Introduction aux Statistiques Bayésiennes

Yann Traonmilin - Arthur Leclaire
(d'après les notes de cours de Charles Dossal, Jérémie Bigot et Adrien Richou)

2019

Introduction

Considérons quatre problèmes d'inférence statistique.

1. Une machine à sous disposant d'un bouton donne 1€ avec une probabilité θ et 0€ sinon. On cherche à estimer cette probabilité.

2. Pour une étude de marché, on cherche à estimer la moyenne du prix de vente d'un produit.

3. Un informateur nous prévient que 30% des machines à sous ont une probabilité θ_1 de donner 1€ , le reste a une probabilité θ_2. On cherche à savoir à quelle type appartient une machine donnée.

4. Une société de conseil nous propose de faire l'étude du prix de vente. Pour un produit, on fait une étude parallèle pour étudier si l'information qu'elle propose est fiable.

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

1 Introduction aux principes de l'inférence bayésienne

1.1 Rappels de probabilités

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

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

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

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

Proof.

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

□

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

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

Proof.

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

□

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

Définition 1.2 (Indépendance conditionnelle).

- *Soit A, B, C des événements tels que $\mathbb{P}(C) \neq 0$. On dit que A est indépendant de B conditionnellement à C si*
$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C). \tag{5}$$
- *Soient X, Y, Z des v.a. On dit que X, Y sont indépendantes conditionnellement à Z si et seulement si pour tous A, B, C tels que $\mathbb{P}(Z \in C) > 0$, les évènements $\{X \in A\}$ et $\{Y \in B\}$ sont indépendants conditionnellement à $\{Z \in C\}$.*

Remarque 1.1. *Le lecteur pourra vérifier que deux événements indépendants A et B (c.à.d. $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$) ne sont pas conditionnellement indépendants en général. Exemple :On lance deux dés équilibrés; on considère les événements : A : le premier dé indique un nb pair B : le second dé indique un nb impair C : la somme des deux dés est paire*

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

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

1.2 Rappels sur l'approche fréquentiste

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

Définition 1.4. *Si $X_1, \ldots, X_n$ sont des variables discrètes i.i.d. selon une loi $\mathbb{P}_\theta$ dépendant d'un paramètre θ, on appelle fonction de vraisemblance la fonction L définie par*

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

Si $X_1, \ldots, X_n$ sont des variables continues i.i.d. dont la loi admet une densité f_θ dépendant d'un paramètre θ, on appelle fonction de vraisemblance la fonction L définie par

$$(\theta; x_1, x_2, \ldots, x_n) \to L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) \tag{8}$$

Dans les deux cas, la valeur de cette fonction au point $(\theta; x_1, x_2, \ldots, x_n)$ est la vraisemblance de θ lorsqu'on observe la réalisation $(x_1, x_2, \ldots, x_n)$.

Exemples :

1. On considère n variables aléatoires $X_1, \ldots, X_n$ i.i.d. de loi gaussienne $\mathcal{N}(\theta, \sigma^2)$ où $\theta \in \mathbb{R}$ et σ^2 est supposé fixé et connu. La vraisemblance s'écrit

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

2. On considère n variables aléatoires $X_1, \ldots, X_n$ i.i.d. de loi de Bernoulli $\mathcal{B}(\theta)$ de paramètre $\theta \in [0, 1]$. La vraisemblance s'écrit

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

où $s = \sum_{i=1}^{n} x_i$.

Dans la suite, on notera $X = (X_1, \ldots, X_n)$ et $x = (x_1, \ldots, x_n)$.
Une large gamme de méthodes d'estimation repose sur la technique du maximum de vraisemblance

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

Dans les deux exemples ci-dessus, on a

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

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

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

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

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

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} s \log(\theta) + (n-s) \log(1-\theta) = \arg\max_{\theta \in [0,1]} G(\theta) \tag{11}$$

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

□

1.3 Le paradigme bayésien.

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

Définition 1.5. *Une loi a priori π est une loi de probabilité sur Θ.*

Définition 1.6. *En adoptant le paradigme bayésien, le modèle statistique définit la loi jointe des observations* conditionnellement *à θ. Ainsi,*

- *dans le cas où les X_i sont discrètes, la loi conditionnelle de X sachant θ s'écrit*

$$x \to (x|\theta) = \mathbb{P}(X = x|\theta) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n|\theta) = \prod_{i=1}^{n} \mathbb{P}(X_1 = x_1|\theta) \dots \mathbb{P}(X_n = x_n|\theta).$$

- *dans le cas où les X_i sont à densité, la loi conditionnelle de X sachant θ admet pour densité*

$$x \to p(x|\theta) = p(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) \ .$$

Dans un modèle bayésien, on s'autorisera à écrire $p(x|\theta)$ la loi conditionnelle des observations sachant θ, mais on prendra garde au fait que cette notation $p(x|\theta)$ désigne une probabilité discrète lorsque X est discrète, et une densité lorsque X est continue.

Définition 1.7 (Modèle Bayésien)**.** *Un* ***modèle bayésien*** *est la donnée d'une loi a priori pour θ et d'une famille $(\mu_\theta)_{\theta\in\Theta}$ de lois conditionnelles pour les X_i. Plus précisément, on supposera que θ suit la loi a priori π, et que* conditionnellement à *θ, les X_i sont indépendantes de loi μ_θ. On pourra s'autoriser à écrire cela de façon condensée sous la forme*

$$\begin{aligned} X_i &\overset{iid}{\sim} \mu_\theta \\ \theta &\sim \pi \end{aligned} \tag{12}$$

où la première équation sous-entend que les X_i sont i.i.d. conditionnellement *à θ.*

À partir d'un modèle bayésien, on peut calculer une loi *a posteriori* sur θ, cette loi n'est rien d'autre que la loi de θ conditionnellement aux observations X. En pratique, on associe une loi *a posteriori* à une réalisation x de X Son calcul se base essentiellement sur la formule de Bayes, que l'on peut écrire de façon résumée en

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

Mais pour poursuivre le calcul rigoureusement, on doit distinguer les différents cas possibles selon le caractère discret/continu des lois en jeu.

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

1. *La loi de X et la loi a priori sont discrètes.*
   C'est le cas des trois exemples de l'introduction.
   Dans cette situation la loi a posteriori s'écrit pour toutes valeurs possibles t ,

$$p(t|x) = \mathbb{P}(\theta = t|X = x) = \frac{\mathbb{P}(X = x|\theta = t)\mathbb{P}(\theta = t)}{\mathbb{P}(X = x)} = \frac{\mathbb{P}(X = x|\theta = t)\mathbb{P}(\theta = t)}{\sum_{u\in\Theta}\mathbb{P}(X = x|\theta = u)\mathbb{P}(\theta = u)}.$$

2. *La loi de X est discrète et la loi a priori est continue de densité notée π. Dans ce cas, la loi a posteriori est une loi continue dont la densité s'écrit*

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

3. *La loi de X est continue et la loi a priori est discrète. La loi a posteriori est une loi discrète qui s'écrit pour toutes valeurs possibles t et x,*

$$p(t|x) = \mathbb{P}(\theta = t|x) = \frac{p(x|t)\mathbb{P}(\theta = t)}{\sum_{u\in\Theta}p(x|u)\mathbb{P}(\theta = u)}$$

4. *La loi de X et la loi a priori sont continues. Dans ce cas, la loi a posteriori est une loi continue dont la densité s'écrit*

$$p(t|x) = \frac{p(x|t)\pi(t)}{\int_{u\in\Theta}p(x|u)\pi(u)du}$$

Ces quatre formulations ne sont que quatre spécifications d'une même égalité que l'on résume souvent sous la forme continue/continue (la quatrième formulation). À noter que dans les égalités suivantes, t désigne une variable muette qui décrit Θ. En pratique, on écrira souvent les densités ou probabilités en utilisant la lettre θ pour cette variable muette.

Remarquons que dans tous les cas, le dénominateur est $p(x)$ et donc ne dépend pas de t (valeur de θ). Dans les problématiques d'inférence sur θ, on aura besoin que des variations relatives de $p(t|x)$. Par exemple la valeur de t qui réalise le maximum de $p(t|x)$ ne dépend de $p(t|x)$ modulo les constantes multiplicatives.

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

$$p(x) = \int_{u\in\Theta} p(x|u)\pi(u)du. \tag{14}$$

Noter qu'elle dépend de la loi a priori sur θ.

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

$$p(\theta|x) \propto p(x|\theta)\pi(\theta).$$

Remarque 1.2. ***Important!*** *Le calcul d'une loi a posteriori mène à une loi. Ainsi, le résultat de l'inférence est beaucoup plus informatif que dans le cas fréquentiste : on a accès beaucoup plus facilement à des intervalles de confiance bayésiens pour une estimation de θ (que l'on peut construire par exemple autour du maximum de la loi a posteriori). Autrement dit, la loi a posteriori fournit naturellement le calcul de la marge d'erreur sur l'estimation d'un paramètre. Il ne faut pas confondre ces intervalles de confiance bayésiens (aussi appelés intervalles de crédibilité) avec les intervalles de confiance rencontrés au premier semestre (dans un cadre fréquentiste).*

2 Comment choisir la loi *a priori* ?

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

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

2.1 Lois subjectives

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

2.2 Approche partiellement informative

2.2.1 Notion de lois conjuguées

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

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

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

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

Quelques exemples de lois conjuguées

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

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

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

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

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

2.2.2 Cas du modèle exponentiel

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

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

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

La densité conditionnelle associée au n-échantillon $x=(x_1,\ldots,x_n)$ s'écrit

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

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

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

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

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

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

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

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

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

Proof.

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

□

Exemple : exercice 4 du TD2.

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

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

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

On parle alors de prior *impropre*.
Important : la distribution *a posteriori* doit être bien définie i.e.

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

Exemples :

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

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

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

- Exercice 5 du TD2.

2.3 Loi *a priori* non informative

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

2.3.1 Lois invariantes

- Soit f une densité sur $\mathbb{R}^d$, la famille de lois $\{p(.|\theta)\}_{\theta\in\mathbb{R}^d}$ avec $p(x|\theta) = f(x-\theta)$ est invariante par translation : en effet, si X suit la loi à densité $p(x|\theta)$ alors $X+\theta_0$ suit la loi à densité $p(x|\theta+\theta_0) = f(x-\theta-\theta_0)$. On dit dans ce cas que θ est un paramètre de position. Comme $\{p(.|\theta)\}_{\theta\in\mathbb{R}^d} = \{p(.|\theta+\theta_0)\}_{\theta\in\mathbb{R}^d}$, il est naturel de demander à la loi *a priori* π d'être invariante par translation, c'est à dire qu'elle satisfasse $\pi(\theta) = \pi(\theta+\theta_0)$ pour tous $\theta_0 \in \mathbb{R}^d$. On trouve alors que π est constante, c'est-à-dire la loi (éventuellement impropre) uniforme sur $\mathbb{R}^d$.

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

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

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

2.3.2 Loi *a priori* de Jeffreys

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

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

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

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

a) Cas unidimensionnel.

On rappelle la définition de l'information de Fischer :

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

qui, sous certaines conditions de régularité, peut se réécrire

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

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

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

Cette loi possède deux intérêts principaux :

- $I(\theta)$ est un indicateur de la quantité d'information apportée par le modèle $p(x|\theta)$. Donc $I(\theta)$ est grand lorsque le modèle varie fortement autour de θ. Par conséquent, au moins à un niveau qualitatif, il paraît intuitivement justifié que les valeurs de θ pour lesquelles $I(\theta)$ est plus grande doivent être plus probables a priori.
- La loi de Jeffreys est invariante par reparamétrisation. En effet soit $\phi = h(\theta)$ avec h un C^1-difféomorphisme. Si on note π la loi *a priori* de θ, alors ϕ est de loi $\tilde{\pi}$ avec $\tilde{\pi}(\phi) = \pi(h^{-1}(\phi))|(h^{-1})'(\phi)|$. De plus on a $\tilde{I}(\phi) = I(h^{-1}(\phi))|(h^{-1})'(\phi)|^2$ donc $\tilde{\pi}(\phi) \propto \sqrt{\tilde{I}(\phi)}$. Ce calcul justifie en particulier la présence de la racine carrée.

b) Cas multi-dimensionnel.

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

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

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

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

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

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

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

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

Exemple : Si $X \sim \mathcal{N}(\mu, \sigma^2)$ et que l'on cherche à estimer $\theta = (\mu, \sigma)$, alors l'information de Fisher sur (m, σ) s'écrit

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix},$$

et donc l'a priori de Jeffreys s'écrit $\pi(m, \sigma) \propto \sigma^{-2}$. On peut remarquer qu'il ne vérifie pas l'invariance par échelle. Néanmoins, on prendra garde aux changements qu'on observe avec la reparamétrisation. En effet, si l'on écrit l'information de Fisher sur (m, σ^2), on obtient alors

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix},$$

et donc l'a priori de Jeffreys sur (m, σ^2) s'écrit $\pi(m, \sigma^2) \propto \sigma^{-3}$. Ceci confirme le calcul théorique effectué ci-dessous : après le changement de variable, il faut corriger par la valeur absolue du déterminant jacobien du changement de variable, qui est ici 2σ.

2.4 Cas particulier du modèle normal

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

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

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

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

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

On peut naturellement se demander ce qui se passe lorsque l'on cherche à estimer l'espérance et la variance en même temps. On a besoin d'une nouvelle loi a priori sur $\theta = (m, \Sigma)$. On commence par le cas unidimensionnel, et on étendra au cas multidimensionnel à la fin de cette partie.

2.4.1 Un premier *a priori*

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

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

La vraisemblance vaut

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

On a vu au dessus que l'a priori de Jeffreys pour (m, σ^2) s'écrit $\pi(m, \sigma) = \frac{1}{\sigma^3}$. On trouve alors

$$p(x|m, \sigma^2) \propto \sigma^{-3} \exp\left(-\frac{n}{\sigma^2}(\bar{x}_n - m)^2\right)(\sigma^{-2})^{\frac{n}{2}} \exp\left(-\frac{s_n^2}{2\sigma^2}\right).$$

On a donc la loi *a posteriori* suivante :

Proposition 2.3.

$$p(m|\sigma^2, x) \sim \mathcal{N}(\bar{x}_n, \frac{\sigma^2}{n})$$
$$p(\sigma^2|x) \sim \mathrm{IG}\left(\frac{n+1}{2}, \frac{s_n^2}{2}\right)$$

où la loi $\mathrm{IG}(\alpha, \beta)$ *est la loi inverse Gamma de densité*

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

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

2.4.2 Loi *a priori* conjuguée

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

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

où

$$\pi_1(m|\sigma^2) \sim \mathcal{N}\left(m_0, \frac{\sigma^2}{n_0}\right) \qquad \text{et} \qquad \pi_2(\sigma^2) \sim \text{IG}\left(\frac{\nu}{2}, \frac{s_0^2}{2}\right).$$

Notons que l'on a 4 hyper-paramètres m_0, n_0, ν et s_0^2. On trouve alors la loi *a posteriori* suivante :

$$\pi(m, \sigma^2|x) \propto \sigma^{-n-\nu-3} \exp\left(-\frac{1}{2\sigma^2}(s_1^2 + n_1(m - m_1)^2)\right)$$

où

$$n_1 = n + n_0, \quad m_1 = \frac{1}{n_1}(n_0 m_0 + n\bar{x}_n)$$
$$s_1^2 = s_n^2 + s_0^2 + (n_0^{-1} + n^{-1})^{-1}(m_0 - \bar{x}_n)^2.$$

Après calculs, on obtient le résultat suivant :

Proposition 2.4.

$$p(m|x, \sigma^2) \sim \mathcal{N}\left(m_1, \frac{\sigma^2}{n_1}\right),$$
$$p(\sigma^2|x) \sim \text{IG}\left(\frac{n + \nu + 1}{2}, \frac{s_1^2}{2}\right).$$

On obtient bien une loi conjuguée. Il reste à savoir comment choisir en pratique les hyper-paramètres m_0, s_0^2, n_0 et ν.

2.4.3 Le cas multidimensionnel

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

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

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

$$\pi(m|\Sigma) \sim \mathcal{N}\left(m_0, \frac{\Sigma}{n_0}\right)$$
$$\pi(\Sigma^{-1}) \sim W(\alpha, V)$$

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

$$p(m|x, \Sigma) \sim \mathcal{N}\left(\frac{n_0 m_0 + n\bar{x}_n}{n_0 + n}, \frac{\Sigma}{n_0 + n}\right),$$
$$p(\Sigma^{-1}|x) \sim W\left(\alpha + n, V^{-1} + S_n + \frac{n n_0}{n + n_0}(\bar{x}_n - m_0)(\bar{x}_n - m_0)^\top\right).$$

3 Estimation bayésienne

On dispose maintenant d'outils pour calculer une loi *a posteriori* sur un paramètre que l'on veut estimer. La quantité à estimer peut être un signal débruité, une quantité d'intérêt en finance ou l'efficacité d'un vaccin en médecine. Dans ce chapitre, nous allons rappeler ce qu'est un estimateur, les différentes notions de risque bayésien qui peuvent lui être associées et leur mise en œuvre dans le cadre des modèles linéaires.

3.1 Estimateurs et risques associés

Un estimateur se définit dans un cadre général.

Définition 3.1 (Estimateur)**.** *Soit un ensemble des paramètres Θ. On appelle estimateur une fonction Δ qui, à une observation $x \in \Omega$, associe un paramètre $\hat{\theta}_{emp} := \Delta(x) \in \Theta$.*

Si on dispose de plusieurs observations $x = (x_1, x_2, \dots, x_n)$ réalisations d'une v.a. X. On a $\hat{\theta}_{emp} = \Delta(x_1, x_2, \dots, x_n)$.
L'objectif d'un estimateur est d'approcher le paramètre θ_0 de la loi de X. Plus $\hat{\theta}_{emp}$ est proche de θ_0, meilleur est l'estimateur Δ.

Remarque 3.1. ***Important!*** *Il ne faut pas confondre Δ qui est une fonction de Ω dans Θ et $\hat{\theta}_{emp} \in \Theta$ qui est une estimation empirique qui dépend des observations x. Dans l'étude d'un estimateur (en particulier dans l'approche fréquentiste), on fait souvent appel à $\hat{\theta} := \Delta(X)$ qui est une variable aléatoire dépendant de la réalisation de la variable aléatoire X qui modélise les expériences sont tirées. Souvent (et dans la suite de ces notes), la notation $\hat{\theta}$ est utilisée à la fois pour l'estimation empirique et l'estimation (v.a.), le contexte permettant de comprendre de quelle notion on parle. On réutilise aussi souvent la notion d'estimateur pour parler de $\hat{\theta}$.*

Il vient tout de suite les questions :

1. Comment définir "$\hat{\theta}$ est proche de θ_0" ?
2. Trouver un bon estimateur Δ ?

Exemples : Comme on l'a vu précédemment (approche fréquentiste), avec la vraisemblance $L(\theta; x_1, x_2, \dots, x_n)$, une manière d'estimer θ_0 à partir de l'échantillon est de choisir l'estimateur Δ qui maximise cette fonction de vraisemblance :

$$\Delta(x) = \hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; x_1, x_2, \dots, x_n). \tag{15}$$

La manière la plus utilisée en réponse à notre première question, est la suivante

Définition 3.2 (Risque ℓ^2 moyen)**.** *Le risque ℓ^2 moyen de l'estimateur Δ (ou $\hat{\theta}$) du paramètre θ_0 qui a engendré les observations est*

$$R(\theta_0, \Delta) := \mathbb{E}_{\theta_0}[\|\hat{\theta} - \theta_0\|_2^2]$$

Ce risque est aussi appelé Erreur Quadratique Moyenne (EQM).

L'espérance est ainsi calculée sur les réalisations, selon la loi paramétrée par θ_0. C'est-à-dire $\mathbb{E}_{\theta_0}$ veut dire espérance à θ_0 fixé. Ainsi c'est le comportement *en moyenne* qui est considéré (d'où le terme fréquentiste)

Remarque 3.2. *Le choix de la norme ℓ^2 est arbitraire. Il existe un ensemble de raisons qui fait que ce choix est souvent pertinent et donne des résultats satisfaisants mais il est tout à fait possible de mesurer la qualité d'un estimateur en utilisant d'autres mesures de distances, par exemple ℓ^1. Nous verrons dans la suite d'autres distances sur les estimateurs.*

Ce risque quadratique moyen peut être impossible à calculer dans certains cas, car il dépend par définition de θ_0 qui est inconnu. Les études théoriques des estimateurs visent à le borner supérieurement par des quantités raisonnables.
On rappelle que la quantité $\mathbb{E}_{\theta_0}[\hat{\theta} - \theta_0]$ est appelé *biais* de l'estimateur et quand celui-ci est nul on dit que l'estimateur est sans biais. Le caractère non biaisé d'un estimateur est important mais dans de nombreuses situations, il peut être important de contrôler sa variance.

3.2 Risque bayésien

En statistiques bayésiennes, on dispose d'une observation x (éventuellement vectorielle i.e. $x = (x_1, x_2, \ldots, x_n)$) et d'une loi *a priori* . Soit C un fonction de coût (une fonction qui mesure une distance entre deux objets). On défini le risque moyen pour ce coût.

Définition 3.3 (Risque moyen)**.** *Le risque moyen pour le coût C de l'estimateur Δ (ou $\hat{\theta}$) du paramètre θ est :*

$$R_C(\theta, \Delta) = \mathbb{E}_\theta[C(\theta, \Delta(X))] := \int_x C(\theta, \Delta(x)) p(x|\theta) dx.$$

Si $C(x, y) = \|x - y\|_2^2$, on retrouve l'erreur quadratique classique. Il est cependant possible d'utiliser d'autres fonctions selon l'application visée.
Comme on dispose d'une loi π *a priori* sur le paramètre θ, on est capable en statistiques bayésiennes de mesurer le risque moyen, dit risque *intégré* selon la loi π.

Définition 3.4 (Risque intégré)**.** *Le risque intégré pour un modèle bayésien défini par $p(\cdot|\theta)$, $\theta \sim \pi$ et pour une fonction de coût C est :*

$$r(\pi, \Delta) = \mathbb{E}_\pi \mathbb{E}_\theta C(\theta, \Delta(X)) = \int_{\theta \in \Theta} R_C(\theta, \Delta) d\pi(\theta). \tag{16}$$

Un estimateur bayésien Δ^π est un estimateur qui minimise le risque intégré :

$$\Delta^\pi = \arg \inf_{\Delta \in \mathcal{D}} r(\pi, \Delta) \tag{17}$$

où $\mathcal{D}$ est l'ensemble des estimateurs possibles.

Un tel estimateur est également lié au risque moyen *a posteriori* .

Définition 3.5 (Risque moyen *a posteriori*)**.** *On appelle risque moyen a posteriori pour un modèle bayésien défini par $p(\cdot|\theta)$ et $\theta \sim \pi$ et pour une fonction de coût C*

$$\rho_C(\pi, \Delta|x) := \mathbb{E}_\pi[C(\theta, \Delta(x))|x] = \int_{\theta \in \Theta} C(\theta, \Delta(x)) p(\theta|x) d\theta. \tag{18}$$

Ce risque mesure, l'erreur moyenne que l'on commet, étant donné les observations $X = x$ en estimant le paramètre θ par $\hat{\theta} = \Delta(x)$, moyenne calculée en fonction de la loi *a priori* π.
Comme dans la pratique, on ne dispose que d'une réalisation $X = x$, il semble naturel de chercher l'estimateur qui minimise un tel risque. Cet estimateur coïncide avec l'estimateur bayésien :

Théorème 3.1. *et* ***!!!définition!!!*** *S'il existe* $\Delta \in \mathcal{D}$ *tel que* $r(\pi, \Delta) < +\infty$ *alors* $\forall x \in X$

$$\Delta^\pi(x) = \arg\min_{\Delta \in \mathcal{D}} \rho(\pi, \Delta | x) \tag{19}$$

Dans le cadre bayésien, c'est toujours une estimation à partir des expériences qui est privilégiée. On définit ainsi ***l'estimateur de Bayes***

Proof. On écrit le risque intégré, applique Fubini, le théorème de Bayes puis faire apparaître le risque *a posteriori* :

$$\begin{aligned} r(\pi, \Delta) &= \int_{\theta \in \Theta} \int_x C(\theta, \Delta(x)) p(x|\theta) dx d\pi(\theta) \\ &= \int_x \int_{\theta \in \Theta} C(\theta, \Delta(x)) p(x|\theta) d\pi(\theta) dx \\ &= \int_x \int_{\theta \in \Theta} C(\theta, \Delta(x)) p(x|\theta) \pi(\theta) d\theta dx \\ &= \int_x \int_{\theta \in \Theta} C(\theta, \Delta(x)) p(\theta|x) p(x) d\theta dx \\ &= \int_x \rho_C(\pi, \Delta|x) p(x) dx \end{aligned} \tag{20}$$

Pour conclure on prend l'inf des deux côtés. Le résultat vient du fait que l'on a de quantités positives et que la loi marginale m ne dépend pas de Δ

□

Remarque 3.3. *L'unicité de l'estimateur de Bayes est garantie si* C *est strictement convexe, mais pas en général.*

Remarque 3.4. *Ce cadre permet de traiter les questions de décisions dans le contexte bayésien, mais cela va au delà de l'objectif du cours.*

3.2.1 Estimateur bayésien et fonction de coût

Fonction de perte quadratique
Si :

$$C(x, y) = \|x - y\|_2^2 \tag{21}$$

alors l'estimateur bayésien est l'espérance de la loi *a posteriori* $p(\theta|x)$:

$$\Delta^\pi(x) = \mathbb{E}_\pi(\theta|x)$$

Voir TD pour la démonstration.

Fonction de perte valeur absolue.
Si (en 1D) :

$$C(x, y) = |x - y| \tag{22}$$

alors l'estimateur bayésien $\Delta^\pi(x)$ est la médiane de la loi *a posteriori* $p(\theta|X = x)$.
Pour le montrer, il faut observer que

$$\rho(\pi, \Delta|X = x) = \int_{-\infty}^{\Delta} (\Delta - \theta) p(\theta|X = x) d\theta + \int_{\Delta}^{+\infty} (\theta - \Delta) p(\theta|X = x) d\theta, \tag{23}$$

intégrer par parties pour faire apparaître la fonction $g(y) = \mathbb{P}(\theta < y|x)$ (commencer par le terme de gauche) puis , puis $\mathbb{P}(\theta \geq y|x)$ dériver par rapport à Δ.

Fonction de coût $0-1$
On appelle la fonction de coût 0-1 la fonction C définie par $C(x,y) = 0$ si $x = y$ et $C(x,y) = 1$ sinon.
Si $\Theta = \{0,1\}$, l'estimateur bayésien est le maximum de la loi *a posteriori* :

$$\Delta^\pi(x) = \arg\max_{\theta\in\Theta} p(\theta|x)$$

Ces trois estimateurs peuvent se confondre dans certains cas mais peuvent différer énormément en général (différence entre une médiane et une moyenne).

Remarque 3.5. *Pour le calcul direct d'un estimateur, le maximum a posteriori (MAP) sera souvent retenu (même dans le cas où θ est continu), surtout en traitement du signal et des images, même si son interprétation dans le cadre bayésien est moins satisfaisante.*

3.2.2 Estimation bayésienne avec modèle linéaire

On commence par rappeler un résultat classique d'estimation. On considère le modèle d'observation suivant :

$$y = M\theta_0 + \epsilon \tag{24}$$

où M est une matrice $m \times d$, $\theta_0 \in \mathbb{R}^d$ et ϵ est une vecteur aléatoire à valeur dans $\mathbb{R}^m$ dont les coordonnées sont i.i.d. de loi $\mathcal{N}(0,\sigma^2)$. On peut noter que y peut être vu comme un vecteur aléatoire gaussien de moyenne $M\theta$ et de variance σ. On peut définir une fonction de vraisemblance

$$L(\theta; y) \propto e^{-\frac{\|y-M\theta\|_2^2}{2\sigma^2}} \tag{25}$$

Calculer le maximum de vraisemblance revient alors à calculer l'estimateur des moindres carrés :

$$\hat{\theta} = \arg\min_\theta \|y - M\theta\|_2^2 \tag{26}$$

On a alors la formule bien connue

$$\hat{\theta} = (M^T M)^{-1} M^T y. \tag{27}$$

si $m \geq d$, ou

$$\hat{\theta} = M^T (MM^T)^{-1} y \tag{28}$$

sinon.

Remarque 3.6. *La matrice M n'est pas forcément inversible. On a donc un estimateur dans le cas où le problème est mal posé.*

Ce résultat est le point de départ d'un large pan des mathématiques appliquées : les méthodes variationnelles pour les problèmes linéaires inverses, on développera ces liens dans les applications au signal et en image.

Dans le cadre bayésien, que se passe-t'il si on a un a priori sur θ_0?

Prenons une loi *a priori* Gaussienne $\mathcal{N}(0, \tau^2)$ sur θ. On a

$$p(\theta|y) \propto p(y|\theta)\pi(\theta) = e^{-\frac{\|y-M\theta\|_2^2}{2\sigma^2}} e^{-\frac{\|\theta\|_2^2}{2\mu^2}} \tag{29}$$

Comme mentionné plus haut beaucoup de méthodes d'estimation consistent à calculer le maximum de cette loi *a posteriori* pour estimer θ_0. On se retrouve donc à calculer

$$\hat{\theta} = \arg\max_{\theta} p(\theta|y) = \arg\min_{\theta} \frac{\|y - M\theta\|_2^2}{2\sigma^2} + \frac{\|\theta\|_2^2}{2\mu^2} \tag{30}$$

Ce qui donne la fameuse régularisation de Tychonov (aussi appelée dans d'autres cadres filtre de Wiener ou bien ridge régression) :

$$\hat{\theta} = (M^T M + \lambda I)^{-1} M^T y \tag{31}$$

où $\lambda = ?$ (Voir TD).

4 Simulation de loi *a posteriori*

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

- L'émergence des probabilités remonte au 17ème siècle tandis que les premiers travaux de statistique datent du 18ème siècle avec Bayes et Laplace. Il s'agit alors de statistique bayésienne.
- Au cours du 19ème siècle et du 20ème siècle les méthodes fréquentistes supplantent largement les méthodes bayésiennes.
- Depuis le début des années 1980, on note un retour très important de la recherche et des applications des méthodes bayésiennes.

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

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

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

En particulier on a

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

avec la constante de renormalisation

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

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

- Inférence : la moyenne a posteriori est donnée par

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

- Région de confiance :

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

- Densités *a posteriori* marginales

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

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

4.1 Méthodes de Monte Carlo

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

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

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

Proposition 4.1. *La quantité*

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

est un estimateur sans biais et fortement consistant de I.

On a également la normalité asymptotique de l'estimateur.

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

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

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

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

Remarque 4.1.

- *La variance, et donc la précision de l'approximation, augmente linéairement avec la dimension. Pour les méthodes d'intégration numérique classique (reposant sur des grilles), la précision augmente exponentiellement avec la dimension : c'est le fléau de la dimension (curse of dimensionality en anglais).*

- *Dans le cas des statistiques bayésiennes, on ne connaît g qu'à une constante de renormalisation près. On ne peut donc pas appliquer ces méthodes directement.*

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

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

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

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

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

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

4.2.1 Généralités sur les chaînes de Markov

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

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

$$f(X_{i+1}|X_i, ..., X_0) = K(X_{i+1}|X_i)$$

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

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

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

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

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

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

Cette matrice doit vérifier

- $0 \leqslant A_{jk} \leqslant 1, \quad 1 \leqslant j \leqslant p, \quad 1 \leqslant k \leqslant p,$

- $\sum_{k=1}^{p} A_{jk} = 1, \quad \forall \quad 1 \leqslant j \leqslant p.$

On note μ_0 la loi initiale de la chaîne, i.e. la loi de X_0.

Proposition 4.3.

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

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

Proposition 4.4.

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

Définition 4.2. *Une distribution g est dite invariante ou stationnaire par rapport à une chaîne de Markov si la chaîne laisse cette distribution invariante, i.e.*

- *Dans le cas fini, $g = gA$,*
- *dans le cas général $g(x') = \int_{\mathcal{X}} g(x) K(x'|x) dx$.*

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

$$g(x) K(x'|x) = g(x') K(x|x'). \tag{32}$$

Preuve.

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

□

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

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

Remarque 4.2.

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

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

4.2.2 Algorithme de Metropolis-Hastings

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

Algorithme 4.1.

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

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

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

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

Preuve. On va montrer que le noyau K de la chaîne de Markov générée par l'algorithme de Metropolis-Hastings vérifie la condition d'équilibre (32). On commence par calculer le noyau $K(\theta'|\theta)$. remarquons que c'est une loi qui n'est ni a densité ($\theta' = \theta$ avec une probabilité potentiellement non nulle) ni discrète. On a

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

Si $\theta' = \theta$, (32) est trivialement vérifiée. On suppose donc $\theta' \neq \theta$. Alors

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

Donc par symétrie on a

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

ce qui prouve le résultat. □

Remarque 4.3.

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

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

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

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

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

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

Dans ce cas le taux d'acceptation est donné par

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

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

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

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

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

- Si σ est faible, l'échantillonneur fait des explorations locales (petits sauts) qui sont presque tous acceptées.
- Si σ est grand, l'échantillonneur fait des grands sauts mais qui sont acceptés avec probabilité faible.

Quelle que soit la valeur de σ, l'algorithme va converger vers la mesure stationnaire. Néanmoins σ influe sur la vitesse de convergence. Empiriquement le taux d'acceptation optimal est entre 0.1 et 0.6, il faut donc choisir σ en fonction.

4.2.3 Algorithme de Gibbs

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

Comment simuler $\theta \sim g$?

On note

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

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

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

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

Fin pour
Fin pour

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

Algorithme 4.3.

- *On définit une valeur initiale* $\theta_0 = (\theta_0^1, ..., \theta_0^p)^\top$,
- *Pour* $i = 1, ..., N$
  1. *Pour* $j = 1, ..., p$ *on propose une nouvelle valeur* $(\theta^j)^* \sim q(.|\theta_i^{-j}, \theta_{i-1}^j)$ *(loi de proposition) avec*
  $$\theta_i^{-j} = (\theta_i^1, ..., \theta_i^{j-1}, \theta_{i-1}^{j+1}, ..., \theta_{i-1}^p)^\top.$$
  2. *On calcule le taux d'acceptation*
  $$\alpha((\theta^j)^*, \theta_{i-1}^j) = \min\left(1, \frac{g(\theta_i^{-j}, (\theta^j)^*) q(\theta_{i-1}^j|(\theta^j)^*, \theta_i^{-j})}{g(\theta_i^{-j}, \theta_{i-1}^j) q((\theta^j)^*|\theta_{i-1}^j, \theta_i^{-j})}\right)$$
  3. *Avec probabilité* α*, on prend* $\theta_i^j = (\theta^j)^*$ *et avec probabilité* $1-\alpha$ *on prend* $\theta_i^j = \theta_{i-1}^j$.

5 Modèles hiérarchiques

5.1 Introduction

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

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

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

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

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

5.2 Justification théorique

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

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

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

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

Remarque 5.1.

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

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

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

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

5.3 Un cas pratique

On considère N élèves du lycée d'une ville appartenant à M lycées différents.
Le niveau d'un élève $y_{i,j}$, $j$ème élève du lycée i peut être modélisé par une variable réelle $y_{i,j} = \theta_i + \varepsilon_{i,j}$ où θ_i est le niveau moyen du lycée indicé par i et $\varepsilon_{i,j}$ est une variable gaussienne $\mathcal{N}(0, 1)$ et où on suppose que les θ_i sont des variables i.i.d. suivant une loi $\mathcal{N}(\mu, \tau^2)$, où τ^2 est connu mais où μ est un hyper-paramètre sur le quel on peut mettre un *a priori* .

Si on met un *a priori* uniforme sur μ, (la mesure de Lebesgue sur $\mathbb{R}$ est impropre mais peu être un *a priori* valide si la loi de θ est gaussienne), on peut estimer la loi *a posteriori* de la moyenne θ_j du lycée j à partir des élèves de ce lycée. Nous allons voir en TD que proposer un modèle hiérarchique c'est à dire, supposer que les différentes moyennes des lycées sont des variables aléatoires suivant une même loi, induit une corrélation sur les $(\theta_j)_{j\leqslant M}$ et que les estimations des différents θ_j vont faire intervenir, le niveau des élèves des autres lycées. Les deux cas extrêmes sont

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

6 Méthodes bayésiennes en traitement du signal et des images

6.1 Introduction

Nous avons vu comment utiliser les méthodes bayésiennes pour estimer des paramètres expliquant un jeu de données. Dans cette partie, on s'intéresse à ces méthodes dans le contexte du traitement du signal et surtout, des images. Dans ce contexte, un objectif très important et d'estimer une image à partir d'une version dégradée. En effet, que ce soit en photographie ou en traitement vidéo, l'image numérisée n'est qu'une version dégradée de la scène naturelle que l'on à voulu capturer. Le champ de lumière à subit un filtrage optique, un sous-échantillonnage et une quantification. De plus, la physique du phénomène rend l'image capturée bruitée. Dans ce contexte de dégradation, le recours à un modèle a priori sur le paramètre à estimer (l'image) prend tout son sens. En effet, comment palier à la perte d'information subie par l'image? Nous allons voir comment les méthodes bayésiennes peuvent servir comme cadre à la résolution de ce problème

6.2 Le modèle bayésien en traitement des images

On considère le modèle d'acquisition d'une image suivant:

$$v = Hu_0 + e \tag{33}$$

où $v \in \mathbb{R}^m$ est l'image acquise numérisée, u_0 est le champ de lumière arrivant sur le capteur, e est le bruit d'acquisition. En pratique, on ne reconstruit pas exactement u_0, on essaie d'en estimer une version échantillonnée de bonne qualité. Ainsi dans la suite on suppose $u \in \mathbb{R}^n$. Comme les effets de quantification ne provoquent pas une trop grande dégradation visuelle, on se limite au cas où H est un opérateur linéaire, ainsi il peut modéliser :

- une transformation géométrique (translation, rotation, etc)
- un sous-échantillonnage (basse- résolution, masquage)
- une convolution par un noyau de flou (flou optique)

Remarque 6.1. *Il faut noter que lorsque H est invariant par translation l'utilisation de la transformée de Fourier est souvent avantageuse pour une utilisation pratique.*

Pour faciliter l'exposition, on considère le cas où l'image est en niveaux de gris (intensité de la lumière). Enfin, e est le bruit d'acquisition qui sera considéré comme i.i.d. gaussien de variance σ. Ce modèle de bruit est une simplification du modèle physique d'acquisition des images où le bruit peut être bien modélisé par un bruit gaussien de variance variable.

L'opérateur u étant un opérateur linéaire de dégradation, son noyau est non trivial, et il existe une infinité de solutions à l'équation $v = Hu$. Dans ce cas là il est impératif d'avoir recours à une information annexe. On utilise alors un modèle bayésien sur le paramètre à estimer u_0 :

$$\begin{aligned} v &\sim f(v|u_0) = \mathcal{N}(Hu_0, \sigma) \\ u_0 &\sim \pi \end{aligned} \tag{34}$$

où π est la la loi a priori sur u_0. Une grande partie des méthodes bayésiennes se contentent d'estimer le maximum a posteriori pour estimer u_0 (la loi a posteriori étant plus difficile à estimer, existe en traitement des images mais sort du cadre introductif de ce cours). On cherche alors

$$u^* = \arg\max f(u|v) \tag{35}$$

avec

$$f(u|v) = f(v|u)\pi(u) \propto e^{-\frac{\|Hu-v\|^2}{2\sigma^2}}\pi(u) = e^{-\frac{\|Hu-v\|^2}{2\sigma^2}+log(\pi(u))} \tag{36}$$

ainsi

$$u^* = \arg\min \frac{\|Hu-v\|^2}{2\sigma^2} - log(\pi(u)) \tag{37}$$

On peut bien sûr se préoccuper de l'existence et de l'unicité de l'infimum. On remarque que u^* est le résultat de la minimisation d'un terme d'attache aux données et d'un terme a priori", que l'on appelle régularisation. Ce terme de régularisation est d'ailleurs ce qui permet en général d'avoir une estimation de u^* satisfaisante lorsque le problème est mal posé (H non inversible). Cette formulation est appelée formulation "variationnelle" (cf autres cours - optimisation et traitement du signal et des images). Un grande partie de la littérature considère des lois a priori de la forme $-log(\pi(u)) = \lambda R(u)$ (souvent $R(u) = \|u\|_p^p$ où $p = 1, 2$)et A est un opérateur linéaire,dit d'analyse). On a alors $\pi(u) = e^{-\lambda R(u)}$. La fonction R est appelée fonction de régularisation. On a vu dans le cas de la régularisation de Tychonov (Modèle bayésien gaussien gaussien) que le **paramètre de régularisation** λ peut être interprété comme un ratio d'énergie entre l'a priori et le bruit d'acquisition. Une grande problématique de traitement des images est de bien estimer ce paramètre λ. On se retrouve donc à résoudre des problèmes de type:

$$u^* = \arg\min \frac{\|Hu-v\|^2}{2\sigma^2} + \lambda R(u) \tag{38}$$

6.3 Quels a priori en traitement des images

Idéalement le choix de la loi a priori devrait être entièrement guidée par la nature de l'image que l'on cherche à estimer. On se heurte à deux problématiques fondamentales :

- la loi a priori doit être suffisamment contraignante pour aider à la résolution du problème. Une loi a priori dans un contexte donné pourrait ne pas fonctionner dans un autre.
- on doit pouvoir effectuer la minimisation pour u^*. Le problème peut être NP-difficile si on prend des a priori de nature combinatoire.

Une grande part des a priori que l'on considère sont des a priori de structure sur l'image. L'image est transformée dans un domaine où elle à une structure remarquable que l'on cherche à exploiter.

6.3.1 Un exemple (trivial) d'a priori exact

Si on sait que u_0 appartient en fait à un sous espace connu E de dimension m. On peut utiliser $R = \iota_E$ la fonction caractéristique de E.

6.3.2 A priori Gaussien

On a vu que les a priori gaussiens donnent une fonction de régularisation qui met en jeu la norme ℓ^2. C'est donc un a priori qui aura tendance à minimiser l'énergie du résultat. En général cet a priori n'a pas de sens intéressant dans le domaine de l'image. Par contre lorsque l'on considère l'énergie du gradient de l'image, on se retrouve à favoriser les estimations qui n'ont des variations faibles, c-à-d des images qui sont assez lisses. Il est aussi envisageable d'aller chercher des dérivées d'ordre supérieur.

6.3.3 A priori Laplacien

Lorsque R est de la forme $R = \lambda\|\cdot\|_1$, on parle d'a priori Laplacien. Il a été montré que ce type d'a priori à tendance à favoriser les estimations parcimonieuses. Il est alors particulièrement intéressant d'obtenir des estimations parcimonieuses dans des espaces transformés particuliers (gradient de l'image: variation totale isotrope et anisotrope, transformée de Fourier/en ondelette...). Pour étudier les formes de la solution, il faut étudier précisément les conditions qui font que u^* est solution de la minimisation. Lors que le bruit d'acquisition est aberrant, il peut être aussi intéressant d'utiliser de considérer une loi laplacienne pour les observations. Ce type d'a priori est plus connu sous le nom de LASSO (nom de l'estimateur) ou de régularisation ℓ^1.

6.3.4 Modèles à patches

Un autre moyen pour modéliser la structure des images est d'utiliser la notion d'auto-similarité. En effet, on remarque souvent dans les images des zones qui se ressemblent. Utiliser ces informations on permit de grandement améliorer les algorithmes de traitement des images. (Voir la partie sur le débruitage)

6.3.5 Apprentissage

Si l'on dispose de suffisamment d'images "nettes", on peut apprendre un espace transformé dans lequel l'image sera parcimonieuse. Cependant ces méthodes renvoient souvent des transformées proches de transformées en ondelettes. Les méthodes d'apprentissage profond (deep neural networks) estiment implicitement ces structures en utilisant de très grandes bases de données d'images.

6.4 Le problème de débruitage

6.4.1 le débruitage multi-image

Lorsque $H = [I; I...; I]$, on parle de débruitage multi-image. Les processus multi-images on pris une grande importance ces dernières années, car c'est un moyen d'améliorer à faible coût matériel la qualité des images produites. On utilise la variabilité du bruit et la redondance de l'image pour produire une image nette. Il faut noter que si suffisamment d'images sont disponibles, une information a priori n'est pas forcément nécessaire. !Attention! il faut que le modèle de bruit Gaussien soit vérifié pour espérer produire un résultat, en effet certains processus peuvent produire des bruits qui n'ont pas de moyenne (v.a. de Cauchy)

6.4.2 le débruitage mono-image

Lorsque $H = I$. On ne peut que se fier aux structures internes à l'image que l'on considère. Si l'image est lisse ont peut utiliser un a priori Gaussien basé sur la dérivée. Dans le cas contraire, on peut utiliser des a priori qui respectent le caractère discontinu des images : la variation totale (Laplacien). Les méthodes d'état de l'art utilisent les ressemblances entre patches de l'image pour débruiter.

6.5 Éléments de théorie de la transformée de Fourier

La transformée de Fourier est un opérateur linéaire qui décompose une fonction sur une base de signaux harmoniques (décomposition en fréquences). Elle est incontournable pour interpréter ou formuler des informations a priori sur des signaux ou des images. Ses 4 propriétés les plus importantes pour nos besoins ici sont la forme de son inverse, la formule de dérivation, la préservation de l'énergie et la formule pour la convolution (circulaire).

6.5.1 Transformée discrète

x est une fonction de $\mathbb{R}$ dans $\mathbb{R}$.

$$\hat{x}(f) = [\mathcal{F}x](f) = \int_{\mathbb{R}} x(t) e^{-2\pi j f t} dt \tag{39}$$

6.5.2 Transformée à temps discret

6.5.3 Transformée discrète

Soit $x \in \mathbb{R}^n$

$$\hat{x}(l) = [Fx](l) = \sum_{k=0,n-1} e^{-2\pi j \frac{kl}{n}} \tag{40}$$

La transformée inverse F^{-1} est donnée par

$$F^{-1} = \frac{1}{n} F^* \tag{41}$$

Préservation de l'énergie

$$\frac{1}{n} \|\hat{x}\|_2^2 = \|x\|_2^2 \tag{42}$$