et seulement si $P(\theta \in \Theta_0|x) > 1/2$.

Quelques estimateurs usuelles

Le tableau ci-dessous représente quelques estimateur de Bayes du paramètre θ sous coût quadratique pour les lois *a priori* conjuguées des familles exponentielles usuelles.

Loi de x	Loi conjuguée	Moyenne a posteriori
Normale $N(\theta,\sigma^2)$	$N(\mu,\tau^2)$	$\frac{\mu\sigma^2+\tau^2x}{\sigma^2}$
Poisson $P(\theta)$	$\Gamma(\alpha,\beta)$	$\frac{\alpha+x}{\beta+1}$
Gamma $\Gamma(\nu,\theta)$	$\Gamma(\alpha,\beta)$	$\frac{\alpha+\nu}{\beta+x}$
Binomiale $B(n,\theta)$	$Be(\alpha,\beta)$	$\frac{\alpha+x}{\alpha+\beta+n}$
Binomiale Négative $Neg(m,\theta)$	$Be(\alpha,\beta)$	$\frac{\alpha+n}{\alpha+\beta+x+n}$
Multinomiale $M_k(\theta_1,...,\theta_k)$	$D(\alpha_1,...,\alpha_k)$	$\frac{\alpha_i+x_i}{(\sum_j \alpha_j)+n}$
Normale $N(\mu,1/\theta)$	$\Gamma(\alpha,\beta)$	$(\frac{\alpha+1}{\beta+(\mu-x)})^2$

Tab1.1-Quelques estimateurs de Bayes usuels

Propriétés de l'estimateur de Bayes

- L'estimateur de Bayes est admissible.
- L'estimateur de Bayes est biaisé.
- L'estimateur de Bayes est convergent en probabilité (quand la taille de l'échantillon$n \longrightarrow +\infty$).
- La loi a posteriori peut étre asymptotiquement (c.a.d. pour de grandes valeurs de n) approximée par une loi normale $N(E[\theta|x],Var[\theta|x])$.

1.3.2 Fonction perte et risque

1.Fonction de perte

Définition 1.3.1. Soit $T = T(x_1,.....x_n)$, un estimateur de θ.
On appelle "fonction de perte" et on note $\ell(t,\theta)$, toute fonction satisfaisant:

- $\ell(t,\theta) \geq 0$ pour toute valeur d'estimateur $t,\forall\theta \in \Theta$

– $\ell(t,\theta) = 0$ si $t = \theta$

Cette fonction mesure la perte occasionnée lorsque on estime θ par t.
Nous représentons quelques fonctions de perte rencontrées dans la littérateur:

1. $\ell_1(t,\theta) = c_1(t-\theta)1\theta \leq t(t) + c_2(\theta - t)1\theta > t(t)$
2. $\ell_2(t,\theta) = \varphi(\theta)\|t-\theta\|^r$ avec $\varphi(\theta) \geq 0$ et $r > 0$
3. $\ell_3(t,\theta) = \begin{cases} A & \text{si } (t-\theta) > \varepsilon \forall \varepsilon > O, A > 0 \\ 0 & \text{sinon} (t-\theta) \leq \varepsilon \end{cases}$

2. Risque Bayésien

Supposon qu'on veut estimer Pour le modèle $X \in \{\chi,\beta,\{P_\theta,\theta \in \Theta\}\}$, on définit D l'ensemble des décisions possibles. c'est-à-dire l'ensemble des fonctions de Θ dans $g(\Theta)$ où g dépend du contexte:

– si le but est d'estimer θ alors D = Θ
– pour un test, D=$\{0,1\}$

La fonction perte est une fonction mesurable de $(\Theta \times D)$ à valeurs positives: $\ell : \Theta \times D \to \mathbb{R}_+$. Elle est définie selon le problème étudié et constitue l'armature du problèm statistique.

Définition 1.3.2. *Risque fréquentiste*

Pour (θ,δ), le risque fréquentiste est défini par

$$R(\theta,\delta) = E_\theta[L(\theta,\delta(x))] = \int_X \ell(\theta,\delta(x)) f(x|\theta) dx$$

C'est une fonction de θ et ne défint donc pas un ordre total sur D et ne permet donc pas de comparer toutes décisions et estimateurs.Donc il n'existe 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 1.3.3. *Risque a posteriori*

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

$$\rho(\pi,\delta|x) = E^\pi[L(\theta,\delta)|x] = \int_\Theta R(\theta,\delta) d\pi(\theta)$$

Définition 1.3.4. *Risque intégré*

A fonction de perte donnée, le risque intégré est défini par :

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

Définition 1.3.5. *Estimateur Bayésien*

Un estimateur Bayésien est un estimateur vérifiant :

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

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. *Méthode de calcul*

Si $\exists\delta \in D,\tau(\pi,\delta) < \infty$ Alors: $\delta^{\pi}(X)$ est un estimateur Bayésien.

1.3.3 Fonctions de coût usuelles

1. Coût quadratique

Introduit par Légende (1805) et Gauss (1810), ce coût est sans contexte le critère d'évaluation le plus commun.

Définition 1.3.6.

La fonction de coût quadratique est la fonction définie par :

$$\ell(\theta,\delta(x)) = (\theta - \delta(x))^2.$$

Une variante de cette fonction de coût est une fonction de coût quadratique pondérée de la forme :

$$\ell(\theta,\delta(x)) = \omega(\theta)(\theta - \delta(x))^2.$$

Le coût quadratique est particulièrement intéressant lorsque l'espace des paramètres est borné et le choix d'un coût plus subjectif.

les estimateurs de Bayes associés au coût quadratique sont les moyennes *a posteriori*. Cependant, notons que le coût quadratique n'est pas le seul coût à avoir cette caract éristique. Les fonctions de coût conduisant à la moyenne *a posteriori* comme estimateur de Bayes sont appelées fonctions de coût propres et ont été identifiées par Lindley (1985).

Proposition 3. *(Christian P.Robert 2006)*

Sous l'hypothèse d'un coût quadratique, l'estimateur de Bayes $\delta^\pi(x)$ de θ associé à la loi *a priori* π est la moyenne *a posteriori* de θ :

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

Preuve:

Comme

$$E^\pi[(\theta-\delta)^2|x] = E^\pi[(\theta^2|x)] - 2\delta E^\pi[(\theta|x)] + \delta^2$$

le minimum du coût *a posteriori* est effectivement atteint par $\delta^\pi(x) = E[(\theta|x)]$
Corollaire 2. *(Christian P.Robert 2006)*

Quand $\Theta \in R^p$, l'estimateur de Bayes δ^p associé à π et au coût quadratique,

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

est la moyenne *a posteriori*, $\delta^\pi(x) = E^\pi[\theta|x]$, pour toute matrice $Q(p*p)$ symetrique définie positive.

L'estimateur de Bayes δ^π associé à π et au coût quadratique ponderé

$$\ell(\theta,\delta) = w(\theta)(\theta-\delta)^2$$

Où $w(\theta)$ est une fonction positive, est

$$\delta^\pi(x) = \frac{E^\pi[w(\theta)\theta|x]}{E^\pi[w(\theta|x)]}$$

Exemple 1.

Si $X \sim P(\theta)$ et si $\pi(\theta)$, la loi *a priori* est une loi Gamma de paramètres (α,β), la loi *a posteriori* $\pi(\theta|x)$ est une loi Gamma de parametre $(x+\alpha,\beta+1)$. Sous l'hypothése d'un coût quadratique, un estimateur de Bayes $\delta^\pi(x)$ de θ sera l'espérance *a posteriori* de θ . Puisque la loi *a posteriori* est une loi Gamma, l'espérence est le rapport des paramètres et on a:

$$\delta^\pi(x) = (x+\alpha)/(\beta+1)$$

– **Quelques estimateurs usuelles**

Estimateurs de Bayes du paramètre θ sous coût quadratique pour les *lois a priori* conjuguées des familles exponentielles usuelles.

Loi de x	Loi conjuguée	Moyenne a posteriori
$\mathcal{N}(\theta,\sigma^2)$	$\mathcal{N}(\mu,\tau^2)$	$\frac{\mu\sigma^2+\tau^2x}{\sigma^2}$
$P(\theta)$	$\Gamma(\alpha,\beta)$	$\frac{\alpha+x}{\beta+1}$
$\Gamma(\nu,\theta)$	$\Gamma(\alpha,\beta)$	$\frac{\alpha+\nu}{\beta+x}$
$B(n,\theta)$	$Be(\alpha,\beta)$	$\frac{\alpha+x}{\alpha+\beta+n}$
$Neg(m,\theta)$	$Be(\alpha,\beta)$	$\frac{\alpha+n}{\alpha+\beta+x+n}$
$M_k(\theta_1,...,\theta_k)$	$D(\alpha_1,...,\alpha_k)$	$\frac{\alpha_i+x_i}{(\sum_j \alpha_j)+n}$
$\mathcal{N}(\mu,1/\theta)$	$\Gamma(\alpha,\beta)$	$(\frac{\alpha+1}{\beta+(\mu-x)})^2$

Tab1.1-Quelques estimateurs de Bayes usuels

2. Coût absolu

Définition 1.3.7. - La fonction de coût absolue es la fonction définie par:

$$\ell(\theta,\delta(x)) = \begin{cases} k_2(\theta-\delta(x)) & \text{si } \theta > \delta(x) \\ k_1(\delta(x)-\theta) & \text{sinon.} \end{cases}$$

Proposition 1. *(Christian P. Robert 2006)*

Un estimateur de Bayes associé à la loi *a priori* π et au coût absolue, est un fractile d'ordre $k_2/(k_1+k_2)$ de $\pi(\theta|x)$

En particulier, si $k_1 = k_2$, dans le cas du coût absolu, l'estimateur de Bayes est la médiane *a posteriori*, qui est l'estimateur obtenu par Laplace.

3. Coût 0-1

Définition 1.3.8. On appelle coût 0-1, l'application L définie par :

$$\ell(\theta,\delta(x)) = \begin{cases} 0 & \text{si } la\ décision\ est\ bonne, \\ 1 & \text{sinon} \end{cases}$$

En utilisant cette fonction de coût, on trouvent les résultats de la théorie des tests d'hypothèses.
Un problème de test est un problème de choix (de prise de décision) entre $H_0 : \theta \in \Theta_0$ et $H_1 : \theta \in \Theta_1$

On définit donc la décision de la manière suivante :
$\delta(X) = 1$: on accepte H_0; $\delta(X) = 0$: on rejette H_0
On a un espace d'actions de la forme : $A = 0{,}1$

Soit W la région de rejet i.e. le sous-ensemble qui conduit à rejeter H_0. On peut construire une fonction de coût de la manière suivante :

supposons $\theta \in \Theta_0$,
Si $X \in W$, on prend la décision de rejeter i.e. $\delta(X) = 0$, mais la décision n'est pas bonne, on va pénaliser et $\ell(\theta,\delta(x)) = 1$.
Si $X \nsubseteq W$,on ne rejette pas, on prend la décision $\delta(X) = 1$, la décision est bonne $\ell(\theta,\delta(x)) = 0$.
Le coût s'écrit donc :

$$l(\theta,\delta(x)) = \{1 - \delta(x) si \theta \in \Theta_0 \delta(x) sinon\}$$

Ce qu'on peut écrire: $l(\theta,\delta(x)) = 1_{x\in W}$ et on calcule la fonction de risque:

$$R(\theta,\delta) = E[l(\theta,\delta(x))] = \int l(\theta,\delta(x)) dP_\theta(x) = P_\theta(x \in W), \theta \in \Theta_0$$

On retrouve le risque de première espèce.

1.3.4 Admissibilité et minimaxité

Définition 1.3.9. Estimateur randomisé

Pour le modèle $X \in \{\chi,\beta,\{P_\theta,\theta \in \Theta\}\}$, un ensemble de décisions D, on définit D^* comme l'ensemble des probabilités sur D. $\delta^* \in D^*$ est appelé estimateur randomisé.
L'idée à l'origine de cette notion est de rendre D convexe pour pouvoir maximiser facilement.

Théorème 1.3.1. (Christian P. Robert 2006)

Pour toute distribution *a priori* π sur Θ, le risque de Bayes pour l'ensemble des estimateurs randomisés est le même que celui pour l'ensemble des estimateurs non randomisés, soit

$$\inf_{\delta \in D} r(\pi,\delta) = \inf_{\delta^* \in D^*} r(\pi,\delta^*) = r(\pi)$$

Minimaxité

Le critère de minimaxité apparaît comme une assurance contre le pire, car il vise à minimiser le coût moyen dans le cas le moins favorable. Il représente aussi un éffort fréquentiste pour éviter de recourir au paradigme Bayésien, tout en engendrant un ordre (faible) sur D^*

Définition 1.3.10.

On appelle risque minimax

$$\overline{R} = \inf_{\delta \in D^*} \sup_{\theta} \mathbb{E}_\theta[L(\theta,\delta(x))]$$

et estimateur minimax tout estimateur δ_0 telque

$$\overline{R} = \sup_{\theta} R(\theta,\delta_0)$$

L'estimateur minimax correspond au point de vue de faire le mieux dans le pire des cas, c'est-à-dire à s'assurer contre la pire. Il est utile dans des cadres complexes mais trop conservateur dans certains cas où le pire est trés probable. Il peut être judicieux de voir l'estimation comme un jeu entre le statisticien (choix de δ) et la Nature (choix de θ), l'estimateur minimax rejoint alors celle de la Théorie des Jeux.

Règle minimax et Stratégie maximin

Une difficulté importante liée à la notion de minimaxité est que les estimateurs minimax n'existent pas nécessairement. En particulier, il existe une stratégie minimax quand Θ est fini et la fonction de coût est continue. Plus généralement, Brown (1976) (voir aussi Le Cam, 1986, et Strasser, 1985) considère l'espace de décision D comme plongé dans un autre espace de manière telle que l'ensemble des fonctions de risque sur D est compact dans ce grand espace. dans cette perspective et sous des hypothèses supplémentaires, il est

alors possible de construire des estimateurs minimax lorsque la fonction coût est continue.

Théorème 1.3.2. *(Christian P. Robert 2006)*

Si $D \subset \mathbb{R}^k$ est convexe et compact et si $L(\theta,d)$ est continue et convexe en tant que fonction de d, pour chaque $\theta \in \Theta$, alors, il existe un estimateur minimax non randomisé.

Lemme 1.3.2. *(Rousseau 2009)*

Le risque de Bayes est toujours plus petit que le risque minimax,

$$\underline{R} = \sup_{\pi} r(\pi) = \sup_{\pi} \inf_{\delta \in D} r(\pi,\delta) \leq \overline{R} = \inf_{\delta \in D^*} \sup_{\theta} R(\theta,\delta)$$

La première valeur est dite *risque maximin* et une distribution π^* telle que $r(\pi^*) = \underline{R}$ est appelée *distribution a priori la moins favorable*, quand de telles distributions existent. En général, la borne supérieure $r(\pi^*)$ est atteinte plutôt par une distribution impropre pouvant s'exprimer comme une limite de distribution *a priori* propre π_n. Mais ce phénomène n'empêche pas nécessairement la construction d'estimateurs minimax. Quand elles existent, les distributions les moins favorables sont celles qui ont le risque de Bayes le plus grand, donc aussi les moins intéressantes en terme de coût lorsqu'elles ne sont pas suggérées par l'information *a priori* disponible. Le résultat ci-dessus est assez logique au sens où l'information *a priori* ne peut qu'améloirer l'erreur d'estimation, même dans le pire des cas.

Définition 1.3.11.

Un problème d'estimation est *dit admettre une valeur* si $\underline{R} = \overline{R}$, c'est-à-dire quand

$$\sup_{\pi} \inf_{\delta \in D} r(\pi,\delta) = \inf_{\delta \in D^*} \sup_{\theta} R(\theta,\delta)$$

Quand le problème admet une valeur, certains estimateurs minimax sont des estimateurs de Bayes correspondant aux lois *a priori* les moins favorables. Cependant, ils peuvent être randomisés. Par conséquent le principe minimax ne fournit pas toujours des estimateurs acceptables.

Lemme 1.3.3. *(Christian P. Robert 2006)*

Si δ_0 est un estimateur de Bayes pour π_0 et si $R(\theta,\delta_0) \leq r(\pi_0)$ pour tout θ dans le support de π_0, δ_0 est minimax et π_0 est la distribution la moins favorable.

Admissibilité

Définition 1.3.12. *Estimateur admissible*

Soit $X \in \{\chi,\beta,\{P_0,\theta \in \Theta\}\}$ un modèle paramétrique et L une fonction de perte sur $\Theta \times D$ où D est l'ensemble des décisions. On dit que $\delta \in \Theta$ est inadmissible si et seulement si $\exists\delta_0 \in D$, $\forall\theta \in \Theta$, $R(\theta,\delta) \geq R(\theta,\delta_0)$ et $\exists\theta_0 \in \Theta$, $R(\theta_0,\delta) \geq R(\theta_0,\delta_0)$. dans le cas contraire, δ est admissible.

Proposition 1.3.1. *(Christian P. Robert 2006)*

S'il existe un unique estimateur minimax, cet estimateur est admissible

Notons que le reciproque de ce résultat est fausse, car il peut exister plusieurs estimateurs minimax admissibles. Par exemple, dans le cas $N_p(\theta,I_p)$, il existe des estimateurs de Bayes réguliers minimax pour $p \geq 5$. Quand la fonction de coût L est absolument convexe (en d), la caractérisation suivante est aussi possible.

Proposition 1.3.2. *(Rousseau 2009)*

Si δ_0 est admissible de risque constant, δ_0 est l'unique estimateur minimax.

Théorème 1.3.3. *Estimateurs Bayésiens admissibles (Rousseau 2009)*

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

Proposition 1.3.3. *(Christian P. Robert 2006)*

Si un estimateur de Bayes, δ^π, associé à une loi a priori (propre ou impropre) π, est tel que le risque de Bayes,

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

soit fini, δ^π est admissible.

Définition 1.3.13. π-admissibilité

Un estimateur δ_0 est π-admissible si et seulement si

$$\forall(\delta,\theta), R(\theta,\delta) \leq R(\theta,\delta_0) : \pi(\{\theta \in \Theta, R(\theta,\delta) < R(\theta,\delta_0\}) = 0$$

Proposition 1.3.4. *(Christian P. Robert 2006)*

Tout estimateur Bayésien tel que $r(\pi) < \infty$ est π-admissible

Théorème 1.3.4. *Continuité et π-admissibilité, (Rousseau 2009)*

Si $\pi > 0$ sur Θ, $r(\pi) < \infty$ pour une fonction perte L donnée, si δ^π estimateur Bayésien correspondant existe et si $\theta \mapsto R(\theta,\delta)$ est continu, alors δ^π est admissible.

1.4 Choix des lois a priori

Le point le plus signifiant dans l'analyse Bayésienne est le choix de la loi *a priori*, sa détermination est donc l'étape la plus importante dans la mise en oeuvre de cette inférence.
En statistique Bayésienne, on considère, en plus des données récoltées dans le cadre d'une expérience, un *a priori* sur le paramètre θ que l'on cherche à estimer. C'est le terme $\pi(\theta)$ Cela peut permettre d'inclure, dans les résultats précédents, formels ou non. En pratique, toute la difficulté consiste à estimer de manière correcte nos *a priori*. Avec beaucoup d'observations, le comportement asymptotique peut guider ce choix mais sinon il est nécessaire de le justifier avec précision.

1.4.1 Approche partiellement informative

Quand on dispose de peu d'information *a priori*, ou quand l'information dont on dispose est trop vague, alors souvent le statisticien ne peut faire une construction subjective complète de l' *a priori*.
De telles situations peuvent obliger le statisticien à avoir recours à des méthodes d'estimation fréquentiste comme: Estimateur du maximum de vraisemblance, estimateur sans biais optimaux, etc.
Cependant, tout en gardant à l'esprit les fondements Bayésiens des critères fréquentistes d'optimalité, il parait don préferable de suivre l'approche Bayésienne, en utilisant un *a priori* dit objectif, c'est-à-dire construit à partir du modèle d'échantillonnage.

Lorsque aucune information *a priori* n'est disponible, ces *a priori* sont dits *non informatifs*.

Maximum d'entropie

Si l'on possède des informations partielles du type $\mathbb{E}^{\pi}[\eth\neg(\theta)] = \mu_k$ où pour chaque $k = 1,...,n$, g_k est une fonction donnée.
Pour $\theta \in \{1,...,n\}$ et $\pi(\theta) = (\pi_1,...,\pi_n)$ tel que $\pi_i > 0$ et $\sum_{i=1}^{n} \_i = 1$, l'entropie de la loi est définie par

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

Ce dernier terme correspond à une répartition uniforme. Pour la masse de Dirac $\delta(j)$ (telle que $\pi_j = 1$ et $\forall i \neq j$, $\pi_i = 0$), $Ent(\delta(j)) = 0$ ce 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 infromative. 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:

$$\arg\max_{\pi} Ent(\pi) \textit{sous la contrainte } \mathbb{E}^{\pi}[\eth\neg(\theta)] = \mu_k$$

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 determine ces valeurs λ à partir des contraintes (systèmes d'équations) comme l'indique l'exemple à suivre.

Exemple 1.3. *Un cas dénombrable*

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

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

Cela conduit à résoudre:

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

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

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:

$$Ent(\pi|\pi_0) = \int_\Theta \pi(\theta) \log(\frac{\pi(\theta)}{\pi(\theta_0)}) d\theta$$

C'est en fait la divergence de Kullback. Dans l'idée π_0 est la plus proche de la repartitition uniforme. L'objectif est donc de maximiser $Ent(\pi|\pi_0)$ sous les contraintes $\mathbb{E}^\pi[g_k(\theta)] = \mu_k$. Là encore, la solution générale est connue:

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

Lois a priori conjuguées

Ce type de lois *a priori* est utilisé quand l'information *a priori* disponible sur le modèle est trop vague ou peu faible. Dans ce cas l'analyste regarde la forme de la fonction de vraisemblance et choisit une famille de lois qui se marie bien avec elle.

Définition 1.4.1. Famille conjuguée

Une famille $\mathcal{F}$ de distributions de probabilité sur Θ est dite conjuguée (ou fermée par échantillonnage) par une fonction de vraisemblance $f(x|\theta)$ si, pour tout $\pi \in \mathcal{F}$, la distribution *a posteriori* $\pi(.|x)$ appartient également à $\mathcal{F}$.

L'avantage des familles conjuguées est avant tout la simplicité des calculs. Avant l'essor du calcul numérique, ces familles étaient pratiquement les seules qui permettaient de faire aboutir des calculs. L'intérêt principal du caractère conjugué se manifeste quand $\mathcal{F}$ est paramétrée.

Effectivement le passage de *distribution a priori* à la *distribution a posteriori* ce n'est dans ce cas qu'une mise à jour des paramètres correspondants. Et par conséquence, les *distributions a posteriori* sont toujours calculables dans ce cas.

L'approche *a priori* conjuguée, introduite par Raiffa et Schlaifer (1961), peut être considérée comme un point de départ pour l'élaboration de distributions *a priori* fondées sur une information *a priori* limitée. On considère une variable x suivant une fonction de densité $f(x|\theta)$

Un exemple trivial d'une famille conjuguée est l'ensemble $\mathcal{F}_0$ de toutes les lois de probabilité sur Θ.
L'avantage des familles conjuguées est avnat tout de simplifier les calculs. Avant l'essor

du calcul numérique, ces familles étaient pratiquement les seules qui permettaient de faire aboutir des calculs.

les lois *a priori* conjuguées sont généralement associées à un type particullier de lois d'échantillonnage qui permet toujours leur obtention; il est même caractéristique des lois *a priori* conjuguées comme nous le verrons ci-dessous. Ces lois constituent ce qu'on appelle *des familles exponentielles.*

Définition 1.4.2. Familles exponentielles

Soient μ une mesure σ-finie sur χ, Θ l'espace des paramètres, C et h des fonctions respectivement de χ et Θ dans $\mathbb{R}_+$, et $\mathcal{R}$ et T des fonctions de Θ et χ dans $\mathbb{R}^k$. La famille des distributions de densité (par rapport à μ)

$$f(x|\theta) = C(\theta)h(x)\exp\{\mathcal{R}(\theta).T(x)\}$$

est dite famille exponentielle de dimension k. Dans le cas particulier où $\Theta \subset \mathbb{R}^k$ et

$$f(x|\theta) = C(\theta)h(x)\exp\{\theta.x\}$$

la famille est dite *naturelle.*
D'un point de vue analytique, les familles exponentielles ont certaines caractéristiques intéressantes. Il existe une statistique exhaustive de dimension constante, en effet, si x_1,...,$x_n \sim f(x|\theta)$, avec f satisfaisant

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \in \mathbb{R}^k$$

est exhaustive pour tout n. La réciproque de ce résultat a été aussi établie par Koopman (1936) et Pitman (1936).

Théorème 1.4.1. *(Lemme de Pitman-Koopman) (Christian P. Robert 2006)*

Si une famille de lois f(.|x) à support constant est telle que, à partir d'une taille d'échantillon suffisament grande, il existe une statistique exhaustive de taille fixe, la famille est exponentielle.

Exemple 1.4.

Soit $x \sim \mathcal{N}_p(\theta,\sigma^2 I_p)$ *alors,*

$$f(x|\theta) = \frac{1}{\sigma^p}\frac{1}{(2p)^{p/2}}\exp\{-\sum_{i=1}^{p}(x_i-\theta_i)^2/2\sigma^2\}$$

$$= C(\theta,\sigma)h(x)\exp\{x.(\theta/\sigma^2) + \|x\|^2(-1/2\sigma^2)\}$$

et la distribution normale appartient à une famille exponentielle de paramètres naturels θ/σ^2 *et* $-1/2\sigma^2$. *De la même façon, si* $x_1,...,x_n \sim \mathcal{N}_p(\theta,\sigma^2 I_p)$, *la distribution jointe satisfait*

$$f(x_1,...,x_n) = C^{'}(\theta,\sigma)h^{'}(x_1,...,x_n) \times \exp\{n\overline{x}.(\theta/\sigma^2) + \sum_{i=1}^{n} \|x_i - \overline{x}\|^2(-1/2\sigma^2)\}$$

et la statistique $\overline{x}$, $\sum_{i=1} \|x_i - \overline{x}\|^2$ *est exhaustive pour tout* $n \geq 2$

Définition 1.4.3.

Soit $f(x|\theta) = C(\theta)h(x)\exp\{\theta.x\}$, une famille exponentielle naturelle. L'espace naturel des paramètres est

$$N = \{\theta; \int_{\chi} e^{\theta.x}h(x)d\mu(x) < +\infty\}$$

La famille est dite régulière si N est un ensemble ouvert et minimale si $dim(N) = dim(K) = k$, où K est la clôture de l'enveloppe convexe du support de μ.

Remarque 1.4.1.

Les familles exponentielles naturelles peuvent aussi être réécrites sous la forme

$$f(x|\theta) = h(x)e^{\theta.x-\psi(\theta)}$$

et $\psi(\theta)$ *est dite fonction cumulante des moments.*

Lois conjuguées des familles exponentielles

Soit $f(x|\theta) = h(x)e^{\theta.x-\psi(\theta)}$ loi générique d'une famille exponentielle. Cette loi admet alors une famille conjuguée.

Proposition 1.4.1.

Une famille conjuguée pour $f(x|\theta)$ *est donnée par*

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)e^{\theta.\mu-\lambda\psi(\theta)}$$

où $K(\mu,\lambda)$ *est la constante de normalisation de la densité. La loi a posteriori correspondante est* $\pi(\theta|\mu + x,\lambda + 1)$.

Remarque 1.4.2.

Seuls les modèles à structure exponentielle admettent une famille conjuguée.

Le tableau suivant contient quelques lois a priori conjuguées usuelles.

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
$N(\theta,\sigma^2)$	$N(\mu,\tau^2)$	$N(\varrho(\sigma^2\mu+\tau^2x),\varrho\sigma^2\tau^2)$ $\varrho^{-1}=\sigma^2+\tau^2$
$P(\theta)$	$\Gamma(\alpha,\beta)$	$\Gamma(\alpha+x,\beta+1)$
$\Gamma(\nu,\theta)$	$\Gamma(\alpha,\beta)$	$\Gamma(\alpha+\nu,\beta+x)$
$B(n,\theta)$	$Be(\alpha,\beta)$	$Be(\alpha+x,\beta+n+x)$
$Neg(m,\theta)$	$Be(\alpha,\beta)$	$Be(\alpha+m,\beta+x)$
$M_k(\theta_1,...,\theta_k)$	$D(\alpha_1,...,\alpha_k)$	$D(\alpha_1+x_1,...,\alpha_k+x_k)$
$N(\mu,1/\theta)$	$\Gamma(\alpha,\beta)$	$\Gamma(\alpha+0.5,\beta+(\mu-x^2)/2)$

Tab1.2-Lois a priori conjuguées usuelles.

1.4.2 Approche non informative

Lorsque aucune information *a priori* n'est disponible, le choix de la loi *a priori* est analytique, puisqu'elles donnent des expressions exactes pour quelques quantités *a posteriori*. Dans de telles situations, il est impossible de justifier le choix d'une loi *a priori* sur des bases subjectives. Plutôt que de revenir aux alternatives classiques, comme l'estimation par maximum de vraisemblance, ou d'utiliser les données pour approcher ces hyperparamètres, comme dans une analyse Bayésienne empirique, il est préférable de faire appel à des techniques bayésiennes, ne serait-ce que parce qu'elles sont à la base des critères classiques d'optimalité. Dans un tel cas, ces lois *a priori* particulières doivent être construites à partir de la distribution d'échantillonnage, puisque c'est la seule information disponible. Pour des raisons évidentes, de telles lois sont dites *non informatives*.
Les lois de probabilités non informatives nous amènent souvent à des résultats qui sont des mesures et non des probabilités qu'on appelle des lois impropre c'est-à-dire:

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

l'ensemble des lois *a priori* impropres constituent un prolongement des lois *a priori* propres. En effet, elles permettent une bonne description du manque d'information *a priori*.
Voici quelques approches pour déterminer des lois non informatives:

Lois a priori de Laplace

Laplace fut le premier à utiliser des techniques non informatives puisque, bien que ne disposant pas d'information, il munit ces paramètres d'une loi *a priori* qui prends en compte son ignorance en donnant la même vraisemblance à chaque valeur à chaque valeur du paramètre, soit donc en utilisant une loi uniforme. Son raisonnement, appelé plus tard principe de la raison insuffisante, se fondait sur l'équiprobabilité des événements élémentaires.
Trois critiques on été plus tard avancées sur ce choix. Premièrement, les lois résultantes sont impropres quand l'espace des paramètres n'est pas compact et certains statisticiens se refusent à utiliser de telles lois, car elles mènent à des difficultés comme le paradoxe de marginalisation.
Deuxièmement, le principe des événements équiprobables de Laplace n'est pas cohérent entermes de partitionnement si: $\Theta = \{\theta_1,\theta_2\}$, la règle de Laplace donne $\pi(\theta_1) = \pi(\theta_2) = 1/2$ mais, si la définition de Θ est plus détaillé, avec $\Theta = \theta_1,\theta_2,\theta_3$ la règle de Laplace mène à $\pi(\theta_1) = 1/3$, ce qui évidemment n'est pas cohérent avec la première formulation, cette cohérence n'est pas un problème important: il peut être évacué en argumentant que le niveau de partionnement doit être fixé à un certain stade de l'analyse et que l'introduction d'un degré plus fin dans le partitionnement modifie le problème d'inférence.
La troisième critique est plus fondamentale, car elle concerne le problème de l'invariance par reparamétrisation. Si on passe de $\theta \in \Theta$ à $\eta = g(\theta)$ par une transformation bijective g, l'information *a priori* reste totalement inéxistante et ne devrait pas être modifiée. Cependant, si $\pi(\theta) = 1$, la loi *a priori* sur η est :

$$\pi^*(\eta) = |\frac{d}{d\eta} g^{-1}(\eta)|$$

Loi a priori de Jeffreys

Les lois *a priori* non informatives de Jeffreys (dans le cas unidimensionnel) sont définies par

$$p(\theta) = I^{\frac{1}{2}}(\theta)$$

où $I(\theta)$ est la quantité de l'information de Fischer

$$I(\theta) = \mathbb{E}_\theta\{\frac{\partial \log p(x|\theta)}{\partial \theta}\}^2,$$

ce qui, sous certaines conditions de régularité, est égale à

$$I(\theta) = -\mathbb{E}_\theta\{\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\}.$$

Dans le cas où $\theta \in \mathbb{R}^k$, on définit la matrice de l'information de Fisher qui a pour éléments

$$I_{ij}(\theta) = -\mathbb{E}_\theta\{\frac{\partial^2 \log p(x|\theta)}{\partial \theta_i \partial \theta_j}\}^2, \qquad i,j = 1,...,k,$$

et la loi non informative de Jeffreys est alors

$$p(\theta) = |I(\theta)|^{\frac{1}{2}}.$$

Notons que, si $p(x|\theta)$ définit une famille exponentielle,

$$p(x|\theta) = h(x)\exp[\theta x - \psi(\theta)],$$

on aura $I(\theta) = \nabla\nabla^t\psi(\theta)$, et

$$p(\theta) = (\prod_{i=1}^{k} \frac{\partial^2\psi(\theta)}{\partial\theta^2})^{\frac{1}{2}}.$$

Exemple 1.5.

Soit $x \sim \mathcal{N}(\mu,\sigma^2)$ avec $\theta = (\mu,\sigma^2)$ inconnu. Dans ce cas,

$$I(\theta) = -\mathbb{E}_\theta \begin{pmatrix} 1/\sigma^2 & 2(x-\mu)/\sigma^3 \\ 2(x-\mu)/\sigma^3 & 3(x-\mu)^2/\sigma^4 - 1/\sigma^2 \end{pmatrix} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix},$$

et la loi non informative associée est

$$P(\theta) = \pi(\mu,\sigma^2) \propto 1/\sigma^2.$$

Deux critiques ont étés faite à cette approche :

- Contradiction avec le principe de vraisemblance, puisque l'information de Fisher dépend des facteurs de proportionalité dans la vraisemblance.
- L'extention multidimensionnelle peut parfois conduire à des incohérences (paradoxe de marginalisation de Stein)

Loi a priori de concordance (matching priors)

Le but est de trouver une loi *a priori* cernant le paramètre θ qui se rapproche le plus possible de la méthode de choix fréquentiste, cela revient à faire en sorte que le tirage x n'influence pas le résultat.
On appelle dans un premier temps des exemples d'une région de confiance ou α-crédible. Elle peut être par exemple un intervalle unilatéral $\{\theta \leq \theta_d^{(x)}\}$ ou bien bilatéral $\{\theta_{\alpha,1} \leq \theta \leq \theta_{\alpha,2}\}$. Il peut s'agir aussi de région HPD, $\theta \in \mathcal{C}_\alpha^\pi$ avec par exemple $\{\log(\widehat{\theta}) - \log(\theta) \leq h_\alpha\}$ tel que $P^\pi(\theta \in \mathcal{C}|x) = 1-\alpha$.
On cherche π tel que $\forall\theta$; $P_\theta(\theta \in \mathcal{C}) = 1-\alpha$, appelé la parfaite concordance. C'est en général impossible. On va chercher r_n le plus petit possible tel que :

$$\forall\theta \in \Theta; \forall\alpha \in]0,1[, P_\theta(\theta \in \mathcal{C}) = 1 - \alpha + o(r_n)$$

La loi a priori est alors dite concordante à l'ordre r_n.

Lois a priori invariante impropres

Définition 1.4.4.

Une loi impropre M sur Θ est invariante par transformation sur le paramètre h de Θ dans Θ si M est identique à son image par h définie par Moh^{-1}.

Remarque 1.4.3.

Si M est caractérisée par sa densité π et h est bijective bidérivable, la densité de Moh^{-1} est égale à $\|\rho^{-1}\|Moh^{-1}$, la condition d'invariance s'exprime par l'égalité :

$$\pi = \|\partial h^{-1}\|\pi oh^{-1}$$

où $\|\partial h^{-1}\|$ est la valeur absolue du déterminant de la matrice des dérivées partielles.

Exemple 1.6. *(Berger et yang 1995)*
Dans le modèle AR(1), en prennant la transformation :

$$h : \rho \mapsto h(\rho) = \frac{1}{\rho}, |\rho| > 1$$

et la loi a priori :

$$\begin{cases} 1/(2\pi\sqrt{(1-\rho^2)}) & \text{si } |\rho| < 1; \\ 1/(2\pi|\rho|\sqrt{(\rho^2-1)}) & \text{si } |\rho| > 1. \end{cases}$$

Dans ce cas, la condition d'invariance par la transformation h sur le paramètre ρ s'exprime par l'égalité

$$\pi(h(p)) = \pi(\rho)|\frac{\partial h(\rho)}{\partial \rho}|^{-1}$$

Exemple 1.7. *(Le choix Bayésien 2002)*

La famille de lois $f(x-\theta)$ est invariante par translation, car y = x - x_0 a une loi de la même famille pour tout x_0, $f(y-(\theta-x_0))$, θ est alors dit paramètre de position et une exigence d'invariance est que la loi *a priori* soit invariante par translation, donc satisfasse

$$\pi(\theta) = \pi(\theta - \theta_0)$$

pour tout θ_0. La solution est $\pi(\theta) = c$ la loi uniforme sur Θ.

Loi de référence

Cette téchinique a été mis au point par Bernardo (1979), c'est une modification de l'approche de Jeffreys dans le cas unidimensionnel, appelée approche de la loi de référence.

La différence qui caractérise cette méthode est que la loi *a priori* résultante par la méthode de référence ne dépend pas seulement de la loi d'échantillonnage, mais aussi du problème inférentiel considéré.
Quand $x \sim f(x|\theta)$ et $\theta = (\theta_1,\theta_2)$, où θ_1 est le paramètre d'intérêt, la loi de référence est obtenue en définissant d'abord $\pi(\theta_2|\theta_1)$ comme la loi de Jeffreys associée à $f(x|\theta)$ pour θ_1 fixé, puis en calculant la loi marginale

$$\widetilde{f}(x/\theta_1) = \int f(x/\theta_1,\theta_2)\pi(\theta_2/\theta_1)d\theta_2$$

et la loi de Jeffreys $\pi(\theta_1)$ associée à $\widetilde{f}(x/\theta_1)$.

1.5 Méthodes de calcul en statistique Bayésien

Cette section développe succinctement les principales approches des méthodes numériques Bayésiennes.

1.5.1 Approches indépendantes

Le problème généralement rencontré est celui du calcul dune intégrale de la forme:

$$I_\pi(h) = \int h(x)\pi(\theta)d\theta$$

Par exemple, le calcul de la loi a posteriori :$\pi(\theta|X) = \frac{f(X|\theta_\pi(\theta)}{m(X)}$ exige le calcul de $m(X)$. De même, $\hat{\theta} = \int \theta_\pi(\theta)d\theta$ peut aussi sécrire $I_\pi(id)$.

Dès que la dimension de l'espace d'intégration est supérieure ou égale à 3 le calcul numérique est délicat, c'est pourquoi on a recours à des méthodes de simulation. La première idée se base sur le principe de Monte Carlo, à l origine développé pour les sciences physique. Pour $\theta^t \sim^{iid}$ où $t = 1,.....T$ avec un T grand devant $1,I_\pi(h)$ est approché par :

$$\widehat{I_\pi}(h) = \frac{1}{T}\sum_{t=1}^{T} h(\theta^t)$$

Notons qu'il peut être difficile de simuler sur π. Une deuxième idée est de baser l'estimation sur l'échantillonnage d'importance. Il s'agit dans un premier temps de définir une densité instrumentale q puis par un jeu d'écriture de noter :

$$I_\pi(h) = \int h(\theta)\frac{\pi(\theta)}{q(\theta)}q(\theta)d\theta$$

Le principe repose sur $\theta^t \sim q$ et l'estimation est alors :

$$\widehat{I}_\pi(h) = \frac{1}{T}\sum_{t=1}^{T} \frac{h(\theta^t)}{q(\theta)}\pi(\theta^t)$$

En se qui concerne la variance :

$$V\widehat{I}_\pi(h) = \int \frac{h^2(\theta)\pi^2(\theta)}{q(\theta)} q(\theta)d\theta - \widehat{I}_\pi(h)^2$$

D'après cette formule, si $\frac{\pi^2}{q}$ est grand alors la variance explose et le calcul devient périlleux, il est nécessaire que, dans une certaine mesure, q domine π. Ainsi le choix de q est critique alors qu'il est arbitraire. Le problème se résout si $q \approx \pi$; seulement trouver un tel q n'est pas évident. Et de toute façon ceci marche très mal en grande dimension.

1.5.2 Méthode classique d'approximation

Point historique

Depuis quelques dizaines années les statisticiens font appel presque automatiquement aux méthodes de simulations MCMC (Markov Chain Monte Carlo) pour simuler des distributions multivariées complexes.
Ces méthodes ont initialement été développées par des physiciens pour répondre à des problématiques bien précises.
Ainsi, l'algorithme de Metropolis remonte aux années cinquante avec l'article de [Metropolis et al., 1953] appliqué aux distributions de Boltzmann et il faudra attendre les années soixante-dix pour une première généralisation de la méthode par [Hastings, 1970] (toujours à partir d'un problème physique sur l'énergie d'un système) et son nom actuel d'algorithme de Metropolis-Hastings . Pour comprendre l'engouement de Hastings, reportons nous à cette citation relevée dans [Rosenthal, 2004].
L'algorithme de Gibbs a été développé dans les années quatre-vingt dans l'article séminal de [Geman et Geman, 1984] dans le cas particulier des distributions de Gibbs.
L'appropriation des méthodes MCMC par les statisticiens et la large diffusion de ces méthodes a attendu la montée en puissance des microprocesseurs.
De nos jours une grande variété de méthodes MCMC est à la disposition du statisticien, ne se réduisant pas aux algorithmes évoqués ci-dessus, même si ceux-ci sont encore largement utilisés dans leur forme originale. Ainsi ces algorithmes peuvent se combiner au besoin dans des algorithmes hybrides ou bien être adaptatifs.
En résumé, le monde des MCMC bouillonne d'algorithmes et d'astuces suivant la libre imagination du statisticien forcé de répondre aux contraintes des applications, pourvu que ces algorithmes satisfassent les critères de convergence vers la bonne distribution.

Intégration numérique

A partir de la simple méthode de Simpson, plusieurs approches ont été conçues en mathématiques appliquées pour l'approximation numérique d'intégrales. Par exemple, la quadrature polynômial est censée approcher les intégrales liées à des distributions proches de la loi normale. L'approximation de base est donnée par

$$\int_{-\infty}^{+\infty} e^{-t^2/2} f(t)dt \approx \sum_{i=1}^{n} w_i f(t_i)$$

où

$$w_i = \frac{2^{n-1} n! \sqrt{n}}{n^2 [H_{n-1}(t_i)]^2}$$

et t_i est le i-ième zéro du n-ième polynôme d'Hermite $H_n(t)$

D'autres approximations d'intégrales reliées à la méthode précédente sont disponibles, qui reposent sur différentes bases orthonormales classiques (voir Abramowitz et Stegun, 1964) mais ces méthodes requièrent généralement des hypothèses de régularité sur la fonction f, ainsi que des études préliminaires pour déterminer quelle base est la plus adéquate et à quel point cette approximation est précise. Par exemple, des transformations du modèle peuvent être nécessaires pour mettre en pratique l'approximation d'Hermite (voir Naylor et Smith, 1982 et Hills et Smith 1992).

Remarque 1.5.1.

Quelle que soit la méthode d'intégration numérique utilisée, sa précision diminue dramatiquement lorsque la dimension de Θ augmente. De façon plus spécifique, l'erreur associée aux méthodes numériques se comporte comme une puissance de la dimension de Θ. En pratique, une règle empirique est que la plupart des méthodes standard ne devraient pas être utilisées pour l'intégration en dimension supérieure à 4. En effet, la taille de la partie de l'espace non pertinente pour le calcul d'une intégration donnée augmente considérablement avec la dimension de l'espace. Ce problème est appelé fléau de la dimension.

Les méthodes de Monte carlo

Dans un problème statistique, l'approximation de l'intégrale

$$\int_{\Theta} g(x) f(x|\theta)\pi(\theta)d\theta$$

doit tirer avantage de la nature particulière, à savoir le fait que π soit une densité de probabilité ou plutôt, que $f(x|\theta)\pi(\theta)$ soit proportionnel à une densité. Une conséquence

naturelle de cette perspective est d'utiliser la méthode de Monte Carlo, introduite par Metropolis et Ulam (1949) et von Neumann (1951). Par exemple, s'il est possible de produire des variables aléatoires θ_1,.....θ_m de loi $\pi(\theta)$, la moyenne

$$\frac{1}{m}\sum_{i=1}^{m} g(\theta_i) f(x|\theta_i)$$

converge (presque sûrement) vers (1.11) lorsque m tend vers 1, selon la Loi des Grands Nombres. De la même façon, si un échantillon iid de θ_i de $\pi(\theta|x)$ peut être simulé, la moyenne

$$\frac{1}{m}\sum_{i=1}^{m} g(\theta_i)$$

converge vers

$$\frac{\int_\Theta g(\theta) f(x|\theta)\pi(\theta)d\theta}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}$$

1.5.3 Méthodes de Monte Carlo par Chaîne de Markov (MCMC)

Les méthodes de Monte-Carlo par chaîne de Markov (MCMC) génèrent une suite de variables aléatoires (θ^1,...,θ^n,...) et, hormis la première à laquelle on donne une valeur arbitraire, chacune d'entres elles dépend uniquement de celle qui la précède, les calculs sont ensuite poursuivis en appliquant à cette séquence une loi des grands nombres pour les chaînes markoviennes ergodiques de forme identique.

L'algorithme de Gibbs

L'algorithme de Gibbs est central en statistique Bayésienne car il permet de réduire un problème complexe de simulation, typiquement la simulation selon la distribution jointe a posteriori des paramètres, en une suite d'étapes simples à simuler. Pour cette raison, l'algorithme de Gibbs est aussi connu sous le nom d'échantillonneur de Gibbs (en anglais, Gibbs sampler). D'un point de vue historique, l'algorithme de Gibbs tire son nom d'un physicien et mathématicien américain du $19^{éme}$ siècle Josiah Willard Gibbs, considéré comme l'un des fondateurs de la thermodynamique moderne et de la mécanique statistique. Josiah Willard Gibbs a donné son nom aux distributions de Gibbs utilisées en traitement d'image. L'association entre le nom et l'algorithme a été réalisée dans l'article de [Geman et Geman, 1984] où cette méthode a été développée pour l'étude Bayésienne des champs de Gibbs en traitraitement d'image, et a perduré depuis.
L'algorithme de Gibbs, sous sa formulation générale, s'écrit
Pour $\theta = (\theta_1,.....,\theta_p)$, on veut simuler $\pi(\theta)$ à partir de $\pi_i(\theta_i|\theta_{(-i)}) = \pi_i(\theta_i|\theta_j, j \neq i)$ pour

tout i. On initialise avec $\theta^{(0)}$ et à l'instant t, on écrit:

$$(\theta_1^{(t)}|\theta^{(t-1)}) \sim \pi_1(\theta_1^{(t)}|\theta_{(-1)}^{(t-1)})$$

$$(\theta_2^{(t)}|\theta^{(t-1)},\theta_1^{(t)}) \sim \pi_2(\theta_2^{(t)}|\theta_3^{(t-1)},.....,\theta_p^{(t-1)})$$

$$(\theta_p^{(t)}|\theta^{(t-1)},\theta_{-p}^{(t)}) \sim \pi_p(\theta_p^{(t)}|\theta_{-p}^{(t)})$$

Dans le cas où une telle loi π existe, $\theta^{(t)}$ issu de cet algorithme est une chaîne de Markov ergodique de loi stationnaire π.

Algorithme Metropolis-Hasting

L'algorithme de Metropolis-Hastings peut être vu comme une alternative à l'algorithme de Gibbs dans le cas où on ne peut pas simuler facilement dans les distributions conditionnelles a posteriori , par exemple lorsque les distributions ne sont pas conjuguées ou lorsque les expressions ne peuvent pas s'exprimer sous forme analytique.
L'algorithme de Metropolis-Hastings est un algorithme d'acceptation/rejet. L'idée de l'algorithme est de simuler selon une autre distribution, plus simple à simuler, appelée la loi de proposition, et d'accepter la valeur simulée avec une certaine probabilité d'être effectivement un tirage selon la loi cible.
Sous les conditions de convergence, les valeurs successivement acceptées de la loi de proposition forment une chaine de Markov convergeant vers la distribution cible.
La différence avec un algorithme acceptation/rejet classique est que tant qu'une nouvelle valeur n'est pas acceptée l'algorithme retourne la dernière valeur acceptée comme nouvelle valeur de la chaîne.
Algorithme Metropolis-Hasting, sous sa formulation générale, s'écrit
Pour $\theta^{(0)}$ est une valeur initiale, on définit par récurrence les valeurs de $\theta^{(t)}$. A l'étape t, à partir de $\theta^{(t-1)}$, $\theta^{(t)}$ est construit en tirant un $\theta^{'}$ à l'aide d'une distribution de probabilité instrumentale : $\theta^{'} \sim q(.|\theta^{(t-1)})$.$\theta^{(t)}$ est alors donné par :

$$\theta^{(t)} = \begin{cases} \theta^{'} & \text{avec une probabilité} \quad \alpha(\theta^{'},\theta^{(t-1)}) \\ \theta^{(t-1)} & \text{avec une probabilité} \quad 1-\alpha(\theta^{'},\theta^{(t-1)}) \end{cases}$$

Où $\alpha(\theta^{'},\theta^{(t-1)}) = \min\left(\frac{\pi(\theta^{'})q(\theta^{(t-1)}\theta^{'}}{\pi(\theta^{(t-1)}q(\theta^{'}|\theta^{(t-1)})},1\right)$.

La loi de densité $\pi(\theta)$ est souvent appelée *loi cible* ou loi objet, tandis que la loi de densité $q(.|\theta)$ est dite *loi de proposition*. Une propriété stupéfiante de cet algorithme est d'autoriser un nombre infini de lois de proposition produisant toute une chaîne de Markov

convergeant vers la loi d'intérêt.

Remarque 1.5.2.

Il est possible suivant cette construction de rester au même endroit après une itération. On peut alors montrer en écrivant la condition de balance, que pour ce choix de α , on obtient une chaîne de Markov de loi stationnaire π. Cette chaîne de Markov est ergodique si et seulement si $(\theta^{(t)})_t$ est irréductible et apériodique.

1.6 Avantages et Inconvénients de l'approche Bayésienne

1.6.1 Avantages

Le premier point fort de la méthode Bayésienne est son élégance philosophique : une probabilité est une mesure d'incertitude subjective et ceci ne préjuge pas de l'existence éventuelle d'un hasard fondamental du monde physique. Il n'est pas nécessaire de supposer l'existence d'expériences aléatoires répétables pour réaliser des inférences inductives.

De plus, les règles de manipulation des probabilités fournissent une méthode automatique, cohérente, exempte de paradoxes et qui ne nécessite pas d'introduction d'autres principes ad-hoc. Les mêmes règles s'appliquent que nous possédions beaucoup ou peu de données et la loi de Bayes fournit une méthode élégante pour combiner nos connaissances a priori avec les informations provenant de ces données.

La règle de Bayes permet aussi des approches incrémentales (l'a posteriori à l'issue d'une expérience peut servir d'a priori pour la prochaine expérience) et hiérarchiques. Les mo- dèles hiérarchiques, considérant les paramètres comme des variables aléatoires, permettent de partager de l'information entre différents sous-modèles. Cette hiérarchisation, qui peut s'étendre sur plusieurs niveaux, est très utilisée en sciences sociales pour modéliser des populations à différentes échelles. Par exemple, pour inférer la taille moyenne des hommes de différentes villes, il est intéressant d'introduire un paramètre de taille pour chaque ville et de les lier par un hyper-paramètre au niveau du pays. Ainsi l'information récoltée en différents lieux peut servir à estimer la taille des hommes d'une ville pour laquelle nous n'avons que peu de données.

Plus techniquement, la règle de la somme permet de traiter aisément les paramètres

de nuisance, en les marginalisant. Il n'existe pas de méthode équivalente pleinement satisfaisante avec les statistiques classiques (Lor99). Lors de la comparaison de modèles, cette marginalisation introduit aussi une pénalité pour les modèles les plus complexes. Cette régularisation est appelée rasoir d'Occam automatique.

1.6.2 Inconvénients

Une des critiques les plus courantes à l'encontre de la méthode Bayésienne est justement sa subjectivité; le but de la science étant d'obtenir des résultats indépendants du scientifique. Nous pensons que cette subjectivité a néanmoins l'avantage d'être explicite, obligeant l'utilisateur à établir clairement quels sont ses a priori. Une fois ces connaissances exprimées, l'application des règles d'inférence est automatique. En revanche, l'approche bayésienne est difficile à suivre d'une façon parfaitement rigoureuse pour deux raisons.

D'abord les connaissances a priori d'un agent sont souvent floues, mal formulées et leur traduction en un modèle probabiliste numérique est très difficile. Les hypothèses posées sont souvent fausses, il est nécessaire daffiner les modèles après avoir vu leurs conséquences. Ce problème est spécialement critique lorsqu'il faut formuler des a priori en grande dimension, car l'intuition y est souvent mise en défaut. C'est pourquoi il est recommandé de mettre en place un cycle modélisation-inférence-évaluation permettant d'améliorer progressivement les modèles.

Il existe aussi le problème dit du "monde fermé" (close world asumption). Si le bon modèle n'a pas été considéré dès le départ, les inférences ne l'inventeront pas. Pour être parfaitement rigoureux, il faudrait considérer tous les modèles possibles avant de voir les données, afin d'être certain de ne pas en avoir oublié. Car oublier de considérer un modèle est équivalent à lui assigner une probabilité a priori nulle et donc une probabilité a posteriori tout aussi nulle, quelle que soit sa vraisemblance.

Le troisième grand problème de l'approche Bayésienne est la difficulté calculatoire des inférences. Il est très tentant de définir un beau modèle génératif pour un problème, mais si il n'existe pas d'algorithme efficace suffisamment précis, ce modèle n'a pas beaucoup de valeur. Tout l'art est de trouver un compromis entre biais du modèle et faisabilité des inférences. Pour des problèmes difficiles, il faudrait en fait penser à la méthode d'inférence dès la phase de modélisation.

1.7 Conclusion

Clairement, l'approche Bayésienne apporte une plus grande souplesse dans la méthodologie statistique de l'analyse des données. D'une part les procédures bayésiennes standard, qui peuvent maintenant être mises en oeuvre aussi facilement que les tests traditionnels, ont le statut privilégié d'objectivité nécessaire à la communication scientifique. D'autre part differentes distributions *a priori*, exprimant des résultats antérieurs ou des opinions d'experts, favorables ou défavorables à la conclusion recherchée, peuvent être utilisées pour éprouver la "robustesse" des conclusions et prendre ainsi des décisions "personnelles".

Chapitre 2

Méthode Bayésienne variationnelle

2.1 Introduction

Une des difficultés de l'approche numérique dans le cadre Bayésien est le temps nécessaire à l'obtention d'échantillons selon la loi a posteriori . De plus, ce temps augmente rapidement avec la dimension de l'espace, ce qui rend cette solution non utilisable en grande dimension. Une méthode d'approximation consiste à trouver une loi analytique approchant la loi a posteriori et à l'utiliser pour obtenir les estimateurs.

Récemment, une approximation déterministe de la loi a posteriori , appelée approximation Bayésienne variationnelle (apprentissage dans un ensemble) a été introduite. L'idée est d'approcher la loi jointe par une loi séparable avec une forme libre. La forme de cette loi approchante est obtenue en minimisant la distance de Kullback-Leibler. Le choix de séparation doit rendre ce calcul plus facile car une des dificultés principales réside dans la dépendance a posteriori entre les paramètres recherchés.

Le principe de cette approximation vient de la physique statistique. Cette méthode a d'abord été introduite dans le domaine de l'inférence Bayésienne pour des applications en réseaux de neurones , puis pour l'apprentissage des modèles graphiques et l'estimation des paramètres des modèles . Son apparition dans le domaine des problèmes inverses est relativement récente avec une première application en restauration d'image , puis dans les problèmes de séparation de sources et, et dans les problèmes d'imagerie hyperspectrale.

L'expression méthode variationnelle vient du calcul des variations : il s'agit d'exprimer comment la valeur de la fonctionnelle se modifie en réponse à d'infimes changements de la fonction d'entrée de la fonctionnelle . Elle fait référence à différents outils mathématiques pour la formulation de problèmes d'optimisation, aussi bien qu'aux techniques associées à leur résolution. L'idée générale est d'exprimer la quantité d'intérêt comme solution d'un problème d'optimisation. Ce problème d'optimisation peut être modifié dans différentes

directions, soit en approchant la fonctionnelle à optimiser, soit en approchant l'ensemble sur lequel est optimisé la fonctionnelle. De telles approximations, à leur tour, donnent un moyen d'approcher la quantité d'intérêt initiale : c'est l'approximation variationnelle .

2.2 Généralités

Dans une approche Bayésienne , on établit une distribution a posteriori des inconnues x sachant les données y et le modèle M, en utilisant la règle de Bayes

$$p(x|\theta,y;M) = \frac{p(y|x,\theta_1;M)p(x|\theta_2 M)}{p(y|\theta;M)}$$

où

- $p(x|\theta,y;M)$ est la loi a posteriori des inconnues x
- $p(y|x,\theta_1;M)$ la vraisemblance des inconnues x et θ du modèle, où θ_1 représente l'ensemble des paramètres qui décrivent cette fonction.
- $p(x|\theta_2;M)$ la loi a priori pour les inconnues x, où θ_2 représente ses paramètres.
- $p(y|M)$ l'évidence du modèle M.

où on suppose implicitement connaître l'ensemble des paramètres $\theta = (\theta_1,\theta_2)$. Mais, dans un cas réel, nous sommes amenés souvent à les estimer aussi. Pour cela, dans l'approche Bayésienne, on leur attribue aussi une loi a priori $p(\theta|M)$, et l'on obtient alors une loi a posteriori conjointe des inconnues x et des hyperparamètres $\theta = (\theta_1,\theta_2)$:

$$\begin{aligned} p(x,\theta|y;M) &= \frac{p(y,x,\theta|M)}{p(y|M)} \\ &= \frac{p(y|x,\theta_1;M)p(x|\theta_2;M)p(\theta|M)}{p(y|M)} \end{aligned} \tag{2.1}$$

Dans cette relation, le dénominateur est la vraisemblance marginale du modèle M dont son logarithme $\ln p(y|M)$ est appellé *evidence du modèle* M s'écrit:

$$p(y|M) = \int\int p(y|x,\theta;M)p(x|\theta;M)p(\theta|M)dxd\theta$$

Afin d'introduire les notions qui vont être utilisées dans la suite de ce travail, il est intéressant de mentionner que, pour n'importe quelle loi de probabilité $p(x,\theta|y;M) = q(x,\theta)$ l'évidence du modèle vérifie:

$$\begin{aligned} \ln p(y|M) &= \ln \int\int p(y|x,\theta;M)p(x|\theta;M)p(\theta|M)dxd\theta \\ &= \ln \int\int q(x,\theta)\frac{p(y,x,\theta|M)}{q(x,\theta)}dxd\theta \\ &\geq \int\int q(x,\theta)\ln\frac{p(y,x,\theta|M)}{q(x,\theta)}dxd\theta \end{aligned}$$

(d'aprés l'inégalité de Jensen: $\ln(Ep|q) \geq E(\ln(p|q))$)

2.3 Principe de l'approche variationnelle

Nous présentons le principe de l'approche variationnelle dans le cadre du problème linéaire, mais le même principe s'applique au cas bilinéaire.

On cherche à approcher la loi *a posteroiri* $p(x,\theta|y;M)$ par une loi plus simple (séparable) $q(x,\theta)$ qui facilite le calcul des estimateurs. on utilise la théorie de l'information pour chercher la forme de q qui minimise la divergence d'information, dite la divergence de Kullback-Leibler $KL(q \parallel p)$ entre q et p

$$KL(q \parallel p) = \int q(u) \ln \frac{q(u)}{p(u|y;M)} du$$

Où $u = (x,\theta)$ regroupe les inconnues.
La divergence de Kullback a les propriétés suivantes :

1. elle est toujours positive $KL(q \parallel p) \geq 0, \forall p,q$, et elle s'annule quand les deux distributions sont identiques;
2. elle est non symétrique $KL(q \parallel p) \neq KL(p \parallel q)$;
3. $KL(q \parallel p) = \infty$ si sur un ensemble d'une mesure positive $q(x,\theta) > 0 \ \ et \ \ p(x,\theta) = 0$;
4. elle est convexe, $KL(\alpha q_1 + (1-\alpha)q_2 \parallel p) \leq \alpha KL(q_1 \parallel p) + (1-\alpha)KL(q_2 \parallel p), \forall \alpha$.

En développant l'expression de la divergence, on trouve:

$$\begin{aligned} KL(q(u) \parallel p(u|y;M) &= \int q(u) \ln \frac{q(u)}{q(u|y;M)} du \\ &= \int q(u) \ln(\frac{q(u)p(y|M)}{p(u,y|M}) du \\ &= \ln(p(y|M)) = -\int q(u) \ln \frac{p(u,y|M)}{q(u)} du \\ &= \ln(p(y|M)) - f(q) \end{aligned}$$

Où $f(q) = \int q(u) \ln(\frac{p(u,y|M)}{q(u)}) du$ est l'énergie libre.

Par la suite, nous allons écrire l'expression de $f(q)$ par

$$f(q) = < \ln p(y,u|M) >_q + H(q)$$

avec $H(q)$ l'entropie de la loi approchante :

$$H(q) = -\int q(u) \ln(q(u)) du$$

On peut faire deux remarques sur la dernière équation :

1. comme la divergence de Kullback-Leiblerest positive,l'énergie libre donne une borne inférieure pour la log-évidence du modèle, $\ln(p(y|M)) \geq f(q)$.

2. on peut chercher la forme de la loi approchante q en maximisant l'énergie libre au lieu de la divergence de Kullback-Leibler, puisque la log-évidence ne dépend pas de la loi approchante.

Supposons que la loi approchante s'écrive sous la forme séparable suivante :

$$q(u) = \Pi_i q_i(u_i)$$

Ce problème d'optimisation fonctionnelle admet une solution sous la forme

$$q_i(u_i) = \frac{1}{K_i} \exp(< \log p(y,u|M) >_{\prod_{j \neq i} q_j(u_j)}) \tag{2.2}$$

où k_i est une constante de normalisation, $\forall i \in 1...N$.

On remarque clairement que cette solution est analytique mais malheureusement qu'elle n'a pas une forme explicite.
Pour obtenir l'optimum de manière pratique, il faut mettre en oeuvre un algorithme itératif de point fixe :

$$q_i^{k+1}(u_i) = \frac{1}{K_i} \exp(< \log p(y,u|M) >_{\prod_{j \neq i} q_j^k(u_j)})$$

La procédure d'obtention de la loi approchante correspond donc à une optimisation alternée par groupe de coordonnées, résumée dans l'algorithme ci-dessous :

Algorithm 1 Algorithme bayésien variationnel classique

Inisialisation (q^0)

repeat

for$i \in 1.....N$**do**

function ESTIMER $q_i^{k+1}(u_i)$ (connaissant $q_j^{k+1}(u_j)$ pour j<i et $q_l^k(u_l)$ pour l>i)

Calcul de $q_i^{k+1}(u_i)$ en utilisant Eq (1.2)

end function

end for

Estimer l'énergie libre $f(q^k)$ **until** Convergence

Pour que la loi approchante soit exploitable en pratique, il est important que le calcul de $< \log p(y,u|M) >_{\prod_{j\neq i} q_j(u_j)}$ donne une expression paramétrique. En utilisant la famille exponentielle conjuguée, la loi approchante aura une forme paramétrique et les paramètres de forme seront mutuellement dépendants. Pour retrouver leurs valeurs finales, ces paramètres peuvent être calculés de façon itérative. le problème d'optimisation fonctionnelle se transforme en un problème de calcul paramétrique itératif où on répète le calcul de l'équation (2.2) pour toutes les composantes de séparation en utilisant les valeurs calculées à l'itération précédente.

Pour estimer la convergence de l'algorithme, on peut évaluer l'énergie libre $f(q)$. À la convergence, on peut utiliser la valeur de l'énergie libre pour avoir une estimation de l'évidence du modèle sous l'hypothèse que la divergence de Kullback soit négligeable, ce qui permet de comparer la qualité entre différentes méthodes d'approximation. De plus, cette valeur permet de choisir entre différents modèles puisque l'évidence mesure l'adéquation aux données. Cela est valable si l'on considère que la divergence de Kullback ne change pas lorsque l'on change de modèle. Néanmoins, cette condition est difficile à vérifier spécialement pour les modèles complexes. L'énergie libre peut être obtenue à partir des paramètres de forme déjà calculés sans coût supplémentaire.

On résume l'approche Bayésienne variationnelle par deux étapes principales :

1. le choix de la séparation dans la loi approchante $q(u) = \Pi_i q_i(u_i)$
2. le calcul des lois approchantes et plus précisément le calcul de $< \log p(y,u|M) >_{\prod_{j\neq i} q_j(u_j)}$ pour chaque loi approchante.

2.3.1 Choix de séparation

La première étape dans l'approche Bayésienne variationnelle consiste à choisir la forme de séparation dans la loi approchante. Il n'y a pas de règle pour ce choix et chaque problème peut avoir sa propre forme de séparation. On essaie généralement de garder les liens forts dans la loi approchante et de négliger les faibles. Il existe deux solutions extrêmes :

1. la première consiste à prendre une loi approchante sans séparation. Ceci n'est pas utile puisque l'on retombe sur la loi a posteriori.
2. la seconde est de choisir une séparation forte. Cela permet de faciliter le calcul des lois approchantes mais élimine la dépendance a posteriori entre les variables et la dépendance se limite aux moments seulement.

2.4 Gradient exponentiel pour le Bayésien variationnel

Le but du gradient exponentiel pour le Bayésien variationnel est de mettre à jour toutes les lois approchantes simultanément. Pour cela, il s'inspire des approches de type gradient à pas optimal, bien connues dans la résolution des problèmes d'optimisation en dimension finie. Toutefois, dans le cas qui nous intéresse ici nous voulons trouver une loi approchante conjointe permettant de minimiser la divergence de Kullback-Leibler.
Nous commencerons par introduire la fonctionnelle à optimiser, puis nous définirons le type de différentielle utilisée ainsi que l'espace fonctionnel dans lequel nous nous placerons. Nous exposerons ensuite la direction de descente utilisée ainsi que la définition du pas de descente.

2.5 Propriétés

2.5.1 Sélection des modèles en utilisant VB

Une propriété particulièrement intéressante de l'approche Bayesien variationnel est la possibilité de sélectionner les modèles durant l'entrainement. L'énergie libre peut être utilisée comme critère de sélection des modèles parce que la distance KL entre les distributions des paramètres a priori et a posteriori agit comme une pénalité comme BIC. Considérons maintenant ce problème de façon plus précise. Soit la densité a posteriori $q(m)$ pour un modèle m. On peut montrer que la densité optimale q(m) peut être écrite :

$$q(m) \propto exp f(\Theta,X,m)p(m)$$

où p(m) est la densité a priori du modèle. En l'absence de toute information a priori sur le modèle, p(m) est uniforme et la densité optimale q(m) dépendra simplement du facteur $F(\Theta,X,m)$ c'est à dire que l'énergie libre peut être utilisée comme critère de sélection. Un avantage important est qu'aucun seuil ne doit être choisi manuellement . Pour le modèle considéré ici, il est possible d'obtenir une forme explicite de l'énergie libre .
Un autre point intéressant dans l'utilisation de l'apprentissage Bayesien variationnelle est la capacité de réduire les degrés de liberté excédentaires. Cela signifie qu'il est possible d'initialiser le système avec un grand nombre de groupes et un grand nombre de gaussiennes par locuteur et de laisser le système éliminer groupes et gaussiennes non utilisés.

2.5.2 Probabilité prédictive

Les méthodes variationnelles Bayesiennes peuvent aussi être utilisées pour faire de la prédiction sur des variables inobservées. Supposons en effet que nous souhaitions prédire des données non-observées xà partir de données observées y . Théoriquement nous devrions utiliser :

$$p(x|y) = \int p(\theta|y)p(x|\theta)d\theta \tag{2.3}$$

Cependant, selon le modèle, l'équation (2.3) n'est pas toujours explicitable et une fois encore, une approximation est nécessaire. Tout d'abord, les distributions a posteriori sur les paramètres θ peuvent être approximées par la distribution variationnelle a posteriori c'est à dire $p(\theta|S) \approx q(\theta|S)$. Cette hypothèse résoud le problème relatif à la distribution des paramètres mais laisse l'intégrale toujours impraticable. Dans ce cas (2.3) peut être approximée à nouveau par l'inégalité de Jensen comme dans l'expression de l'energie libre.

Une autre solution très simple pour calculer la probabilité des données non-observées est d'utiliser des moments de distributions variationnelles comme paramètres du modèle; même si c'est une approximation très grossière, les résultats sont cependant comparables avec ceux obtenus par l'estimation des paramètres MAP .

2.6 Avantages et limites de l'approche variationnelle

2.6.1 Avantages

• La solution calculée est une loi de probabilité, ce qui permet d'en déduire immédiatement tous les estimateurs habituels (maximum a posteriori , moyenne a posteriori , médiane a posteriori).

•On peut aussi obtenir des variances a posteriori très utile, pour l'exploitation des résultats.

•l'approche Bayésienne variationnelle permet d'avoir une approximation analytique de la loi a posteriori . Ceci transforme le calcul de l'estimateur en un problème de calcul paramétrique itératif, ce qui réduit le temps de calcul d'une manière importante par rapport aux méthodes fondées sur l'échantillonnage.

• l'approche Bayésienne variationnelle permet d'obtenir une estimation de l'évidence du modèle via l'énergie libre sans coût de calcul important. Ceci est intéressant pour répondre au problème de sélection de modèles (comme par exemple, le choix du nombre de

classes).

• On peut aussi obtenir numériquement, la vraisemblance du modèle. C'est à dire la probabilité des données sahant toutes les hypothèses ayant conduit à la résolution du problème inverse. À l'aide de cette vraisemblance on peut choisir entre différents modèles a priori ou différents modèles physiques, par exemple on peut déterminer le nombre de sources dans le cas d'un problème de séparation de sources.

• Cette approche permet aussi de faire très facilement un bon compromis entre approximation et coût de calcul. En effet, on peut choisir la granularité de la séparabilité des lois approchantes. Plus la loi est séparable plus le nombre de paramètres à estimer est faible, donc plus l'approche itérative est efficace.

• Elle se révèle bien adaptée à des problèmes de grande dimension.

• Un autre avantage important que l'on peut mentionner est l'existance d'un critère (l'énergie libre) que l'on peut utiliser : comme critère d'arrêt pour l'algorithme et comme un critère de choix de modèle. En effet, comme nous l'avons vu, grâce à la relation 9, minimiser KL(q : p) est équivalent à maximiser F(q) et sa valeur optimale est un bon indicateur pour ln p(y|M). Ainsi, les valeurs relatives de F(q) au cours des itérations de l'algorithme peuvent être utilisées comme un critère d'arrêt et sa valeur op- timale atteint pour un modèle M_1 peut être comparée à sa valeur optimale atteint pour un autre modèle M_2 comme un critère de préférence entre les deux modèles.

2.6.2 Limites

• La difficulté principale liée à cette approche est la nécessite d'avoir des formes conjuguées des lois pour pouvoir exploiter les lois approchantes. Pour contourner ce problème, une approche sous-optimale consiste à chercher une loi approchante avec une forme paramétrique fixée minimisant la divergence de Kullback-Leibler. Cependant, cette approche est considérée comme une approche locale qui risque de se bloquer dans des maxima.

• la complexité de calculs lorsque que le nombre de fonctions de base devient élevé (> 10 000).

• l'approche variationnelle est locale; il faut partir d'une bonne solution de d'epart.
• le conditionnement et la dimension des variables de contrôle joue un rôle essentiel.

2.7 Conclusion

Nous avons présenté dans ce chapitre l'approche Bayésienne variationnelle qui permet d'approximer analytiquement la loi a posteriori par une loi séparable, ce qui a permis d'obtenir un estimateur simple de l'objet inconnu et les autres paramètres du modèle.Cette approche bayésienne est très prometteuse car elle permet d'étendre les approches non supervisées exploitant de la parcimonie aux problèmes de très grande dimension.

Chapitre 3

Approximation variationnelle et application en neuroimagerie

3.1 Introduction

Les applications de neuroimagerie utilisent des modèles Bayésiens, dont les grandeurs d'intérêt (évidence, lois a posteriori) posent des problèmes de calcul : ceci peut être dû à la forme trop complexe des lois a posteriori. Il est alors nécessaire d'avoir recours à des méthodes permettant d'approcher ces grandeurs. La méthode variationnelle en est un exemple, relevant du champ de l'approximation déterministe, et son utilisation s'est récemment répandue dans la communauté de neuroimagerie. Le succès de l'approximation variationnelle est en effet dû à sa facilité d'utilisation et sa rapidité d'exécution dans des cas d'estimation qu'il peut être difficile de traiter avec les outils classiques (méthodes de Monte Carlo par Chaîne de Markov (MCMC) par exemple). Au lieu de calculer la grandeur exacte d'intérêt, elle maximise une fonctionnelle la minorant, obtenant ainsi une valeur approchée minorant la grandeur exacte à déterminer.

3.2 l'approximation variationnelle

Un modèle statistique Bayésien est composé d'un modèle statistique paramétrique $p(y|\theta)$ et d'une loi *a priori* pour le paramètre $\pi(\theta)$ modélisant son incertitude. Rappelons que Le théorème de Bayes est donné par :

$$p(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{p(y)} \tag{3.1}$$

Si la variable cachée x est présente , la formule (1) est donnée par:

$$p(x,\theta|y) = \frac{p(y|x,\theta)p(x|\theta)\pi(\theta)}{p(y)} \tag{3.2}$$

les lois marginales *a posteriori* sont données par:

$p(\theta|y) = \int p(x,\theta|y)dx \qquad$ et $\qquad p(x|y) = \int p(x,\theta|y)d\theta$

L'utilisation des méthodes de monte Carlo par Chaine de Markov (MCMC) pour calculer ces lois marginales n'est pas toujours simple et peut s'avérer impossible, en particulier si la structure cachée est de grande dimension. Même chose pour le calcul de la vraisemblance marginale(ou l'évidence)

$$p(y) = \int p(x,\theta,y)dxd\theta = \int p(y|x,\theta)p(x|\theta)\pi(\theta)dxd\theta \tag{3.3}$$

Dans ce cas le choix Bayésien est basé sur la loi *a posteriori* des modèles définit par:

$$p(m|y) \propto p(m)p(y|m)$$

avec:

- $p(m)$: les probabilités des différents modèles pour m.

- $p(y|m)$: la vraisemblance marginale qu'on calcule comme suit:

$$p(y|m) = \int p(y|x,\theta,m)p(x|\theta,m)\pi(\theta,m)dxd\theta \tag{3.4}$$

Pour se sortir on utilise la méthode variationnelle qui permet de Transformer le calcul de l'intégrale (1.4) en résolution d'un problème d'optimisation, en remarquant que la vraisemblance est un majorant d'une quantité appelé *énergie libre*

$$f(q_{x,\theta}) = \int q_{x,\theta}(x,\theta)\log\frac{p(x,y,\theta)}{q_{x,\theta}(x,\theta)}d\theta dx \tag{3.5}$$

avec $q_{x,\theta}$: fonction d'une distribution libre.
l'inégalité de Jensen permet d'écrire:

$$\log p(y) \geqslant f(q_{x,\theta}) \qquad avec \qquad q_{x,\theta} = p(x,\theta|y)$$

si $q_{x,\theta}$ n'est pas limité alors $p(x,\theta|y)$ est la distribution libre qui maximise $f(q_{x,\theta})$ et on écrit:

$$p(x,\theta|x) = \arg\max_{q_{x,\theta}} f(q_{x,\theta})$$

la valeur de l'énergie libre au maximum est alors:

$$\log p(y) = \max_{q_{x,\theta}} f(q_{x,\theta}) = f(p(x,\theta|y))$$

Les méthodes d'approximation variationnelle nous permettre de chercher la solution d'un problème approché, en modifiant la fonctionnelle à optimiser, ou en approchant l'ensemble des distributions libres sur lequel est optimisé la fonctionnelle : dans ce dernier cas, on doit chercher une forme approchée $q_{x,\theta}(x,\theta)$ de $p(x,\theta|y)$ dans un ensemble de fonctions dans lequel les calculs sont aisés, et d'en déduire une approximation de la log-évidence comme le majorant de l'énergie libre sur cette ensemble de fonctions.

$$\log p(y) \geq \max_{q_{x,\theta}} f(q_{x,\theta})$$

$$\begin{aligned}\log p(y) &= \int q_{x,\theta}(x,\theta) \log \frac{p(x,y,\theta)}{q_{x,\theta}(x,\theta)} d\theta dx + \int q_{x,\theta}(x,\theta) \log \frac{q_{x,\theta}(x,\theta)}{p(\theta,x|y)} d\theta dx \\ &= f(q_{x,\theta}) + D(q_{x,\theta} \parallel p(\theta,x|y))\end{aligned}$$

on a :

$$\begin{aligned}f(q_{x,\theta}) &= \int q_{x,\theta}(x,\theta) \log \frac{p(y,x,\theta)}{q_{x,\theta}(x,\theta)} d\theta dx \\ &= E_{q_{x,\theta}}[\log \frac{p(y,x,\theta)}{q_{x,\theta}(x,\theta)}]\end{aligned}$$

et

$$KLD(q_{x,\theta} \parallel p(x,\theta|y) = -\int q_{x,\theta}(x,\theta) \log \frac{p(x,\theta|y)}{q_{x,\theta}(x,\theta)} d\theta dx$$

$$= E_{q_{x,\theta}}[\log \frac{q_{x,\theta}(x,\theta)}{p(x,\theta|y)}]$$

l'erreur d'approximation entre la log-évidence et l'énergie libre s'écrit alors :

$$KLD(q_{x,\theta} \parallel p(\theta,x|y))$$

cette quantité appelée la divergence de Kullback entre la distribution libre et la loi jointe *a posteriori*.

Remarque

La maximisation de l'énergie libre est équivalente à la minimisation de la divergence de Kullback et on écrit :

Maximisé $f(q_{x,\theta})$ =: Minimisé KLDivergence

donc: $f(q_{x,\theta})$ optimisé quand $q_{x,\theta} = p(x,\theta|y)$.

Pour une meilleure approximation de $\log p(y)$ on présente deux types d'approximation:

1. Approximation en champ moyen (factorisation)
2. Approximation de Laplace

3.2.1 Approximation en champ moyen (factorisation)

Cette méthode permet de rechercher une distribution libre factorisée, par exemple en séparant les variables cachées des paramètres et on écrit :

$$q_{x,\theta} = q_x(x)q_\theta(\theta) \tag{3.6}$$

On a deux cas de figure:

- Cas ou les x sont indépendantes conditionnellement à θ, l'approximation optimale de q_x se factorise simplement,

$$q_{x,\theta}(x,\theta) = q_{x_1}(x_1)....q_{x_n}(x_n)q_\theta(\theta)$$

- Cas ou les x sont dépendantes: il est possible de restreindre l'epace des distributions libres a celles se factorisant suivant les variable cahées et imposer :

$$q_x(x) = q_{x_1}(x_1)q_{x_2}(x2).....q_{x_n}(x_n) = \prod_i q_{x_i}(x_i)$$

Algorithme Bayésien variationnel (VBEM):

Cette algorithme maximise d'une façon répétitive l'énergie libre $f(q_x,q_\theta)$ par rapport aux distributions libres q_x et q_θ respectivement en deux étapes

- Etape **VBE**: estimer la loi approchée des variables cachées
- Etape **VBM**: maximisation pour obtenir la loi *a posteriori* des paramètres

Le théorème général suivant fournit le cadre général des équations de mise à jour pour l'apprentissage Bayésien variationnel $(VBEM)$:

Théorème 3.2.1. soit un modèle de paramètre θ, dont on observe un n-échantillon i.i.d, $y = y_1,....,y_n$, avec des variables cachées correspondantes $x = x_1,....,x_n$. une borne inférieure de la vraisemblance marginal est:

$$f(q_x,q_\theta) = \int q_x(x)q_\theta(\theta)\log\frac{p(x,y,\theta)}{q_x(x)q_\theta(\theta)}d\theta dx$$

qui peut être optimisée itérativement en effectuant les mises à jour suivantes, l'indice (t) indiquant le numéro d'itération:

étape VBE : $q_{x_i}^{(t+1)}(x_i) = \frac{1}{Z_i^{(t+1)}}\exp\left[\int q_\theta^{(t)}(\theta)\log p(x_i,y_i|\theta)d\theta\right]\forall i$

avec:

$$q_x^{(t+1)}(x) = \prod_{i=1}^{n} q_{x_i}^{(t+1)}(x_i)$$

étape VBM : $q_\theta^{(t+1)}(\theta) = \frac{1}{Z_\theta^{(t+1)}}p(\theta)\exp\left[\int q_x^{(t+1)}(x)\log p(x,y|\theta)dx\right]$

où

$$\bullet Z_i^{(t+1)} = \int \exp\left[\int q_\theta^t(\theta)\log p(x_i,y_i|\theta)d\theta\right]dx_i$$

$$\bullet Z_\theta^{(t+1)} = \int p(\theta)\exp\left[\int q_x^{t+1}(x)\log p(x,y|\theta)dx\right]d\theta$$

avec $Z_i^{(t+1)}$ et $Z_\theta^{(t+1)}$ sont des constantes de normalisation. De plus, l'algorithme converge vers un maximum local de $f(q_x,q_\theta)$.

Il y a bien symétrie de l'écriture des deux étapes, qui s'observe en réécrivant $q_x^{(t+1)} et q_\theta^{(t+1)}$ sous la forme suivante:

$$q_x^{(t+1)}(x) = \frac{1}{\tilde{Z_x^{t+1}}}\exp\left[E_{q_\theta}\log p(x,y,\theta)\right] = \prod_i \frac{1}{\tilde{Z_x^{t+1}}}\exp[E_{q_x}\log p(x,y,\theta)]$$

$$q_\theta^{(t+1)}(\theta) = \frac{1}{\tilde{Z_\theta^{t+1}}}\exp\left[E_{q_x}\log p(x,y,\theta)\right]$$

où E_{q_θ} et E_{q_x} désignent l'espérance sous les lois q_{θ^t} et $q_{x^{t+1}}$ respectivement, et où $\tilde{Z_x^{t+1}}$, $\tilde{Z_\theta^{t+1}}$ sont les constantes de normalisation indépendantes de x et θ respectivement.

3.2.2 Approximation de Laplace

Une autre méthode pour le calcul de l'ividence est l'approximation de laplace, qui approche l'integrale en utilisant un développement de Taylor du logarithme de la fonction à intégrer autour de son maximum. Soit alors $g(\theta)$ une densité non normalisée ayant un mode en θ^* et pour laquelle on cherche à calculer

$$Z_p = \int g(\theta) d\theta.$$

le développement de taylor de $logg(\theta)$ autour du maximum θ^* permet d'écrire

$$g(\theta) \simeq g(\theta^*) \exp -\frac{1}{2}(\theta - \theta^*)' H(\theta^*)(\theta - \theta^*)$$

où $H(\theta^*)$ est l'opposé du hessien de g en θ^*, d'où

$$\int g(\theta) d\theta \simeq \frac{g(\theta^*)(2\pi)^{\frac{d}{2}}}{\mid H(\theta^*) \mid}$$

Dans le cadre Bayésien, $g(\theta) = \pi(\theta)p(y|\theta)$, et l'approximation de laplace fait une approximation gaussienne locale autour de l'estimateur du Maximum A Posteriori (MAP) :

$$\theta^* = \widehat{\theta}_{MAP} = \arg\max_{\theta} p(\theta)p(y|\theta)$$

La validité de cette approximation est basée sur des propriétés de comportement gaussien asymptotique de la loi de l'échantillon et quelques conditions de régularité :

• La loi a posteriori ainsi approchée peut être mauvaise pour de petits jeux de données.

• Cette approximation peut être mauvaise pour des paramètres bornés, contraints ou positifs, comme des proportions de mélange ou des précisions ou si le maximum n'est pas proche de la masse principale de la probabilité.

• à cause des problèmes d'identifiabilité, la loi a posteriori peut ne plus être unimodale pour des vraisemblances avec données cachées, et dans ce cas, les conditions de régularité pour la convergence ne sont pas vérifies.

Remarque: Les deux méthodes -approximation en champ moyen, approximation de Laplace- sont concurrentes, mais elles peuvent aussi être combinées dans certains cas.

3.3 Exemple d'application

La localisation des zones réactives du cerveau dépend du stimulus auquel est soumis un individu. En imagerie par résonance magnétique fonctionnelle (IRMf), cette réaction est mesurée par l'augmentation de l'oxygénation du sang dans la zone activée (signal BOLD), apportant le "carburant" nécessaire à la réaction. Il est ainsi possible de dresser des cartes d'activation cérébrale, enregistrant pour chaque volume élémentaire (voxel) l'intensité de la réponse BOLD constatée lors du stimulus. Le but de l'analyse des cartes d'activation est de détecter les zones activées lors d'un stimulus, de les comparer avec les zones anatomiques, et de confronter ces résultats pour une meilleure connaissance du cerveau. À terme, il est envisageable d'utiliser cette méthode pour prédire une pathologie donnée.

L'IRMf pose le problème délicat de la sélection de variables en grande dimension. En effet, on observe en général quelques dizaines d'images 3D (scans), alors qu'il y a des dizaines, voire des centaines de milliers de voxels. Nous présentons ici une étude de données IRMf(Friston et al. 2008).Cette étude propose des modèles de sélection alternatif, nécessitant l'utilisation de l'approximation bayésienne variationnelle.

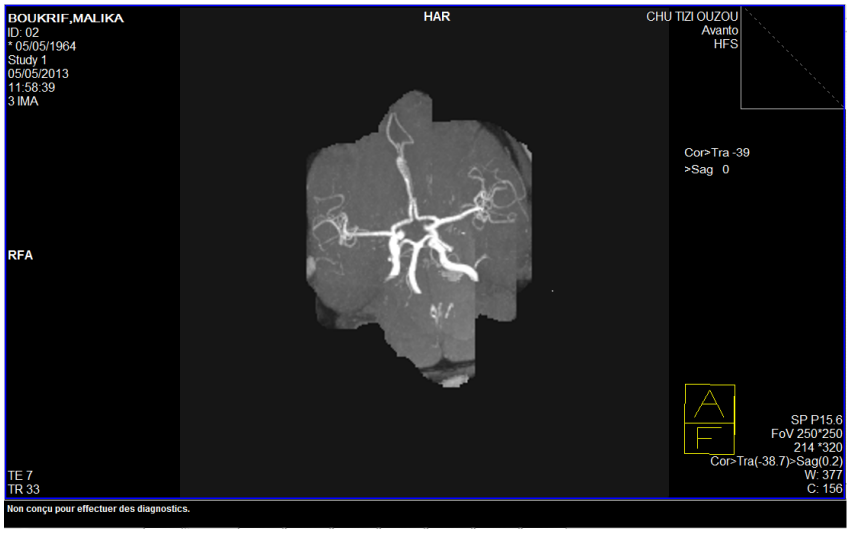

FIG. 3.1 – une image 3D (scan) de IRMF .

3.3.1 Modèle bayésien hiérarchique de défnition de motifs (Friston et al. 2008)

Friston et al utilisent un modèle bayésien hiérarchique dans lequel les voxels sont regroupés dans des sous-ensembles (motifs) successifs (peu nombreux) de poids crois- sants. Les loi a posteriori sont impossibles à calculer explicitement et l'estimateur bayésien variationnel se calculé en utilisant l'approximation en champ moyen et l'approximation de Laplace.

A partir de s images 3D (scans) de résonance magnétique fonctionnelle (IRMf) comportant n volumes élémentaires de mesure (voxels), Friston et al proposent le schéma suivant de décodage des états du cerveau.
Soit le modèle

$$M : WX = RY\beta + \zeta$$

avec

X est un vecteur de dimension s représentant la mesure de l'état comportemental, perceptif ou cognitif pour chacun des s scans de l'expérience.

Y est une matrice de taille $s \times n$ représentant la mesure du flux de la réponse hémodynamique en chaque voxel.
W et **R** sont des matrices données et connues de dimension $s \times s$.

β est un vecteur de dimension n représentant l'influence de chaque voxel dans la mesure de l'état comportemental.

ζ est un vecteur aléatoire corrélé de dimension s.

la covariance de ζ est:

$$cov(\zeta) = \Sigma^{\zeta}(\lambda) = \exp(\lambda^{\zeta})Q_0.$$

Les paramètres à estimer sont β et λ^{ζ}. En général, il y a peu de scans par rapport au nombre de voxels, donc le problème est mal posé Ainsi, l'estimation de β nécessite des a priori, ce qui est traité en invoquant un second niveau dans le modèle. Soit U la matrice d'une fonction de base spatiale (U peut éventuellement être l'identité), de dimension $R^{n\times u}$, et $\eta \in R^u$ les poids (inconnus) des vecteurs de la base dans la définition de β est:

$$\beta = u\eta$$

La contrainte de parcimonie est définie sur la covariance de η, qui est posée diagonale et dont les coefficients de la diagonale ne peuvent prendre que m valeurs distinctes :

$$cov(\eta) = \Sigma^{\eta}(\lambda) = \exp(\lambda_1^{\eta} I^{(1)}) + ... + \exp(\lambda_m^{\eta} I^{(m)})$$

Les matrices $I^{(i)}$ sont des matrices diagonales permettant de coder l'appartenance du poids d'une colonne de U à un ensemble de poids de même variance, formant une suite emboîtée de sous ensembles de poids $s^{(1)} \supset s^{(2)} \supset ..... \supset s^{(m)}$ de plus en plus petits regroupant des éléments de variance de plus en plus grande. La variance du poids d'une colonne dans un sous-ensemble $s^{(I)}$ est toujours supérieure à celle du poids d'une colonne dans sur-ensemble $s^{(i-1)}$.

Cette modélisation permet de séparer la spécification de l'a priori spatial, codé par la base spatiale U, des variances codées comme un mélange de composantes de covariance dans $\Sigma^{\eta}(\lambda)$. Le modèle est donc défini par:

$$WX = L\eta + \zeta$$

où

- ζ est un bruit gaussien inconnu corrélé, de loi normale $p(\zeta|\lambda) \sim Ns(0,\Sigma^{\zeta}(\lambda))$
- η est un effet aléatoire de loi normale $p(\eta|\lambda) \sim Nu(0,\Sigma^{\zeta}(\lambda))$
- $\lambda = (\lambda^{\zeta},\lambda_1^{\eta},....,\lambda_m^{\eta})$ est le paramètre à estimer de dimension m+1, inférieure à s.

Sous cette forme, les seules quantités inconnues sont les paramètres λ, contrôlant la covariance $\Sigma(\lambda)$

$$\begin{aligned} cov(WX) = \Sigma(\lambda) = \Sigma^{\zeta}(\lambda) + L\Sigma^{\eta}L' &= \exp(\lambda^{\zeta})Q_0 + \sum_{i=1}^{m} \exp(\lambda_i^{\eta} Q_i) \\ &= \sum_{i=0}^{m} \exp(\lambda_i) Q_i \end{aligned} \tag{3.7}$$

La loi a priori sur λ est choisie gaussienne $N_{m+1}(\pi,\Pi^{-1})$, ses hyperparamètres sont pris tels que $\pi_1 = ... = \pi_{m+1} = 32$, et $\prod = \frac{1}{256} I_{m+1}$, ce qui permet une loi a priori relativement peu informative (très grande variance) et de faible espérance pour $\exp(\lambda)$.

Friston trouve que les calcul de la loi a posteriori $p(\lambda|WX) = \frac{p(WX|\lambda)p(\lambda)}{p(WX)}$, et de l'évidence $p(WX) = \int_{\lambda} p(WX,\lambda)d\lambda$ ne sont pas explicites à cause de la forme de $\Sigma(\lambda)$.

De plus,Friston et al souhaitent calculer une estimation des données cachées η afin de déterminer progressivement la suite des sous-ensembles.

Pour une suite donnée de sous-ensembles ($I^{(1)}$,...$I^{(m)}$), ils ont procède en deux étapes :

- l'estimation des paramètres λ
- l'estimation les données manquantes η

Les résultats sont alors utilisés pour créer un nouveau sous-ensemble emboîté de plus grande variance, et le processus est itéré jusqu'à ce que l'énergie libre ne s'accroisse plus.

3.4 Calcul de l'évidence

on a trouvé précédament que Le logarithme de l'évidence est la somme de l'énergie libre$f(q_\lambda)$ et de la distance de Kullback entre la loi a posteriori $p(\lambda|WX)$ du paramètre et son approximation q_λ Donc friston a calcule d'abord l'energie libre.

3.4.1 Calcul de l'énergie libre:

Notons E_{q_λ} l'espérance sous la loi q_λ. L'énergie libre à maximiser est

$$f(q_\lambda) = E_{q_\lambda}(\log p(WX,\lambda)) - E_{q_\lambda}(q_\lambda)$$

L'approximation de Laplace permet d'approcher quadratiquement la log-vraisemblance complète $\log p(WX,\lambda)$ autour de son maximum μ^λ est

$$\log p(WX,\lambda) \simeq \log p(WX,\mu^\lambda) - \frac{1}{2}(\lambda - \mu^\lambda)' H(\mu^\lambda)(\lambda - \mu^\lambda)$$

telleque $H(\mu^\lambda) = -\partial^2_{\lambda=\mu^\lambda} p(WX,\lambda)$ est l'opposé du hessien évalué en μ^λ

Comme $p(\lambda|WX) \propto p(WX,\lambda)$,on maximise f en q_λ sur l'ensemble des gaussiennes $N(\mu^\lambda,\Sigma^\lambda)$. Si q_λ est gaussienne, on a alors:

$$E_{q_\lambda}(\log q_\lambda) = -\frac{1}{2}((m+1)\log(2\pi) + \log|\Sigma^\lambda| + m + 1)$$

D'ou

$$E_{q_\lambda}((\lambda - \mu^\lambda)' H(\mu^\lambda)(\lambda - \mu^\lambda)) = trace(H(\mu^\lambda)\Sigma^\lambda)$$

et

$$f(q_\lambda) = \log p(WX,\mu^\lambda) - \frac{1}{2}\text{trace}Z(H(\mu^\lambda)\Sigma^\lambda) + \frac{1}{2((m+1))}\log(2\pi) + \log|\Sigma^\lambda| + m + 1$$

La dérivé de l'énergie libre par rapport à Σ^λ , donne la forme de la covariance conditionnelle approchée

$$\frac{\partial f(q_\lambda)}{\partial \Sigma^\lambda} = -(\Sigma^\lambda)^{-1} + H(\mu^\lambda) = 0 \Leftrightarrow \Sigma^\lambda = H(\mu^\lambda)^{-1}$$

Donc $trace(\Sigma^{\lambda}H(\mu^{\lambda})) = m+1$, et l'énergie libre se simplifie comme suit:

$$f(q_{\lambda}) = \log p(WX,\mu^{\lambda}) + \frac{1}{2}\Big((m+1)\log(2\pi) + \log|\Sigma^{\lambda}|\Big)$$

Enfin, soit w le rang de W, la loi de $WX|\lambda$ est gaussienne centrée:

$$\log p(WX|\mu^{\lambda}) = -\frac{1}{2}\Big(\log|\Sigma(\mu^{\lambda})| + w\log(2\pi) + (WX)'\Sigma(\mu^{\lambda})^{-1}WX\Big)$$

tandis que la loi a priori de λ est gaussienne $N(\pi,\Pi)$:

$$\log p(\mu^{\lambda}) = -\frac{1}{2}\Big(\log|\Pi^{-1}| + (m+1)\log(2\pi) + (\mu^{\lambda}-\pi)'\Pi(\mu^{\lambda}-\pi)\Big)$$

D'où l'approximation de la log-évidence $\log(WX)$ par l'énergie libre calculée de la façon suivante :

$$f(q_{\lambda}) = -\frac{1}{2}\Big[(WX)'\Sigma(\mu^{\lambda})^{-1}WX + \log|\Sigma(\mu^{\lambda})| + w\log(2\pi) - \log|\Sigma^{\lambda}\Pi| + (\mu^{\lambda}-\pi)'\Pi(\mu^{\lambda}-\pi)\Big]$$

Les deux premiers termes reflètent la précision du modèle, les deux derniers sa complexité, $w\log(2\pi)$ étant un terme constant. Cette expression dépend des paramètres (donnés) de la loi a priori π, Π, et de ceux la loi a posteriori μ^{λ}, Σ^{λ} calculés par l'itération d'un schéma de Newton (M-Step). Soit L_{λ} et $L_{\lambda\lambda}$ le gradient et le Hessien de l'énergie libre, les étapes suivantes sont répétées jusqu'à la convergence

3.4.2 Éléments de calcul des étapes du M-step

Grâce à la forme particulière(3.7) de la covariance $\Sigma(\lambda)$ on a :

$$\frac{\partial\Sigma(\lambda)}{\partial\lambda_i} = \exp(\lambda_i)Q_i$$

d'ou la forme de la dérivée partielle de la matrice de précision $\Sigma(\lambda)^{-1}$

$$\begin{aligned} P_i &= \frac{\partial}{\partial\lambda_i}\Sigma(\lambda)^{-1} \\ &= -\Sigma(\lambda)^{-1}\frac{\partial\Sigma(\lambda)}{\partial\lambda_i}\Sigma(\lambda)^{-1} \\ &= -\exp(\lambda_i)\Sigma(\lambda)^{-1}Q_i\Sigma(\lambda)^{-1}, \end{aligned}$$

et

$$\begin{aligned} \frac{\partial}{\partial\lambda_i}\Big((WX)'\Sigma(\lambda)^{-1}WX\Big) &= (WX)'P_iWX \\ &= trace\big[(WX)'P_iWX\big] \\ &= trace\big[P_iWX(WX)'\big] \end{aligned}$$

On en déduit l'expression suivante de la dérivée de l'énergie libre $f(q_\lambda)$, Π_i étant la $i^{ème}$ ligne de Π

$$L_{\lambda_i} = -\frac{1}{2}\Big[trace(P_i(WXX'W' - \Sigma(\lambda))\Big] - \Pi_i(\lambda - \pi)$$

La dérivée seconde s'obtient grâce à quelques manipulations élémentaires de matrices :

$$\begin{aligned} L_{\lambda\lambda_{ij}} &= -\frac{1}{2}trace\Big(P_i(-\exp(\lambda_i)Q_j)\Big) - \Pi_{ij} \\ &= -\frac{1}{2}trace\Big(P_i\Sigma(\lambda)(-\exp(\lambda_i))\Sigma(\lambda)^{-1}Q_j\Sigma(\lambda)^{-1}\Sigma(\lambda)\Big) - \Pi_{ij} \\ &= -\frac{1}{2}trace\Big(P_i\Sigma(\lambda)P_j\Sigma(\lambda)\Big) - \Pi_{ij} \end{aligned}$$

3.5 Estimation de η

Étant données les espérances conditionnelles des hyper-paramètres obtenus à l'étape précédente, il est maintenant possible d'obtenir analytiquement les moyennes a posteriori des états cachés η.

3.5.1 Éléments de calcul de l'étape E-step

L'utilisation de l'approximation de Laplace permet d'écrire

$$\log(WX,\eta,\lambda) = \log p(WX,\mu^\eta,\mu^\lambda) - \frac{1}{2}\Big(\lambda - \mu^\lambda)'(\eta - \mu^\eta)'\Big)H\begin{pmatrix} \lambda - \mu^\lambda \\ \eta - \mu^\eta \end{pmatrix}$$

où $H = H(\mu^\eta,\mu^\lambda)$ est l'opposée de la matrice des dérivées secondes de $\log p(WX,\eta,\lambda)$ calculée au maximum (μ^η,μ^λ)
Or

$$q_\eta(\eta) \propto \int q_\lambda(\lambda)\log p(WX,\eta,\lambda)d\lambda$$

ce qui revient à prendre pour q_η une gaussienne $N(\mu^\eta,\Sigma^\eta)$

$$q_\eta(\eta) \propto \exp -\frac{1}{2}(\eta - \mu^\eta)'(\Sigma^\eta)^{-1}(\eta - \mu^\eta)$$

avec

$$\Sigma^\eta = -\left(\frac{\partial^2}{\partial\eta^2}\log(WX,\eta,\mu^\lambda)_{|\eta=\mu^\eta}\right)^{-1}$$

La dérivée du logarithme de la vraisemblance complète amène directement à

$$\mu^\eta = \Sigma^\eta L' \Sigma^\zeta(\mu^\lambda)^{-1} WX$$

et la variance a posteriori Σ^η s'écrit, en utilisant le lemme d'inversion matricielle de Woodbury pour supprimer les matrices de grande taille

$$\begin{aligned}\Sigma^\eta &= (L'\Sigma^\zeta(\mu^\lambda)^{-1}L + \Sigma^\eta(\mu^\lambda)^{-1})^{-1} \\ &= \Sigma^\eta(\mu^\lambda) - \Sigma^\eta(\mu^\lambda)L'(\Sigma^\zeta(\mu^\lambda) + L\Sigma^\eta(\mu^\lambda)L')^{-1}L\Sigma^\eta(\mu^\lambda) \\ &= \Sigma^\eta(\mu^\lambda) - \Sigma^\eta(\mu^\lambda)L'\Sigma(\mu^\lambda)^{-1}L\Sigma^\eta(\mu^\lambda).\end{aligned}$$

L'énergie libre est maximum pour la loi jointe a posteriori $P(\eta,\lambda|WX)$.
L'hypothèse de champ moyen permet de chercher une approximation factorisée en η et λ de la loi a posteriori :

$$P(\eta,\lambda|WX) \simeq q_\lambda(\lambda)q_\eta(\eta)$$

et la solution optimum en q_η est

$$q_\eta(\eta) = \frac{1}{K_\eta}\exp\int q_\lambda \log p(WX,\eta|\lambda)d\lambda$$

En utilisant l'approximation de Laplace, on cherche q_η sous forme d'une gaussienne $N(\mu^\eta,\Sigma^\eta)$ dont on calcule explicitement l'espérance et la variance

$$\Sigma^\eta(\mu^\lambda) - \Sigma^\eta(\mu^\lambda)L'\Sigma(\mu^\lambda)^{-1}L\Sigma^\eta(\mu^\lambda).$$

$$\mu^\eta = \Sigma^\eta L'\Sigma^\zeta(\mu^\lambda)^{-1}WX$$

S'en déduisent immédiatement l'estimation des paramètres de l'approximation de la loi a posteriori de $\beta = U\eta$

$$\Sigma^\beta = U\Sigma^\eta U'$$

$$\mu^\beta = U\mu^\eta$$

On résout ainsi facilement le problème d'inférence dans un modèle mal conditionné. Il reste cependant une diffficulté à lever : le choix de la partition codée par les matrices diagonales $I^{(i)}$.

3.6 Recherche des sous-ensembles de poids

La solution proposée est de faire une recherche progressive ascendante. Partant de l'hypothèse d'égalité de variance $I^1 = I$ pour tous les poids, les espérances conditionnelles des poids sont utilisées pour créer un sous-ensemble avec les plus hauts poids (en prenant

les poids du sous espace précédent de $\|\mu^\eta\|$ supérieurs à la médiane par exemple). Puis l'EM est réexécuté, et le sous-ensemble de plus haut poids est à nouveau scindé. La procédure est répétée jusqu'à ce que la log-évidence cesse de croître. L'algorithme peut ne converger que vers un maximum local.

3.7 résultat (Friston et al. 2008)

Friston et al illustrent d'abord leur méthode sur un jeu de données simulées. Ils montrent qu'elle permet de distinguer un modèle nul (c'est à dire sans motif ni influence quelconque de voxels), d'un modèle défini avec un codage parcimonieux. Cependant, seuls les voxels ayant un poids important sont bien récupérés. Et il faudrait sans doute étudier l'influence du signal sur bruit sur ces résultats.

Dans le cas d'un jeu de données réelles (étude de l'attention dans un mouvement visuel), la distribution des poids pour le modèle optimum affirme bien leur parcimonie. Cependant, le graphique de l'évidence en fonction du nombre de partitions présente une croissance jusqu'à 4 partitions, puis une décroissance, ce qui est en contradiction avec le fait que l'énergie libre ne peut qu'augmenter ou se stabiliser avec des composants supplémentaires (cf 4.1.3). Ceci est vraisemblablement du au fait que l'algorithme a convergé vers un minimum local quand le nombre de partitions devient important. Enfin, la méthode est utilisée pour comparer des modèles avec des codages spatiaux différents ou des régions d'intérêt différentes. Les résultats sont positifs, mais effectués sur des régions très différentes et demanderaient à être précisés.

Friston et al utilisent l'énergie libre pour sélectionner un nombre de composantes de variance, la recherche s'arrêtant quand l'énergie libre n'augmente plus à l'ajout d'une composante : la solution est donc théoriquement sensible à la variation de la proximité entre l'énergie libre et l'évidence suivant les modèles. Pour valider leur approche, ils estiment l'évidence par MCMC (échantillonnage de Metropolis-Hasting et identité de la moyenne harmonique) d'une part, et la calculent par approximation variationnelle d'autre part, dans des modèles ayant de plus en plus de composantes de variance. Ils montrent alors que l'évidence estimée par MCMC et celle calculée par approximation variationnelle varient dans les mêmes proportions en fonction du nombre de composantes, donnant, pour cet exemple, l'avantage à l'approximation variationnelle, qui est beaucoup plus rapide d'exécution.

Conclusion et Perspectives

Même si les méthodes variationnelles Bayésiennes sont récemment utilisées dans les applications de neuroimagerie avec des résultats prometteurs, il convient d'être attentif à la qualité de l'approximation obtenue, dont on ne peut quantifier la précision. Ainsi, notre objectif est double : présenter la méthode de l'approximation variationnelle (principalement dans le cadre bayésien) et montrer son utilisation dans un exemple d'application de neuroimagerie.

En Perspectives:

– approche bayésienne variationnelle : qualité d'approximation, convergence.

Enfin, Il serait intéressent de reprendre ce travail avec des données censurées ou avec des données manquantes et de les appliquer.

Annexes

Quelques résultats de calcul matriciel

Lemme 1

Si $X \sim N(\mu,V), alors E(X'AX) = \mu' A\mu + trace(AV)$

Lemme 2

Siot Σ et H deux matrices de taille compatible. Alors

$$\frac{\partial}{\partial \Sigma} trace(\Sigma H) = H$$

Lemme 3

Soit Σ une matrice inversible. Alors,

$$\frac{\partial \log |\Sigma|}{\partial \Sigma} = \Sigma^{-1}$$

Lemme 4

Soit $P_i = \frac{\partial}{\partial \lambda_i} \Sigma(\lambda)^{-1}$, les dérivées partielles de la précision $\Sigma(\lambda)^{-1}$
On a :

$$\frac{\partial}{\partial \lambda_i} \Sigma(\lambda) = -\Sigma(\lambda) P_i \Sigma(\lambda)$$

$$\frac{\partial}{\partial \lambda_i} \log |\Sigma(\lambda)| = trace(P_i \Sigma(\lambda))$$

Lemme 5: Lemme d'inversion de Woodbury

Soient les matrices A de dimension nxn, U de dimension $n \times k$, C de dimension $k \times k$ inversible et V de dimension $k \times n$ telles que la matrice suivante

$$\begin{pmatrix} C & V \\ -U & A \end{pmatrix}$$

soit inversible. On a :

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}$$

Bibliographie

[1] Ali Mohammad-Djafari. (2009). *Approche variationnelle pour le calcul bayésien dans les problèmes inverses en imagerie.*

[2] Berger J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* 2éme édition Springer-Verlag.

[3] BELKACEM N. (2012).*Modèles d'incertitude appliqués au probléme de management de l'eau* Memoire de magister (Ecole doctorale) en statistique, U.M.M.T.O.

[4] Christine Keribin.(2010).*Méthodes bayésiennes variationnelles : concepts et applications en neuroimagerie*Journal de la Société Française de Statistique.

[5] Christian P.Robert.(2006).*Le choix bayésien Principes et pratique* Springer.

[6] DJOWYDA.G. *Aspects de la robustesse Bayésienne dans les modèles AR(1)* Memoire de magister (Ecole doctorale) en statistique, U.M.M.T.O, .

[7] Éric Parent, Jacques Bernier.(2007). *Le raisonnement bayésien Modélisation et inférence* Springer.

[8] François ORIEUX.(2009).*Inversion bayésienne myope et non-supervisée pour l'imagerie sur-résolue. Application à l'instrument SPIRE de l'observatoire spatial Herschel.* Thèse de doctorat en phisique, université Paris-sud 11.

[9] Hartigan J.A. (1983).*Bayes Theory.* Springer-Verlag.

[10] H.FELLAG. (2012-2013) *Introduction aux statistiques bayesiennes.* Cours de 2éme année Master, U.M.M.T.O Tizi ouzou.

[11] Hacheme AYASSO.(2010) *Une approche bayésienne de l'inversion. Application à l'imagerie de diraction dans les domaines micro-onde et optique* THÈSE DE

DOCTORAT Spécialité: Physique.

[12] Jean-Jacques Boreux, Eric Parent.(2010) *Pratique du calcul bayésien*. Springer.

[13] J.J.Droesbeke, J.Fine, G.Saporta.(2002) *Méthodes bayésiennes en statisque*. Technip.

[14] Larbi L.(2011)*Sur la décision statistique dans le contexte Bayesien*, Memoire de master en statistique, U.M.M.T.O.

[15] Michel Lejeune.(2010) *Statistique La théorie et ses applications*. Springer, 2éme édition (2010).

[16] Pierre-Charles.(2007)*Fondations, méthode et applications de l'apprentissage bayésien.*, Memoire DOCTEUR DE L'INPG Spécialité : Imagerie, Vision.

[17] Robert C.P.(1992) *L'Analyse Statistique Bayésienne, Economica*.

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

[19] Séverine DEMEYER.(2011) *Approche bayésienne de l'évaluation de l'incertitude de mesure : application aux comparaisons interlaboratoires*, Memoire de doctorat en informatique et statistique, École Doctorale EDITE .

[20] Thomas RODET, Yuling ZHENG. *Approche bayésienne variationnelle: appliation à la déonvoluti cononjointe d'une soure pontuelle dans une soure étendue*,

[21] Thomas RODET.(2012) *Inversion Bayésienne : illustration sur des problèmes tomographiques et astrophysiques*, Mémoire pour obtenir L'habilitation à diriger des recherches, Université Paris-Sud 11 Faculté des sciences d'Orsay.