République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université AMO de Bouira
Faculté des Sciences et des Sciences Appliquées
Département du Mathématiques

Mémoire de Master

Filière : Mathématiques

Spécialité : Recherche Opérationnelle

Thème

Estimation fonctionnelle par l'approche bayèsienne non paramétrique

Présenté par :
- **BAAZIZ Leila**
- **SAIDI Naima**

Devant le jury composé de :

Président	M^r **BOUDREF Mohamed**	**MCA**	**U. A/M/O Bouira.**
Encadreur	M^r **BEDDEK Said**	**MAA**	**U. A/M/O Bouira.**
Examinateur	M^r **BOUGHANI L'hadi**	**MAA**	**U. A/M/O Bouira.**

2019/2020

Remerciements

Nous remercions, en premier lieu, notre ***Dieu*** *le tout puissant pour la volonté, la santé et la patience qu'il nous a donné pour effectuer le présent travail.*

En second lieu, Nous tenons à adresser nos remerciements à notre encadreur M^r SAID BEDDEK *pour son aide et ses conseils avec beaucoup de patience et d'encouragements.*

Nos remerciement vont également aux membres de jury, M^r BOUDREF MOHAMED *et* M^r BOUGHANI L'HADI *qui ont accepté d'examiner notre modeste travail.*

Nos sincères remerciements s'adressent enfin à tous ceux qui nous ont soutenu de prés ou de loin.

Merci à Tous

Dédicaces

Je dédie ce modeste travail

À mes très chers parents qui ont toujours été là pour moi, qui n'ont jamais cessés, de formuler des priéres à mon égard. À leurs amours, confiance, soutiens et sacrifices. Que Dieu vous préserve et vous accorde santé et longue vie.

À mon chèr frère Hamadache, À ma chère soeur Kahina et son époux Hamza, Pour leurs soutiens et leurs conseils précieux tout au long de mes études.

À ma chère niéce Mastina notre petite ange.

À mes grands parents, que Dieu vous protége et vous prête bonne santé et longue vie.

À mes chérs amis Mechekak Cylia et Rezkallah Mohand, pour leurs aides et présence dans les bons et les mauvais moments.

À mon ancle et ma tante et leurs enfants "Thilleli, Assirem, Youfrar".

À mon binhôme Saidi Naima et sa famille.

À toute ma famille.

.

De Leila.

Dédicaces

Je dédie ce modeste travail

À l'étre le plus chèr de ma vie, ma mére. Que Dieu te protége.

À mon chèr frére "Farid", et mes chères soeurs "Souhila" et "Yasmina".

À tous mes amis de promotion 2^{me} année Master mathématique .

À mon binhôme Baaziz Leila et sa famille.

À mes chères amies : Amina, Hanane, Katiba, Lina.

À tous les membres de ma famille et toutes personne qui porte le nom Saidi, et à tout ceux qui ont participé à ma réussite.

De Naima.

Résumé

Ce travail consiste à présenter l'approche bayésien qui complète la démarche inférentielle classique. Alors que la statistique classique repose sur la loi des observations, la statistique bayésienne repose sur la loi à posteriori. La loi à posteriori 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 à priori par l'observation x, au travers de $\pi(\theta|x)$. On distingue deux approches, l'une paramétriqe où le nombre des paramétre est finis, et l'autre non paramétrique définissant de ce fait une distribution de probabilité sur des espaces fonctionnels (de dimension infinie). Dans ce travail, on s'est focalisé sur l'estimation de f, une fonction de densité de probabilité par l'approche bayésienne non paramétrique. Dans ce contexte, on a utilisé le processus de Dirichlet en tant que loi à priori de ce modéle.

Mots clés inférence bayésienne, statistique inférentielle, Processus de Dirichlet, loi a priori, loi a postériori.

Abstract

This work consists in presenting the Bayesian approach which complements the classical inferential approach. Whereas classical statistics are based on the law of observations, Bayesian statistics are based on the law of observations. is based on the posterior law. The posterior law can be interpreted as a summary (in a sense probabilistic) information available on θ. Once x is observed. The Bayesian approach achieves in a way the updating of information the prior by observation x, through the use of $\pi(x|\theta)$. We distinguish two approaches, one parametric where the number of parameters is finite, and the other non-parametric, thus defining a probability distribution over functional spaces (of infinite dimension). In this work, we focused on the estimation of f, a probability density function by the nonparametric Bayesian approach. In this context, the Dirichlet process was used as the prior law of this model.

Key words Bayesian inference, inferential statistics, Dirichlet's process, the prior law, the posterior law.

Notations :

Ω :	Ensemble des résultats possibles.
P(Ω) :	Ensemble constitué de tout les sous ensembles (parties) de Ω.
$\mathbb{R}$:	Est l'ensemble des nombres réels.
$\mathbb{N}$:	Est l'ensemble des nombres entiers naturels.
f $\propto$ g :	$f$est proportionnel à g.
μ :	Mu.
ξ :	Xi.
γ :	Gamma.
ω :	Omega.
ψ :	psi.
ι :	Iota.
η :	Eta.
δ :	Delta.
ϱ :	Rho.
$\Gamma(.)$:	Fonction Gamma $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$; $(\Gamma(n+1) = n!)$.
$\simeq$:	Est asymptotiquement égal à (approximation,équivalence d'homotopie).
$\equiv$:	Identique à.
$\sim$:	Équivalence en loi de probabilité.
ν :	Nu.
χ :	Khi.
Δ :	Delta.
$< ., . >$:	Produit scalaire.
$\bigotimes$:	Produit tensoriel.

.

Abréviation :

v.a :	Variable aléatoire.
v.a.r :	Variable aléatoire reélle.
i.i.d :	Indépendant identiquement distribué.
ind :	Indépendant.
CDF :	Fonction de repartition.
IFR :	Increasing Failur Rate.
DFR :	Decreasing Failure Rate.

Table des matières

Introduction Générale

Christian Robert [44] à affirmé en 2006 que : «L'objet principal de la statistique est de mener, grâce à l'observation d'un phénomène aléatoire, une inference sur la distribution probabiliste à l'origine de ce phénomène, c'est-à-dire de fournir une analyse (ou une description) d'un phénomène passé ou une prédiction d'un phénomène à venir de nature similaire ». En pratique ceci se traduit par l'estimation de cette distribution de probabilité et de ses paramètres. Ce qui fait que la théorie de l'estimation est devenue aujourd'hui l'une des préoccupations majeurs des statisticiens. On trouve dans la littérature deux approches d'estimations : l'approche paramétrique et l'approche nonparamétrique.
Dans ce travail, on s'intéresse aux modéles bayésiens qui sont utilisés dans différents thèmes de recherche pour résoudre des problèmes d'estimation et d'inférence. Contrairement à l'approche déterministe, le paradigme bayésien considère les paramètres du modèle comme des variables aléatoires.

Soit Θ l'espace des paramètres et X l'espace des observations. Dans l'approche bayésienne paramétrique le nombre des paramètres à estimer est finis. Par contre, pour le cas non paramétrique, l'espace des paramètres est un espace vectoriel de dimension infinie, c'est à dire un espace fonctionnel. Par conséquent, dans le cas non paramétrique, le but est l'estimation fonctionnelle (estimation d'une fonction de densité de probabilité par exemple).

L'objectif de ce mémoire est l'estimation fonctionnelle dans un cadre bayésien. En particulier, on s'est focalisé sur l'estimation de la densité de probabilité f ou une fonctionnelle de celle-ci. Dans ce cas la fonction f est vue comme la réalisationn d'un processus stochastique.

Le mémoire est organisé de la manière suivante :

Dans le premier chapitre, nous avons donné un rappel sur la théorie d'estimation et les principaux outils mathématiques nécessaires à l'accomplissement de ce travail.

Le second chapitre est consacré a la présentation de la méthode bayésienne. Tout d'abord, on a présenté la méthode bayésienne paramétrqiue et ensuite on a introduit la méthode non paramétrique.

Enfin, dans le dernier chapitre on traite un exemple pratique sur l'estimation d'une densité de probabilité par l'approche bayésienne non paramétrique en utulisant le processus stochastique de Dirichlet.

Chapitre 1

Notions de bases sur la théorie de l'estimation

Ce chapitre introduit la notion fondamentale de modèle statistique. Une introduction aux différentes méthodes statistiques d'estimation déterministes sera aussi exposée. Ainsi que quelques cas particuliers importants que nous retrouverons dans les développements ultérieurs.

1.1 Variables Aléatoires :

1.1.1 Définitions :

Définition 1.1. *Un espace de probabilité est un espace mesurable $(\Omega, \mathcal{F})$ muni d'une mesure de probabilité P, c'est à dire une mesure de masse totale 1 :$P(\Omega) = 1$.*
Les ensembles mesurables $A \in \mathcal{F}$ sont appelés les évènements (ou observables)[12].

Définition 1.2. *(variable aléatoire)*

- *On appelle variable aléatoire (v.a.) toute application mesurable X d'un espace de probabilité $(\Omega, \mathcal{F}, P)$ dans $\mathbb{R}$ muni de la tribu borélienne $\mathcal{B}(\mathbb{R})$.*
- *On appelle loi de X la mesure de probabilité P_X définie sur $\mathbb{R}$ par*

$$P_X = P(X \in A) = P(\omega \in \Omega | X(\omega) \in A), \;\; A \in \mathcal{B}(\mathbb{R})$$

- *La v.a. X est discrète si elle est à valeurs dans un ensemble au plus dénombrable (en bijection avec une partie de $\mathbb{N}$) : $X(\Omega)$ est fini ou dénombrable (on peut compter ses éléments).*
- *La v.a. X est continue si elle peut prendre toutes les valeurs dans un intervalle donné (borné ou non borné). En régle générale, toutes les variables qui resultent d'une mesure sont de type continue [12].*

1.1.2 Fonction de répartition :

Soit $F(x)$ la fonction de répartition de la variable aléatoire X. Il s'agit de la probabilité que X soit plus petite ou égale à une valeur donnée, c'est-à-dire

$$F(x) = P[X \leq x].$$

De manière formelle, la fonction $F(x)$ est une fonction de répartition si et seulement si les trois conditions suivantes sont respectées :

1. $F(x)$ est non-décroissante.

2. $F(x)$ est continue à droite, c'est-à-dire que pour tout point x_0, la valeur limite de $F(x)$ lorsque x s'approche de x_0 par la droite est $F(x_0)$.
3. $\lim_{x\longrightarrow-\infty} F(x) = 0$ et $\lim_{x\longrightarrow\infty} F(x) = 1$ [4].

1.1.3 Densité et fonction de probabilité :

Soit $f(x)$ la fonction de densité de probabilité (fdp) de la variable aléatoire continue X. De manière formelle, $f(x)$ est une densité de probabilité si et seulement si elle satisfait les deux conditions suivantes :

1. $f(x) \leq 0$ pour toutes les valeurs de $X \in \Omega$.
2. $\int_\Omega f(x) = 1$.

Pour les variables aléatoires discrètes, on définit plutôt $P[X = x]$ la fonction de masse de probabilité (fmp). Elle doit satisfaire les deux conditions suivantes :

1. $P[X = x] \geq 0$.
2. $\sum_\omega P[X = x] = 1$.

Dans le cas continu, la fonction de densité est la dérivée de la fonction de répartition [4].

Définition 1.3. *La fonction caractéristique d'une v.a. X est la fonction de $\mathbb{R}$ dans $\mathbb{C}$*

$$\phi_X(t) = \mathbb{E}[e^{itX}] = \int_\Omega e^{itX} dP = \int_\Omega e^{itx} dP_X(x). \tag{1.1}$$

Il s'agit de la transformée de Fourier de la loi P_X de X. Cette fonction caractérise la loi de X.[12]

1.2 Théorème de Bayes :

Définition 1.4. *(Probabilité conditionnelle). Soit A et B deux événements tels que $P(B) \neq 0$, alors [7]*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

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

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

Théorème 1.2. *[7] (Bayes). Soit A et B deux événements tels que $P(B) \neq 0$, alors*

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

Définition 1.5. *(Probabilité conditionnelle). Soit A, B, C des événements tels que $P(C) \neq 0$ alors [7] A est indépendant de B conditionnellement à C si*

$$P(A \cap B|C) = P(A|C)P(B|C).$$

Remarque : Le lecteur pourra vérifier que deux événements indépendants A et B (c.à.d. $P(A \cap B) = P(A)P(B)$ ne sont pas conditionnellement indépendants en général [7].

Définition 1.6. *(Densité conditionnelle). Soit X et Y deux variables aléatoires de loi jointe $f(x, y)$ sous réserve de non négativité du dénominateur on définit la densité conditionnelle [7] :*

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

1.3 Modèle statistique :

On appelle modèle statistique, la donnée d'un espace d'observations E, d'une tribu A d'événements sur E et d'une famille de probabilités $\mathcal{P}$ sur l'espace probabilisable (A, E). On le note $(E, A, \mathcal{P})$ ou, quand il n'y a pas de risque de confusion, plus simplement $\mathcal{P}$.
On note X la v.a qui modélise le phénomène aléatoire que l'on étudie. Autrement dit la v.a X engendre les observations dont on dispose. Elle est à valeurs dans (E, A) et sa loi de probabilité $\mathcal{P}$ inconnue est dans la famille P. On appellera parfois X v.a générique du modèle statistique [5].

Définition 1.7. *On dit qu'un modèle statistique est paramétrique s'il existe un entier d et un sous ensemble Θ de $\mathbb{R}^d$ tels que la famille de probabilités $\mathcal{P}$ puisse être paramétrée par Θ, i.e. tels que l'application :*

$$\Theta \longmapsto \mathcal{P}$$
$$\theta \longmapsto \mathcal{P}_\theta$$

est surjective.
On note $\mathcal{P} = \{\mathcal{P}_\theta : \theta \in \Theta\}$. Dans le cas contraire on parle de modèle non-paramétrique [5].

1.4 Modèle d'échantillonnage :

Pour étudier un phénomène aléatoire, on a souvent intérêt à observer plusieurs réalisations indépendantes de celui-ci. On parle alors d'échantillon ou d'échantillonnage.

Définition 1.8. *On appelle n-**échantillon** de la loi P_θ, la donnée d'un vecteur $\underline{X} = (X_1, ..., X_n)$ constitué de n v.a indépendantes et identiquement distribuées (i.i.d.) de loi P_θ.*
*On appelle **modèle d'échantillonnage**, le modèle*

$$(E^n, A^{\otimes n}, \mathcal{P}^n = \{P_\theta^{\otimes n}, \theta \in \Theta\})$$

où $A^{\otimes n}$ est la tribu produit (engendrée par les pavés) sur A^n et $P_\theta^{\otimes n} = P_\theta \otimes ... \otimes P_\theta$ est la probabilité produit sur $(E^n, A^{\otimes n})$ qui est la loi du vecteur $\underline{X} = (X_1, ..., X_n)$.

Toutes les v.a ont la même loi, donc la même valeur de θ. Un échantillon est un vecteur aléatoire. Sa réalisation, fruit de n observations indépendantes du même phénomène, est notée $\underline{x} = (x_1, .., x_n)$. On fera toujours cette distinction entre v.a. et sa réalisation en utilisant majuscules ou minuscules. Un modèle d'échantillonnage est donc un modèle statistique particulier, où l'espace des observations est de la forme (E^n muni de sa tribu produit classique et de probabilités de la forme $P_\theta^{\otimes n}$. Aussi parfois on parlera dans ce cas simplement de modèle statistique. L'important est de bien avoir en tête quelle est la nature des observations, par exemple variable aléatoire réel, vecteur aléatoire (mais avec composantes non nécessairement indépendantes, ni de même loi) ou encore échantillon.

La densité de l'échantillon sous la loi P_θ est donnée :

$$\underline{x} = (x_1, ..., x_n) \longmapsto \prod_{i=1}^{n} f_\theta(x_i);$$

pour tout x de E^n. Si on considère le produit de droite non plus comme une fonction de x mais comme une fonction du paramètre θ, pour un $\underline{x} = (x_1, ..., x_n)$ fixé, on parle de vraisemblance [5].

1.5 Vraisemblance :

Définition 1.9. *Dans un modèle statistique paramétrique (E, A, P), on appelle* ***vraisemblance*** *de l'observation x la fonction*

$$\begin{aligned} \mathcal{L}(x, .) : \Theta &\longrightarrow \mathbb{R}^+ \\ \theta &\longmapsto \mathcal{L}(x, \theta) = f_\theta(x) \end{aligned}$$

Bien sûr, dans le cas d'un modèle d'échantillonnage, la vraisemblance de l'échantillon observé $\underline{x} = (x_1, .., x_n)$ s'écrit sous la forme :

$$\mathcal{L}(x_1, .., x_n; \theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

C'est donc la loi conjointe du n-échantillon évaluée aux valeurs observées et considérées comme fonction du paramètre θ [5].

1.6 Statistique et Estimateur :

Définition 1.10. *Soit $(E^n, A^{\otimes n}, \mathcal{P}^n = \{P_\theta^{\otimes n}, \theta \in \Theta\})$ un modèle d'échantillonnage. On appelle statistique la v.a. $T(\underline{X}) = T(X_1, ..., X_n)$ où T est une fonction mesurable connue de $(E^n, A^{\otimes n}, \mathcal{P}^n = \{P_\theta^{\otimes n}, \theta \in \Theta\})$ vers un espace probabilisable $(F, \mathcal{F})$:*

$$T : \left\{ \begin{array}{ll} E^n & \longrightarrow F \\ \underline{x} = (x_1, .., x_n) & \longmapsto T(x_1, .., x_n) \end{array} \right.$$

Définition 1.11. *On appelle estimateur de $g(\theta)$, toute statistique $T(X)$ de $(E^n, A^{\otimes n})$ à valeurs dans $g(\Theta)$ [5].*

Notation : Quand il s'agit d'estimer le paramètre θ on note souvent $\hat{\theta}$ son estimateur et $\hat{\theta_n}$ quand on souhaite préciser la taille n de l'échantillon. Pour l'estimation de $g(\theta)$ on utilise parfois aussi la notation $\widehat{g(\theta)}$ [5].

1.7 Construction d'estimateurs :

1.7.1 Estimateurs empiriques (des moments) :

Soit donc X une v.a. générique d'un modèle d'échantillonnage $(E^n, A^{\otimes n}, \mathcal{P}^n = \{P_\theta^{\otimes n}, \theta \in \Theta\})$. C'est à dire que $X_1, ..., X_n$ est un échantillon de même loi que X. Notons $\mathbb{E}_\theta(.)$ et $Var_\theta(.)$ respectivement les opérateurs espérance et variance sous la loi P_θ, en supposant que ces quantités sont bien définies. Pour simplifier les notations, on notera $m_\theta = \mathbb{E}_\theta(X)$ et $\sigma_\theta^2 = Var_\theta(X)$ [5].

Définition 1.12. *On appelle moyenne empirique, la statistique $\bar{X}$ définie, pour une taille n d'échantillon, par :*

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Quand on peut écrire l'espérance de la v.a. générique X en fonction du paramètre du modèle, i.e. quand il existe une fonction g telle que $m_\theta = g(\theta)$ (ce qui est souvent le cas), alors on pourra donner le titre d'estimateur à $\bar{X}$. On dira alors qu'il estime m_θ [5].

Proposition 1.1. *La moyenne empirique est telle que*

$$\mathbb{E}_\theta(\bar{X}_n) = m_\theta$$
$$Var(\bar{X}_n) = \frac{\sigma_\theta^2}{n}$$

Preuve. Immédiate par linéarité de l'espérance et grâce à l'indépendance entre les termes pour le calcul de la variance.
Le premier point de la proposition montre que l'estimateur $\bar{X}$ est, dans un certain sens, un "bon" estimateur de l'espérance m_θ puisqu'il est égal en espérance à ce qu'il cherche à estimer. On parlera d'estimateur sans biais.
Une généralisation évidente de ce qui précède est donnée par l'estimation empirique d'un moment de X d'ordre quelconque. Notons $m_\theta(p) = E_\theta(X^p)$ le moment d'ordre p de X sous la loi P_θ, en supposant que celui-ci existe. Par analogie avec ce qui précède, on peut définir l'estimateur empirique du moment d'ordre p [5].

Définition 1.13. *On appelle estimateur empirique du moment d'ordre p, la statistique*

$$\hat{m}_\theta(p) = \frac{1}{n}\sum_{i=1}^{n} X_i^p,$$

On peut aussi s'intéresser à l'estimation de la variance σ_θ^2. Le raisonnement est le même. On sait que l'on peut écrire :

$$\sigma_\theta^2 = E_\theta(X^2) - E_\theta^2(X) = m_\theta(2) - (m\theta(1))^2.$$

D'où l'idée d'estimer σ_θ^2 par :

$$S_n^2 = \hat{m}_\theta(2) - (\hat{m}_\theta(1))^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2.$$

Un calcul élémentaire montre que S_n^2 s'écrit aussi sous la forme :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

C'est sous cette forme qu'est plus connu cet estimateur.[5]

Définition 1.14. *On appelle estimateur de la variance empirique, la statistique S_n^2 définie pour une taille n de l'échantillon par :*

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Cette méthode d'estimation empirique des moments est très générale. Elle peut, par exemple, s'appliquer pour l'estimation de la fonction de répartition. Il suffit en effet de remarquer que l'on peut écrire :

$$F_\theta(x) = P_\theta(X <= x) = E_\theta(\mathbf{1}_{\{X<=x\}}) = E(Y),$$

avec $Y = \mathbf{1}_{]-\infty,x]}(X)$.On peut donc estimer $F_\theta(x)$

$$\hat{F}_\theta(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{]-\infty,x]}(X_i),$$

et on retrouve l'estimateur de la fonction de répartition empirique[5].

Méthode des moments :

Soit X le vecteur formé par un n-échantillon $(X_1, ..., X_n)$. Les X_i sont à valeurs dans un ensemble E. Soit $f = (f_1, ..., f_k)$ une application de E dans $\mathbb{R}^k$ telle que l'application :

$$\varphi : \begin{cases} \Theta \to \mathbb{R}^k \\ \theta \mapsto E_\theta[f(X_i)] \end{cases}$$

soit injective. On définit l'estimateur $\hat{\theta_n}$ comme la solution dans Θ (quand elle existe) de l'équation [2] :

$$\varphi(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

La méthode des moments consiste alors à estimer θ par [5] :

$$\hat{\theta_n}(X) = \varphi^{-1}(\frac{1}{n}\sum_{i=1}^{n} f(X_i)).$$

1.7.2 Maximum de vraisemblance :

Définition 1.15. *Soit $(E^n, \mathcal{E}^{\otimes n}, \mathcal{P}^n = \{P_\theta^{\otimes n}, \theta \in \Theta\})$ un modèle statistique paramétrique et X sa v.a générique. On appelle estimateur du maximum de vraisemblance la statistique $\hat{\theta}(X)$ où $\hat{\theta}$ est une application :*

$$\hat{\theta} : \begin{cases} E & \longrightarrow \Theta \\ x & \longmapsto \hat{\theta}(x) \end{cases}$$

telle que :

$$\mathcal{L}(x, \hat{\theta}(x)) \geq \mathcal{L}(x, \theta)$$

Pour tout $\theta \in \Theta$. On note :

$$\hat{\theta}(x) = Arg\max_{\theta} \mathcal{L}(x, \theta)$$

Dans le cas d'un modèle d'échantillonnage la variable générique $\underline{X} = (X_1, ..., X_n)$ et l'estimateur du Maximum de Vraisemblance est

$$\hat{\theta}(\underline{X}) = Arg\max_{\theta} \mathcal{L}(\underline{X}, \theta)$$

Il est bien évident que d'une part l'estimateur du maximum de vraisemblance n'existe pas toujours et que, d'autre part, s'il existe rien ne garantie qu'il soit unique. Si la fonction vraisemblance est concave, on sait que le maximum est unique et atteint en la valeur qui annule la dérivée première (cas unidimensionnel) ou le gradient (cas multidimensionnel). Cette méthode ne peut être utilisée que si l'hypothèse de concavité est vérifiée [5].
Comme la vraisemblance est souvent sous la forme d'un produit (modèle d'échantillonnage) il est généralement plus aisé (pour les dérivations) de travailler avec la logvraisemblance qui est definée comme suit [5] :

Définition 1.16. *On appelle fonction de log-vraisemblance pour $(x_1, ..., x_n)$ la fonction de θ définie par :*

$$l(x_1, ..., x_n; \theta) = \ln(\mathcal{L}(x_1, ..., x_n; \theta))$$

Elle n'a de sens que si θ vérifie $\mathcal{L}(x_1, ..., x_n; \theta) > 0$.
La fonction logarithme népérien étant croissante, l'estimateur de maximum de vraisemnlance $\hat{\theta}$ de θ pour $(x_1, ..., x_n)$ vérifie [13] :

$$\hat{\theta} \in \arg\max_{\theta} \mathcal{L}(x_1, ..., x_n; \theta) = \arg\max_{\theta} l(x_1, ..., x_n; \theta).$$

1.7.3 Qualité d'un estimateur :

On a vu quelques techniques pour construire des estimateurs, abordons maintenant le problème de l'évaluation de la qualité d'un estimateur et la comparaison d'estimateurs entre-eux. Le but étant bien sûr de prendre le meilleur (s'il existe). Naturellement, on voudra qu'un estimateur possède quelques unes (à défaut de toutes) des qualité suivantes [5].

- Quand la taille d'échantillon augmente, l'estimateur a tendance à se rapprocher (dans un sens à définir) de la valeur $g(\theta)$ qu'il estime. On parlera dans ce cas d'estimateur convergent ou consistant.
- Même si l'estimateur commet une erreur d'estimation à chaque fois, "en moyenne" (en fait en espérance) il ne se trompe pas. On dira dans un tel cas que l'estimateur est sans biais.
- L'estimateur doit être le plus précis possible : les variations de l'estimateur autour de $g(\theta)$ doivent être réduites, voir les plus petites possible. La fonction de risque nous permet de pallier à ce problème [5].

Il y aurait d'autres critères, mais nous n'aurons pas le temps de les étudier.

Estimateur convergent [5] :

Lorsque l'on augmente la taille de l'échantillon, on augmente la quantité d'information dont on dispose sur le phénomène aléatoire que l'on étudie. Aussi, il est assez naturel de souhaiter qu'un estimateur ait tendance à s'approcher de la valeur qu'il estime, lorsque la taille de l'échantillon croît.

Définition 1.17. *Un estimateur $T(X) = (T_n(X))_{n\in\mathbb{N}}$ de $g(\theta)$ est dit (faiblement) convergent ou consistant si la suite $(T_n(X))_{n\in\mathbb{N}}$ converge en probabilité (sous la loi P_θ) vers $g(\theta)$, i.e.*

$$T_n(X) \underset{n\longrightarrow+\infty}{\xrightarrow{P_\theta}} g(\theta)$$

quand $n \longrightarrow +\infty$.

Si $T(X)$ et $g(\theta)$ sont dans $\mathbb{R}$, la définition de la convergence de l'estimateur signifie que l'on a, pour tout $\varepsilon > 0$:

$$P(\mid T_n(X) - g(\theta) \mid > 0) \underset{n\longrightarrow+\infty}{\longrightarrow} 0.$$

Si $T(X)$ et $g(\theta)$ sont dans $\mathbb{R}^p$, la définition de la convergence de l'estimateur s'écrit à partir de la notion précédente sous la forme :

$$\parallel T_n(X) - g(\theta) \parallel \underset{n\longrightarrow+\infty}{\xrightarrow{P_\theta}} 0.$$

Estimateur sans biais[5] :

Définition 1.18. *Le biais d'un estimateur $T(X) = (T_n(X))_{n\in\mathbb{N}}$ de $g(\theta)$ est la fonction b_T définie sur Θ par :*

$$b_T(\theta) = E_\theta(T(X)) - g(\theta),$$

pour tout θ dans Θ et à condition que $E_\theta(T(X))$ existe. Il est dit sans biais si cette fonction est identiquement nulle, i.e. si l'on a :

$$E_\theta(T(X)) = g(\theta).$$

Le biais est généralement une fonction de la taille n de l'échantillon et on peut, si nécessaire la noter $b_{n,T}$ dans ce cas. Aussi, si certains estimateurs se trouvent être biaisés pour toute taille finie d'échantillon, on peut espérer qu'ils soient non biaisés asymptotiquement, c'est à dire quand n tend vers $+\infty$.

Définition 1.19. *Un estimateur $T(X) = (T_n(X))_{n\in\mathbb{N}}$ de $g(\theta)$, où $T_n(X)$ est intégrable pour tout n, est dit asymptotiquement sans biais si l'on a :*

$$b_{n,T} = E_\theta(T_n(X)) - g(\theta) \underset{n\longrightarrow+\infty}{\longrightarrow} 0 \qquad \text{pour tout } \theta \text{ dans } \Theta.$$

Risque d'un estimateur :

Une autre manière de mesurer la qualité d'un estimateur est d'évaluer sa précision [5].

Définition 1.20. *On appelle risque d'un estimateur $T(\underline{X}) = (T_n(\underline{X}))_{n\in\mathbb{N}}$ de $g(\theta)$,la fonction R de Θ définie par :*

$$R(T(\underline{X}), \theta) = E_\theta(T(\underline{X}), \theta) \qquad \text{pour tout } \theta \text{ de } \Theta,$$

sous réserve que cette espérance existe [2].

Proposition 1.2. *Soit $T(\underline{X}) = (T_n(\underline{X}))_{n\in\mathbb{N}}$ de $g(\theta) \in \mathbb{R}$, de carré intégrable pour la loi P_θ. Dans le cas d'un risque quadratique on a :*

$$R(T(X), \theta) = Var_\theta(T(X)) + b_T^2(\theta),$$

Pour un estimateur sans biais, le risque quadratique est donc égal à sa variance [2].

Définition 1.21. *Soient $S(\underline{X})$ et $T(\underline{X})$ deux estimateurs de $g(\theta)$. On dit que $T(\underline{X})$ est préférable à $S(\underline{X})$ si l'on a :*

$$R(T(\underline{X}), \theta) \leq R(S(\underline{X}), \theta) \qquad \text{pour tout } \theta \text{ de } \Theta \ [2].$$

Information de Fisher [13] :

Définition 1.22. *on appelle information de Fisher fournit par $(X_1, ..., X_n)$ sur le réel :*

$$I_n(\theta) = nI_1(\theta)$$

Avec :

$$I_1(\theta) = E_\theta((\frac{\partial}{\partial\theta}\ln\mathcal{L}(X,\theta))^2)$$

Dans cette expression, $\frac{\partial}{\partial\theta}(\ln\mathcal{L}(X,\theta))$ désigne la dérivée partielle de $\mathcal{L}(X,\theta)$ en θ.

On peut établir une autre écriture de l'information de Fisher. L'information de Fisher est aussi égale à :

$$I(\theta) = -E(\frac{\partial^2}{\partial\theta^2}\ln\mathcal{L}(X,\theta))$$

$$I(\theta) = V(\frac{\partial}{\partial\theta}\ln\mathcal{L}(X,\theta))$$

1.8 Processus aleatoire :

Un processus aléatoire est défini par la mise en correspondance des résultats d'une experience avec une fonction du temps (ou de plusiurs autres variables aléatoires).

Définition 1.23. *Un processus stochastique $X = (X_t)_{t\in T}$ est une famille de variables aléatoires X_t indéxée par un ensemble T.*
En général $T = \mathbb{R}$ ou $\mathbb{R}_+$ et on considère que le processus est indéxé par le temps t.
Si T est un ensemble fini,le processus est un vecteur aléatoire. Si $T = \mathbb{N}$ alors le processus est une suite de variables aléatoires. Plus généralement quand $T \subset \mathbb{Z}$, le processus est dit discret.
Un processus dépend de deux paramètres : $X_t(\omega)$ dépend de t (en général le temps) et de l'aléatoire $\omega \in \Omega$.
Pour $t \in T$ fixé, $\omega \in \Omega \longrightarrow X_t(\omega)$ est une variable aléatoire sur l'espace de probabilité $(\Omega, \mathcal{F}, \mathcal{P})$.
Pour $\omega \in \Omega$ fixé, $t \in T \longrightarrow X_t(\omega)$ est une fonction à valeurs réelles, appelée trajectoire du processus. C'est un enjeu que de savoir si un processus admet des trajectoires mesurables, continues et dérivables ou encore plus régulières. On prend $T = \mathbb{R}$ ou $\mathbb{R}_+$ [12].

1.8.1 Processus gaussien :

Variables gaussiennes :

Définition 1.24. *Une v.a. X suit la loi normale standard $N(0,1)$ si elle admet pour densité :*

$$t \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-t^2/2}$$

De façon générale, une v.a X suit la loi normale $N(m,\sigma^2)$ si elle admet pour densité :

$$t \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} exp(\frac{-(t-m)^2}{2\sigma^2}).$$

$\sigma^2 = 0$, la loi est dégénérée, v.a X est égale à m [12].

Proposition 1.3. *Une v.a X de loi $N(m,\sigma^2)$ a pour [12] :*
- *Espérance : $\mathbb{E}[X] = m$.*
- *Variance : $Var(X) = \sigma^2$.*
- *Fonction caractéristique : $\phi_X(t) = exp(imt - \sigma^2 t^2/2)$.*

Vecteur gaussien :

Définition 1.25. *Un vecteur aléatoire $X = (X_1, ..., X_n)$ est gaussien ssi toutes les combinaisons linéaires de ses coordonnées $\langle a, X\rangle = a_1X_1 + ... + a_nX_n$ suivent une loi gaussienne dans $\mathbb{R}$ (pour tout $a = (a_1, ..., a_n) \in \mathbb{R}^n$) [12].*

Définition 1.26. *La matrice de covariance d'un vecteur aléatoire $X = (X_1, ..., X_n)$ est la matrice carrée symétrique, positive*

$$K = (Cov(X_i; X_j))_{1\leq i,j\leq n}.$$

L'espérance de $X = (X_1, ..., X_n)$ est le vecteur des espérances de ses marginales

$$E[X] = (E(X_1), ..., E(X_n)).$$

Si $E[X] = 0$, le vecteur X est dit centré [12].

Proposition 1.4. *La fonction caractéristique d'un vecteur gaussien $X = (X_1, ..., X_n)$ est donnée par [12]*

$$\begin{aligned}\phi_X(x) &= exp(i\,\langle x, E(X)\rangle - \frac{1}{2}(x^t)Cov(X)x) \\ &= exp(i\,\langle x, E(X)\rangle - \frac{1}{2}\,\langle x, Cov(X)x\rangle).\end{aligned}$$

Remarque [12] :

— La loi d'un vecteur gaussien est connue dès qu'on a le vecteur moyenne $E[X]$ et la matrice de covariance $Cov(X)$.

— On parle du vecteur gaussien standard en dimension n lorsque $E[X] = 0$ et $Cov(X) = I_n$. Sa fonction caractéristique est alors

$$\phi_X(x) = exp(-\langle x, x/2\rangle = exp \parallel x \parallel^2 /2).$$

Densité gaussienne en dimension n [12] : Soit $X \simeq N(0, I^n)$ un vecteur gaussien standard en dimension n La densité d'un vecteur gaussien standard en dimension n est

$$f_X(x) = \frac{1}{\sqrt{2\pi}^2} exp(-(x_1^2 + ... + x_n^2)/2).$$

Proposition. *[12] La densité d'un vecteur gaussien $X \simeq N(m, K)$ non dégénéré est*

$$f_X(x) = \frac{exp(-\langle (x-m), K^{-1}(x-m)\rangle /2)}{((2\pi)^n det(K)^{1/2}}.$$

Processus Gaussien :

Définition 1.27. *Un processus est dit gaussien si toutes ses lois fini dimensionnelles $L(X_{t_1}, .., X_{t_n})$ sont gaussiennes ($\forall n \in \mathbb{N}, \forall t_1, ..., t_n \in T$). Autrement dit $X = (X_t)_{t\in T}$ est gaussien si toute combinaison linéaire $a_1X_{t_1} + ... + a_nX_{t_n}$ suit une loi gaussienne (pour tout $n \in N, t_1, ..., t_n \in T$ et $a_1, ..., a_n \in \mathbb{R}$) [12].*

Il est connu que la loi d'un vecteur gaussien $(X_{t_1}, .., X_{t_n})$ est connue (via sa fonction caractéristique) par le vecteur moyenne $(E[X_{t_1}], ..., E[X_{t_n}])$ et la matrice de covariance $(Cov(X_{t_i}, X_{t_j})_{1\leq i;j\leq n})$. On comprend dès lors que toute la loi d'un processus gaussien est connue dès qu'on se donne la fonction moyenne $a(t) = E[X_t]$ et l'opérateur de covariance $K(s,t) = Cov(X_s, X_t)$. En effet, la loi finie multidimensionnelle de $(X_{t_1}, .., X_{t_n})$ est alors la loi normale de dimension n $N(a_n, K_n)$ avec $a_n = (a(t_1), ..., a(t_n))$ et $K_n = (K(t_i; t_j))_{1\leq i;j\leq n}$. Les fonctions a et K définissent donc toutes les lois fini multidimensionnelles de X [12].

Remarque [12] :

— Toutes les lois marginales d'un processus gaussien sont bien sûr gaussiennes.

— Toute combinaison linéaire des lois marginales d'un processus gaussien est encore gaussienne.

Théorème 1.3. *[12] La donnée de la fonction d'espérance m et de la fonction de covariance C suffit à caractériser un processus gaussien, plus précisement, si on suppose*

1. *$X = (X_t)_{t\in T}$ et $Y = (Y_t)_{t\in T}$ sont deux processus gaussiens (indexès par T).*
2. *Pour tous $t \in T$, $E[X_t] = m(t) = E[Y_t]$.*
3. *Pour tous $t, s \in T$, $Cov(X_t, X_s) = C(t,s) = Cov(Y_t, Y_s)$, alors les processus X et Y ont même loi.*

1.8.2 Processus de Dirichlet :

Loi de Dirichlet [15] :

Avant de parler du processus de Dirichlet, rappelons tout d'abort la définition et les propriétés de la distribution de Direchlet. La distribution de Dirichlet est une généralisation de la loi Bêta. Pour rappel, les familles de lois Bêta est multinomiales. On peut définir la famille des lois de Dirichlet comme la conguguée de la famille des lois multinomials.

Définition 1.28. *Un vecteur aléatoire $Y = (y_1, ..., y_K)$ à valeur dans l'ensemble $\Delta_K \subset \mathbb{R}^K$:*

$$\Delta_K = \left\{ (y_1, ..., y_K); y_k > 0, k = 1, 2...K, \quad \sum_{i=1}^{K} y_i = 1 \right\}$$

suit une loi de Dirichlet de patamètres $\alpha_1, ..., \alpha_K > 0$, si sa densité de probrbilité par rapport à la mesure de lebesgue s'écrit :

$$\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} y_k^{\alpha_k - 1}, \tag{1.2}$$

Remarque :

- Si $Y = (y_1, ..., y_K)$ suit une loi de Dirichlet, notée $Y \sim Dir(\alpha_1, ..., \alpha_K)$, alors les $k-1$ premiére composantes de y possédent la distribution définie précédemment y vérifie $y_K = 1 - y_1 - ... - y_{K-1}$
- Quand k vaut 2, on retrouve une loi Bêta de paramétre (α_1, α_2) sur léquation (1.2). C'est pour quoi on parle d'une généralisation de la loi Bêta.

Propriété 1.1. *[15] Soit $y = (y_1, ..., y_K)$ une v.a distribuée suivant une loi de Dirichlet de paramétre $(\alpha_1, ..., \alpha_K)$, notons $\alpha = \sum_{i=1}^{k} \alpha_i$ alors : $E[y_k] = \frac{\alpha_k}{\alpha}$, $Var(y_k) = \frac{\alpha_k(\alpha - \alpha_k)}{\alpha^2(\alpha + 1)}$*

Processus de Dirichlet [15] :

Le procesus de Dirichlet peut s'interpréter comme une généralisation de la loi de Direchet où k est infini. Il permet de définir une distribution sur un ensemble de distributions de probabiltés. Ce processus est défini à partir d'un coefficient de précision (paramétre d'échelle) $\alpha > 0$ et d'une mesure de base G_0 qui est une loi de probabilité. Nous le notons par $DP(G_0, \alpha)$.

Définition 1.29. *Soit α un réel positif. Soit (Θ, I) un espace mesurable et G_0 une mesure de probabilité sur (Θ, I). On dit q'une distribution de probabilité G est distribuée selon un procesus de Dirichlet de distribution de base G_0 et de facteur d'échelle $\alpha > 0$ si pour toute partition mesurable $(A_1, ..., A_k)$ de Θ, le vecteur de probabilité aléatoire $(G(A_1), ..., G(A_k))$ suit une distribution de Dirichlet standard : on le note $G \sim DP(G_0, \alpha)$ [15].*

Propriété 1.2. *[15] Soit G une mesure aléatoire distribuée suivant un processus de Dirichlet. Alors pour tout A élément de la tribu $\mathcal{A}$, la moyenne et la variance de $G(A)$ sont les suivantes :*

$$E[G(A)] = G_0(A).$$

$$Var(G(B)) = \frac{G_0(B)(1 - G_0(B))}{\alpha_0 + 1}.$$

Chapitre 2

Inférence statistique par l'approche bayésienne

Introduction :

L'approche bayésienne fournit un cadre naturel pour résoudre les problèmes d'inférence statistique. Elle se distingue de la statistique classique parce qu'elle considère les paramètres du modèle comme des variables aléatoires. Dans le cas où le nombre de paramètres du modèle étudié est connu ou fini, on parle d'approche paramétrique. Dans le cas contraire où le nombre de paramètres est inconnu ou infini, on parle alors d'approche non paramétrique.

2.1 Cadre général de la méthode bayésienne paramétrique :

Considérons un modèle statistique $(E; A; P)$, où E est l'espace des observations, A une tribu sur E et P une famille de mesures de probabilité connu sur $(E; A)$. On écrit généralement la famille P sous la forme :

$$P = \{P_{\theta,\theta\in\Theta}\}$$

Le but de l'analyse statistique paramétrique est de faire de l'inférence sur $g(\theta)$, avec g une fonction définie sur Θ à valeurs dans Θ. En général, on définit g comme étant la fonction identité sur Θ. L'idée centrale de l'analyse bayésienne est de considérer le paramètre inconnu θ comme aléatoire : l'espace des paramètres Θ est muni d'une probabilité π tel que $(\Theta; A; \pi)$ est un espace probabilisé. Nous noterons $\theta \sim \pi$ et π est appelée loi à priori. Intuitivement et en termes informationnels, elle détermine ce qu'on sait et ce qu'on ne sait pas avant d'observer X [45].

Définition 2.1. *(Modèle dominé [45]) Le modèle est dit dominé s'il existe une mesure commune dominante μ. C'est-à- dire, pour tout θ, P_θ admet une densité par rapport à μ définit par*

$$f(X|\theta) = \frac{dP_\theta}{d\mu}.$$

Le modèle s'écrit alors sous la forme suivante : $\{f(x|\theta), \theta \in \Theta\}$.

Considérons les informations $X_1, ..., X_n \sim f(x|\theta)$, où $X = (X_1, ..., X_n)$ un échatillon de taille n issu de la variable aléatoire X et $f(.|\theta)$ sa densité de probanilité. La fonction de vraisemblance associée est donée par :

$$\mathcal{L}_n(\theta) = \prod_i f(X_i|\theta)$$

donc le modéle est resumé par :

$$\begin{aligned} \theta &\sim \pi \\ X_1, ..., X_n|\theta &\sim f(x|\theta) \end{aligned} \tag{2.1}$$

2.1.1 Distribution de la loi à priori :

La loi à priori est la clé de l'inférence bayésienne et sa détermination est donc l'étape la plus importante dans la mise en oeuvre de cette inférence[44]. Les informations à priori sur le paramètre θ s'expriment à travers une loi de probabilité appelée loi à priori sur l'espace des parametres Θ. En d'autres termes, Cette loi nous permet de traduire les connaissances dont on disposent sur le paramètre avant l'expérience. On l'interprète comme la représentation formelle sous forme probabiliste de ces informations. Ainsi, Une loi a priori π est une loi de probabilité (densité de probabilité) sur Θ.
On distingue deux types de lois à priori : les lois informatives et les lois non informatives.

Les lois non informatives :

Dans le cas où on dispose que de peu d'informations sur θ, on peut choisir des loi à priori dites peu ou non informatives. On souhaite que l'à priori intervienne de façon minimale dans la loi à posteriori, i.e. que les donnèes parlent d'elles même[7].

Définition 2.2. *Une loi non informative est une loi qui porte une information sur le paramètre à estimer dont le poids dans l'inférence est réduit [3].*

Remarque 2.1. *Certains auteurs définissent la loi non informative comme une loi à priori qui ne contient aucune information sur le paramètre à estimer ou encore comme une loi qui ne donne pas davantage de poids à telle ou telle valeur du paramètre [3].*

Distributions à priori impropres [44] : Lorsque le paramètre θ peut être traité comme une variable aléatoire avec une distribution de probabilité π connue, la distribution à priori est déterminée par des critères subjectifs ou théoriques qui conduisent à une mesure σ-finie sur l'espace des paramètres Θ plutôt qu'à une mesure de probabilité, c'est-à-dire une mesure π telle que

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

Dans de tels cas, on dit que la distribution à priori est impropre (ou généralisée).
Quand une telle loi à priori impropre a été obtenue par des méthodes automatiques, à partir de la densité $f(x|\theta)$, elle paraît plus susceptible aux critiques. Ces approches automatiques sont souvent la seule façon d'obtenir une distribution à priori dans un cadre non informatif. Dans certains cas, l'unique information disponible (ou retenue) est la connaissance de la distribution d'échantillon $f(x|\theta)$.

Il est donc important de prendre plus de précautions quand on a affaire à des lois impropres, afin d'éviter les distributions mal définies. La difficulté pratique est de vérifier la condition d'intégrabilité :

$$\int f(x|\theta)\pi(\theta)d\theta < \infty$$

La loi π est alors un moyen de résumer l'information disponible sur ce phénomène.

Loi à priori Uniforme [3] : La densité à priori non informative la plus simple et la plus communément utilisée est la densité Uniforme. En effet, ce choix repose sur l'équiprobabilité des valeurs possibles du paramètre θ sur son domaine de définition, et donc n'apporte aucune information supplémentaire sur θ. Ainsi, la densité est définie par

$$\pi(\theta) = k, k \text{ est une constante positive.}$$

Loi à priori invariant : Si on passe d'un paramètre $\theta \in \Theta$ au paramètre $\eta = h(\theta) \in h(\Theta)$ par une transformation bijective h, l'information à priori n'est pas modifiée, puisqu'elle est toujours inexistante, donc on devrait aussi utiliser une loi à priori non informative pour η. Il s'agit de l'invariance par reparamétrisation [3].

Invariance pour le paramètre de position : Soit X est une v.a de densité ne dépendant que de $x - \theta$. Dans ce cas, θ est appelé paramètre de position et $X - \theta$ le pivot.

Un invariant à priori par rapport à un paramètre de position est un invariant à priori par rapport au choix de l'origine (position). L'absence d'information à priori pour θ implique nécessairement que le décalage de la position n'a pas d'influence sur l'état de connaissance. On a donc,
$\theta' = \theta + b$ où b est une constante.
La loi à priori doit donc vérifier :$\pi(\theta + b) = \pi(\theta)$.
La seule solution est donnée par la loi Uniforme sur Θ (qui est une loi impropre) [3].

Invariance pour le paramètre d'échelle : Soit X est une v.a de densité ne dépendant que de $\frac{1}{\theta} f(x|\theta)$. Dans ce cas, θ est un paramètre d'échelle.

Un à priori invariant par rapport à un paramètre d'échelle est un à priori qui est invariant pour la mesure d'échelle. L'absence d'information à priori pour θ implique nécessairement l'invariance par changement d'échelle $\theta' = a\theta$ où a est une constante.
La loi à priori doit donc vérifier :$\pi(a\theta) = a\pi(\theta)$
La seule solution est de prendre une densité à priori telle que $\pi(\theta) = \frac{k}{\theta}$ (impropre) [3].

Règle de Jeffreys [3] : Jeffreys en 1961 [29] propose une méthode de construction de lois à priori non informative, en se basant sur le principe d'invariance par transformation, utilisant l'information de Fisher Euler qui représente une mesure de la quantité d'informations sur θ contenue dans l'observation .

Définition 2.3. *Soit θ un paramètre réel. On appelle loi à priori non informative de Jeffreys la loi (éventuellement impropre) de densité :*

$$\pi_j(\theta) \propto \sqrt{I_X(\theta)} 1_\Theta(\theta),$$

au encore, $\pi_j(\theta) = C\sqrt{I_X(\theta)}$, C est une constante et $I_X(\theta)$ l'information de Fisher apportée par X sur le paramètre θ, définie par :

$$I_X(\theta) = E[(\frac{\partial}{\partial\theta} \ln f(x|\theta))^2] \tag{2.2}$$

si le domaine de X est indépendant de θ alors

$$I_X(\theta) = -E[\frac{\partial^2}{\partial\theta^2} \ln f(x|\theta)] \tag{2.3}$$

Remarque :

1. Plus $I_X(\theta)$ est grande, plus l'observation apporte de l'information. Il est donc naturel de favoriser les valeurs de θ pour lesquelles $I_X(\theta)$ est grande, ce qui rend la loi $\pi(\theta)$ plus probable et qui minimise l'influence de la loi à priori au profit de l'observation.
2. Ce type de construction de lois à priori non informatives conduit très souvent à des lois à priori impropres (appelées aussi quasi à priori).
3. Dans le cas d'un échantillon de taille n, $\pi_J(\theta) \propto (I_n(\theta))^{\frac{1}{2}}$.
4. L'à priori de Jeffreys offre une méthode automatisée pour obtenir une loi à priori non informatif pour n'importe quel modèle paramétrique.
5. L'à priori de Jeffreys est invariant par transformation bijective.

Prior informatif :

L'un des critéres de choix pour ces lois est de simplifier les calculs, d'où l'utilisation répendue de distributions naturelles conjuguées à un modéle d'échantillonnage [16]. Ces familles étaient pratiquement les seules qui permettaient de faire aboutir des calculs [7].

Définition 2.4. *(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 à posteriori $\pi(.|x)$ appartient également à $\mathcal{F}$[6].*

Les lois à priori conjuguées sont généralement associées à un type particulier de lois d'échantillonnage qui permettent toujours leur obtention, il est mème caractéristique des lois à priori conjuguées comme nous le verrons ci dessous. Ces lois constituent ce qu'on appelle des familles exponentielles [6].

Définition 2.5. *Soient μ une mesure σ-finie sur X, Θ l'espace des paramètres, C et h des fonctions respectivement de X et Θ dans $\mathbb{R}^+$, et R et T des fonctions de Θ et X dans $\mathbb{R}^k$. La famille des distributions de densité (par rapport à μ)*

$$f(x|\theta) = C(\theta)h(x)exp\{R(\theta).T(x)\} \tag{2.4}$$

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

$$f(x|\theta) = C(\theta)h(x)exp\{\theta.x\}, \tag{2.5}$$

la famille est dite naturelle [16].

Notons qu'un changement de variable de x en $z = T(x)$ et une reparamétrisation de θ en $\eta = R(\theta)$ nous permettent de considérer principalement la forme naturelle [8], bien que les espaces $T(X)$ et $R(\theta)$ puissent être difficiles à décrire et à utiliser.

2.1.2 Distribution à posteriori :

La distribution à posteriori est la quantité la plus importante dans l'inférence bayésienne. Elle contient toutes les informations disponibles sur le paramètre θ inconnu après avoir observé les données $X = x$. Soit $X = x$ la réalisation observée d'une variable aléatoire X (éventuellement multivariée) avec la fonction de densité $\pi(x|\theta)$ [31].
Soit $\pi(\theta)$ la distribution à priori qui nous permet de calculer la fonction de densité $\pi(\theta|x)$ de la distribution à posteriori en utilisant le théorème de Bayes [31] :

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

où :

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

est la loi marginale de X.

Le raisonnement proportionnel [19] :

Il est parfois possible d'éviter le calcul de l'intrégrale : $\int_{\Theta} f(x|\theta)\pi(\theta)d\theta$ en raisonnant proportionellement.

Définition 2.6. *Soient deux fonctions réelles f et g définies sur le même espace $\mathcal{Y}$. On dit que f et g sont proportionnelles, ce qu'on note $f \propto g$, s'il existe une constante a tel que $f(y) = ag(y)$ pour tout $y \in \mathcal{Y}$. Il est clair que la relation $\propto$ est une relation d'équivalence. En particulier : $f \propto g$ et $g \propto h$ entrainent $f \propto h$.*

Remarques :

1. Soit $f(y)$ la densité d'une variable aléatoire Y de loi inconnue. Si $f \propto u_1 ... \propto u_k \propto g$, où $u_1 ... u_k$ désigent des fonctions réelles et $g(y)$ est la densité d'une loi de probabilité P, alors $Y \sim P$.
2. Dans un contexte bayésien on a : $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$. En tant que fonctions de θ, les deux expressions $\pi(\theta|x)$ et $f(x|\theta)$ sont effectivement proportionnelles ; la constante a qui apparaît dans la définition est égale ici à $1/m(x)$; à noter que cette quantité est bien une constante, au sens où elle ne dépend pas de θ. L'écriture $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$ est souvent reformulée de la façon suivante :

$$\pi(\theta|x) \propto \mathcal{L}(\theta, x)\pi(\theta)$$

2.1.3 La Théorie de la Décision bayésienne :

La recherche d'estimateurs de Bayes peut se faire dans le cadre de la théorie de la décision. La démarche consiste alors à se fixer des critères pour evaluer les décisions et à chercher un estimateur optimal au sens de cette règle de préférence. On défine ce critère appelé fonction de coût où perte. L'ensemble des décisions possibles D est appelé espace de décision et la plupart des exemples théoriques se concentrent sur le cas $D = \Theta$, qui représente le cadre d'estimation standard [44].

Définition 2.7. *La fonction de perte est une fonction mesurable de $(\Theta \times D)$ à valeurs réelles positives :*

$$L : \Theta \times D \longmapsto \mathbb{R}^+.$$

La fonction de coût est censée évaluer la pénalité (ou l'erreur) $L(\theta, \delta)$ associée à la décision δ quand le paramètre prend la valeur θ[44].
Une base fondamentale de la Théorie de la Décision bayésienne est que l'inférence statistique devrait commencer par la détermination de trois facteurs [44] :

(1) la famille des distributions pour les observations $f(x|\theta)$.
(2) la distribution à priori pour les paramètres $\pi(\theta)$.
(3) la fonction de perte associé aux décisions $\mathcal{L}(\theta, \delta)$.

D'autre part, pour obtenir un critère de comparaison utilisable à partir d'une fonction de coût dans un contexte aléatoire, l'approche fréquentiste propose de considérer plutôt le coût moyen (ou risque fréquentiste) qui est une fonction du paramètre θ :

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_{\theta}[L(\theta, \delta(x))] \\ &= \int L(\theta, \delta(x)) f(x|\theta)dx \end{aligned}$$

où $\delta(x)$ est la règle de décision, soit l'attribution d'une décision à chaque résultat $x \sim f(x|\theta)$ de l'expérience aléatoire. La fonction δ, de X dans D, est habituellement appelée estimateur, tandis que la valeur $\delta(x)$ est appelée estimation de θ. Nous noterons aussi D l'ensemble des estimateurs. D'autre part, on intègre sur l'espace Θ car θ est inconnu, plutôt que de le faire sur l'espace X , x étant connu. Donc on a le coût moyenne à posteriori :

$$\varrho(\pi, \delta|x) = \mathbb{E}[L(\theta, \delta)|x] = \int_\Theta L(\theta, \delta)\pi(\theta|x)d\theta$$

qui moyenne le coût selon la distribution à posteriori du paramètre θ, conditionnellement à la valeur observée x.
En se donnant une distribution à priori π, il est aussi possible de définir le risque intégré, qui est le risque fréquentiste moyenné sur les valeurs de θ selon leur distribution à priori [44].

$$r(\pi, \delta) = \mathbb{E}^\pi[R(\theta, \delta)] = \int_\Theta \int_X L(\theta, \delta(x))f(x|\theta)dx\pi(x)d\theta$$

Un intérêt particulier de ce deuxième concept est qu'il associe un nombre réel à chaque estimateur, et non une fonction de θ.

Théorème 2.1. *[44] Un estimateur minimisant le risque intégré $r(\pi, \delta)$ est obtenu par sélection, pour chaque $x \in X$, de la valeur $\delta(x)$ qui minimise le coût moyen à posteriori, $\varrho(\pi, \delta|x)$, puisque*

$$r(\pi, \delta) = \int_X \varrho(\pi, \delta|x)m(x)dx \tag{2.6}$$

en effet : L'égalité (2.6) découle directement du Théorème car, comme $L(\theta, \delta) \geq 0$

$$\begin{aligned} r(\pi, \delta) &= \int_\Theta \int_X L(\theta, \delta(x))f(x|\theta)dx\pi(\theta)d\theta \\ &= \int_X \int_\theta L(\theta, \delta(x))f(x|\theta)\pi(\theta)d\theta dx \\ &= \int_X \int_\Theta L(\theta, \delta(x))\pi(\theta|x)d\theta m(x)dx \end{aligned}$$

Ce résultat mène à la définition suivante d'un estimateur de Bayes.

Définition 2.8. *Un estimateur de Bayes associé à une distribution à priori π et une fonction de perte L, est un estimateur δ^π minimisant $r(\pi, \delta)$. Pour chaque $x \in X$, ce dernier est donné par*

$$\delta^\pi(x) = arg\min_d \varrho(\pi, d|x)$$

La valeur $r(\pi) = r(\pi, \delta^\pi)$ est alors appelée risque de Bayes.
Le Théorème fournit ainsi un outil constructif pour la détermination des estimateurs de Bayes.

2.1.4 Les fonctions de perte :

L'estimateur de Bayes de θ dépend de la fonction de perte choisie.

Fonction de perte quadratique [19] :

La fonction de perte quadratique est la fonction définie par :

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

L'estimateur de Bayes $\delta^\pi(x)$ de θ associé à la loi à priori π est la moyenne à posteriori de θ :

$$\delta^\pi(X) = E^\pi(\theta|X).$$

En effet, Comme la norme au carré est une fonction convexe deux fois dérivable sur Θ, alors pour trouver δ^π l'estimateur bayésien, il suffit de déterminer les points critiques du risque à posteriori. i,e : $\rho(\pi|\delta|X) = E^\pi(L(\theta,\delta(X))|X) = \int_\Theta L(\theta,\delta(X)d\pi(\theta|X)$

$$\begin{aligned} \rho(\pi|\delta|X) &= E^\pi((\theta-\delta)^2|X) \\ \frac{\partial\rho(\pi|\delta|X)}{\partial\delta} &= -2\int_\Theta(\pi-\delta(X))d\pi(\theta,X) \\ \text{donc} \quad \frac{\partial\rho(\pi|\delta|X)}{\partial\delta} &= 0 \Leftrightarrow \delta(X) = E^\pi(\theta|X) \end{aligned}$$

D'aprés l'inégalité de Jensen, $\delta \longmapsto \rho(\pi|\delta|X)$ est aussi convexe. D'où l'estimateur bayésienne vaut $\delta^\pi(X) = E^\pi(\theta|X)$.

Fonction de perte absolue [19] :

La fonction de perte absolue est la fonction définie par :

$$L(\theta,\delta) = \sum_{i=1}^{d} \mid \theta_i - \delta_i \mid$$

L'estimateur bayésienne associé pour le cas simple où $\delta = 1$:

$$\rho(\pi|\delta|X) = \int_\Theta(\theta-\delta)\pi(\theta|x)d\theta$$

Comme $\delta \longmapsto \rho(\pi|\delta|X)$ est convexe et dérivable, il suffit là encore de déterminer les points critiques :

$$\begin{aligned} \frac{\partial\rho(\pi|\delta|X)}{\partial\delta} &= \int_{-\infty}^{\delta}\pi(\theta|x)d\theta - \int_\delta^\infty \pi(\theta|x)d\theta \\ \text{Donc} \quad \frac{\partial\rho(\pi|\delta|X)}{\partial\delta} &= 0 \Leftrightarrow P^\theta(0\le\delta|X) = P^\theta(0\ge\delta|X) \end{aligned}$$

Fonction de perte de Dirac [19] :

Cette fonction de perte est utilisée dans le contexte des tests. Un test est la donnée d'une partition de Θ en Θ_0 et Θ_1. $\theta \in \Theta_i$ correspond à l'hypothèse H_i et H_0 est appelée l'hypothèse nulle. Le principe du test (décisions) δ est défini comme suit :

$$\delta = \begin{cases} 0 & si \quad \theta \in \Theta_1 \\ 1 & si \quad \theta \in \Theta_0 \end{cases}$$

La fonction de perte correspondant au test est définie par :

$$L(\theta,\delta) = 1_{\theta\in\Theta_1} \times 1_{\delta=0} + 1_{\theta\in\Theta_0} \times 1_{\delta=1}$$

Le risque à posteriori est alors le suivant :

$$\rho(\pi|\delta|X) = 1_{\delta=0}P^\pi(\Theta_1|X) + 1_{\delta=1}P^\pi(\Theta_0|X)$$

Ainsi :

$$\delta^\pi = 1 \Leftrightarrow P^\pi(\Theta_0|X) \le P^\pi(\Theta_1|X)$$

C'est-à-dire que l'estimation permet d'accepter H_0 si c'est l'hypothèse la plus probable à posteriori, ce qui est une réponse naturelle.
Une variante du test 0 si la pénalité associée à un estimateur et 1 si nn sont des tests proposée par Neyman et Pearson qui permet de distinguer le risques de première et de deuxième espèce :

$$L(\theta,\delta) = \begin{cases} 0 & si \quad \theta \in \Theta_1 \\ a_0 & si \quad \theta \in \Theta_0 \\ a_1 & si \quad \theta \in \Theta_0 \end{cases}$$

Le risque à posteriori est donné par :

$$\rho(\pi|\delta|X) = a_0 \delta P^{\pi}(\Theta_1|X) + a_1(1-\delta)P^{\pi}(\Theta_0|X)$$

Ainsi :

$$\delta^{\pi} = 1 \Leftrightarrow a_0 P^{\pi}(\Theta_0|X) \leq a_1 P^{\pi}(\Theta_1|X)$$

Ceci permet de modéliser des raisonnements de la forme « j'ai plus ou moins tort » en jouant sur le rapport $\frac{a_1}{a_0}$ et accorder plus ou moins d'importance relative à Θ_0 selon des arguments à priori.

Il existe d'autres fonction de pertes comme [31] :

— **Fonction de perte de DeGroot** : la fonction est définie par

$$L(\theta,\delta(x)) = (\frac{\theta-\delta(x)}{\delta(x)})^2$$

sous cette fonction l'estimateur de bays est :

$$\delta_{\pi}(x) = \frac{E^{\pi}(\theta^2|x)}{E^{\pi}(\theta|x)}$$

— **La fonction de perte Linex** :la fonction de perte linex pour θ est définie par

$$L(\Delta) \propto e^{a\Delta} - a\Delta - 1, a \neq 0$$

Où : $\Delta = (\delta(x) - \theta)$ où $\delta(x)$ est un estimateur de θ.
l'estimateur de Bayes sous la fonction de perte Linex est :

$$\delta(x) = -\frac{1}{a}\log(E_{\theta}(e^{-a\theta}))$$

étant donné que $E_{\theta}(e^{-a\theta})$ existe et finie.

— **Fonction de perte d'entropie** : la fonction de perte entropie est définie par

$$L_E(\theta,d) \propto (\frac{d}{\theta})^p - p\ln(\frac{d}{\theta}) - 1$$

L'estimateur de Bayes de paramètre θ sous cette fonction de perte est

$$\delta(x) = (E_{\theta}(\theta^{-p}))^{\frac{-1}{p}}$$

a)Lorsque p = 1, l'estimateur de Bayes coïncide avec l'estimateur de Bayes sous la fonction de perte quadratique pondéré $\frac{(d-\theta)^2}{\theta}$.
b)Lorsque p = -1, l'estimateur de Bayes coïncide avec l'estimateur de Bayes sous la foncion de perte quadratique.

2.2 Le cadre générale de l'approche bayésienne nonparamétrique :

Considérons un modèle stochastique $(E; A; P)$, où E est l'espace des observations, A une tribu sur E et $P = \{P_\theta, \theta \in \Theta\}$ une famille de mesures de probabilité sur $(E; A)$. En inférence bayésienne non paramétrique, le problème posé est celui d'estimation de $g(P_\theta)$, c'est à dire l'objet sur lequel on infère une fonctionnelle de la distribution du modèle considéré[1]. En d'autres termes, on remplace le modèle paramétrique à dimension fini (2.1) par le modèle à dimension infini[30] :

$$\mathcal{F} = \{f : \int (f''(x))^2 < \infty\}$$

alors l'à postériori peut être trouvé par le théorème de Bayes :[30]

$$\pi_n(A) \equiv \mathbb{P}(f \in A|Y) = \frac{\int_A \mathcal{L}_n(f)d\pi(f)}{\int_{\mathcal{F}} \mathcal{L}_n(f)d\pi(f)}$$

où $A \subset \mathcal{F}$,$\mathcal{L}_n(f) = \prod_i f(X_i)$ est est la fonction de vraisemblance et π la fonction à priori associée sur $\mathcal{F}$.
Pour trouver la loi à priori sur un espace dimentionnel infini, supposons que nous observions $X_1, ..., X_n \sim F$ où F est une fonction inconnue. Nous mettrons la loi à priori sur l'espace fonctionnel $\mathcal{F}$, il y a des cas où on ne peut pas écrire explicitement une formule pour π comme on fait dans le modèle paramétrique.
Le modèle bayésien peut s'écrire comme suit :

$$F \sim \pi$$
$$X_1, ..., X_n|F \sim F$$

Le modèle et la loi à priori induisent une distribution marginale m pour $X_1, ..., X_n$ [30]

$$m(A) = \int \mathbb{P}_F(A)d\pi(F)$$

Où :

$$\mathbb{P}_F(A) = \int I_A(x_1, ..., x_n)dF(x_1)...dF(x_n)$$

aprés avoir observer les informations $X_1, ..., X_n$ on aura la loi à posteriori associée [30] :

$$\pi_n(A) \equiv \pi(F \in A|X_1, ..., X_n)$$

Pour définir le modèle bayésienne non paramétrique on doit définir la distribution à priori pour l'espace de dimension infini, cette distribution est un processus stochastique.

2.2.1 Processus de Dirichlet et mélanges :

Les processus de Dirichlet définissent une mesure de probabilité sur l'espace des mesures de probabilité. Ils permettent donc de définir, dans le cadre de l'estimation bayésienne, un à priori sur une distribution de probabilité inconnue. On pourra se référer aux articles de Ferguson ([23]) et ([25]), Sethuraman ([46]), Teh ([47]) pour de plus amples détails. Une propriété importante est que les réalisations d'un processus de Dirichlet sont discrètes, avec une probabilité égale à 1 [14].

Si $G \sim DP(\alpha_0, G_0)$, alors la représentation Stick-Breaking de G introduite par Sethuraman ([46]) est simplement de la forme suivante :

$$\begin{cases} V_k \overset{iid}{\sim} Beta(1,\alpha) & \theta_k \overset{iid}{\sim} G_0 \\ P_k = V_k \prod_{j=1}^{k-1}(1-V_j) & G(.) = \sum_{k=1}^{\infty} P_k \delta_{\theta_k(.)}. \end{cases}$$

où $\delta_{\theta_k}(.)$ est la mesure de Dirac en θ_k. Il est important de noter que la suite $p = (p_k)_{k\in N^*}$ satisfait que les p_k sont des variables aléatoires indépendantes de θ_k tel que $0 \leq P_k \leq 1$ et que $\sum_{k=1}^{\infty} P_k = 1$. Le processus de Dirichlet peut donc être vu comme un mélange infini dénombrable de mesures de Dirac. Une motivation fondamentale pour l'utilisation du processus de Dirichlet est que la loi à posteriori est également un processus de Dirichlet.
Le processus de Dirichlet dans un mélange peut intervenir comme une loi mélangeante. Nous parlerons indifféremment de PDM, aussi de PDH (Process de Dirichlet Hiérarchique). Une des plus importantes applications du processus de Dirichlet est son utilisation comme loi à priori dans les composantes d'un modèle de mélanges non paramétrique. Muller et Quintana ([40]) discutent de l'analyse bayésienne non paramétrique et comparent différentes variantes ou généralisations du processus de Dirichlet. Soit les observations y_i sont définies comme suit :

$$\begin{aligned} y_i|\theta_i &\sim k(.|\theta_i) \\ \theta_i|G &\sim G \end{aligned}$$

où $k(.|\theta_i)$ dénote la loi des observations y_i conditionnellement à θ_i.

En choisissant comme loi à priori pour G, le processus de Dirichlet, on peut reformuler le problème d' estimation de la densité selon le modèle hiérarchique (MH) suivant et connu sous le nom de MDP [46].

$$\begin{aligned} y_i|\theta_i &\sim k(.|\theta_i) \\ \theta_i|G &\sim G \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

Cette écriture peut-être reformulée comme l'équation suivante avec $G \sim Dir(\alpha, G_0)$[46] :

$$f(.) = \int_{\Theta} k(.|\theta) dG(.) \tag{2.7}$$

ou :
- $f(.)$ une mélange (Mixture).
- $k(.|\theta)$ une loi mélangée (Mixture law).
- $G(\theta)$ loi mélangeante (Mixing law).

2.2.2 Processus Stick-Breaking et extensions :

Ishwaran et James ([28]) ont introduit une classe de processus appelée Stick-Breaking Priors(SBP) construite de la façon suivante :

$$\begin{cases} V_k \overset{ind}{\sim} Beta(a_k, b_k) & \theta_k \overset{i.i.d}{\sim} G_0 \\ p_k = V_k \prod_{j=1}^{k-1}(1-V_j) & G(.) = G_{\infty}(1,\alpha,.) = \sum_{k=1}^{\infty} p_k \sigma_{\theta_k}(.). \end{cases} \tag{2.8}$$

Les processus appartenant à cette classe diffèrent dans la façon d'obtenir les coefficients V_k. On peut citer trois exemples de processus :

1. le processus de Dirichlet vu à la partie précédente : $V_k \sim Beta(1, \alpha)$.
2. le processus de Pitman-Yor encore appelé processus de Poisson-Dirichlet à deux paramètres qui a été développée récemment par Pitman et Yor en 1997 [43] :$V_k \sim Beta(a_k, b_k)$ avec

$$\begin{cases} a_k = 1_a, \quad 0 \leq a < 1 \\ \\ b_k = b + ka, \quad b >_a . \end{cases}$$

3. le processus de Beta à deux paramètres :$V_i \sim Beta(a, b)$.
   Dunson et Park ([18]) ont généralisé cette classe de processus stick-breaking en introduisant les processus stick-breaking à noyaux (KSBP) :

$$G_x = \sum_{k=1}^{\infty} \pi_k(x, V_k\Gamma_k)G_k^*$$

$$\pi_k(x, V_k\Gamma_k) = \pi_k(x, V_k\Gamma_k) \prod_{\iota<k} (1 - \pi\iota(x, V_\iota, \Gamma_\iota))$$

$$\pi_k(x, V_k\Gamma_k) = V_k.K(x, \Gamma_k) \tag{2.9}$$

$$v_k \overset{ind.}{\sim} Beta(a_k, b_k)$$

$$\Gamma_k \overset{i.i.d}{\sim} H$$

$$G_k^* \sim Q$$

   où $K \longrightarrow [0; 1]$ est une fonction de noyau bornée qui est initialement supposée connue, et x un point quelconque de l'espace étudié. Ce modèle est similaire au (2.7), mais maintenant le SBP est augmenté en employant une fonction de noyau qui quantifie l'à priori associé. Il est employé pour la segmentation d'images en imposant la condition que les pixels voisins dans l'espace sont plus probablement associés dans la même classe .

2.2.3 Processus gamma :

Soit $G(\alpha, \beta)$ la distribution gamma avec $\alpha > 0$ un paramètre de forme et $\beta > 0$ un paramètre d'echelle et soit $\alpha(t), t > 0$ une fonction croissante et continue à gauche tel que $\alpha(0) = 0$ [42].

Définition 2.9. *Soit $Z_t, t \leq 0$ un processus stochastique tel que :*
- *$Z_0 = 0$,*
- *Z_t est un processus à accroissements indépendants et stationnaires.*
- *Pour $t > s$, l'incrément $Z_t - Z_s$ suit une distribution gamma $G(c(\alpha(t) - \alpha(s)), c)$, où $c > 0$ est une constante. Alors Z_t est un processus de gamma avec les paramètre $c\alpha(t)$ et c et on le note $\mathcal{G}(c\alpha(t), c)$.*

le processus stochastique étudié est caractérisé par des incréments independants uniquement positifs , avec la fonction génératrice des moments donnée par :

$$log(E[\exp^{-\theta Z_t}]) = -\theta b(t) + \int_0^{\infty} (e^{-\theta s}) dN_t(s).$$

et la mesure de Lévy sous forme : $dN_t(s) = \alpha(t)e^{-cs}ds/s$ et $b(t) \equiv 0$.

La distribution à posteriori :

Considérons $H \sim \mathcal{G}(cH_0, c)$, où H_0 est la loi à priori supposée sur H. Pour determiner la distribtion à posteriori, on doit résoudre un problème de points fixes à priori liés aux discontinuités du processus à incréments indépendants. Kalbfleisch [32] a supposé que H_0 est absolument continu, auquel cas il n'y a pas de points de discontinuité. Par conséquent, Il montre que la distribution à postériori de $H(t)$ est à nouveau un processus d'incrémentation indépendant [42].
Etant donné un échantillon de F, la distribution à postériori est dérivée en spécifiant la distribution à postériori des incréments, une stratégie utilisée par Doksum([17]). Une démonstration pour un échantillon de taille 1a été exposé par cet auteur. Une application répétée de cette procédure donne la solution pour toute taille d'échantillon. Pour une observation $X = x$ telle que $x \in [t_{i-1}, t_i)$, r_i est la somme des trois composantes indépendantes : U, l'incrément à gauche de x, J, le saut de x et V, l'ncriment à droite de x. U et V sont des variables gamma. La distribution de V et les incréments ultérieure en H restent inchangée. Tant que a distribution U et tout les incrément a priori à x ont des distributions gamma avec le paramètre d'échelle est passé de c à $c+1$, ou par $c+e^{\beta w}$ par rapport à l'observation si on a considéré un modèle de regression, donc $r_j \sim G(cH_0(x) - H_0(t_{i-1}), c+1)$,$j = 1, ..., i-1$, et $U \sim G(cH_0(x) - H_0(t_{i-1}), c+1)$. La distribution à posteriori du saut j est considéré a être une distribution avec la densité :

$$f_j(s) = \frac{e^{-sc} - e^{-s(c+1)}}{s(log(c + 1/c))} \tag{2.10}$$

et MGF $M_j(\theta) = log((c+1-\theta)/(c-\theta))$. On pose toutes les variables indépendantes, l'à posteriori de H donnée par $X = x$ est dérivable[42].

Processus Gamma étendu (Extended Gamma Process) :

Soit F une foncion de répartition CDF continue à gauche avec $F(x) = 0$ pour $x \leq 0$, $S(x) = 1 - F(x)$, $H(x) = -lnS(x)$. Si $r(t)$ est une fonction continue à droite telle que $H(x) = \int_{[0,x)} r(t)dt$, alors $r(t)$ est connue comme une fonction de risque. Soit $\alpha(0) = 0$;$\beta(t),\ t \geq 0$ une fonction réelle positive continue à droite, limité par 0 avec des limites à gauche existantes, et enfin soit $Z(t),\ t \geq 0$ un processus gamma avec des incréments indépendants correspondant à $\alpha(t)$ [42].

Définition 2.10. *(Dykstra and Laud [20]) Soit $Z(t) \in \mathcal{G}(\alpha(t), 1)$. Un processus stochastique défini par*

$$r(t) = \int_{[0,t)} \beta(s)dZ(s) \tag{2.11}$$

est appelé un processus gamma étendu et noté par $r(t) \sim \Gamma(\alpha(.), \beta(.))$.
$\Gamma(\alpha(.), \beta(.))$ est également connue comme processus gamma pondéré ou mélange de processus gamma. Evidement si $r(t)$ est aléatoire alors, en conséquence, $F(x) = 1 - exp\{-\int_{[0,x)} r(t)dt\}$ sera aussi aléatoire.
D'aprés Doksum ([17]), $F(x)$ sera neutre à droite que si $H(x) = \int_{[0,x)} r(t)dt$, a des incréments indépendants. $H(x)$ ne le sera pas, et donc les resultats distributionnels de Doksum ne sont pas applicables [42].

Propriétés :

1. La fonction génératrice des moments (MGF) $M_{r(t)}(\theta) = \exp\{-\int_{[0,t)} log(1 - \beta(s)\theta)d\alpha(s)\}$.
2. $\mu(r(t)) = E(r(t)) = \int_{[0,t)} \beta(s)d\alpha(s)$.
3. $\sigma^2(r(t)) = Var(r(t)) = \int_{[0,t)} \beta^2(s)d\alpha(s)$.
4. La loi marginale et les fonctions de survie joints sont annoncées dans ce théorème.

Théorème 2.2. *(Dykstra and Laud [20]) Si le taux de risque $r(t)$ a la distribution à priori $\Gamma(\alpha(.), \beta(.))$, alors la fonction de survie marginale d'une observation X est donné par*

$$S(t) = P(X \geq t) = exp\left\{ - \int_{[0,t)} log(1 + \beta(s)(t-s) d\alpha(s) \right\}$$

et la fonction de survie jointe étant donné n observations $X_1, ..., X_n$ est

$$\begin{aligned} S(t_1, ..., t_n) &= P(X_1 \geq t_1, ..., X_n \geq t_n) \\ &= exp\left\{ - \int_{[0,t)} log(1 + \beta(s) \sum_{i=1}^{n} (t_i - s)^+) d\alpha(s) \right\} \end{aligned}$$

ou $a^+ = max(a, 0)$.
il est facile de dériver la distribution à posteriori pour $X \geq x$.

5. Dans le cas du processus de Dirichlet, le paramétre F_0, a été interprêté comme une distribution à priori de la fonction inconnue F, et M comme le paramétre de concentration ou le poids lié a la distribution a priori. De même, en définissant $\mu(t)$ et $\sigma^2(t)$ comme des fonctions croissantes, les auteurs jugent raisonable d'inteprêter $\mu(t)$ comme une meilleure estimation de taux de risque et $\sigma^2(t)$ comme une meilleure mesure d'incertitude ou de la variation de risque au point t. Ensuite si $\mu,\ \sigma^2$ sont supposés differentiables, $\alpha(.)$ et $\beta(.)$ peuvent être spécifiés d'une manière appropriée en terme de $\mu(.)$ et $\sigma(.)$ comme suit :

$$\begin{aligned} \beta(t) &= \left(\frac{d\sigma^2(t)}{dt} \right) / \left(\frac{d\mu(t)}{dt} \right) \qquad et \\ \frac{d\alpha(t)}{dt} &= \left[\frac{d\mu(t)}{dt} \right]^2 / \left(\frac{d\sigma^2(t)}{dt} dt \right). \end{aligned}$$

2.2.4 La distribution à posteriori :

La propriété de conjugaison pour cet à priori n'est valable que dans le cas de données censurées à droite. Pour les observations exactes, la distribution à postériori est un mélange de processus gamma étendus.

Théorème 2.3. *(Dykstra and Laud [20]) Soit la loi à priori sur les taux de risque $\Gamma(\alpha(.), \beta(.))$, puis la loi à posteriori sur les taux de risque.*

(i) etant donné m observations censurées sous la forme $X_1 \geq x_1, ..., X_m \geq x_m$, est $\Gamma(\alpha(.), \beta^*(.))$ où

$$\beta^*(t) = \frac{\beta(t)}{1 + \beta(t). \sum_{i=1}^{m} (x_i - t)^+};$$

(ii) Etant donné m observations de la forme $X_1 = x_1, ..., X_m = x_m$, mélange de processus gamma étendu

$$\begin{aligned} &P(r(t) \in B | X_1 = x_1, ..., X_m = x_m) \\ &\qquad = \frac{\int_{[0,x_m)} \cdots \int_{[0,x_1)} \prod_{i=1}^{m} \beta^*(z_i) \psi(B; Q) \prod_{i=1}^{m} d[\alpha + \sum_{j=i+1}^{m} I_{(x_j, \infty)}](z_i)}{\int_{[0,x_m)} \cdots \int_{[0,x_1)} \prod_{i=1}^{m} C(z_i) \prod_{i=1}^{m} [d\alpha + \sum_{j=i+1}^{m} I_{(x_j, \infty)}](z_i)}, \end{aligned}$$

où $\psi(B; Q)$ indique la probabilité de l'ensemble $B \in \mathcal{B}$ sous un processus stochastique distribuée comme $Q = \Gamma(\alpha + \sum_{i=1}^{m} I_{(x_j, \infty)}, \beta^*)$.

L'effet des observations censurées est donc de diminuer la pente de l'échantillon chemins à gauche des points de censure tout en le laissant inchangé à droite de points de censure.
La distribution à postériori par rapport aux observations exactes est quelque peu compliquée. Cependant, les méthodes de Kalbfleisch ([32]) et Ferguson et Phadia ([26]) peuvent être utilisées pour l'exprimer en termes de MGFs. Ammann ([10],[9]) généralise cette approche en refondant le taux de risque comme une fonction des chemins d'échantillonnage de processus non négatifs avec incréments indépendants qui comprennent une composante croissante ainsi qu'une composante décroissante. Cette façon dont il est capable de définir une large classe de l'à priori sur un espace de distributions qui incluent les fonctions de survie du taux de défaillance croissant (IFR), et décroissant (DFR) et en forme de U [42].

Le processus ponctuel de Poisson :

En introduisant le processus de Dirichlet, Ferguson ([23]) était motivé par le fait que la distribution de Dirichlet est conjuguée par rapport à une distribution multinomiale. Lo ([36]) a reconnu que la distribution gamma est conjuguée par rapport à une distribution de Poisson. Ainsi, comme le Processus de Dirichlet, il devrait être possible de définir un processus gamma pour résoudre le problème d'inférence pour le processus ponctuel de Poisson à partir d'un point de vue bayésien non paramétrique. Lo ([36]) a montré que cela est possible grâce au processus gamma pondéré introduit auparavant et a établi les propriétés de conjugaison suivantes :

Théorème 2.4. *Lo ([36]) Si la distribution a priori de mesure d'intensité γ d'un processus ponctuel de Poisson est $\Gamma(\alpha, \beta)$ et étant donné un échantillon $N_1, ..., N_n$ de la taille n d'un Processus ponctuel de Poisson avec mesure d'intensité γ, alors la distribution a postériori de γ est $\Gamma(\alpha + \sum_{j=1}^{m} N_j, \beta/(1 + n\beta))$ [42].*

Fonction de distribution pondérée :

Dans l'analyse bayésienne de modèle d'échantillonnage pondéré (où la probabilité d'inclure une observation dans l'échantillon est proportionnelle à une fonction de pondération), Lo ([39]) montre que le processus de gamma pondéré normalisé peut être utilisé comme la loi à priori conjugué pour l'échantillonnage d'une distribution pondérée. La distribution pondérée est définie comme suit

$$F(dx|G) = \frac{w(x)G(dx)}{\int w(x)G(dx)},$$

où $w(x)$ est une fonction de poids connue, $0 < w(x) < \infty$, et G un paramétre inconnu. Le processus gamma pondéré normalisé est défini comme $\gamma(.) = r(.)/r(+\infty)$ et est noté par $\Gamma^*(\alpha(.), \beta(.)$, où α et β sont des paramètres de forme et de poids, respectivement. Supposons que nous avons un échantillon aléatoire $X_1, ..., X_n|G \overset{iid}{\sim} F(dx|G)$, i.e la probabilité d'inclure une observation dans l'échantillon est proportionnelle à la fonction de pondération w. Il est ensuite établi que si la loi à prior de G est $\Gamma^*(\alpha, 1/w)$, alors la distribution à postériori de $G|X$ est $\Gamma^*(\alpha + \sum_1^n \delta_{x_i}, 1/w)$ [42].

2.2.5 Processus gaussien :

Dans l'approche bayésienne non paramétrique, le paramètre à estimer est une fonction (par exemple la fonction de régression aléatoire). Pour définir une distribution a priori, il faut définir une distribution de probabilité sur un espace T approprié de fonctions que nous considérons comme des solutions viables. Le processus Gaussien est l'une de ces distribution, c'est la distribution la plus simple que l'on puisse espérer définir sur des fonctions continues [41].
Soit T un espace de fonctions à partir d'un ensemble $S \subset \mathbb{R}^d$ à $\mathbb{R}$. Si Θ est un élément aléatoire de T et $s \in S$ un point fixe, alors $\Theta(s)$ est une variable aléatoire dans $\mathbb{R}$. Plus généralement, si on fixe n points $s_1, ..., s_n \in S$ alors $(\Theta(s_1), ..., \Theta(s_n))$ est un vecreur aléatoire dans $\mathbb{R}^n$ [41].

Processus Gaussien à priori et à posteriori :

Dans un problème de régression, la solution expliquant les données est une fonction, de sorte que nous pouvons en principe utiliser un processus gaussien comme loi à priori dans un modèle de régression bayésien. Pour que cela soit possible, nous devons cependant être en mesure de calculer une distribution à postériori.

Nous devons d'abord définir un modèle d'observation : supposons que nos données soient de la forme $(s_1, x_1), ..., (s_n, x_n)$, où $s_i \in S$ sont des points d'observation (covariables) et $x_i \in \mathbb{R}$ est la valeur observée à s_i (la réponse). Nous supposons qu'il y a une fonction $\theta : S \longrightarrow \mathbb{R}$ la fonction de régression, à partir de laquelle les x_i sont générés sous forme d'observations bruitées :

$$X_i = \Theta(s_i) + \varepsilon_i \quad avec \quad \varepsilon_i \sim N(0, \sigma^2) \tag{2.12}$$

Nous supposons, bien sûr que les contributions des bruits $\varepsilon_1, \varepsilon_2, ...$ sont i.i.d. La distribution à postériori que nous recherchons est la distribution

$$P[\Theta \in \bullet | X_1 = x_1, ..., X_n = x_n], \tag{2.13}$$

et donc une mesure sur l'espace de fonction T. Comme nous savons qu'une distribution sur T est uniquement déterminée par les distributions marginales à dimension infinie, il est judicieux de déterminer les distributions

$$\mathcal{L}(\Theta(s_{n+1}), ..., \Theta(s_{n+m}) | X_1 = x_1, ..., X_n = x_n) \tag{2.14}$$

pour tout ensemble de nouvelles observations locales $s_{n+1}, ..., s_{n+m}$. Pour garder la notations on donne l'abréviation

$$A := \{n+1, ..., n+m\} \qquad et \qquad B := \{1, ..., N\} \tag{2.15}$$

et on écrit :

$$\Theta(s_A) := (\Theta(s_{n+1}), ..., \Theta(s_{n+m})) \quad et \quad X_B := (X_1, ..., X_n) \tag{2.16}$$

Pour conditionner les variables X_i, nous devons d'abord examiner de plus près leur distribution : dans (2.11), $\Theta_{(s_i)}$ est la somme de deux variables gaussiennes indépendantes de variance $k(s_i, s_i)$ et σ^2, et donc à nouveau gaussiennes avec variance $k(s_i, s_i) + \sigma^2$. Pour $i \neq j$, seules les contributions $\Theta_{(s_i)}$ et $\Theta_{(s_j)}$ forment un couple (puisque le bruit est indépendant). Ainsi, X_B a une matrice de covariance

$$\begin{pmatrix} k(s_1, s_1) + \sigma^2 & \dots & k(s_1, s_n) \\ \vdots & \ddots & \vdots \\ k(s_n, s_1) & \dots & k(s_n, s_n) + \sigma^2 \end{pmatrix}$$

La covariance conjointe de vecteur $(\Theta(s_A), X_B)$ de dimension $(n+m)$ est alors

$$Cov[(\Theta(s_A), X_B)] = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{AB}^t & \Sigma_{BB} \end{pmatrix} \tag{2.17}$$

Là encore, puisque le bruit est indépendant, les seules contributions à la covariance proviennent du GP, et donc

$$\Sigma_{AB} = (k(s_i, s_j))_{i \in A, j \in B} \qquad et \qquad \Sigma_{AA} = (k(s_i, s_j))_{i,j \in A}\,. \tag{2.18}$$

La détermination du GP à postériori revient donc à un conditionnement dans un gaussien multi-dimensionnel. Le lemme simple suivant explique comment cela fonctionne [41] :

Lemme 2.1. *(Conditionnement dans les distributions gaussiennes). Soit (A, B) une partition de l'ensemble $\{1, ..., d\}$ et que $X = (X_A, X_B)$ soit un vecteur aléatoire gaussien dans $\mathbb{R}^d = \mathbb{R}^A \times \mathbb{R}^B$, avec*

$$E(X) = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \qquad et \qquad Cov[X] = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{AB}^t & \Sigma_{BB} \end{pmatrix}$$

Ensuite, la distribution conditionnelle de $X_A|(X_B = x_B)$ *est à nouveau gaussienne, avec une moyenne*

$$E[X_A|(X_B = x_B)] = \mu_A \Sigma_{AB} \Sigma_{BB}^{-1}(x_B - \mu_B) \tag{2.19}$$

Et la covariance

$$Cov[X_A|X_B = x_B] = \Sigma_{AA} - \Sigma_{AB}^t \Sigma_{BB}^{-1} \Sigma_{AB} \tag{2.20}$$

Nous pouvons maintenant lire la partie à postériori du processus gaussien, simplement en substituant dans le lemme. Comme les distributions marginales de dimension fini sont toutes gaussiennes, la partie à postériori est une GP.

Théorème 2.5. *[41] La partie à postériori d'un* $GP(0, k)$ *sous observation (2.11) est un processus gaussien. Ses distributions marginales de dimension fini à n'importe quel ensemble fini* $\{s_{n+1}, ..., s_{n+m}\}$ *est le gaussien avec vecteur moyen :*

$$E[\Theta(s_A)|X_B = x_B] = \Sigma_{AB}(\Sigma_{BB} + \sigma^2 \mathbf{I})^{-1} x_B \tag{2.21}$$

et la matrice des covariances

$$Cov[\Theta(s_A)|X_B = x_B] = \Sigma_{AA} - \Sigma_{AB}^t (\Sigma_{BB} + \sigma^2 \mathbf{I})^{-1} \Sigma_{AB}. \tag{2.22}$$

Chapitre 3

Estimation fonctionelle par le processus de Dirichlet :

3.1 Introduction :

Dans le chapitre précédent, nous avons présenté la méthode bayésienne non paramétrique. Dans ce chapitre nous introduisons un exemple pratique de l'estimation bayésienne non paramétrique où nous allons estimer la fonction de densité de probabilité dans le cadre bayésien, dans ce cas f est la réalisation d'un mélange des procédés de Dirichlet.

3.2 L'estimation de la fonction de densité de probabilité :

Soit $X_1, ..., X_n$ un échantillon de taille n de la fonction de densité $f(x)$ par rapport à une mesure finie de $\mathbb{R}$. Considérons le problème d'estimation de $f(x)$ en un point fixe x, ou une fonction de $f(x)$, telle que la moyenne $\int xf(x)dx$. Pour le traitement bayésien, nous devons attribuer une loi a priori sur l'espace de toutes les fonctions de densité et être capable de traiter analytiquement la distribution à postériori. Pour que la distribution à postériori soit gérable, il serait préférable de trouver une famille a priori conjuguée. On sait que cela est difficile. Lo ([37],[38]) aborde ce problème en utilisant une approche à noyau de la fonction de densité, et en attribuant Dirichlet a priori G. Ses résultats sont présentés ici.
Soit G une fonction de distribution sur $\mathbb{R}$ et α une mesure finie sur $(\mathbb{R}, B)$. Soit $K(x, u)$ un noyau défini sur $(\mathcal{X}, \mathbb{R})$ dans $\mathbb{R}^+$ tel que pour chaque $u \in \mathbb{R}$, $\int_X K(x, u)dx = 1$ et pour chaque $x \in \mathcal{X}$, $\int_{\mathbb{R}} K(x, u)\alpha(du) < \infty$ (Lo ([37],[38]) prend X et $\mathbb{R}$ comme des sous-ensembles borélien d'espaces euclidiens). La distribution a postériori de $G|X$ a été obtenu par Antoniak ([11],) comme indiqué précédemment. Pour chaque $G \in \mathcal{F}$, définir $f(x|G) = \int_{\mathbb{R}} K(x, u)G(du)$, ensuite $f(.|G)$ est une représentation en noyau de la fonction de densité f et G est connue comme une distribution de mélange. Lo ([37],[38]) définit une distribution a priori pour l'inconnue f en laissant G une distribution aléatoire avec un processus de Dirichlet a priori à $D(\alpha)$. On a $G \in D(\alpha)$, on peut voir que pour chaque $x \in \mathcal{X}$, la densité marginale de X est $f_0(x) = \int_{\mathcal{F}} f(x|G)D_\alpha(dG) = \int_{\mathbb{R}} K(x, u)\alpha(du)/\alpha(\mathbb{R})$. La distribution a postériori de G pour les information X peuvent être considérées comme

$$P(G \in B|X) = \frac{\int_B \prod_{i=1}^n \int_R K(x_i, u_i)G(du_i)D_\alpha(dG)}{\int_{\mathcal{F}} \prod_{i=1}^n \int_R K(x_i, u_i)G(du_i)D_\alpha(dG)}, \tag{3.1}$$

pour tout $B \in \mathcal{F}$.

Par l'application répétée de son lemme [37] (en changeant l'ordre des l'intégration),

$$\int_{\mathcal{F}} \int_{\mathbb{R}} h(u, G)G(du)D_\alpha(dG) = \int_{\mathbb{R}} \int_{\mathcal{F}} h(u, G)D_{\alpha+\delta_u}(dG)\alpha(du)/\alpha(R), \tag{3.2}$$

il montre que

$$P(G \in B|X) = \frac{\int_{\mathbb{R}^n} D_\alpha + \sum \delta_{u_i}(B)\mu_{n,k,\alpha}(du)}{\int_{\mathbb{R}^n} \mu_{n,k\alpha}(du)} \tag{3.3}$$

où :

$$\mu_{n,k,\alpha}(C) = \int_C \prod_{i=1}^{n} K(x_i, u_i) \prod_{i=1}^{n} \left(\alpha + \sum_{j=1}^{i-1} \delta_{u_j} \right) (du_i), \tag{3.4}$$

où $C \in \mathcal{B}^n$, $du = \prod_{i=1}^n du_i$ et $u \in \mathbb{R}^n$, pour chaque fonction mesurable g, cela donne

$$E(g(G)|X) = \frac{\int_{\mathbb{R}^n} g(G) D_\alpha + \sum \delta_{u_j}(dG)\mu_{n,k,\alpha}(du)}{\int_{\mathbb{R}^n} \mu_{n,k,\alpha}(du)} \tag{3.5}$$

Maintenant, en prenant $g(G) = f(x|G)$ et par la simplification, l'a posteriori $\widehat{f}(x|G)$ de $f(x|G)$ est dérivé comme :

$$\widehat{f_a}(x|G) = E(f(x|G)|X) = p_n f_0(x) + (1 - p_n)\widehat{f_n}(x), \tag{3.6}$$

qui est une combinaison convexe de la loi à priori $f_0(x)$ définie ci-dessus, et d'une quantité $\widehat{f_n}(x)$ à définir ci-dessous.
Soit $N(\underline{P})$ le nombre des cellules dans la partition $\underline{P}$ de $1, 2, ..., m$; C_i la i-ème cellule de $\underline{P}$ avec m_i un élément dedans, $i = 1, ..., N(\underline{P})$; $g_i(u), i = 1, ..., m$ sont m fonctions positives ou α-intégrables ;

$$\varphi(\underline{P}) = \prod_{i=1}^{N(\underline{P})} \left\{ (m_i - 1)! \int_{\mathbb{R}} \prod_{l \in C_i} g_l(u)\alpha(du) \right\} \tag{3.7}$$

et finalement $w(\underline{P}) = \varphi(\underline{P}) / \sum_{\underline{P}} \varphi(\underline{P})$. Ensuite $\widehat{f_n}(x)$ est donnée par

$$\widehat{f_n}(x) = \frac{1}{n} \sum_{\underline{P}} \varphi(\underline{P}) \sum_{i=1}^{N(\underline{P})} m_i \left\{ \frac{\int_{\mathbb{R}} K(x, u) \prod_{l \in C_i} K(x_l, u)\alpha(du)}{\int_{\mathbb{R}} \prod_{l \in C_i} K(x_l, u)\alpha(du)} \right\}, \tag{3.8}$$

où la somme est prise sur toutes les partitions $\underline{P}$ de $\{1, 2, ..., m\}$. f sert a une estimation de Bayes sous la fonction de perte $L(f, \widehat{f}) = \int \left| f(x|G) - \widehat{f}(x|G) \right|^2 W(dx)$, où W est une fonction de pondération.
Lo discute du choix du noyau K et du paramètre α de la loi a priori, et donne plusieurs exemples de $K(x, u)$ et α et calcule les estimateurs de Bayes. Ces exemples de noyaux comprennent l'histogramme, la loi normale avec les paramètres de localisation et/ou d'échelle, les densités symétriques et unimodales, les densités décroissantes, etc. Par exemple, si K est choisi pour refléter le modèle de l'histogramme, l'estimateur réduit aux estimations de Bayes habituelles des probabilités de cellules. La méthode de Monte Carlo de Kuo ([34],[33]) peut être adaptée pour effectuer les calculs. Vous trouverez plus de détails dans son article. Lavine ([35]) utilise des mélanges d'arbres Polya dans l'estimation de la densité.
Ghorai et Susarla ([27]) ont considéré une approche empirique de Bayes pour le problème ci-dessus. En supposant que $\alpha(R)$ soit connu, ils ont obtenu un estimateur de $f_0(x) = \int_{\mathbb{R}} K(x, u)\alpha(du)$ $/\alpha(\mathbb{R})$ basé sur les n copies précédentes et substitué dans l'estimateur bayésien $\widehat{f}(x|G)$ au $(n+1)$-éme stade. Sous certaines conditions, ils prouvent l'optimalité asymptotique de l'estimateur résultant.

Ferguson ([24]) a considéré une formulation différente de la fonction de densité. Il l'a modélisé comme un mélange dénombrable de densités normales : $f(x) = \sum_{i=1}^{\infty} p_i h(x|\mu_i, \sigma_i)$ où $h(x|\mu, \sigma)$ est la densité normale avec la moyenne μ et la variance σ^2. Cette formulation a un nombre infini de

paramètres, $(p_1, p_2, ..., \mu_1, \mu_2, ..., \sigma_1, \sigma_2, ...)$. Comme le but est d'estimer $f(x)$ à un point x, et non d'estimer les paramètres eux-mêmes, on peut écrire $f(x) = \int h(x|\mu, \sigma)dG(\mu, \sigma)$, où G est la mesure de probabilité sur le demi-plan est la mesure de probabilité sur le demi-plan $\{(\mu, \sigma) : \sigma > 0\}$ qui donne du poids p_i au point $(\mu_i, \sigma_i), i = 1, 2, ...$ Alors que Lo suppose un processus de Dirichlet à priori pour l'inconnu G, Ferguson définit l'a prior via la représentation Sethuraman[46] de G. ll a définit la distribution à priori pour le vecteur paramètre $(p_1, p_2, ..., \mu_1, \mu_2, ..., \sigma_1, \sigma_2)$ comme suit : les vecteurs $(p_1, p_2, ...)$ et $(\mu_1, \mu_2, ..., \sigma_1, \sigma_2, ...)$ sont indépendants $p_1, p_2, ...$ sont les poids avec le paramètre M dans la représentation de Sethuraman, et $\xi_i = (\mu_i, \sigma_i)$ sont iid avec la loi a priori conjuguée de gamma normale. Cela montre que G est un processus de Dirichlet avec le paramétre $\alpha = MG_0$, où $G_0 = E(G)$ est la loi a priori conjuguée pour (μ, σ^2), et sa représentation de somme infinie est $G = \sum_{i=1}^{\infty} p_i \delta_{\xi_i}$ où comme $(p_1, p_2, ...)$ et $(\xi_1, \xi_2, ...)$ sont indépendants et $\xi_i \sim G_0$. Maintenant on donne un échantillon $(x_1, ..., x_n)$ de taille n à partir d'une distribution avec densité $f(x) = \int h(x|\xi)dG(\xi)$ la distribution postériori de G donné $(x_1, ..., x_n)$ a été obtenu par Antoniak ([11]) comme mélange des procédés de Dirichlet

$$G|x_1, ..., x_n \sim \int ... \int D(\alpha + nG_n)dH(\xi_1, ..., \xi_n|x_1, ..., x_n)$$

avec $nG_n = \sum_{i=1}^{n} \delta_{\xi_i}$. $H(\xi_1, ..., \xi_n|x_1, ..., x_n)$ est la distribution a postériri de $\xi_1, ..., \xi_n$ donné $x_1, ..., x_n$.

Depuis $\mathcal{E}(D(\alpha + nG_n)) = (MG_0 + nG_n)/(M + n)$.

$$E(G(\xi)|x_1, ..., x_n) = p_n G_0(\xi) + (1 - p_n) \int ... \int G_n(\xi)dH(\xi_1, ..., \xi_n|x_1, ..., x_n) \tag{3.9}$$

et

$$\widehat{f}(x) = E(f(x)|x_1, ..., x_n) = p_n f_0(x) + (1 - p_n)\widehat{f_n}(x) \tag{3.10}$$

où : $p_n = M/(M + n)$ comme avant, $f_0(x) = E(f(x)) = \sum_{i=1}^{\infty} E(p_i)E(h(x|\mu_i), \sigma_i) = E(x|\mu, \sigma)$ et $\widehat{f_n}(x)$ est donnée par

$$\widehat{f_n}(x) = \frac{1}{n} \sum_{i=1}^{n} \int ... \int h(x|\xi_i)dH(\xi_1, ..., \xi_n|x_1, ..., x_n) \tag{3.11}$$

Suivant Lo, $\widehat{f_n}(x)$ peut écrire comme $h(x, x_1, ..., x_n)/h(x_1, ..., x_n)$, où

$$h(x_1, ..., x_n) = \frac{1}{M^{(n)}} \int ... \int \left(\prod_{n=1}^{n} (h(x_i|\xi_i)) \right) \prod_{n=1}^{n} d \left(MG_0 + \sum_{j=1}^{i-1} \delta_{\xi_i} \right) (\xi_i) \tag{3.12}$$

et les calculs sont effectués par la méthode de Monte Carlo de Kuo ([34],[33]). Des mélanges normaux apparaissent également dans Escobar ([21]) et Escobar et West ([22]).

La configuration d'Escobar est la suivante. Soit $Y_i|\mu_i \sim N(\mu_i, 1)$, $\mu_i|G \overset{iid}{\sim} G$, μ_i et G sont inconnus. Contrairement aux objectifs de Ferguson et de Lo, son objectif est d'estimer μ_i, (la variance étant connue comme étant de 1) sur la base des Y_i observés en utilisant l'approche bayésienne non paramétrique. Lorsque G est connu, l'estimateur bayésien est la moyenne postériori

$$E(\mu_i|Y_i) = \frac{\int \mu_i \phi(Y_i - \mu_i)dG(\mu_i)}{\int \phi(Y_i - \mu_i)dG(\mu_i)} \tag{3.13}$$

où ϕ est la densité de la fonction de distribution normale standard. Lorsque G est inconnu, les méthodes empiriques de Bayes sont généralement utilisées. Escobar utilise plutôt un processus de Dirichlet a prior pour G. Antoniak a montré que si le processus de Dirichlet a prior est utilisé

pour G, alors la distribution postériori de μ_i est un mélange de processus de Dirichlet. Ainsi, il a été difficile à calculer. Kuo ([33]) et Lo ([37]) ont développé des algorithmes d'intégration de Monte Carlo, mais Escobar souligne qu'ils sont inefficaces car ils n'échantillonnent pas les valeurs de manière conditionnelle en fonction des données. Il présente une nouvelle méthode semblable à celle de l'échantillonneur Gibbs qui a remédié à ce problème.

Escobar et West ([22]) décrivent un modèle de mélange normal, similaire à celui de Ferguson ([24]), en termes de distribution prédictive d'une observation future. Pour leur modèle, donné (μ_i, σ_i^2), nous avons un échantillon aléatoire, disons $Y_1, ...Y_n$, tel que $Y_i|(\mu_i, \sigma_i^2) \sim N(\mu_i, \sigma_i^2), i = 1, ..., n$ et l'objectif est de trouver la distribution prédictive de la prochaine observation Y_{n+1} qui est un mélange de normales, $Y_{n+1}|Y_1, ..., Y_n \sim N(\mu_{n+1}, \sigma_{n+1}^2)$. Une pratique habituelle consiste à mettre l'a priori paramétrique sur le vecteur $v = (\mu_1, ..., \mu_n, \sigma_i^2, ..., \sigma_n^2)$. Ferguson modélise le priori commun pour $v_i = (\mu_i, \sigma_i^2)$ comme un processus Dirichlet a priori. Ainsi, les données sont considérées comme provenant d'un mélange de Dirichlet de normales, contrairement à Antoniak où les processus du processus de Dirichlet ont été mélangés en fonction d'une distribution paramétrique $H(\theta), \alpha_\theta \sim H(\theta)$. Un cas particulier de $(\mu_i, \sigma_i^2) = (\mu_i, \sigma^2)$ a été étudié (voir West [48],[49]) dans lequel la distribution de μ_i est modélisée comme un processus de Dirichlet avec une mesure de base normale.

Compte tenu de la discrétion du processus de Dirichlet qui induit des multiplicités d'observations, $v_{n+1}|v_1, ...v_n$ aura la distribution . Puis ils procèdent sur la ligne de Ferguson, en déduisent le conditionnel distribution de $Y_{n+1}|v_1, ...v_n$ qui est un mélange de la distribution t d'un étudiant et de n normales $N(\mu_i, \sigma_i^2)$, et il est ensuite démontré que la distribution prédictive inconditionnelle est donnée par $Y_{n+1}|Y_1, ..., Y_n \sim \int P(Y_{n+1}|v)dP(v|Y_1, ..., Y_n)$. Puisque le l'évaluation de $P(v|Y1, ..., Yn)$ est difficile, même pour de petits échantillons, ils utilisent Monte Approximation Carlo utilisant des extensions de la technique itérative développée par Escobar ([21]).

Conclusion Générale

Dans ce présent travail qui s'inscrit dans le cadre de notre projet de fin d'études, nous avons abordé l'approche bayésienne qui est une des méthodologie les plus importante et qui présente un grand avantage dans l'inférence statistique. Cette approche est basée sur deux méthodes : paramétrique et non paramétrique. Elle est inductive et part des données pour estimer la distribution du paramètre inconnu.

Le premier chapitre de ce mémoire résume les concepts fondamentaux de l'estimation classique. C'est dans le deuxième chapitre que nous avons introduit la méthode bayésienne paramétrique et la méthode bayésienne non paramétrique qui constitue la partie la plus conséquente de notre travail. Nous avons terminé notre travail par une application où nous nous sommes intéressé à l'estimation fonctionnelle dans un cadre non paramétrique en abordant le probléme d'estimation de la fonction de densité de probabilité par le processus de Dirichlet.

Annexe

Carré intégrable pour la loi P_0 :

On dit qu'un estimateur $T(X)$ est de carré intégrable c'est-à dire $\int [T(X)]^2 dP\theta < \infty$.

σ-fini :

Soit (Ω, B, μ) un espace mesuré. On dit que la mesure μ est σ-finie s'il existe une suite croissante (pour l'inclusion) (E_n) d'éléments de B, toutes de mesures finies, et telle que

$$\Omega = \bigcup_n E_n$$

. Autrement dit, Ωest la réunion dénombrable d'ensembles de mesures finies pour μ.

Fonction indicatrice :

Soit Ω un ensemble. La fonction indicatrice d'un sous-ensemble A de l'ensemble Ω notée est la fonction définie sur Ω qui vaut 1 sur A et 0 à l'extérieur de A :

$$1_A(x) = \begin{cases} 1 & si \quad x \in A. \\ 0 & si \quad x \in \Omega \setminus A. \end{cases}$$

Processus à accroissement independant :

Un processus stochastique $X = \{X_t : t \geq 0\}$ est appelé processus de Lévy ou un processus dont les accroissements sont stationnaires et indépendants si :

1. $X_0 = 0$.
2. Accroissements indépendants : Pour tout $0 \leq t_1 < t_2 < ... < t_n < \infty, X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, ..., X_{t_n} - X_{t_{n-1}}$ sont indépendants
3. Accroissements stationnaires : Pour tout $s < t, X_t - X_s$, est égale en loi à X_{t-s}.
4. $t \mapsto X_t$ est presque sûrement continue à droite et limitée à gauche.

Somme pondéré :

Le modèle de somme pondérée (WSM), également appelé combinaison linéaire pondérée (WLC) est la méthode d'analyse décisionnelle multicritères (MCDA) est donée par :

$$A_i = \sum_{j=1}^{n} w_j a_{ij}, \quad for \quad i = 1, 2, 3, ..., m.$$

ou on suppose ce probléme soit defini sur m alternatives et n critères de décisions ou : w_j désigne le poind relatif d'importance du critère C_j et que a_{ij} est la valeur de performance de la variante A_i lorsqu'elle est évaluée en fonction du critère C_j.

La distribution multinomiale :

Dans un cadre statisque bayésien, la distribution de Dirichlet est souvent associé à des ensembles des données multinomiale pour la distribution à priori des paramétres de brobabilités.

$(u_1, ..., u_k)$ suit une distribution multinomiale avec le paramètre $(N, p_1, ..., p_k)$

$$(u_1, ..., u_k) \sim Mu(N, p_1, ..., p_k)$$

si les conditions suivantes sont satisfaites :

1. $p_1 + p_2 + ... + p_k = 1$ et tous les p_i sont non négatifs.
2. $u_1 + u_2 + ... + u_k = N$ et tous les u_i sont des entiers non negatifs.
3. $u_1, u_2, ..., u_{k-1} = \dfrac{\Gamma(N+1)}{\prod_{i=1}^{k} \Gamma(u_i + 1)} (\prod_{i=1}^{k-1} p_i^{u_i})(p_k)^{N - \sum_1^{k-1} u_i}$.

Bibliographie

[1] *Analys bayesienne non paramétrique.http ://web.univ-ubs.fr/lmba/gouno/BAYES/COURS/Cours9.pdf.*

[2] *Etimation paramétrique www.math.univ-toulouse.fr.*

[3] *Statistique Bayésienne "statistique bayésienne elearning".*

[4] *G Vincent .Modélisation des distributions de sinistres avec R,2013.*

[5] *D Jean-Yves.CTU, Master Enseignement des Mathématiques Statistique Inférentielle,(2011-2012).*

[6] *D Marie* . méthodes de modélisation bayésienne et applications en recherche clinique, l'université montpellier1. 2010.

[7] *T Yann - R Adrien . Introduction aux Statistiques Bayésiennes, 2018.*

[8] *Intoduction à la statistique bayésienne*, 2018-2019.

[9] L P Ammann. Conditional laplace transforms for bayesian nonparametric inference in reliabity theory. *Stochastic processes and their applications*, 20(2) :197–212, 1985.

[10] L P Ammann et al. Bayesian nonparametric inference for quantal response data. *The Annals of Statistics*, 12(2) :636–645, 1984.

[11] C E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

[12] J-C Breton. Processus gaussiens. *Université de La Rochelle (version de décembre 2006)*, 2006.

[13] C Chesneau. Sur l'estimateur du maximum de vraisemblance (emv). 2017.

[14] F Chimard and J Vaillant. Processus stick-breaking et extensions pour le traitement bayésien de processus ponctuels. 2010.

[15] H-Phuong Dang. *Approches bayésiennes non paramétriques et apprentissage de dictionnaire pour les problèmes inverses en traitement d'image*. PhD thesis, 2016.

[16] M Denis. *Méthodes de modélisation bayésienne et applications en recherche clinique*. PhD thesis, UM1, 2010.

[17] K Doksum. Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, pages 183–201, 1974.

[18] D B Dunson and J-H Park. Kernel stick-breaking processes. *Biometrika*, 95(2) :307–323, 2008.

[19] J Dupuis. Statistique bayésienne et algorithmes mcmc, 2007.

[20] RL Dykstra and P Laud. A bayesian nonparametric approach to reliability. *The Annals of Statistics*, pages 356–367, 1981.

[21] M D Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425) :268–277, 1994.

[22] M D Escobar and M West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430) :577–588, 1995.

[23] T S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[24] T S Ferguson. Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics*, pages 287–302. Elsevier, 1983.

[25] T S Ferguson et al. Prior distributions on spaces of probability measures. *The annals of statistics*, 2(4) :615–629, 1974.

[26] T S Ferguson and E G Phadia. Bayesian nonparametric estimation based on censored data. *The Annals of Statistics*, pages 163–186, 1979.

[27] JK Ghorai and V Susarla. Empirical bayes estimation of probability density function with dirichlet process prior. In *Probability and statistical inference*, pages 101–114. Springer, 1982.

[28] Ht Ishwaran and L F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453) :161–173, 2001.

[29] H Jeffreys. Theory of probability, clarendon, 1961.

[30] Han L John L and Larry W. *(Chapter 27) Nonparametric Bayesian Methods.*

[31] Boudjerda K. Etude de l'estimateur de bayes sous différentes fonctions de perte, université badji mokhtar annaba, année :2016-2017.

[32] J D Kalbfleisch. Non-parametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society : Series B (Methodological)*, 40(2) :214–221, 1978.

[33] L Kuo. Computations of mixtures of dirichlet processes. *SIAM Journal on Scientific and Statistical Computing*, 7(1) :60–71, 1986.

[34] L Kuo. A note on bayes empirical bayes estimation by means of dirichlet processes. *Statistics & probability letters*, 4(3) :145–150, 1986.

[35] M Lavine et al. Some aspects of polya tree distributions for statistical modelling. *The annals of statistics*, 20(3) :1222–1235, 1992.

[36] Albert Y Lo. Bayesian nonparametric statistical inference for poisson point processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 59(1) :55–66, 1982.

[37] Albert Y Lo. On a class of bayesian nonparametric estimates : I. density estimates. *The annals of statistics*, pages 351–357, 1984.

[38] Albert Y Lo. Bayesian statistical inference for sampling a finite population. *The Annals of Statistics*, pages 1226–1233, 1986.

[39] Albert Y Lo et al. A bayesian method for weighted sampling. *The Annals of Statistics*, 21(4) :2138–2148, 1993.

[40] P Müller and F A Quintana. Nonparametric bayesian data analysis. *Statistical science*, pages 95–110, 2004.

[41] P Orbanz. Lecture notes on bayesian nonparametrics. *Journal of Mathematical Psychology*, 56 :1–12, 2012.

[42] G Phadia, E. *Prior processes and their applications.* Springer, 2015.

[43] J Pitman and M Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

[44] C Robert. *Le choix bayésien : Principes et pratique.* Springer Science & Business Media, 2005.

[45] J Rousseau. Statistique bayésienne notes de cours. *ENSAE ParisTech troisieme année*, 2010, 2009.

[46] J Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[47] Y Whye Teh, M I Jordan, M J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476) :1566–1581, 2006.

[48] M West. *Bayesian kernel density estimation.* Institute of Statistics and Decision Sciences, Duke University, 1990.

[49] M West. Modeling with mixtures,"(with discussion) in bayesian statistics 4, 1992.