Modélisation et statistique bayésienne computationnelle

Notes de cours

nicolas.bousquet@sorbonne-universite.fr

8 février 2022

Master 2, Sorbonne Université, 2022

Résumé

Ce cours a pour objectif de présenter d'une part les principales méthodologies de modélisation bayésienne appliquées à des problèmes d'aide à la décision en univers risqué sur des variables scalaires et fonctionnelles, et d'autre part des méthodes avancées de calcul inférentiel permettant l'enrichissement de l'information utile, en fonction de l'emploi et de la nature des modèles. Il nécessite les pré-requis suivants : notions fondamentales de probabilités et statistique, introduction aux statistiques bayésiennes, méthodes de Monte-Carlo, calcul scientifique en `R` ou/ et en `Python`.

Tout en évoluant au fil du temps, il considère plus spécifiquement les méthodes et outils (théoriques et pratiques) suivants :

- Formalisation et résolution de problèmes d'aide à la décision en univers risqué
- Représentation probabiliste des incertitudes (Cox-Jaynes, de Finetti)
- Maximum d'entropie, familles exponentielles, modélisation par données virtuelles
- Modèles hiérarchiques
- Règles d'invariance, de compatibilité et de cohérence pour les modèles bayésiens
- Méthodes d'échantillonnage (rejet, importance, Gibbs, MCMC, MCMC adaptatives, méthodes de filtrage
- Quelques perspectives : quadrature bayésienne, modèles hiérarchiques de haute dimension, modélisation bayésienne fonctionnelle, processus gaussiens, calibration par expériences numériques, critères d'enrichissement bayésiens

Tout au long du cours, des liens avec l'apprentissage statistique (*machine learning*) sont présentés.

Table des matières

1 Notations

La définition des notations suivantes sera rappelée à leur première occurrence dans le document, et elles seront réutilisées par la suite sans rappel obligatoire. D'une manière générale, les variables aléatoires (v.a.) seront notées en majuscules, les réalisations de ces variables en minuscules. Les vecteurs et matrices sont indiqués en gras, à la différence des scalaires.

NOTATIONS GÉNÉRALES

X	variable aléatoire d'étude, unidimensionnelle ou multidimensionnelle
$\mathbb{P}(.)$	mesure de probabilité usuelle
$\mathcal{B}(A)$	tribu ($\sigma-$algèbre) des boréliens sur un espace A
$P(A)$	ensemble des parties de A
$\mathbb{1}_{\{x\in A\}}$	fonction indicatrice
$\varnothing$	ensemble vide
F_X	fonction de répartition de X
f_X	fonction de densité de probabilité de $\mathbf{X}$
$F_X(.\|\theta)$	fonction de répartition de X, paramétrée par le vecteur θ
$f_X(.\|\theta)$	fonction de densité de probabilité de X, paramétrée par θ
$\ell(x_1,\ldots,x_n\|\theta)$	vraisemblance statistique des observations conditionnelle au vecteur $\boldsymbol{\theta}$
$\pi(\theta)$	densité *a priori* (bayésienne) sur le vecteur θ
$\pi(\theta\|x_1,\ldots,x_n)$	densité *a posteriori* (bayésienne) sur le vecteur θ sachant un échantillon d'observations $x_1,\ldots,x_n$
$\Pi(\theta)$	fonction de répartition *a priori*
$\Pi(\theta\|x_1,\ldots,x_n)$	fonction de répartition *a posteriori*
sign(x)	signe de x
Supp(f)	support de la densité f
X^T	transposée de X
$\lfloor x \rfloor$	partie entière de X

$\mathbb{E}_X[.]$	espérance selon la loi de X (le X peut être ôté si pas d'ambiguïté)
$\mathbb{V}_X[.]$	variance selon la loi de X
$\mathbb{C}ov_{\mathbf{X}}[.]$	matrice de covariance selon la loi de $\mathbf{X}$
$\mathbb{R}$	ensemble des réels
$\mathbb{N}$	ensemble des entiers naturels
L^2	espace des fonctions de carré intégrable
$\mathcal{C}$	notation générique pour une classe de régularité fonctionnelle
$< ., . >$	produit scalaire canonique
A^T	transposée de A
tr(A)	trace de A
diag(A)	vecteur diagonal de A
$\|A\|$	déterminant de A
∇X	gradient de X
$\mathbf{0}_d$	vecteur nul de dimension d
$X_1 \vee X_2$	vecteur de composantes maximales deux à deux
$\xrightarrow{\mathcal{L}}$	convergence en loi
$\xrightarrow{\mathbb{P}}$	convergence en probabilité
$\xrightarrow{p.s.}$	convergence presque sûre
log	logarithme népérien (ln)
exp$(.)$	exponentielle
cste	valeur constante
resp.	respectivement

Bernoulli $\mathcal{B}(p)$	$\mathbb{P}(X=1) = 1 - \mathbb{P}(X=0) = p$
Binomiale $\mathcal{B}(N,p)$	$\mathbb{P}(X \leq k) = \sum_{i=0}^{k} \frac{i!(n-i)!}{n!} p^i (1-p)^{n-i}$
Poisson $\mathcal{P}(\lambda)$	$\mathbb{P}(X \leq k) = \sum_{i=0}^{k} \frac{\lambda^i}{i!} \exp(-\lambda)$
Normale centrée réduite $\mathcal{N}(0,1)$	$F_X(x) = \Phi(x)$
Gaussienne $\mathcal{N}(\mu, \sigma^2)$	$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$
Exponentielle $\mathcal{E}(\lambda)$	$F_X(x) = 1 - \exp(-\lambda x)$
Bêta $\mathcal{B}_e(a,b)$	$\mathbb{P}(X \leq x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \mathbb{1}_{\{0 \leq x \leq 1\}}$
Gamma $\mathcal{G}(a,b)$	$F_X(x) = \frac{\gamma(a,bx}{\Gamma(a)}$ avec $\gamma(a,x) = \int_0^x t^{a-1} \exp(-t)\, dt$
Inverse gamma $\mathcal{IG}(a,b)$	$F_X(x) = \frac{\Gamma(a,b/x}{\Gamma(a)}$ avec $\Gamma(a,x) = \int_x^\infty t^{a-1} \exp(-t)\, dt$
χ^2_k (Chi-2)	$F_X(x) = \frac{\gamma(k/2, x/2}{\Gamma(k/2)}$
Student $\mathcal{S}_t(k)$	$F_X(x) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2}}{\frac{k}{2}} \int_{-\infty}^{x} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} dt$

Voir également l'Annexe B pour des précisions sur les modèles fréquemment rencontrés durant le cours.

2 Introduction et rappels

2.1 Modélisation, inférence et décision statistique

Afin d'aborder sereinement ce cours, rappelons que la Statistique (avec une majuscule) peut être vue comme une théorie de la description d'un phénomène incertain, perçu au travers de données $x_n = (x_1, \ldots, x_n)$, décrites comme des observations d'une variable X vivant dans un espace Ω. Cette incertitude du phénomène est fondamentalement supposée aléatoire ; c'est-à-dire que l'incertitude sur les valeurs que prend X ne peut pas être réduite à 0 même si le nombre n d'observations tend vers $+\infty$.

La distribution probabiliste à l'origine de ce caractère aléatoire est notée $\mathcal{P}$, et l'objectif premier de la Statistique est donc d'inférer sur $\mathcal{P}$ à partir de x_n.

Le second objectif est de pouvoir mener une prévision (ou "prédiction") d'une répétition future du phénomène. Le troisième objectif est de prendre une décision ayant des conséquences mesurables, sur la base de l'étude du phénomène.

Remarque 1 *Une intelligence artificielle (IA) dite connexioniste (qui se fonde sur l'exploitation des structures de corrélation dans des données) agglomère ces trois objectifs en fournissant une réponse finale à la prise de décision (troisième objectif). Comprendre le comportement d'une telle IA (par exemple en vue de l'étude de sa robustesse puis sa certification) nécessite donc de comprendre les fondations en modélisation et en inférence de la Statistique, et ses liens avec la théorie de la décision.*

La modélisation du phénomène consiste en une interprétation réductrice faite sur $\mathcal{P}$ par le biais d'une approche statistique qui peut être :

- non-paramétrique, qui suppose que l'inférence doit prendre en compte le maximum de complexité et à minimiser les hypothèses de travail, en ayant recours le plus souvent à l'estimation fonctionnelle ;
- paramétrique, par laquelle la distribution des observations x_n est représentée par une fonction de densité $f(x|\theta)$ où seul le paramètre θ (de dimension finie) est inconnu.

Ce cours s'intéresse uniquement au cas de l'approche statistique paramétrique. On considèrera en effet en permanence un nombre n fini (et parfois restreint) d'observations, qui ne peut en théorie servir qu'à estimer un nombre fini de paramètres. L'évaluation des outils inférentiels paramétriques peut d'ailleurs être faite avec un nombre fini d'observations.

La section suivante résume brièvement le cadre de la statistique paramétrique. Une revue des concepts fondamentaux de l'aléatoire est donnée en Annexe A, ceux-ci n'étant pas rappelés durant le cours.

2.2 Cadre statistique paramétrique

Pour formaliser la description faite précédemment, et fixer les notations pour le reste du cours, on décrit X comme une variable évoluant dans un espace mesuré et probabilisé

$$(\Omega, \mathcal{A}, \mu, \mathcal{P})$$

où :

1. Ω est l'espace d'échantillonnage des $X = x$, soit l'ensemble de toutes les valeurs possibles prises par X ;
2. la tribu (ou $\sigma-$algèbre) $\mathcal{A}$ est la collection des événements (sous-ensembles de Ω) mesurables par μ ;
3. μ est une mesure positive dominante sur $(\Omega, \mathcal{A})$.
4. $\mathcal{P}$ est une famille de distributions de probabilité dominée par μ, que suit X.

Définition 1 (Domination) *Le modèle $P \in \mathcal{P}$ est dit dominé s'il existe une mesure commune dominante μ tel que P admet une densité par rapport à μ* [1]

$$f(X) \quad = \quad \frac{d\mathcal{P}(X)}{d\mu}.$$

De manière générale, on travaillera avec $\Omega \subset I\!R^d$ avec $d < \infty$ et des échantillons de réalisations $x_n = (x_1, \ldots, x_n)$ de X. La mesure dominante μ sera Lebesgue (cas continus) ou Dirac (cas discrets). Enfin, $\mathcal{A}$ sera très généralement / classiquement choisie comme la tribu des boréliens

$$\mathcal{A} = \mathcal{B}(I\!R^d) = \sigma\left(\left\{\otimes_{i=1}^{d}]a_i, b_i];\ a_i < b_i \in I\!R\right\}\right).$$

Dans le cadre paramétrique, on supposera que $\mathcal{P}$ peut se définir par

$$\mathcal{P} = \{\mathbb{P}_\theta;\ \theta \in \Theta \subset I\!R^p\}$$

où $p < \infty$. De plus, on notera généralement $f(.|\theta)$ la densité (ou fonction de masse) induite par la dérivée de Radon-Nikodym de Pp_θ :

$$\frac{d\mathbb{P}_\theta}{d\mu} \quad = \quad f(X|\theta)$$

et parfois, lorsque X sera unidimensionnelle ($d = 1$), nous utiliserons aussi la notation classique $F(x|\theta)$ pour désigner la fonction de répartition $Pp_\theta(X \leq x)$. Par la suite, on parlera indifféremment de la variable aléatoire

$$X \quad \sim \quad f(x|\theta)$$

ou de son observation $x \sim f(x|\theta)$, et on parlera plus généralement de loi en confondant Pp_θ et $f(.|\theta)$. Enfin, la notation μ sera généralement induite dans les développements techniques :

$$\mathbb{P}_\theta(X < t) \quad = \quad \int_\Omega f(x)\mathbb{1}_{\{x<t\}}\, dx.$$

Remarque 2 *Suivant l'usage classique, les variables et processus aléatoires sont décrits par des majuscules, tandis que leurs réalisations sont décrits par des minuscules. On notera souvent v.a. pour variable aléatoire.*

Nous retrouverons et utiliserons abondamment la notion de *vraisemblance* statistique $f(\mathbf{x_n}|\theta)$, définie dans un cadre paramétrique comme la densité jointe des observations $\mathbf{x_n} = (x_1, \ldots, x_n)$ sachant le paramètre θ. Lorsque les données sont *indépendantes et identiquement distribuées* (iid) selon $f(|\theta)$, alors

$$f(\mathbf{x_n}|\theta) \quad = \quad \prod_{i=1}^{n} f(x_i|\theta).$$

D'autres formes de vraisemblance existent, notamment lorsque les données sont bruitées, censurées, etc. Voir Annexe A pour des rappels sur ces principaux concepts.

Remarque 3 (Statistique bayésienne non paramétrique) *Jusqu'à présent, θ est considéré comme appartenant à un espace Θ de dimension finie. On peut étendre la statistique bayésienne à Θ un ensemble comme $[0,1]^{I\!R}$ (l'ensemble des distributions sur $[0,1]$) ou encore l'ensemble des probabilités sur $I\!R$. Ces deux espaces ne sont pas dominés par μ. C'est le principe fondateur de la statistique non paramétrique (au sens où le paramètre n'a pas de dimension finie).*

2.3 Estimation statistique classique ("fréquentiste")

(ou fréquentielle *en meilleur français)*

1. Pour des mesures $\sigma-$finies et de part le théorème de Radon-Nykodim, ceci est équivalent à être absolument continue par rapport à μ

2.3.1 Rappel des principes

L'inférence statistique consiste à estimer "les causes à partir des effets". Ces *cause* sont réduites, dans le cadre paramétrique, au paramètre θ du mécanisme générateur des données que représente la distribution $\mathbb{P}_\theta$. Les *effets* sont naturellement les données observées $\mathbf{x_n} = (x_1, \ldots, x_n)$. De ce fait, dans un cadre paramétrique, l'inférence consiste à produire des règles d'estimation de θ à partir de $\mathbf{x_n}$. Dans ce cadre classique, θ **est supposé inconnu, mais fixe** (et à Θ n'est pas conféré la structure d'un espace probabilisé).

Les règles d'estimation les plus courantes, fondées sur de l'optimisation de critère ($M-$estimation, telles la *maximisation de la vraisemblance*

$$\hat{\theta}_n(\mathbf{x_n}) = \arg\max_\theta \log f(\mathbf{x_n}|\theta)$$

ou les *estimateurs des moindres carrés*), par *moments*, par des combinaisons linéaires de statistiques d'ordre ($L-$estimation, en général moins robuste), etc. sont nombreuses et doivent faire l'objet d'une sélection. Voir Annexe A.6 pour quelques rappels.

Pour mener cette sélection, les estimateurs sont comparés en fonction de différents critères, comme le biais, la rapidité de convergence vers la valeur supposée "vraie" θ_0 du paramètre, et d'autres différentes propriétés asymptotique (telle la nature de la loi d'un estimateur $\hat{\theta}_n(\mathbf{X_n})$, qui est une variable aléatoire dont la loi dépend de celle des X.

D'une manière générale, si l'on note $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X_n})$ tout estimateur classique de θ, à de rares exceptions près la validité de ce choix d'estimateur est dépendante du caractère *reproductible* et *échangeable* des données $x_1, \ldots, x_n$ conditionnellement à θ.

Définition 2 (Échangeabilité.) *Les données $x_1, \ldots, x_n$ sont dites échangeables si, pour toute permutation σ : $\mathbb{N}^n \to \mathbb{N}^n$, la loi jointe $f(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ est indépendante de σ.*

Cette validité, donc en général fondée sur des critères asymptotiques ($n \to \infty$), s'exprime en termes de *région de confiance* (cf. Annexe -16)

$$\mathbb{P}\left(\hat{\theta}_n - \theta \in A_\alpha\right) \quad = \quad 1 - \alpha.$$

En général, la distribution $\mathbb{P}$ de l'estimateur est inconnue pour $n < \infty$, elle est le plus souvent approximée asymptotiquement via un théorème de convergence en loi, tel que :

$$\text{si } x_1, \ldots, x_n \text{ sont iid} \quad \Sigma_n^{-1/2}\left(\hat{\theta}_n - \theta_0\right) \quad \xrightarrow{\mathcal{L}} \quad \mathcal{Q}$$

où $\mathbb{E}_\mathcal{Q}[X] = 0$ et $\mathbb{V}_\mathcal{Q}[X] = 1$. Ici Σ_n est lui-même un estimateur consistant de la matrice de covariance de $\hat{\theta}_n$, et le résultat précédent est issu de l'usage de la méthode Delta de dérivation des lois d'estimateur, ainsi que du théorème de Slutstky de composition des convergences.

Remarque 4 *On utilise souvent le terme d'*inférence *en machine learning pour désigner la tâche de prévision (*prediction*) d'un modèle appris, et l'entraînement la phase d'estimation de ce modèle. En ce sens, le mot* inférer *est tout aussi valide, car il signifie "aller des principes vers la conclusion".*

2.3.2 Difficultés pratiques, théoriques et conceptuelles

Ce *paradigme* [2] forme, depuis les travaux de Fisher, Neyman et Pearson dans la première moitié du XXème siècle, le socle théorique de la majeure partie des études statistiques. Il n'est pas cependant sans poser quelques problèmes :

(a) Tout d'abord, les difficultés rencontrées sont **pratiques** : face à de petits échantillons, le cadre asymptotique ne tient plus : la comparaison des estimateurs doit alors reposer sur des critères non asymptotiques [3], et on perd l'usage des résultats de la convergence en loi et ses dérivées (ex : production des régions de confiance). De même, la plupart des résultats utiles pour mener des tests statistiques (voir Annexe A.2) deviennent inutilisables.

2. Modèle censé être cohérent d'un univers scientifique, faisant l'objet d'un consensus.
3. Parmi ces critères, les inégalités de concentration (Markov, Bienaymé-Chebychev, Bernstein, etc.) se révèlent fondamentales.

(b) Des difficultés peuvent aussi être **théoriques**.

1. Ainsi, pour de nombreux modèles complexes, tels les modèles à espace d'états (qui font partie des modèles à données latentes), tels que les modèles de population, la dimension de Θ peut augmenter linéairement avec le nombre de données. Dans ce cas, la théorie asymptotique classique n'a plus de sens.

   EXEMPLE 1. *On considère une population suivie annuellement, n étant le nombre d'années de mesure. A chaque année est associée un paramètre spécifique de renouvellement de la population. La dimension augmente donc linéairement avec le nombre de donnée, si aucune réduction de dimension (par exemple via des covariables connues) n'est effectuée.*

2. Plus fondamentalement, l'utilisation d'un estimateur fréquentiste peut contredire le principe fondamental de la statistique inférentielle :

   Définition 3 (Principe de vraisemblance) *L'information (= l'ensemble des inférences possibles) apportée par une observation x sur θ est entièrement contenue dans la fonction de vraisemblance $\ell(\theta|x) = f(x|\theta)$. De plus, si x_1 et x_2 sont deux observations qui dépendent du même paramètre θ, telle qu'il existe une constante c satisfaisant*

   $$\ell(\theta|x_1) \quad = \quad c\ell(\theta|x_2) \quad \forall\theta \in \Theta,$$

   alors elles apportent la même information sur θ et doivent conduire à la même inférence.

   Exercice 1 (Adapté de [27]) *Soient (x_1, x_2) deux réalisations aléatoires. Nous disposons de deux candidats pour la loi jointe de ces observations : $x_i \sim \mathcal{N}(\theta, 1)$ ou encore*

   $$g(x_1, x_2|\theta) \quad = \quad \pi^{-3/2}\frac{\exp\left\{-(x_1 + x_2 - 2\theta)^2/4\right\}}{1 + (x_1 - x_2)^2}.$$

   Quel est l'estimateur du maximum de vraisemblance de θ dans chacun des cas ? Que constate-on ?

3. Citons également le fait que l'estimateur du maximum de vraisemblance (EMV), considéré généralement comme le plus efficace (atteignant la borne de Cramer-Rao et asymptotiquement sans biais dans la plupart des cas), peut ne pas exister ou être unique.

   EXEMPLE 2. *Modèles à paramètre de position, modèles de mélange...*

   Par ailleurs, l'usage de l'EMV pose un autre problème, qui contredit le principe de vraisemblance : es régions de confiance de la forme *(test du rapport de vraisemblance)*

   $$\mathcal{C} \quad = \quad \left\{\theta;\ \frac{\ell(\theta|x)}{\ell(\hat{\theta}|x)} \geq c\right\}$$

   qui sont les plus petites asymptotiquement, ne dépendront pas uniquement de la fonction de vraisemblance si la borne c doit être choisie de manière à obtenir un niveau de confiance α.

4. Une dernière difficulté théorique posée par les estimateurs fréquentiels apparaît lorsqu'on cherche à mener une *prévision*. Considérons en effet Soit $\mathbf{X_n} = (X_1, \ldots, X_n) \overset{iid}{\sim} f(.|\theta)$. On cherche à prévoir le plus précisément possible ce que pourrait être le prochain tirage X_{n+1}. Dans l'approche classique, on utilise

   $$f(X_{n+1}|X_1, \ldots, X_n, \hat{\theta}_n) \quad = \quad \frac{f(X_1, \ldots, X_n, X_{n+1}|\hat{\theta}_n)}{f(X_1, \ldots, X_n|\hat{\theta}_n)}$$

   et ce faisant on utilise deux fois les données et on risque de sous-estimer les incertitudes (intervalles de confiance) en renforçant arbitrairement la connaissance.

(c) Enfin, les difficultés peuvent être **d'ordre conceptuel**. En effet, le sens donné à une probabilité est, dans la statistique bayésienne, celui d'une *limite de fréquence*, et la notion de *confiance* est uniquement fondée sur la répétabilité des expériences peut ne pas être pertinente.

EXEMPLE 3. *Le premier pari d'une course de chevaux ?*

En prévision, nous souhaiterions connaître parfaitement l'incertitude sur le mécanisme générateur de X, mais c'est une tâche impossible en pratique. Dans de nombreux contexte, toute variable aléatoire est la représentation mathématique d'une grandeur soumise à deux types d'incertitude :

1. $\mathbb{P}_\theta$ représente la partie *aléatoire* du phénomène considéré ;
2. l'estimation de θ souffre d'une incertitude *épistémique*, réductible si de l'information supplémentaire (données) est fournie (typ. : données).

L'approche classique des statistiques souffre donc de difficultés qui limitent son usage à des situations généralement restreintes à l'asymptotisme. Elle constitue en en fait une *approximation* d'un paradigme plus vaste, celui de la *statistique bayésienne*, qui permet notamment de *correctement appréhender la gestion des incertitudes en estimation, prévision, et en aide à la décision*.

Remarque 5 (Écriture fiduciaire) *L'écriture fiduciaire $\ell(\theta|x) = f(x|\theta)$ a été proposée au début du XXème siècle pour témoigner du fait qu'on cherche à* mesurer *l'éventail des valeurs possibles de θ sachant l'observation des x_i. Toutefois, il s'agissait d'une confusion entre la définition d'un estimateur statistique et celle d'une véritable variable aléatoire nécessitant l'ajout d'une mesure dominante sur θ. Il vaut mieux ne pas l'utiliser pour ne pas oublier le sens statistique d'une vraisemblance (loi jointe des données).*

2.4 Principes de la statistique bayésienne

2.4.1 Paradigme

Le paradigme de la statistique bayésienne paramétrique part du principe que le **vecteur θ est une variable aléatoire**, vivant dans un espace probabilisé (on utilisera généralement $(\Theta, \Pi, \mathcal{B}(\Theta))$).

En reprenant la formulation *L'inférence statistique consiste à estimer "les causes à partir des effets"* au § 2.3.1, cela revient à associer X aux effets, et θ aux causes, et d'"estimer ces causes" par la mise à jour de la distribution (mesure) $\Pi(\Theta)$ via la *règle de Bayes* :

Si C (cause) et E (effet) sont des évènements tels que $P(E) \neq 0$, alors

$$\begin{aligned} P(C|E) &= \frac{P(E|C)P(C)}{P(E|C)P(C) + P(E|C^c)P(C^c)} \\ &= \frac{P(E|C)P(C)}{P(E)} \end{aligned}$$

Il s'agit d'un principe d'*actualisation*, décrivant la mise à jour de la vraisemblance de la cause C de $P(C)$ vers $P(C|E)$.

Ce paradigme a historiquement été proposé par Bayes (1763) puis Laplace (1795), qui ont supposé que l'*incertitude sur θ* pouvait être décrite par une distribution de probabilité Π de densité $\pi(\theta)$ sur Θ, appelée *loi a priori*. On notera en général

$$\theta \sim \pi$$

Formulation en densité. Sachant des données $\mathbf{x_n}$, la mise à jour de cette loi *a priori* s'opère par le conditionnement de θ à $\mathbf{x_n}$; on obtient la *loi a posteriori*

$$\pi(\theta|\mathbf{x_n}) = \frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{\int_\Theta f(\mathbf{x_n}|\theta)\pi(\theta)\, d\theta} \tag{1}$$

Définition 4 *Un modèle statistique bayésien est constitué d'un modèle statistique paramétrique (ou vraisemblance) $f(x|\theta)$ et d'une mesure a priori $\pi(\theta)$ pour les paramètres.*

En conséquence, là où la statistique classique s'attache à définir des procédures d'estimation ponctuelle de θ, la statistique bayésienne va s'attacher à définir des procédures d'estimation de la loi *a posteriori* $\pi(\theta|\mathbf{x_n})$.

Exercice 2 (Bayes (1763)) *Une boule de billard Y_1 roule sur une ligne de longueur 1, avec une probabilité uniforme de s'arrêter n'importe où. Supposons qu'elle s'arrête à la position θ. Une seconde boule Y_2 roule alors n fois dans les mêmes conditions, et on note X le nombre de fois où Y_2 s'arrête à gauche de Y_1. Connaissant X, quelle inférence peut-on mener sur θ ?*

Exercice 3 (Loi gaussienne / loi exponentielle) *Soit une observation $x \sim \mathcal{N}(\theta, \sigma^2)$ où σ^2 est connu. On choisit a priori*

$$\theta \quad \sim \quad \mathcal{N}(m, \rho\sigma^2)$$

Quelle est la loi a posteriori de θ sachant x ? Même question en supposant que $X \sim \mathcal{E}(\lambda)$ et

$$\lambda \quad \sim \quad \mathcal{G}(a, b).$$

Définition 5 (Loi impropre) *Une "loi impropre" est une mesure a priori $\sigma-$finie qui vérifie $\int_\Theta \pi(\theta)\, d\theta = \infty$.*

La mesure de Lebesgue sur un ouvert est un exemple de loi impropre. Le choix de manier ce type de mesure peut sembler étrange, mais ce choix peut s'avérer en fait particulièrement intéressant. Par exemple, travailler avec une loi normale centrée à grande variance pour approcher une "loi uniforme sur $\mathbb{R}$" peut être précieux. Une telle loi *a priori* n'a cependant d'intérêt que si la loi *a posteriori* correspondante existe. On se limitera donc aux lois impropres telles que la *loi marginale* soit bien définie :

$$m_\pi(x) \quad = \quad \int_\Theta f(x|\theta)\, d\pi(\theta) \; < \; \infty$$

Exercice 4 (Loi uniforme généralisée) *Soit $X \sim \mathcal{N}(\mu, \sigma^2)$ et $d\pi(\mu) = d\mu$ (mesure de Lebesgue). Que vaut $m_\pi(x)$?*

Exercice 5 (Loi d'échelle) *Soit $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ et $\pi(\mu, \sigma) = 1/\sigma$ avec $\Theta = \mathbb{R} \times \mathbb{R}_*^+$. Que vaut $m_\pi(x_1, \ldots, x_n)$? La mesure $\pi(\mu, \sigma)$ peut-elle être utilisable ?*

Dans le cas où π est une mesure impropre $\sigma-$finie, on considère $\pi^*(\theta) = c\pi(\theta)$ où c est une constante arbitraire. Elle doit être sans influence pour l'usage du modèle bayésien. On peut facilement voir que c'est bien le cas dans le calcul *a posteriori* (exercice), puisqu'elle apparaît aussi bien au numérateur qu'au dénominateur de l'expression (1) : on a bien

$$d\pi^*(\theta|X) \quad = \quad d\pi(\theta|X).$$

Ainsi, l'usage de lois impropres *a priori* est justifié si la loi *a posteriori* est propre [4] car cette dernière ne dépend pas de la constante multiplicative c inconnue. C'est à rapprocher du principe de vraisemblance énoncé précédemment.

2.4.2 Fondations théoriques

Les fondations théoriques de la statistique bayésienne seront progressivement investiguées durant le cours, notamment en lien avec la section consacrée à la théorie de la décision (§ 3), mais il est important de connaître un premier résultat, dû originellement à De Finetti. Il s'agit d'un *théorème de représentation*, c'est-à-dire un théorème qui permet de justifier un choix de représentation probabiliste des variations de θ dans Θ.

4. C'est-à-dire intégrable : une loi de probabilité qui *mesure* les informations une fois les données connues.

Théorème 1 (De Finetti (1931)) *Soit $X_1, \ldots, X_n, \ldots$ une séquence* échangeable *de variables aléatoires binaires (0-1) de probabilité jointe P. Alors il existe une mesure de probabilité unique $\pi(\theta)$ telle que*

$$P(X_1 = x_1, \ldots, X_1 = x_n, \ldots) \quad = \quad \int_\Theta f(x_1, \ldots, x_n, \ldots|\theta)\pi(\theta)\, d\theta$$

où $f(x_1, \ldots, x_n|\theta)$ est la vraisemblance d'observations iid *de Bernoulli (également notée $\ell(\theta|x_1, \ldots, x_n, \ldots)$.*

Nous admettrons ce théorème ainsi que ses nombreux dérivés. En effet, il a été généralisé successivement par Hewitt, Savage (1955), Diaconis, Freedman (1980) pour l'ensemble des distributions discrétisées puis continues.

Selon ce théorème, la modélisation bayésienne apparaît comme une modélisation statistique naturelle de *variables corrélées mais échangeables*. L'existence formelle d'une mesure *a prior* (ou *prior* dans la suite de ce cours) $\pi(\theta)$ est assurée en fonction du mécanisme d'échantillonnage, qui apparaît dès lors comme une simplification d'un mécanisme par essence mal connu ou inconnu.

Un autre théorème fondamental qui nous permet de justifier l'usage du cadre bayésien est le *théorème de Cox-Jaynes*, qui sera introduit plus tard dans le cours (Section 5). Il est fondé sur une *axiomatique de la représentation de l'information* et il constitue aujourd'hui à la fois une autre façon de défendre le choix de la théorie des probabilités pour le théorème fondamental de l'inte.

Le prior correspond donc à une mesure d'information incertaine à propos de θ, et (comme on le verra) un *pari probabiliste* pour certains théoriciens des probabilités. Cette probabilisation de θ va permettre de répondre de façon pratique :

- à la nécessité de *satisfaire le principe de vraisemblance* ;
- à la nécessité de *tenir compte de toutes les incertitudes épistémiques* s'exprimant sur θ, en particulier dans un objectif de *prévision* ;
- de distinguer ces incertitudes de l'incertitude *aléatoire*, intrinsèque au modèle $f(.|\theta)$;
- à la possibilité d'intégrer de la connaissance *a priori* sur le phénomène considéré, autre que celle apportée par les données $\mathbf{x_n}$;
- à la nécessité de faire des choix de modèles en évitant les difficultés des tests statistiques classiques ;
- l'invariance $\pi(\theta|\mathbf{x_n}) = \pi(\theta)$ permet en outre d'identifier des problèmes d'*identifiabilité* du modèle d'échantillonnage $X \sim f(x|\theta)$

2.4.3 Plan du cours

Ce cours va considérer successivement plusieurs aspects du choix et de la mise en œuvre du cadre statistique bayésien. Il cherche à fournir les éléments nécessaires pour répondre aux questions fondamentales suivantes :

(a) Quant le paradigme bayésien est-il préférable ? Hors du contexte spécifique des petits échantillons, pour lesquels la statistique classique apporte des réponses limitées, cette question revient d'abord à comprendre que la statistique bayésienne est d'abord une *théorie de la décision, centrale en apprentissage statistique* et dans la formalisation du travail du statisticien. Le cadre décisionnel proposé par la statistique bayésienne améliore la vision fréquentielle du monde, et s'accorde avec elle lorsque l'information apportée par les données augmente. Ces deux aspects sont considérés dans les Sections 3 et 4.

(b) Comment construire une ou plusieurs mesures *a priori* $\pi(\theta)$ **?** Cette partie importante du cours est traitée plusieurs sections. La section 5 propose d'abord de formuler les principes généraux de compréhension et de représentation probabiliste de l'information incertaine. Sur la base de ces principes, issus d'une axiomatique, la section 6 proposera un panorama des méthodes et outils de la modélisation bayésienne.

(b) Comment faire du calcul bayésien ? La mise en oeuvre concrète des outils et méthodes de la statistique bayésienne suppose de pouvoir manipuler les lois *a posteriori* $\pi(\theta|\mathbf{x_n})$. Les méthodes par simulation (échantillonnage) et les approches par approximation variationnelle font aujourd'hui partie des outils courants pour ce faire. Elles seront abordées dans la section 7.

2.5 Liens avec le *machine learning*

Dans une optique de *régression supervisée*, le paradigme du *machine learning* propose de produire un estimateur (ou *prédicteur*) de la fonction inconnue $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ (plus généralement vers un espace euclicien de dimension d_2) telle que

$$Y \quad = \quad g(X)$$

à partir de couples connus $\mathbf{z_n} = (x_i, y_i)_{1 \leq i \leq n}$, où chaque x_i est un ensemble de d_1 *covariables* et chaque y_i est un *label* de dimension d_2. La recette est la suivante :

1. Faire un choix g_θ pour "mimer" g ;
   - Dans un problème de régression linéaire, θ (noté généralement β) est le vecteur des coefficients de la régression).
   - g_θ est un réseau de neurones d'architecture choisie, alors θ constitue un vecteur de paramètres structurant pour ce réseau (poids, biais, nombre de neurones par couche, éventuellement les choix de fonctions d'activation, etc.).
2. Décider d'une *fonction de coût*[5] souvent définie comme la somme d'un regret quadratique et d'une pénalité
$$L(\theta|\mathbf{z_n}) \quad = \quad \sum_{i=1}^{n} \|y_i - g_\theta(x_i))\|_2^2 + \text{pen}(\theta) \tag{2}$$
où pen(θ) dépend de la complexité du problème.
3. Définir l'estimateur $\hat{\theta}_n$ par
$$\hat{\theta}_n \quad = \quad \arg\min_{\theta \in \Theta} L(\theta|\mathbf{z_n}) \tag{3}$$
et choisir une méthode pour minimiser la fonction de coût (exemple : rétropropagation du gradient).

On peut alors réécrire l'équation (3) de la façon suivante :

$$\begin{aligned}
\hat{\theta}_n &= \arg\max_{\theta \in \Theta} \{-L(\theta|\mathbf{z_n})\}, \\
&= \arg\max_{\theta \in \Theta} \log \{f_g(\mathbf{z_n}|\theta)\pi(\theta)\}, \\
&= \arg\max_{\theta \in \Theta} \log \pi(\theta|\mathbf{z_n}), \\
&= \arg\max_{\theta \in \Theta} \pi(\theta|\mathbf{z_n})
\end{aligned}$$

où $f_g((\mathbf{z_n}|\theta)$ est une vraisemblance de forme gaussienne de $\mathbf{z_n}$ et

$$\pi(\theta) \quad \propto \quad \exp(-2\text{pen}(\theta)).$$

Le cadre bayésien explique le sens d'une pénalisation comme celui d'une transformation d'une mesure *a priori*, et l'optimisation en *machine learning* consiste à estimer le mode d'une distribution *a posteriori* (calcul simplificateur de la véritable inférence, qui serait celle de la loi $\pi(\theta|\mathbf{z_n})$ toute entière).

EXEMPLE 4. *La régression* lasso *propose un choix de pénalisation pen$(\theta) = \lambda\|\theta\|_1$, qui correspond à l'action d'un prior $\pi(\theta) \propto \exp(-2\lambda\|\theta\|_1)$. De même, la régularisation* ridge *est similaire à l'action d'un prior $\pi(\theta) \propto \exp(-2\lambda\|\theta\|_2^2)$.*

5. On retrouvera ce terme plus tard dans la partie du cours consacré à la théorie de la décision (Section 3).

2.6 Quelques lectures conseillées

Ce cours s'inspire de plusieurs ouvrages et résultats publiés ces dernières années. L'étudiant intéressé par une vision générale du cadre pourra approfondir les aspects théoriques à partir de l'ouvrage de référence [27]. Une démarche plus appliquée de la statistique bayésienne bénéficie d'une présentation pédagogique dans l'ouvrage [23]. Les aspects computationnels historiques sont au coeur des ouvrages de référence [28, 19]. Le cadre décisionnel de la théorie bayésienne, dans un contexte d'usage concret (et relié à l'industrie), fait l'objet de l'article (français) [12].

L'article de revue récent [34] offre enfin une vision générale du cadre statistique bayésien, et complète utilement les lectures précédentes.

3 Éléments de théorie de la décision

L'objectif général de la plupart des études inférentielles est de fournir une *décision* au statisticien (ou au client) à partir du phénomène modélisé par $X \sim f(x|\theta)$ (dans le cadre paramétrique). Il faut donc exiger un *critère d'évaluation* des procédures de décision qui :

- prenne en compte les conséquences de chaque décision
- dépende des paramètres θ du modèle, càd du *vrai état du monde (ou de la nature)*.

Un autre type de décision est d'*évaluer* si un nouveau modèle descriptif est compatible avec les données expérimentales disponibles (*choix de modèle*). Le critère en question est habituellement nommé **fonction de coût**, **fonction de perte** ou **utilité** (opposé du coût).

EXEMPLE 5. *Acheter des capitaux selon leurs futurs rendement θ, déterminer si le nombre θ des SDF a augmenté depuis le dernier recensement...*

Formellement, pour le modèle $X \in \{\Omega, \mathcal{B}, \{\mathbb{P}_\theta,\ \theta \in \Theta\}$ on définit donc trois espaces de travail :

- $\Omega =$ espace des observations x ;
- $\Theta =$ espace des paramètres θ ;
- $\mathcal{D} =$ espace des décisions possibles d.

En général, la décision $d \in \mathcal{D}$ demande d'évaluer (*estimer*) une *fonction d'intérêt* $h(\theta)$, avec $\theta \in \Theta$, estimation fondée sur l'observation $x \in \Omega$. On décrit alors $\mathcal{D}$ comme l'ensemble des fonctions de Θ dans $h(\Theta)$ où h dépend du contexte :

- si le but est d'estimer θ alors $\mathcal{D} = \Theta$;
- si le but est de mener un test, $\mathcal{D} = \{0, 1\}$.

3.1 Existence d'une fonction de coût

La *théorie de la décision* suppose alors que :

- chaque décision $d \in \mathcal{D}$ peut être évaluée et conduit à une *récompense* (ou *gain*) $r \in \mathcal{R}$
- l'espace $\mathcal{R}$ des récompenses peut être *ordonné totalement* :

  (1) $r_1 \preceq r_2$ ou $r_2 \preceq r_1$;

  (2) si $r_1 \preceq r_2$ et $r_2 \preceq r_3$ alors $r_1 \preceq r_3$;
- l'espace $\mathcal{R}$ peut être étendu à l'espace $\mathcal{G}$ des distributions de probabilité dans $\mathcal{R}$;
  - les décisions peuvent être alors partiellement aléatoires
- la relation d'ordre $\preceq$ peut être étendue sur les **moyennes** des récompenses aléatoires *(et donc sur les distributions de probabilité correspondantes)* ;
  - il existe au moins un ordre partiel sur les gains (même aléatoires) et un gain optimal.

Ces axiomes expriment une certaine **hypothèse de rationalité du décideur**. Ils impliquent l'existence d'une **fonction d'utilité** $U(r)$ permettant de trier les gains aléatoires. Cette utilité ne dépend en fait que de θ et de d : on la note donc $U(\theta, d)$ Elle peut être vue comme une *mesure de proximité* entre la décision proposée d et la vraie valeur (inconnue) θ.

Définition 6 *On appelle* fonction de coût *ou* fonction de perte *une fonction L mesurable de $\Theta \times \mathcal{D}$, telle que*

$$L(\theta, d) \quad = \quad -U(\theta, d),$$

à valeurs réelles positives :

$$L : \Theta \times \mathcal{D} \longrightarrow I\!R^+.$$

La fonction de coût est définie selon le problème étudié et constitue l'armature d'un problème de décision statistique (qui comprend notamment les problèmes d'estimation).

EXEMPLE 6. *On considère le problème de l'estimation de la moyenne θ d'un vecteur gaussien*

$$x \quad \sim \mathcal{N}_p(\theta, \Sigma)$$

où Σ est une matrice diagonale connue avec pour éléments diagonaux σ_i^2 $(i = 1, \ldots, p)$. Dans ce cas $\mathcal{D} = \Theta = I\!R^p$ et d représente une évaluation de θ. S'il n'y a pas d'information additionnelle disponible sur ce modèle, il paraît logique de choisir une fonction de coût qui attribue le même poids à chaque composante, soit un coût de la forme

$$\sum_{i=1}^{p} L\left(\frac{x_i - \theta_i}{\sigma_i}\right) \quad \textit{avec } L(0) = 0.$$

Par normalisation, les composantes avec une grande variance n'ont pas un poids trop important. Le choix habituel de L est le coût **quadratique** $L(t) = t^2$.

Dans un contexte de gain aléatoire, l' *approche fréquentiste* propose de considérer le *coût moyen* ou *risque fréquentiste*. Pour une fonction de coût quadratique, le risque fréquentiste est souvent appelé *risque quadratique*. On appelle $\delta : \Omega \mapsto \mathcal{D}$ minimisant un risque un estimateur et $\delta(x)$ une estimation.

Définition 7 (Risque fréquentiste) *Pour $(\theta, \delta) \in \Theta \times \mathcal{D}$, le risque fréquentiste est défini par*

$$R(\theta, \delta) \quad = \quad \mathbb{E}_\theta \left[L(\theta, \delta(x))\right] \; = \; \int_\Omega L(\theta, \delta(x)) f(x|\theta) \, dx$$

où $\delta(x)$ est la règle de décision = attribution d'une décision connaissant l'observation x.

Cette définition du risque n'est pas sans poser problème. En effet :

- le critère évalue les procédures d'estimation selon leurs *performances à long terme* et non directement pour une observation donnée ;
- on suppose tacitement que le problème sera rencontré de nombreuses fois pour que l'évaluation en fréquence ait un sens

$$R(\theta, \delta) \quad \simeq \quad \text{coût moyen sur les répétitions;}$$

- ce critère n'aboutit pas à un *ordre total* sur les procédures de construction d'estimateur.

Exercice 6 *Soient x_1 et x_2 deux observations de la loi définie par*

$$P_\theta(x = \theta - 1) \quad = \quad P_\theta(x = \theta + 1) \; = \; 1/2 \quad \textit{avec } \theta \in I\!R$$

Le paramètre d'intérêt est θ (donc $\mathcal{D} = \Theta$) et il est estimé par δ sous le coût

$$L(\theta, \delta) \quad = \quad 1 - \mathbb{1}_\theta(\delta)$$

appelé coût 0-1*, qui pénalise par 1 toutes les erreurs d'estimation quelle que soit leur magnitude (grandeur). Soit les estimateurs*

$$\begin{aligned}\delta_1(x_1, x_2) &= \frac{x_1 + x_2}{2},\\ \delta_2(x_1, x_2) &= x_1 + 1,\\ \delta_3(x_1, x_2) &= x_2 - 1.\end{aligned}$$

Calculez les risques $R(\theta, \delta_1)$, $R(\theta, \delta_2)$ et $R(\theta, \delta_3)$. Quelle conclusion en tirez-vous ?

L'approche bayésienne de la théorie de la décision considère que *le coût $L(\theta, d)$ doit plutôt être moyenné sur tous les états de la nature possibles*. Conditionnellement à l'information x disponible, ils sont décrits par la loi *a posteriori* $\pi(\theta|x)$. On définit donc le coût moyenné *a posteriori*, ou *risque a posteriori*, qui est l'erreur moyenne résultant de la décision d pour un x donné.

Définition 8 (Risque *a posteriori***)**

$$R_P(d|\pi, x) = \int_\Theta L(\theta, d)\pi(\theta|x)\, d\theta.$$

On peut enfin définir le risque fréquentiste intégré sur les valeurs de θ selon leur distribution *a priori*. Associant un nombre réel à chaque estimateur δ, ce risque induit donc une *relation d'ordre total* sur les procédures de construction d'estimateur. Il permet donc de définir la notion d'estimateur bayésien (ou estimateur de Bayes).

Définition 9 (Risque intégré) *À fonction de coût (perte) donnée, le risque intégré est défini par*

$$R_B(\delta|\pi) = \int_\Theta \int_\Omega L(\theta, \delta(x)) f(x|\theta)\, dx\, \pi(\theta)\, d\theta.$$

Définition 10 (Estimateur bayésien et risque de Bayes) *Un* estimateur de Bayes *associé à une distribution a priori π et une fonction de coût L est défini par*

$$\delta^\pi = \arg\min_{\delta \in \mathcal{D}} R_B(\delta|\pi)$$

la valeur $r(\pi) = R_B(\delta^\pi|\pi)$ est alors appelée **risque de Bayes***.*

Le résultat suivant peut être obtenu par interversion d'intégrales (théorème de Fubini). *Modulo* un peu de machinerie technique, on peut montrer que celui-ci reste vrai même si $\int_\Theta \pi(\theta) d\theta = \infty$ (mesure *a priori* non informative) à condition que $\int_\Theta \pi(\theta|x) d\theta = 1$.

Théorème 2 *Pour chaque $x \in \Omega$,*

$$\delta^\pi(x) = \arg\min_{d \in \mathcal{D}} R_P(d|\pi, x). \tag{4}$$

Un corollaire est le suivant : s'il existe $\delta \in \mathcal{D}$ tel que $R_B(\delta|\pi) < \infty$, et si $\forall x \in \Omega$ l'équation (4) est vérifiée, alors $\delta^\pi(x)$ est un estimateur de Bayes.

3.2 Supériorité des estimateurs de Bayes sur les estimateurs fréquentistes

Le risque minimax est le coût fréquentiste minimum dans le cas le moins favorable (l'écart entre θ et δ, càd l'*erreur d'estimation*, est maximal(e)).

Définition 11 (Risque minimax) *On définit le risque minimax pour la fonction de coût L par*

$$\bar{R} \quad = \quad \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta) \; = \; \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[L(\theta, \delta(x)) \right].$$

Théorème 3 *Le risque de Bayes est toujours plus petit que le risque minimax*

$$\mathrm{R} = \sup_{\pi} r(\pi) = \sup_{\pi} \inf_{\delta \in \mathcal{D}} R_B(\delta|\pi) \quad \leq \bar{R}.$$

Si elle existe, une distribution *a priori* π^* telle que $r(\pi^*) = R$ est appelée *distribution a priori la moins favorable*. Ainsi, l'apport d'information *a priori* $\pi(\theta)$ ne peut qu'améliorer l'erreur d'estimation, même dans le pire des cas.

Définition 12 (Inadmissibilité d'un estimateur) *Un estimateur δ_0 est dit* inadmissible *s'il existe un estimateur δ_1 qui* **domine** *δ_0 au sens du risque fréquentiste, càd si*

$$R(\theta, \delta_0) \quad \geq \quad R(\theta, \delta_1) \quad \forall \theta \in \Theta$$

et $\exists \theta_0$ tel que $R(\theta_0, \delta_0) > R(\theta_0, \delta_1)$. Sinon, il est dit admissible.

Théorème 4 *Si un estimateur de Bayes δ^π associé à une mesure a priori π (probabiliste ou non) est tel que le risque $R(\theta, \delta^\pi) < \infty$ et si la fonction $\theta \mapsto R(\theta, \delta)$ est continue sur Θ, alors δ^π est admissible.*

Théorème 5 *Si un estimateur de Bayes δ^π associé à une mesure a priori π (probabiliste ou non) et une fonction de coût L est unique, alors il est admissible.*

Notons que les critères de minimaxité et d'admissibilité sont éminement *fréquentistes* (car construits à partir du risque fréquentiste). Selon ces critères fréquentistes, les estimateurs de Bayes font mieux ou au moins aussi bien que les estimateurs fréquentistes :

- leur risque minimax est toujours égal ou plus petit ;
- ils sont tous admissibles (si le risque de Bayes est bien défini).

Les estimateurs de Bayes, plus généralement, sont souvent optimaux pour les concepts fréquentistes d'optimalité et devraient donc être utilisés même lorsque l'information *a priori* est absente. On peut ignorer la signification d'une distribution *a priori* tout en obtenant des estimateurs corrects d'un point de vue fréquentiste.

3.3 Choix d'une fonction de coût

La fonction de coût L est l'élément fondamental du choix d'un estimateur. Le choix dépend du contexte décisionnel et s'écrit souvent sous la forme

$$L \quad = \quad \text{Coût financier, etc.} - \text{Bénéfice.}$$

Une alternative, lorsqu'il est difficile de la construire, est de faire appel à des *fonctions de coût usuelles, mathématiquement simples et de propriétés connues*. L'idée est simplement de construire une "distance" usuelle entre $\theta \in \Theta$ et $d \in \mathcal{D}$ permettant une bonne optimisation (convexe par exemple).

EXEMPLE 7.**Fonction de coût quadratique** *Soit $\mathcal{D} = \Theta$. On pose*

$$L(\theta, \delta) \quad = \quad \|\theta - \delta\|^2. \tag{5}$$

Cette fonction de coût constitue le critère d'évaluation le plus commun. Elle est convexe (mais pénalise très (trop) fortement les grands écarts peu vraisemblables). Elle est justifié par sa simplicité, le fait qu'elle permet de produire des estimateurs de Bayes intuitifs, et qu'elle peut être vue comme issue d'un développement limité d'un coût symétrique complexe.

Proposition 1 *L'estimateur de Bayes associé à toute loi a priori π et au coût (5) est l'espérance (moyenne) de la loi a posteriori $\pi(\theta|\mathbf{x_n})$*

La fonction de coût absolu, également convexe, croît plus lentement que le coût quadratique et ne surpénalise pas les erreurs grandes et peu vraisemblables.

EXEMPLE 8.**Fonction de coût absolu (Laplace 1773)** *Soit $\mathcal{D} = \Theta$ et* $\dim \Theta = 1$. *On pose*

$$L(\theta, \delta) \quad = \quad |\theta - \delta| \tag{6}$$

ou plus généralement une fonction linéaire par morceaux

$$L_{c_1,c_2}(\theta, \delta) \quad = \quad \left\{ \begin{array}{ll} c_2(\theta - \delta) & \textit{si } \theta > \delta \\ c_1(\delta - \theta) & \textit{sinon} \end{array} \right. \tag{7}$$

Proposition 2 *L'estimateur de Bayes associé à toute loi a priori π et au coût (7) est le fractile $c_1/(c_1 + c_2)$ de la loi a posteriori $\pi(\theta|\mathbf{x_n})$. En particulier, la médiane de la loi a posteriori est l'estimateur de Bayes lorsque $c_1 = c_2$ (qui sont donc des coûts associés à la sous-estimation et la surestimation de θ).*

La fonction de coût 0-1, non quantitative, est utilisé dans l'approche statistique classique pour construire des test d'hypothèse.

EXEMPLE 9.**Fonction de coût 0-1**

$$L(\theta, \delta) \quad = \quad \left\{ \begin{array}{ll} 1 - \delta & \textit{si } \theta \in \Theta_0 \\ \delta & \textit{sinon} \end{array} \right. \tag{8}$$

Le risque fréquentiste associé est

$$R(\theta, \delta) \quad = \quad \mathbb{E}_\theta[L(\theta, \delta(x))] \; = \; \left\{ \begin{array}{ll} P_\theta(\delta(x) = 0) & \textit{si } \theta \in \Theta_0 \\ P_\theta(\delta(x) = 1) & \textit{sinon} \end{array} \right.$$

Proposition 3 *L'estimateur de Bayes associé à toute loi a priori π et au coût 0-1 est*

$$\delta^\pi \quad = \quad \left\{ \begin{array}{ll} 1 & \textit{si } \Pi(\theta \in \Theta_0|\mathbf{x_n}) > \Pi(\theta \notin \Theta_0|\mathbf{x_n}) \\ 0 & \textit{sinon} \end{array} \right.$$

Ainsi, l'estimation bayésienne permet d'accepter une hypothèse (nulle) $H_0 : \theta \in \Theta_0$, si c'est l'hypothèse la plus probable *a posteriori*, ce qui est une réponse intuitive.

Une variante du test 0-1 est le test de Neyman-Pearson qui permet de distinguer risques de première et de deuxième espèce :

$$L(\theta, d) = \left\{ \begin{array}{ll} 0 & \text{si } d = \mathbb{1}_{\Theta_0} \\ a_0 & \text{si } \theta \in \Theta_0 \text{ et } d = 0 \\ a_1 & \text{si } \theta \notin \Theta_0 \text{ et } d = 1 \end{array} \right. \tag{9}$$

qui donne l'estimateur bayésien

$$\delta^\pi(x) = \left\{ \begin{array}{ll} 1 & \text{si } \Pi(\theta \in \Theta_0|x) > a_1/(a_0 + a_1), \\ 0 & \text{sinon.} \end{array} \right.$$

Ainsi, l'hypothèse nulle est rejetée quand la probabilité *a posteriori* de H_0 est trop petite. Il est cependant délicat de choisir les poids a_0 et a_1 sur des considérations d'utilité.

Plus généralement, ce résultat permet d'illustrer une différence majeure entre statistique classique et statistique bayésienne. L'approche classique (dite de Fisher-Neyman-Pearson) suppose qu'on puisse définir une statistique de test dont la loi, sous l'hypothèse nulle H_0, est indépendante du paramètre estimé sous H_0. Ce faisant, la seule décision que l'on prendre avec une bonne certitude est de refuser H_0. Cette dissymétrie entre H_0 et toute autre hypothèse alternative H_1 n'existe pas dans le cadre bayésien : celui-ci émet un prior sur chaque modèle en compétition, puis compare les modèles selon leur probabilité d'explication des données disponibles *a posteriori*. Cette approche semble plus séduisante d'un point de vue intuitif et opérationnel. Voir également § **??** pour plus de détails.

3.4 Coûts intrinsèques

On peut enfin chercher à à trouver des fonctions de coûts qui restent invariantes par *transformation monotone inversible* sur les données (action d'un C^1-difféomorphisme sur Ω). On obtient ce faisant des fonctions de coûts définies à partir de *distances* ou de *divergences* D entre distributions

$$L(\theta, d) \quad = \quad D(f(.|\theta) \parallel f(.|d)).$$

Ci-dessous, quelques distances usuelles entre des densités $(f_\theta, f_{\theta'})$ de fonctions de répartition $(F_\theta, F_{\theta'})$, qui induisent des fonctions de coût intrinsèques, sont présentées.

1. *Distance de Kolmogoroff-Smirnoff :*
$$d_{KS}(f_\theta, f_{\theta'}) = \sup_x |F_\theta(x) - F_{\theta'}(x)|$$
2. *Distance L^1 :*
$$\begin{aligned} d_1(f_\theta, f_{\theta'}) &= \int |f_\theta(x) - f_{\theta'}(x)|\, dx & (2.1)\\ &= 2 \sup_A |P_\theta(A) - P_{\theta'}(A)| & (2.2)\end{aligned}$$
3. *Distance de Hellinger :*
$$d_H(f_\theta, f_{\theta'}) = \left(\int (\sqrt{f_\theta(x)} - \sqrt{f_{\theta'}(x)})^2 dx \right)^{\frac{1}{2}}$$
4. *Pseudo-distance*[8] *de Kullback-Liebler :*
$$K(f_\theta, f_{\theta'}) = \int f_\theta(x) \log \frac{f_\theta(x)}{f_{\theta'}(x)} dx$$
Avec l'inégalité de Jensen, on prouve l'inégalité $K(f_\theta, f_\delta) \geq 0$. De plus, $K(f_\theta, f_\delta) = 0$ si et seulement si $f_\theta = f_{\theta'}$ μ-presque sûrement.
5. *Distance L^2 :*
$$d_2(f_\theta, f_{\theta'}) = \int (f_\theta(x) - f_{\theta'}(x))^2 dx$$
Ceci peut s'utiliser si les densités sont de carré intégrable.

Exercice 7 *Lorsqu'on fait un choix de fonction de coût $L(\theta, \delta)$ dans un ensemble $U : \Theta \times \mathcal{D} \to \Lambda \in I\!R^+$, on commet une erreur par rapport à la meilleure fonction de coût possible pour le problème. On peut donc proposer un estimateur bayésien de cette fonction de coût en introduisant une fonction de coût sur les fonctions de coût $L(\theta, \delta)$:*

$$\begin{aligned} \tilde{L} : \Theta \times U \times \mathcal{D} &\to I\!R^+ \\ (\theta, \ell, \delta) &\to \tilde{L}(\theta, \ell, \delta). \end{aligned}$$

Quel est l'estimateur bayésien de $\tilde{L}(\theta, \ell, \delta)$ sous un coût quadratique, lorsque $L(\theta, \delta)$ est elle-même quadratique ?

3.5 Mode *a posteriori* (MAP)

L'estimateur du mode *a posteriori*, ou MAP, est défini par

$$\delta^{\pi}(\mathbf{x_n}) \quad = \quad \arg\max_{\theta\in\Theta} \pi(\theta|\mathbf{x_n}).$$

Cet estimateur, contrairement aux précédents, n'est pas issu de la minimisation d'une fonction de coût (il n'est donc pas bayésien *stricto sensu*) mais peut être vu comme la limite d'estimateurs bayésiens.

Il correspond à un maximum de vraisemblance (MV) pénalisé (voir § 2.5) et souffre donc en général des mêmes inconvénients que le MV, en particulier une certaine instabilité d'estimation ponctuelle. Par ailleurs, à la différence du MV, il est en général non invariant par reparamétrisation. Cette gêne décisionnelle mène à le déconseiller formellement, ou du moins à s'en méfier, même si ce type d'estimateur est couramment privilégié par les praticiens du *machine learning*.

3.6 Sélection de modèle et facteur de Bayes

La sélection de modèle bayésien est un choix particulier de décision. Supposons qu'on cherche à mener le test d'une *hypothèse nulle* $H_0 : \theta \in \Theta_0$. Soit $H_1 : \ \theta \in \Theta_1$ une hypothèse alternative.

Le *facteur de Bayes* est une transformation bijective de la probabilité *a posteriori*, qui a fini par être l'outil le plus utilisé pour choisir un modèle bayésien.

Définition 13 *Facteur de Bayes Le facteur de Bayes est le rapport des probabilités a posteriori des hypothèses nulle et alternative sur le rapport a priori de ces mêmes hypothèses*

$$B_{01}(x) \quad = \quad \left(\frac{\Pi(\theta\in\Theta_0|x)}{\Pi(\theta\in\Theta_1|x)}\right) \Big/ \left(\frac{\Pi(\theta\in\Theta_0)}{\Pi(\theta\in\Theta_1)}\right) \tag{10}$$

qui se réécrit comme le pendant bayésien du rapport de vraisemblance en remplaçant les vraisemblances par les marginales (les vraisemblances intégrées sur les a priori) sous les deux hypothèses

$$B_{01}(x) \quad = \quad \frac{\int_{\Theta_0} f(\mathbf{x_n}|\theta)\pi_0(\theta)\,d\theta}{\int_{\Theta_1} f(\mathbf{x_n}|\theta)\pi_1(\theta)\,d\theta} \quad = \quad \frac{f_0(\mathbf{x_n})}{f_1(\mathbf{x_n})}$$

Sous le coût généralisé (9), en posant

$$\gamma_0 = \Pi(\theta\in\Theta_0) \quad \text{et} \quad \gamma_1 = \Pi(\theta\in\Theta_1).$$

Ainsi l'hypothèse H_0 est acceptée si

$$B_{01}(x) > (a_1\gamma_1)/(a_0\gamma_0).$$

(i) si $\Lambda = \log_{10} B_{10}(\mathbf{x_n})$ varie entre 0 et 0.5, la certitude que H_0 est *fausse est faible* ;

(ii) si $\Lambda \in [0.5, 1]$, cette certitude est *substantielle* ;

(iii) si $\Lambda \in [1, 2]$, elle est *forte* ;

(iv) si $\Lambda > 2$, elle est *décisive*.

Malgré le côté heuristique de l'approche, ce genre d'échelle reste très utilisé.

Remarque 6 *Le calcul du facteur de Bayes n'est pas évident et demande le plus souvent de savoir simuler a posteriori.*

3.6.1 Cas de l'estimation ponctuelle et des tests de significativité en régression

Dans la définition (10), on sous-entend que chaque alternative $\Pi(\theta \in \Theta_i) > 0$ sinon le facteur de Bayes n'est pas défini. Cela exclurait les situations fréquentes où Θ_i est de mesure nulle. Par exemple lorsqu'on veut tester $\Theta_0 = \{\theta_0\}$, ou pour mener un test de significativité pour les modèles de régression.

Explications à venir en cours...

3.7 TP : Création d'un système d'alerte pour la circulation routière

On s'intéresse à un évènement routier $X = x$ relevé par un système de détection vivant dans l'espace χ de dimension finie. Ce système de détection peut prédire des évènements répétés du type "un animal sur la voie", "accrochage", "accident", "bouchon"... La question est de déterminer si, à chaque fois qu'un événement routier x est collecté, il est utile qu'une intervention de secours soit menée.

Nommons θ une variable indiquant la gravité de l'évènement. Cette variable a des valeurs dans les ensembles disjoints Θ_0 (incidents sans gravité) et Θ_1 (accidents nécessitant possiblement une intervention). On suppose disposer d'un échantillon labélisé $\mathbf{e_n} = (\mathbf{x_n}, \theta_\mathbf{n})$.

Questions.

1. Lorsqu'une observation x apparaît, comment prévoir θ ?
2. Comment peut-on en déduire une alarme efficace ?

4 Propriétés fondamentales du cadre bayésien

4.1 Prédiction (prévision)

Le contexte du problème de la prédiction est le suivant : les observations X sont identiquement distribuées selon P_θ, qui est absolument continue par rapport à une mesure dominante μ. Il existe donc une fonction de densité conditionnelle $f(|\theta)$. Par ailleurs on suppose que θ suit une loi a priori π. *Mener une prévision* consiste alors, à partir de n tirages observés $x_1, \ldots, x_n$, de déterminer le plus précisément possible ce que pourrait être le tirage suivant $X_n + 1$.

Dans l'approche fréquentiste, on calcule dans les faits $f(x_{n+1}|x_1, , x_n, \hat{\theta}_n)$, puisqu'on ne connaît pas θ et qu'on doit l'estimer : on utilise donc deux fois les données (une fois pour l'estimation de θ, et une nouvelle fois pour la prévision). En règle générale, ceci amène à sous-estimer les intervalles de confiance.

La stratégie du paradigme bayésien consiste à intégrer la prévision suivant la loi courante *a posteriori* sur θ et ce, afin d'avoir la meilleure prévision compte-tenu à la fois de notre savoir et de notre ignorance sur le paramètre. La loi prédictive s'écrit ainsi :

$$f(X_{n+1}|x_1, \ldots, x_n) \quad = \quad \int_\Theta f(X_{n+1}|x_1, \ldots, x_n, \theta)\pi(\theta|x_1, \ldots, x_n)\, d\theta$$

qui s'écrit plus simplement, lorsque *sachant* θ les tirages sont iid :

$$f(x_{n+1}|x_1, \ldots, x_n) \quad = \quad \int_\Theta f(x_{n+1}|\theta)\pi(\theta|x_1, \ldots, x_n)\, d\theta.$$

Ainsi, le prédicteur de X_{n+1} sous le coût quadratique est

$$\mathbb{E}[X_{n+1}|x_1, \ldots, x_n] \quad = \quad \int_\Omega x f(x|x_1, \ldots, x_n)\, dx.$$

4.2 Propriétés asymptotiques

Les approches classique et bayésienne de la modélisation et de la décision statistique aboutissent à des résultats similaires à l'asymptotisme, et les principaux théorèmes classiques connaissent leur pendant bayésien. Ainsi, le théorème central limite "classique" devient le théorème de Bernstein-von Mises dans le cadre bayésien (on l'appelle également *théorème central limite bayésien* par abus de langage). Afin de comparer les deux approches, on doit d'abord définir ce que signifie "vraie valeur θ_0 du paramètre θ".

Notons $\tilde{f}(x)$ la "vraie loi" inconnue des données, que l'on notera. Si on fait maintenant le choix d'une loi paramétrique $X \sim f(x|\theta_{\theta_0})$ (ou mécanisme génératif), alors la loi $f(x|\theta_0)$ doit être la plus proche possible de $\tilde{f}(x)$. Cette notion de proximité est généralement définie de la façon suivante.

Définition 14 *Soit $\tilde{f}(x)$ la loi inconnue des données. On définit θ_0 par*

$$\theta_0 \quad = \quad \arg\min_{\theta\in\Theta} KL\left(\tilde{f}(x)||f(x|\theta)\right)$$

où KL est la divergence de Kullback-Leibler. On notera par la suite plus simplement ce terme $KL(\theta)$.

Théorème 6 *Consistance Si $f(.|\theta)$ est suffisamment régulière et identifiable, soit si $\theta_1 \neq \theta_2 \Rightarrow f(x|\theta_1) \neq f(x|\theta_2)\ \forall x \in \Omega$, alors pour tout échantillon $\mathbf{x_n}$ iid*

$$\pi(\theta|\mathbf{x_n}) \quad \xrightarrow{p.s.} \quad \delta_{\theta_0}.$$

Par ailleurs, si $g : \Theta \to I\!R$ est mesurable et telle que $\mathbb{E}[g(\theta)] < \infty$, alors sous les mêmes hypothèses

$$\lim_{n\to\infty} \mathbb{E}\left[g(\theta)|X_1, \ldots, X_n\right] \quad = \quad g(\theta) \quad p.s.$$

Un résultat utile, intermédiaire entre la consistance et la convergence en loi (Théorème 9), est la convergence en probabilité.

Théorème 7 *Si Θ est fini et discret et $\Pi(\theta = \theta_0) > 0$, alors pour tout échantillon iid $X_1, \ldots, X_n|\theta \sim f(X|\theta)$,*

$$\Pi\left(\theta = \theta_0 | X_1, \ldots, X_n\right) \quad \xrightarrow[\mathbb{P}]{n\to\infty} \quad 1.$$

Si Θ est continu, alors $\pi(\theta_0|x)$ vaut toujours 0 pour tout échantillon fini x, et on ne peut appliquer les outils menant au résultat précédent. Pour adapter cette preuve, il faut définir un voisinage V_{θ_0} qui est un ensemble ouvert de points de Θ à une distance maximum fixée de θ_0 (Θ étant un espace métrique).

Théorème 8 *Si Θ est un ensemble compact et si V_{θ_0} est tel que $\Pi(\theta \in V_{\theta_0}) > 0$ avec*

$$\theta_0 \quad = \quad \arg\min_{\theta\in\Theta} KL(\theta)$$

alors

$$\Pi(\theta \in V_{\theta_0}|X) \quad \xrightarrow[\mathbb{P}]{n\to\infty} \quad 1.$$

Le théorème de Bernstein-von Mises suppose l'existence de l'information de Fisher I_θ. Il n'existe pas d'ensemble de conditions de régularité minimal nécessaire pour l'existence de I_θ ; cependant, la plupart des auteurs s'accordent sur les conditions suffisantes suivantes d'existence, de positivité et de continuité dans un sous-espace de Θ :

- $f(x|\theta)$ est absolument continue en θ ;
- sa dérivée doit exister pour tout $x \in \Omega$.

Alors

$$I_\theta \quad = \quad \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right] \; = \; -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$$

si $\log f(x|\theta)$ est deux fois différentiable en θ.

Théorème 9 *Normalité asymptotique (**Bernstein-von Mises**) Soit I_θ la matrice d'information de Fisher du modèle $f(.|\theta)$ et soit $g(\theta)$ la densité de la gaussienne $\mathcal{N}(0, I_{\theta_0}^{-1})$. Soit $\hat{\theta}_n$ le maximum de vraisemblance. Alors, dans les conditions précédentes,*

$$\int_\Theta \left|\pi\left(\sqrt{n}\left\{\theta - \hat{\theta}_n\right\}|\mathbf{x_n}\right) - g(\theta)\right| \, d\theta \quad \to \quad 0.$$

4.3 Régions de crédibilité

Soit $x \sim f(.|\theta)$ une (ou plusieurs) observations.

Définition 15 *Région $\alpha-$crédible Une région A de Θ est dite $\alpha-$crédible si $\Pi(\theta \in A|x) \geq 1-\alpha$.*

Notons que le paradigme bayésien permet une nouvelle fois de s'affranchir d'un inconvénient de l'approche fréquentiste. Rappelons qu'au sens fréquentiste, A est une région de confiance $1-\alpha$ si, en refaisant l'expérience (l'observation d'un $X \sim f(.|\theta)$) un nombre de fois tendant vers ∞,

$$P_\theta(\theta \in A) \quad \geq \quad 1-\alpha.$$

En refaisant l'expérience un grand nombre de fois, la probabilité que θ soit dans A est plus grande que 1α. Une région de confiance n'a donc de sens que pour un très grand nombre d'expériences tandis que la définition bayésienne exprime que la probabilité que θ soit dans A au vue des celles déjà réalisées est plus grande que 1α. Il n'y a donc pas besoin ici d'avoir recours à un nombre infini d'expériences pour définir une région α-crédible, seule compte l'expérience effectivement réalisée.

Remarque 7 *On distingue bien ici la probabilité "fréquentiste" P_θ de la probabilité bayésienne Π. Dans le premier cas, l'aléatoire concerne la région A, qui est un estimateur statistique dépendant d'un estimateur classique $\hat{\theta}(X_1, \ldots, X_n)$ et θ est considéré comme fixe. Dans le second cas, c'est bien θ qui est une aléatoire.*

Il y a une infinité de régions αcrédibles, il est donc logique de s'intéresser à la région qui a le volume minimal. Le volume étant défini par $\text{vol}(A) = \int_A d\mu(\theta)$, si $\pi(\theta|x)$ est absolument continue par rapport à une mesure de référence μ.

Définition 16 Région HPD. *$A_{\alpha,\pi}$ est une région HPD (highest posterior density) si et seulement si*

$$A_{\alpha,\pi} \quad = \quad \{\theta \in \Theta,\ \pi(\theta|x) \geq h_\alpha\}$$

où h_α est défini par

$$h_\alpha \quad = \quad \sup\left\{\Pi\left(\theta|\pi(\theta|x) \geq h, X\right) \geq 1-\alpha\right\}.$$

$A_{\alpha,\pi}$ est parmi les régions qui ont une probabilité supérieure à $1-\alpha$ de contenir θ (et qui sont donc α-crédibles) et sur lesquelles la densité *a posteriori* ne descend pas sous un certain niveau (restant au dessus de la valeur la plus élevée possible).

Théorème 10 *$A_{\alpha,\pi}$ est parmi les régions α-crédibles celle de volume minimal si et seulement si elle est HPD.*

Exercice 8 *Soit $x_1, \ldots, x_n$ des réalisations iid de loi $\mathcal{N}(\mu, \sigma^2)$. On choisit la mesure a priori (non probabiliste) jointe*

$$\pi(\mu, \sigma^2) \quad \propto \quad 1/\sigma.$$

1. *Déterminez la loi a posteriori jointe $\pi(\mu, \sigma^2|x_1, \ldots, x_n)$*
2. *Déterminez la loi a posteriori marginale $\pi(\mu|x_1, \ldots, x_n)$*
3. *Calculez la région HPD de seuil α pour μ et comparez-la à la région de confiance fréquentiste, de même seuil, qu'on pourrait calculer par l'emploi du maximum de vraisemblance.*

Les régions HPD sont à manier avec précaution, car elle ne sont pas indépendantes de la paramétrisation.

Exercice 9 *Soit $A_{\alpha,\pi} = \{\theta \in \Theta\ ,\ \pi(\theta|x) \geq h_\alpha\}$ une région HPD et soit*

$$\eta = g(\theta)$$

un C^1-diffémorphisme (bijection). On définit alors la région HPD correspondante pour $\pi(\eta|x)$:

$$\tilde{A}_{\alpha,\pi} \quad = \quad \left\{\theta \in \Theta\ ,\ \pi(\eta|x) \geq \tilde{h}_\alpha\right\}$$

- *Sous quelle condition peut-on écrire que $\tilde{A}_{\alpha,\pi} = g\left(A_{\alpha,\pi}\right)$?*
- *Illustrons cela en supposant $X \sim \mathcal{N}(\theta, 1)$ et $\pi(\theta) \propto 1$, puis en posant $\eta = \exp(\theta)$.*

Nous pouvons comprendre pourquoi une région de confiance n'est pas invariante par reparamétrisation. En effet, cette région se définit comme une solution du problème de minimisation suivant :

$$A_{\alpha,\pi} = \arg \min_{A,\Pi(A|X)\geq 1-\alpha} \text{Vol}(A)$$

où $\text{Vol}(A) = \int_A d\mu(\theta)$. Or la mesure de Lebesgue n'est pas invariante par reparamétrisation. Une idée pour lever cette difficulté est donc logiquement d'abandonner la mesure de Lebesgue et de considérer pour une mesure s :

$$A_{\alpha,\pi,s} = \arg \min_{A,\Pi(A|X)\geq 1-\alpha} \int_A ds(\theta).$$

4.3.1 Calcul de régions HPD

Pour calculer les régions HPD, il y a plusieurs méthodes :

1. *Méthode analytique et numérique* : c'est ce qui a été fait lors de l'exemple précédent. Précisons une nouvelle fois que cette méthode ne peut s'appliquer que dans des cas assez rares.

2. *Méthode par approximation* : cette méthode peut être appliquée si le modèle est régulier. L'usage du théorème de Bernstein-von Mises permet d'approximer la loi *a posteriori* par une gaussienne. On retombe peu ou prou sur des régions HPD proches de celles du maximum de vraisemblance.

3. *Méthode par simulation*. En effet, une région $\alpha-$crédible peut génériquement être estimée par les quantiles empiriques de la simulation *a posteriori* (voir plus loin).

Théorème 11 *Supposons avoir un échantillon iid $\theta_1, \ldots, \theta_m \sim \pi(\theta|x_1, \ldots, x_n)$ avec $\theta \in \mathbb{R}$. Alors les intervalles de quantiles empiriques de la forme $[\theta^{(\alpha/2)}, \theta^{(1-\alpha/2)}]$ sont tels que*

$$\Pi\left(\theta \in \left[\theta^{(\alpha/2)}, \theta^{(1-\alpha/2)}\right] |x_1, \ldots, x_n\right) \quad \xrightarrow{m\to\infty} 1-\alpha.$$

Il n'est cependant pas garanti qu'une telle région soit HPD. Pour m grand, $\theta^{(\alpha/2)}$ s'approche du quantile d'ordre $\alpha/2$ de la loi *a posteriori*. Cette région n'est pas nécessairement HPD mais reste $\alpha-$crédible. Cette méthode est particulièrement adaptée lorsque la loi *a priori* est unimodale. Il est toujours utile de représenter graphiquement les sorties pour fixer les idées. Enfin, il est aussi envisageable d'avoir recours à une estimation non paramétrique par noyaux.

5 Compréhension et représentation de l'information incertaine

5.1 Une vision subjectiviste de la théorie bayésienne

La fonction de coût et le processus décisionnel permettent de proposer une interprétation importante de la distribution *a priori*. Elle peut être comprise comme pari (personnel) fait sur l'éventualité d'un évènement, et notamment un gain conditionné par l'occurence du phénomène modélisé par $f(x|\theta)$. Cette interprétation subjective, proposée par de Finetti (1948), est certainement le point le plus critiqué de la démarche bayésienne, mais c'est aussi celui qui permet à l'application de cette théorie d'être ancrée dans le réel.

On peut en fait mieux appréhender cette interprétation subjectiviste en la reliant à l'histoire récente de l'axiomatique de la connaissance incertaine.

5.2 Théories de la connaissance incertaine

Dans l'histoire des théories de représentation mathématique de la connaissance incertaine, il existe essentiellement deux grandes écoles de pensée :

1. des **théories de la représentation qui s'adaptent** aux moyens variés, pour un humain, d'exprimer son opinion personnelle sur le comportement d'une variable d'ancrage X ou d'un paramètre perceptible θ (plus rare) ;
    - *Exemples* : théories extra-probabilistes : Dempster-Schafer, possibilités, logique floue ...
2. des **théories qui visent à établir des axiomes de rationalité** à propos des décisions sous-tendant l'expression d'une opinion : un expert est perçu comme un preneur de décision selon ces axiomes.

Une vision axiomatique de la représentation mathématisée de la connaissance incertaine, qui permet d'interpréter la théorie des probabilités comme une "bonne" façon de représenter cette connaissance (ou plutôt cette information), a été construite par Cox et Jaynes. Elle s'incarne dans le théorème de Cox-Jaynes, dont les versions successives, au cours du temps, sont devenus les théorèmes fondamentaux de l'intelligence artificielle. Ce théorème permet de donner un cadre plus robuste à la vision subjectiviste du sens d'un modèle *a priori*.

5.3 Une vision plus claire de la statistique bayésienne

En définitive, il apparaît après ces premiers chapitres que la statistique bayésienne est à la fois :

- une théorie de la description d'un phénomène incertain, où "incertitude" signifie "mélange d'aléatoire (incertitude non-réductible) et d'épistémique (incertitude réductible) ;
- une théorie de la décision, sous certains axiomes de rationalité.

Sachant un modèle $f(x|\theta)$, le travail bayésien consiste donc à :

1. déterminer le coût associé aux décisions, $L(\theta, \delta)$;
2. éliciter ("construire") une loi *a priori* $\pi(\theta)$;
3. réaliser l'inférence *a posteriori* et produire un ou plusieurs estimateurs, voire faire un choix de modèle.

Notons qu'il y a redondance entre les deux premières étapes : présupposer l'existence d'une fonction de coût implique qu'une certaine information *a priori* sur le problème considéré est disponible.

6 Modélisation *a priori*

7 Méthodes de calcul bayésien

ANNEXES

A Rappels : concepts et outils fondamentaux de l'aléatoire

Remarque 8 *Pour faciliter la lecture et l'appropriation, cette annexe de rappels est illustrée par de nombreux exemples de phénomènes naturels dits* extrêmes*, telles des pluies diluviennes, des vents forts, etc. dont on cherche à modéliser le comportement.*

La modélisation probabiliste d'un aléa X repose sur le caractère de *variable aléatoire* conféré à X, évoluant dans un ensemble d'échantillonnage Ω de dimension d. Puisque $\Omega \neq \varnothing$, les sous-ensembles de valeurs $\mathcal{A} \subset \Omega$ que peut parcourir X sont non vides, et ils présentent une certaine stabilité : l'union dénombrables de plusieurs $\mathcal{A}_i$ est encore dans Ω, de même que le complémentaire de tout sous-ensemble $\mathcal{A}$.

Ces propriétés fondamentales permettent de "paver" (*mesurer*) l'ensemble Ω de façon à associer à toute observation (survenue) d'un *événement* $A \in \mathcal{A}$ une valeur numérique $\mathbb{P}(A)$. L'ensemble de ces valeurs numériques vit dans l'intervalle $[0, 1]$, et est tel que

$$\mathbb{P}(\Omega) = 1.$$

On parle alors, pour désigner $\mathbb{P}$, de *mesure de probabilité*.

La théorie des probabilités nomme le triplet $(\Omega, \mathcal{A}, \mathbb{P})$ *espace probabilisé*, l'ensemble Ω *univers* et $\mathcal{A}$ *tribu* (ou $\sigma-$algèbre). En général, le choix de $\mathcal{A}$ est l'ensemble des parties de Ω dont la mesure de Lebesgue peut être définie (cf. § A.1). Il n'est donc usuellement pas donné de précision, dans les problèmes appliqués, sur $\mathcal{A}$.

A.1 Problèmes unidimensionnels

Considérons tout d'abord le cas où $d = 1$. Si Ω est *discret* (par exemple si $\Omega = \{1, 2, 3, \dots, \}$) ou *catégoriel*, et plus généralement si Ω est *dénombrable*, la distribution de probabilité est dite discrète et est déterminée par la *fonction de masse* probabiliste

$$f(x) = \mathbb{P}(X = x)$$

pour toute valeur $x \in \Omega$. Cependant, la très grande majorité des variables aléatoires considérées dans cet ouvrage présente un caractère *continu*. En particulier, les valeurs prises par X (vitesse du vent, température, débit d'une rivière...) évoluent continûment – ce qui est indispensable pour appliquer la théorie des valeurs extrêmes – et Ω constitue généralement un sous-ensemble continu de $\mathbb{R}^d$, même si le dispositif de mesure est nécessairement limité, en pratique, par une précision donnée. Cette précision ne joue pas de rôle dans la construction du modèle probabiliste mais dans celui du modèle *statistique*, qui englobe le modèle probabiliste en établissant un lien direct avec des observations bruitées (voir § A.5). Dans la pratique, les deux modèles sont confondus quand le bruit d'observation est considéré comme négligeable.

Dans le cas continu, c'est-à-dire lorsque Ω n'est plus dénombrable, la distribution de probabilité peut être spécifiée par la *fonction de répartition*

$$F_X(x) = \mathbb{P}(X \leq x)$$

pour toute valeur $x \in \Omega$. Afin de satisfaire les axiomes des probabilités [14], cette fonction doit être croissante, et telle que, lorsque la dimension $d = 1$,

$$\begin{aligned} \lim_{x \to x_{\inf}} F_X(x) &= 0, \\ \lim_{x \to x_{\sup}} F_X(x) &= 1 \end{aligned}$$

où $(x_{\inf}, x_{\sup})$ sont les bornes inférieure et supérieure (éventuellement infinies) de Ω. Le cas multidimensionnel où $d > 1$ est précisé au § A.3. Toujours pour $d = 1$, l'équivalent de la probabilité discrète $f(x)$ dans le cas continu est fourni par la probabilité que X se situe entre les valeurs $x - a$ et $x + b$ (avec $a, b \geq 0$) :

$$\mathbb{P}(x - a \leq X \leq x + b) = F_X(x + b) - F_X(x - a).$$

Cette propriété pousse à définir, dans les cas où F_X est dérivable, la dérivée de F_X (dite *de Radon-Nikodym-Lebesgue*) définie comme le cas-limite $a = b = \epsilon \to 0$

$$f_X(x) = \frac{dF_X}{dx}(x),$$

appelée *densité de probabilité* de X, qui est donc telle que

$$F(x) = \int_{-\infty}^{x} f_X(u)\, du$$

et

$$\mathbb{P}(x - a \leq X \leq x + b) = \int_{x-a}^{x+b} f_X(u)\, du.$$

Nécessairement, $\int_\Omega f_X(u)\, du = 1$. Ainsi, toute distribution de probabilité continue, en dimension $d = 1$ (c'est aussi le cas en dimension $d > 1$) peut être représentée de façon équivalente (sous réserve de dérivabilité [6]) par sa fonction de répartition ou sa densité (figure 2).

Informellement, f_X peut être vue comme la limite de l'histogramme en fréquence des valeurs possibles de X, pour des classes de valeurs étroites (figure 2). Plus formellement, fonction de répartition et densité de probabilité doivent être interprétées comme des outils permettant d'opérer une *mesure* de la distribution des X relativement à une mesure de l'espace Ω.

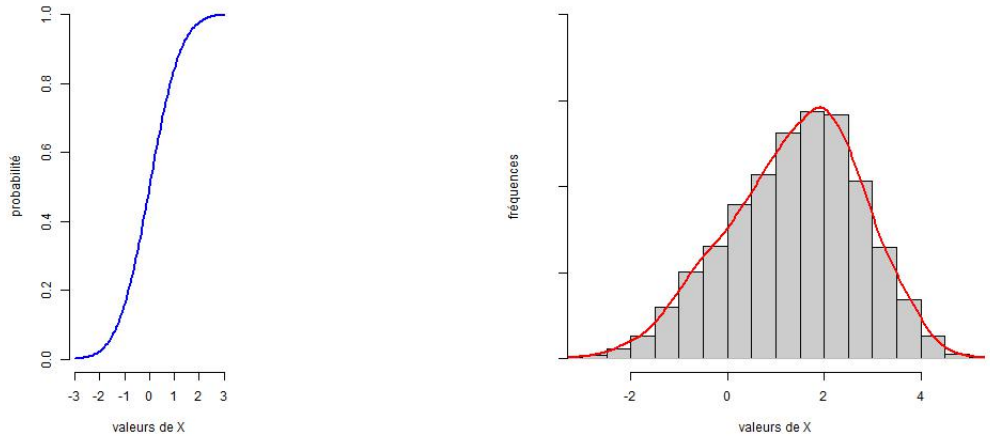

FIGURE 1 – Gauche : exemple de fonction de répartition. Droite : histogramme en fréquence de valeurs de X et densité de probabilité correspondante (courbe).

Considérons par exemple que $\Omega = I_1 \times I_2 \times \ldots \times I_d$, où chaque I_k est un intervalle de $\mathbb{R}$ (fermé, ouvert ou semi-ouvert), l'ensemble constituant un parallélépipède contenant toutes les valeurs de X pouvant être observées. Ce solide (ou cet espace) peut être décrit par un ensemble de mesures, par exemple son volume. La mesure de Lebesgue [16], notée μ_L, a été construite comme une mesure de référence permettant de décrire ce type d'espace de façon universelle et uniforme. Comme le volume, elle prend une valeur finie si Ω est compact. La densité f_X définit une autre mesure sur Ω, qui spécifie la forme de la distribution des X et permet de la différencier de l'uniformité. Il faut donc l'interpréter comme une mesure *relative* à celle de Lebesgue (ou *dominée* par la mesure de Lebesgue). Au lecteur intéressé par une introduction détaillée à la théorie de la mesure, nous suggérons les ouvrages [4] (pour une approche "ingénieure") et [15] (pour une vision plus mathématique).

6. Plus généralement de *différentiabilité* en dimension quelconque.

L'information incertaine transportée par les distributions de probabilité est très souvent résumée par des indicateurs statistiques particuliers : les *moments* d'ordre $k \in I\!N$, définis comme l'ensemble des valeurs moyennes de la variable X^k :

$$M_k = \mathbb{E}[X^k] = \int_\Omega x^k f_X(x)\, dx.$$

Si ceux-ci existent pour $k = 1$ et $k = 2$, ils permettent de définir l'*espérance* $\mathbb{E}[X]$ et la *variance*

$$\mathbb{V}[X] \quad = \quad \int_\Omega (x - \mathbb{E}[X])^2 f_X(x)\, dx.$$

L'espérance fournit une mesure de localisation moyenne de X dans la distribution f_X, tandis que $\mathbb{V}[X]$ est une mesure de la variabilité (ou dispersion) de f_X. L'*écart-type* de f_X, homogène à X, est défini par

$$\sigma_X = \sqrt{\mathbb{V}[X]}.$$

Alternativement, le *cœfficient de variation* de X

$$\mathrm{CV}[X] = \frac{\sigma_X}{\mathbb{E}[X]},$$

fournit une autre mesure *relative* de la variabilité ou dispersion de f_X (plus usuelle pour les ingénieurs). Enfin, on parlera de variable centrée-réduite si X est transformée en

$$X' = \frac{X - \mathbb{E}[X]}{\sigma_X},$$

d'espérance nulle et de variance unitaire.

A.2 Familles de modèles paramétriques

Rappelons quelques modèles probabilistes ou statistiques fondamentaux, qui interviennent très souvent dans les constructions plus élaborées qui seront décrites dans cet ouvrage. Ces modèles seront, dans le cadre de cet ouvrage, considérés *paramétriques*, c'est-à-dire descriptibles de façon exhaustive par un ensemble fini de paramètres.

La première raison de ce choix est liée au cadre d'étude : le comportement des extrêmes d'un échantillon aléatoire suit, sous certaines conditions théoriques, des lois paramétriques. C'est aussi le cas du comportement des estimateurs statistiques (§ A.6) obéissant à une loi des grands nombres.

Cet argument fondamental se renforce de la constatation suivante : lorsqu'on s'intéresse à ces comportements extrêmes, le nombre d'observations disponibles devient faible. Expliquer la production de ces observations par un mécanisme aléatoire déterminé par un nombre infini ou même simplement grand de paramètres (c'est-à-dire plus grand que le nombre de données) semble déraisonnable car la majeure partie de ces paramètres resteront inconnus, ou possèderont plusieurs valeurs possibles, et le modèle ainsi créé ne serait pas identifiable et utilisable.

Dans ce document, on notera très généralement θ ce vecteur de paramètres, qui évoluera donc dans un espace θ de dimension finie. Le conditionnement à θ du mécanisme de production aléatoire sera rappelé dans les notations des densités et fonctions de répartition : $f_X(x) = f(x|\theta)$ et $F_X(x) = F(x|\theta)$.

Lois

Dans un cadre discret, on peut s'intéresser à la survenue d'un événement ponctuel $Z > z_0$, où Z est, par exemple, un niveau d'eau maximal mensuel, et z_0 une hauteur de digue de protection. Supposons disposer d'un échantillon d'indicateurs $(\delta_1, \ldots, \delta_n) \in \{0, 1\}^n$ valant chacun 1 si la crue ainsi définie survient, et 0 sinon. Faisons l'hypothèse que les δ_i sont indépendants et correspondent chacun au résultat d'un "essai de submersion"

réussissant avec une même probabilité p. Si l'on note $X_n = \sum_{i=1}^{n} \delta_i$ le nombre total de "succès" parmi ces n essais, alors la fonction de masse probabiliste de X_n s'écrit

$$f(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$$

pour $x \in \Omega = \{0, 1, 2, \ldots, n\}$, et où

$$\left(\begin{array}{c} n \\ x \end{array} \right) = \frac{n!}{x!(n-x)!}.$$

La variable aléatoire X_n est alors dite suivre la *loi binomiale* $\mathcal{B}(n, p)$ (figure 2). Dans le cadre d'une étude de risque, on s'attachera à estimer la probabilité de surverse p à partir de la statistique observée x_n.

La variable X_n dite de *comptage* définie ci-dessus peut être généralisée dans une perspective d'estimer l'occurence d'événements survenant de façon aléatoire durant un laps de temps fixé (par exemple une année). Si on suppose que ces événements surviennent avec une fréquence moyenne unique $\lambda > 0$ dans cet intervalle de temps, alors la probabilité qu'il survienne exactement $X_n = x \in \Omega = \{0, 1, \ldots, \infty\}$ occurences est

$$f(x) = \frac{\lambda^x}{x!} \exp(-\lambda),$$

qui définit la fonction de masse probabiliste de la *loi de Poisson* d'espérance λ (figure 2). Celle-ci joue notamment un grand rôle dans l'établissement des lois statistiques associées aux observations historiques car elle permet de modéliser la survenue du nombre d'événements situés entre deux dates (par exemple séparés par plusieurs dizaines d'années) et non observés directement. Le lien technique entre la loi binomiale et la loi de Poisson s'exprime dans le lemme suivant :

LEMME 1. *Si X_n suit une loi binomiale $\mathcal{B}(n, p)$ avec $p \ll 1$, alors la loi de X_n peut être approximée par la loi de Poisson d'espérance np lorsque $n \to \infty$.*

Rappelons enfin, dans le cas continu, l'importance fondamentale de la *loi normale* $X \sim \mathcal{N}(\mu, \sigma^2)$, d'espérance μ et de variance σ^2, et de densité de probabilité (pour $d = 1$ et $\Omega = I\!R$)

$$f_X(x) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right).$$

Celle-ci modélise un grand nombre de phénomènes, en particulier celui de la répartition de la moyenne d'un échantillon aléatoire (loi des grands nombres). La convergence en loi normale d'un estimateur statistique (cf. § A.6) constitue un type de résultat très classique (théorème de la limite centrale). La variable $(X - \mu)/\sigma$ suit la loi normale dite *centrée réduite* $\mathcal{N}(0, 1)$ (figure 2). On note usuellement par $\phi(.)$ et $\Phi(.)$ les densité et fonction de répartition de cette loi centrée réduite.

Tests statistiques

La démarche générale des tests consiste à rejeter ou ne pas rejeter (sans forcément accepter) une hypothèse statistique H_0, dite *nulle*, en fonction d'un jeu de données $\mathbf{x_n}$. Par exemple, dans un cadre paramétrique cette hypothèse peut correspondre au choix spécifique d'une valeur $\theta = \theta_0$ dans une même famille $f(x|\theta)$ ou d'un domaine $\theta \in \theta_0$. Définir un test revient à définir une statistique

$$R_n \quad = \quad R(X_1, \ldots, X_n)$$

qui est une variable aléatoire dont la loi $\mathcal{F}_{R_n}$ est connue (au moins asymptotiquement, c'est-à-dire quand $n \to \infty$) lorsque l'hypothèse H_0 est vraie, et cette loi est indépendante de la *valeur de l'hypothèse*. (ex : indépendante de θ). Plus précisément, dans un cadre paramétrique où θ est testé, la loi $\mathcal{F}_{R_n}$ ne doit pas dépendre de θ, et la variable R_n est dite *pivotale* . Lorsque R_n est défini indépendamment de θ, cette statistique est dite également *ancillaire*.

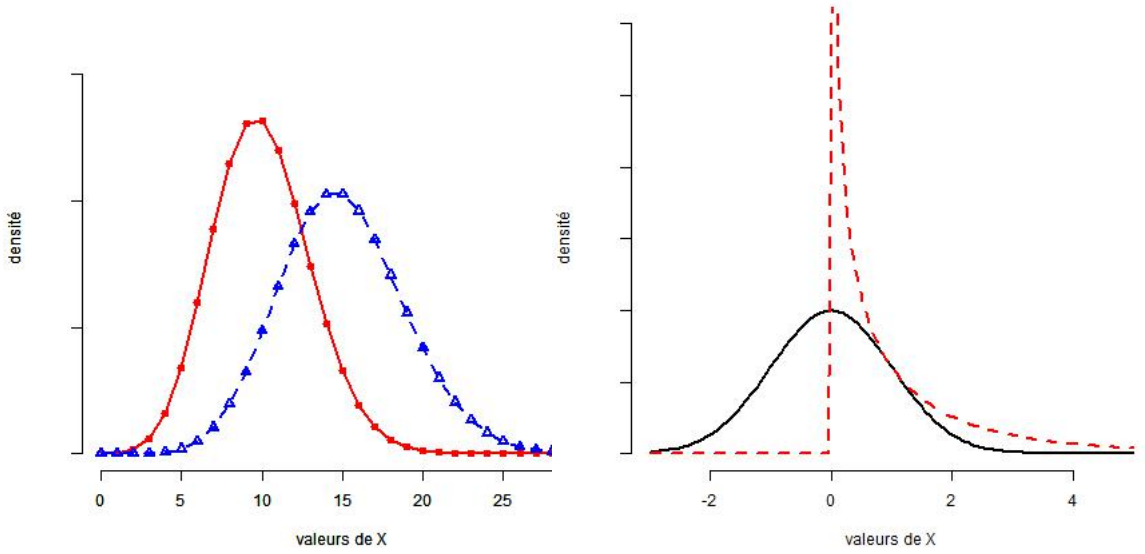

FIGURE 2 – Gauche : fonction de masse des lois discrètes binomiale $\mathcal{B}_n(100, 0.1)$ (carrés) et Poisson $\mathcal{P}(15)$ (triangles). Droite : densités de probabilité continues de la loi normale centrée réduite $\mathcal{N}(0, 1)$ (courbe pleine) et χ^2_1.

Le positionnement de la statistique *observée* $r_n = r(x_1, \ldots, x_n)$ dans la loi $\mathcal{F}_{R_n}$ a été définie par Fisher (1926 ; [7]) comme la probabilité p_{r_n} (dite $p-$valeur ou $p-$*value*) d'observer un évènement plus "extrême" (plus petit ou plus grand) que r_n. Plus cette probabilité est faible, plus l'événement r_n est "loin" des valeurs de R_n de plus haute densité, et moins H_0 est probable (rappelons que la $p-$valeur n'est pas la probabilité que H_0 soit vraie). En d'autres termes, si H_0 est fausse, r_n devrait être une valeur extrême de $\mathcal{F}_{R_n}$.

L'approche courante des tests, dite de *Neyman-Pearson* (1928 ; [17]), impose de fixer un *seuil de significativité* $\alpha \ll 1$ définissant l'extrêmalité et de comparer le quantile $q_{1-\alpha}$ de la loi $\mathcal{F}_{R_n}$ avec p_{r_n} ; si $p_{r_n} < q_{1-\alpha}$, l'événement r_n est encore moins probable que α, et l'hypothèse H_0 doit être rejetée. Dans le cas contraire, cette hypothèse est plausible (mais pas forcément validé). La pratique courante dans l'ensemble des sciences expérimentales, là encore, est de fixer $\alpha = 5\%$ ou $\alpha = 1\%$, mais ces seuils arbitraires sont de plus en plus critiqués [21, 6], et il est actuellement recommandé [11, 3] de mener plusieurs tests et de tester des seuils α très faibles (ex : $\alpha \in [1‰, 5‰]$)

Dans de nombreux cas, la statistique R_n est choisie positive, afin de pouvoir définir simplement la $p-$valeur $p_{r_n} = \mathbb{P}(R_n > r_n)$.

EXEMPLE 10. **Test de Kolmogorov-Smirnov [33].** *Disposant de l'estimateur empirique classique (cf. § A.6)* $x \mapsto \hat{F}_n(x)$ *de la fonction de répartition* F *d'un échantillon iid unidimensionnel* $x_1, \ldots, x_n$*, défini par*

$$\hat{F}_n(x) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{} x_i \leq x \},$$

et d'un candidat F_0 *pour* F*, on souhaite tester l'hypothèse* $H_0 : F = F_0$*. La statistique de test est définie par*

$$R_n \quad = \quad \sqrt{n} \sup_{x \in I\!R} \left\| \hat{F}_n(x) - F_0(x) \right\|.$$

Sous H_0 et pour n grand, R_n suit approximativement la loi de Kolmogorov, définie par sa fonction de répartition

$$F_{KS}(x) \quad = \quad 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp\left(-2k^2x^2\right) \quad \textit{pour } x \in I\!R^+,$$

qui est généralement tabulée au sein des outils logiciels classiques.

Pour une classe importante de tests, dits du χ^2 (Chi-2), la statistique R_n est construire de façon à suivre *loi du χ^2 avec $q \geq 1$ degrés de liberté*

$$R_n \sim \chi^2_q$$

dont la densité est tracée sur la figure 2 pour $q = 1$. Les lois du χ^2 sont intrinsèquement liées aux lois normales par une relation quadratique. Par exemple, la somme des carrés de n variables $\mathcal{N}(0,1)$ indépendantes suit une loi du χ^2_n à n degrés de liberté. Les quantiles de cette loi sont fournis en pratique par des tables ou algorithmes spécifiques.

Puissance d'un test. Rappelons que deux procédures testant une même hypothèse H_0 ne sont pas forcément aussi pertinentes l'une que l'autre ; elles peuvent être comparées par leur *puissance*, c'est-à-dire leur probabilité respective de rejeter l'hypothèse nulle H_0 sachant qu'elle est incorrecte. Lorsqu'on utilise un test, il convient toujours de s'assurer que sa puissance est élevée, voire la meilleure possible [32]. Elle est définie par

$$1 - \beta$$

où β est nommée *erreur* ou risque *de deuxième espèce* - c'est-à-dire le risque d'accepter à tort l'hypothèse H_0. L'erreur de deuxième espèce est équivalente à un *taux de faux positifs* dans une procédure de détection. Un exemple classique de test le plus puissant entre deux hypothèses simples $H_0 : \mathbb{P} = P_0$ et $H_1 : \mathbb{P} = P_1$ est le *test de rapport de vraisemblance* (Théorème de Neyman-Pearson), dit aussi test LRT (*likelihood ratio test*).

EXEMPLE 11. **Test d'adéquation du χ^2 (cas discret) [36].** *Soit $\mathbf{x_n} = (x_1, \ldots, x_n)$ un échantillon de réalisations de X supposées iid dans un ensemble fini de valeurs $\{1, \ldots, M\}$. On souhaite tester l'hypothèse nulle H_0 selon laquelle les probabilités que X prenne les valeurs 1 à M sont respectivement $p_1, \ldots, p_M$ avec $\sum_{k=1}^{M} p_k = 1$. On note alors*

$$\hat{p}_k \quad = \quad \frac{1}{n}\sum_{j=1}^{n}\delta_{\{x_j=k\}}$$

où $\delta_{\{x_j=k\}} = 1$ si $x_j = k$ et 0 sinon. On définit alors

$$R_n \quad = \quad \sqrt{n\sum_{k=1}^{M}\frac{\left(\hat{p}_k - p_k\right)^2}{p_k}} \tag{-15}$$

qui suit, sous l'hypothèse H_0, une loi χ^2_{M-1}.

Théorème 12 Test LRT (rapport de vraisemblance). *Soit $\mathbf{X_n} = (X_1, \ldots, X_n)$ un échantillon de variables aléatoires indépendantes et de même loi $\mathbb{P}$ de densité f. On souhaite tester $H_0 : \mathbb{P} = P_0$ contre $H_1 : \mathbb{P} = P_1$. On nomme $L_i(\mathbf{X_n}) = \prod_{k=1}^{n} f_i(X_k)$ la vraisemblance statistique maximisée sous l'hypothèse $i \in \{0, 1\}$ (voir § A.6 pour une définition détaillée de la vraisemblance et sa maximisation). Soit*

$$R_n \quad = \quad 2\log\frac{L_1(\mathbf{X_n})}{L_0(\mathbf{X_n})}.$$

Alors, si P_0 désigne un modèle paramétré par θ tel que $\theta \in \theta_0$ et P_1 est spécifié par $\theta \notin \theta_0$, alors R_n suit asymptotiquement un mélange de mesures de Dirac et de lois du χ^2 dont le degré de liberté est égal ou inférieur au nombre de contraintes q imposées par l'hypothèse nulle.

De nombreuses précisions sur les mécanismes, les spécifications et les mises en garde sur l'interprétation des tests statistiques (tests paramétriques, non paramétriques, tests de conformité, d'adéquation, d'homogénéité, d'indépendance, d'association...) sont fournis dans [30] et [10]. Le cas spécifique des tests LRT est particulièrement détaillé dans [9]. Appliqués au cas spécifique des modèles d'extrêmes, le lecteur intéressé par une revue générale pourra consulter avec profit l'article [20].

EXEMPLE 12.**Test LRT.** *Dans le cas spécifique où θ_0 est dans l'intérieur strict de θ, alors*

$$R_n \overset{n\to\infty}{\sim} \chi^2_q. \tag{-15}$$

Considérons ainsi une loi normale $\mathcal{N}(\mu,\sigma)$ avec $\theta=(\mu,\sigma)\in I\!R\times I\!R_*^+$. On souhaite tester $H_0:\mu=0$ contre $H_1:\mu\neq 0$. Une seule contrainte différencie les deux hypothèses, et $0\in I\!R$. Donc $q=1$ et le résultat (-15) s'applique. Si on souhaite tester $H_0:\mu=0$ contre $H_1:\mu>0$, le domaine θ est alors restreint à $I\!R^+\times I\!R_*^+$, et (-15) doit être remplacé par

$$R_n \overset{n\to\infty}{\sim} \frac{1}{2}\delta_0+\frac{1}{2}\chi^2_1.$$

A.3 Cas multidimensionnels

L'étude d'aléas conjoints nécessite de pouvoir généraliser les principaux concepts et notions décrits au § A.1. Soit $\mathbf{X}=(x_1,\ldots,x_d)^T$ le vecteur des aléas considérés. La *fonction de répartition jointe* est définie par

$$F_X(\mathbf{x})=\mathbb{P}\left(X_1\leq x_1,\ldots,X_d\leq x_d\right)$$

où $\mathbf{x}=(x_1,\ldots,x_d)$. Lorsque les X_i sont des variables aléatoires continues, et en supposant F_X différentiable, la densité de probabilité jointe s'écrit

$$f_X(\mathbf{x})=\frac{\partial^d F_X}{\partial x_1\ldots\partial x_d}(\mathbf{x}).$$

Alors, pour tout ensemble $\mathcal{A}\subset\Omega\subset I\!R^d$

$$\mathbb{P}\left(\mathbf{X}\in\mathcal{A}\right)=\int_{\mathcal{A}} f_X(\mathbf{u})\,d\mathbf{u}.$$

En particulier, si $\Omega=I\!R^d$:

$$F_X(\mathbf{x})=\int_{-\infty}^{x_1}\ldots\int_{-\infty}^{x_d} f_X(\mathbf{u})\,du_1\ldots du_d.$$

Chaque *densité marginale*, caractérisant X_i indépendamment des autres variables, s'obtient par intégration sur les autres composantes : si $\Omega=\bigotimes_{i=1}^d\Omega_i$, alors

$$f_{X_i}(x_i)=\iint_{\bigotimes_{j\neq i}\Omega_j} f_X(u_1,\ldots,u_{i-1},x_i,u_{i+1},\ldots,u_d)\,du_1\ldots du_d.$$

La notion de *covariance* permet de résumer la dépendance entre les X_i deux à deux :

$$\mathbb{C}ov(X_i,X_j)=\int_{\Omega_i}\int_{\Omega_j}\left(x_i-\mathbb{E}[X_i]\right)\left(x_j-\mathbb{E}[X_j]\right)f_{X_i,X_j}(x_i,x_j)\,dx_i dx_j$$

où $\mathbb{E}[X_i]$ est l'espérance marginale de X_i et f_{X_i,X_j} est la densité jointe bivariée de X_i et X_j, définie comme la marginale

$$f_{X_i,X_j}(x_i,x_j)=\int_{\bigotimes_{k\neq i,j}\Omega_k} f_X(\ldots,u_{i-1},x_i,\ldots,u_{j-1},x_j,\ldots,u_d)\,du_1\ldots du_d.$$

La covariance généralise la notion de variance : $\mathbb{C}ov(X_i,X_i)=\mathbb{V}[X_i]$ (variance de la loi marginale de X_i). Dans la pratique, la loi multivariée est souvent résumée par son vecteur d'espérances $\mathbb{E}[\mathbf{X}]=(\mathbb{E}[X_1],\ldots,\mathbb{E}[X_d])^T$ et sa matrice de variance-covariance

$$\Sigma=(\mathbb{C}ov(X_i,X_j))_{i,j}$$

ou sa *matrice de corrélation* $\Sigma' = (\rho_{i,j})_{i,j}$ définie par

$$\rho_{i,j} = \frac{\mathbb{C}ov(X_i, X_j)}{\sqrt{\mathbb{V}[X_i]\mathbb{V}[X_j]}}. \tag{-15}$$

Chaque $\rho_{i,j}$ évolue entre -1 et 1 et fournit une information sur la dépendance *linéaire* entre les variables X_i et X_j. Toutefois, ce résumé est en général très incomplet. Par exemple, s'il y a indépendance entre X_i et X_j, alors $\mathbb{C}ov(X_i, X_j) = 0$, mais la réciproque n'est pas toujours vraie. La matrice des cœfficients de corrélation Σ' n'apporte une information exhaustive sur la structure de dépendance que dans des cas très précis, notamment lorsque $\mathbf{X}$ est un vecteur gaussien, mais ne fournit pas en général une mesure réellement pertinente de cette dépendance. Il faut donc combattre la pratique bien établie d'accorder une confiance importante à cet indicateur [37].

Un cours spécifique doit préciser ce qui est entendu par *information exhaustive sur la structure de dépendance*, et fournir des outils plus adaptés au maniement des lois multivariées. Les premiers de ces outils sont les **copules**.

A.4 Processus aléatoires et stationnarité

Les lois apparaissant dans cet ouvrage constituent un cas particulier des *processus aléatoires* (ou stochastiques) en temps discret [7], qui définissent le comportement général d'une suite de variables aléatoires $X_1, \ldots, X_n$. Ces variables ne sont plus obligatoirement considérées comme indépendantes et identiquement distribuées (*iid*). La loi f_{X_i} de chaque X_i peut varier selon i. Il peut aussi y avoir dépendance entre les X_i tout en conservant l'hypothèse d'une loi similaire pour chaque X_i. Dans ce dernier cas, le processus est alors dit *stationnaire*.

Définition 17 Stationnarité d'un processus. *Un processus aléatoire $X_1, \ldots, X_n$ est dit stationnaire si, pour tout ensemble d'entiers $\{k_1, \ldots, k_s\}$ et pour tout entier m, les distributions de probabilité jointes de $(X_{k_1}, \ldots, X_{k_s})$ et $(X_{k_1+m}, \ldots, X_{k_s+m})$ sont identiques.*

Cette définition permet par exemple de caractériser les séries temporelles de façon plus appropriée que la mention *iid*.Le mécanisme stochastique définissant le processus aléatoire nécessite parfois d'être précisé. C'est en particulier vrai lorsqu'on étudie si ce processus converge vers un processus stationnaire lorsque n grandit.

On peut ainsi imaginer que $X_1, \ldots, X_n, \ldots$ représentent des observations d'une température à des pas de temps très courts, et qu'il est souhaitable de pouvoir sélectionner des valeurs de températures stabilisées afin de calculer des grandeurs représentatives. Pour cela, il est nécessaire de pouvoir spécifier la distribution de probabilité de X_k conditionnelle à $X_{k-1}, X_{k-2}, \ldots, X_1$ et d'utiliser une représentation par *chaîne de Markov*.

Définition 18 Chaîne de Markov. *Un processus aléatoire $X_1, \ldots, X_n, \ldots$ est une chaîne de Markov d'ordre $r \in \mathbb{N}^*$ si, pour tout $i \geq r$,*

$$\mathbb{P}(X_i | X_{i-1}, \ldots, X_1) \quad = \quad \mathbb{P}(X_i | X_{i-1}, \ldots, X_{i-r}).$$

Si, de plus, $r = 1$ et que cette probabilité de transition ne dépend pas de i, le processus est dit homogène.

Les chaînes de Markov d'ordre 1 sont donc les plus aisées à spécifier, et constituent un outil de généralisation important des cas *iid* (un exemple est tracé sur la figure 3). Elles jouent également un grand rôle dans des cadres d'inférence et d'échantillonnage. Ainsi, un processus $\theta_1, \ldots, \theta_n$ peut être construit comme un mécanisme d'exploration de l'espace θ, par exemple dans un cadre bayésien, et ce mécanisme d'exploration est très souvent construit en produisant une chaîne de Markov d'ordre 1, qui possède des propriétés de convergence vers un processus-limite stationnaire [8], dont les propriétés (espérance, variance, etc.) peuvent être estimées. Nous suggérons l'ouvrage de référence [28] au lecteur désireux d'explorer ce champ de la théorie des probabilités.

7. Les processus aléatoires à temps continu seront présentés dans cet ouvrage.
8. On parle aussi de *distribution stationnaire*.

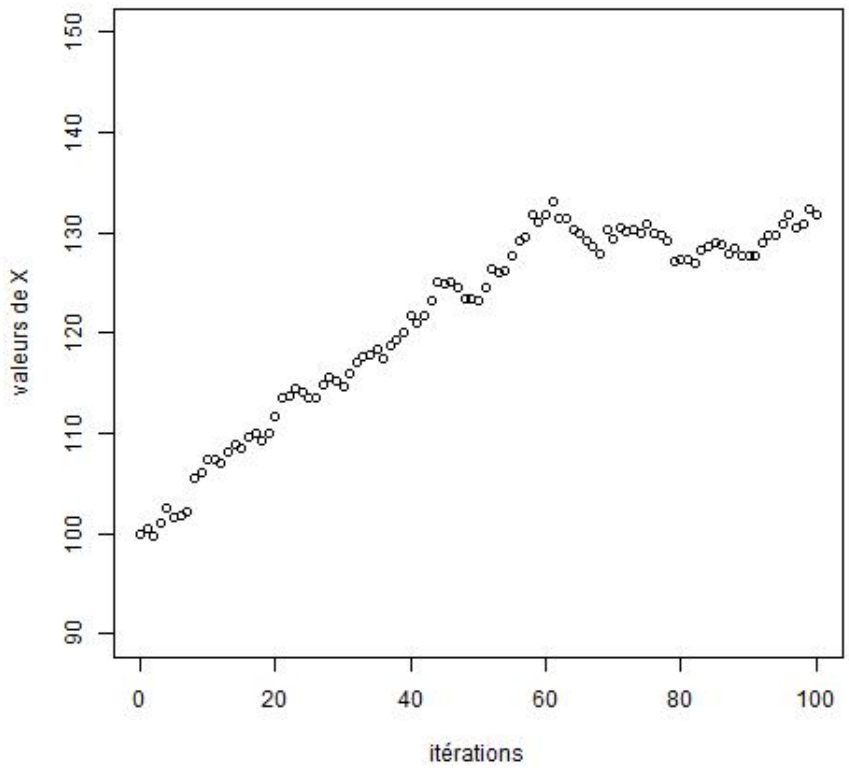

FIGURE 3 – Exemple d'une chaîne de Markov d'ordre 1 non stationnaire.

A.5 Modélisations probabiliste et statistique

Les termes de modélisations probabiliste et statistique sont souvent confondus, en particulier dans la littérature d'ingénierie. Cependant, ils possèdent des sens différents. Un modèle probabiliste décrit par sa densité de probabilité f_x est voué à représenter un phénomène (ex : physique) réel :

$$X \sim f_x$$

tandis que le modèle statistique traduit le fait qu'une ou plusieurs observations de X, notée(s) x^*, sont reliés à une réalisation réelle x de X par un dispositif de mesure : par exemple

$$x^* \quad = \quad x + \epsilon \tag{-15}$$

où ϵ est un *bruit d'observation* dont la nature est aléatoire et qui est souvent supposé gaussien. On notera f_ϵ sa densité, qui est en général connue [9]. La connaissance de la relation (-15) permet de définir la distribution de probabilité de la variable aléatoire X^* de réalisation x^* comme une *loi de convolution* de densité f_{x^*}

$$f_{x^*}(u) = \int f_x(u+y) f_\epsilon(y)\, dy$$

et cette loi détermine la *vraisemblance statistique* de l'observation x^* (cf. § A.6). Toutefois, il est essentiel pour la détermination de f_x, enjeu majeur de l'étude, que cette loi explique la majeure partie de la variabilité et

9. Notamment *via* les spécifications des constructeurs des dispositifs de mesure, ou par des tests répétés dans des conditions contrôlées.

des valeurs observées (celles de X^*). Souvent, on fera l'hypothèse que l'influence de ϵ est négligeable, ce qui revient à écrire

$$f_{x^*}(u) \simeq f_x(u) \quad \forall u \in \Omega,$$

et à confondre modèles probabiliste et statistique. Cette hypothèse n'est cependant pas toujours vérifiée en pratique, en particulier pour les observations historiques [26]. Le bruit affectant une mesure peut être important car cette dernière peut :

- ne pas être directe (par exemple, les mesures de pluie torrentielles ancestrales peuvent être reconstituées à partir d'études stratigraphiques [8]) ;
- être très imprécise (exemple : une crue datant du Moyen Âge a fait l'objet d'une chronique en termes qualitatifs (elle a emporté un pont, recouvert des champs...) ou quantitatif avec beaucoup d'incertitude (marque sur un mur de maison démolie depuis) [5, 24] ;
- souffrir d'un biais inconnu lié à un dispositif de mesure mal calibré (ou abîmé par l'aléa lui-même, surtout s'il est extrême) [2].

Même certaines mesures récentes peuvent souffrir d'un bruit potentiellement fort, car elles sont issues d'un calcul - et non d'une mesure directe - soumis à certaines incertitudes (voir également § **??**).

A.6 Contrôle de l'erreur de modélisation

Convergence des modèles

La fiabilité des modèles probabilistes et statistiques repose sur une approximation du réel dont l'erreur peut être encadrée sous certaines hypothèses techniques. On distingue dans cet ouvrage deux types d'approximation :

1. une approximation du comportement inconnu d'une grandeur X considérée comme aléatoire (par exemple le maximum d'un échantillon sur un intervalle de temps donné) par un comportement théorique (par exemple issu de la théorie statistique des valeurs extrêmes) permettant de quantifier et d'extrapoler ;
2. une approximation d'un modèle probabiliste théorique par un modèle statistique *estimé*, au sens où ce modèle théorique implique des paramètres *a priori* inconnus θ, qui seront quantifiés grâce aux observations réelles ; puisque ces observations $x_1, \ldots, x_n$ sont considérées comme des réalisations d'une variable aléatoire X, le paramètre *estimé* est également considéré comme une réalisation d'une autre variable aléatoire $\hat{\theta}_n$, définie comme un *estimateur statistique*.

La suite $(\hat{\theta}_n)_n$ constitue donc un premier processus stochastique, dont on souhaite qu'il approxime θ (paramètre fixe mais inconnu). L'ensemble des variables aléatoires $(X_n)_n$ produites alors par le modèle estimé forme un deuxième processus stochastique, dont on souhaite qu'il approxime le comportement réel X (variable aléatoire de loi inconnue).

Il est donc indispensable de vérifier que ces deux types d'approximation n'empêchent pas les *modèles statistiques estimés* - les outils concrets de l'étude - de fournir un diagnostic pertinent en termes de reproductibilité des observations, et n'entravent pas significativement leur emploi dans des études prévisionnelles. Une condition indispensable est d'avoir *convergence* entre modélisation théorique et réalité, puis entre modèle estimé et modélisation théorique. Cette convergence s'exprime sous la forme d'un écart entre les protagonistes, qui doit nécessairement diminuer lorsque la quantité d'information (c'est-à-dire le nombre d'observations n) s'accroît jusqu'à devenir nul lorsque $n \to \infty$.

Dans le monde probabiliste, cet écart est aléatoire, et il est donc possible qu'un écart soit nul sauf en un nombre k de situations données, formant un sous-ensemble de l'espace des événements Ω de mesure nulle. Typiquement, cet ensemble peut être formé d'un nombre fini de valeurs ponctuelles, ou d'éléments appartenant à la frontière de Ω ; en effet, dans le monde continu on sait que (sous des conditions d'indépendance)

$$\mathbb{P}(X \in \{x_1, \ldots, x_m\}) \quad = \quad \sum_{i=1}^{m} \mathbb{P}(X = x_i)$$

et que $\mathbb{P}(X = x_i) = 0$ pour tout x_i (puisque X est continu). On parlera dans ce cas de nullité *presque sûre*.

La notion de *convergence presque sûre* s'en déduit assez naturellement : il s'agit de vérifier que la probabilité que la limite d'un processus stochastique $\hat{\theta}_n$ (ou X_n) corresponde à la cible θ (ou X) vaut 1 ; ou de façon équivalente, que l'écart entre la limite de ce processus et θ (ou X) est nul presque sûrement :

$$X_n \xrightarrow{p.s.} X \quad \Leftrightarrow \quad \mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Cette notion de convergence est la plus forte et la plus courante en pratique pour démontrer le comportement attendu d'un processus aléatoire vers une variable aléatoire, éventuellement réduite à un vecteur (ou un scalaire). On parle également de *consistance forte* [1].

D'autres notions de convergence moins fortes, au sens où elles sont entraînées par la convergence presque sûre, sans réciprocité assurée, sont également très utilisées :

1. la convergence *en probabilité*

$$X_n \xrightarrow{\mathbb{P}} X \quad \Leftrightarrow \quad \forall \epsilon > 0, \ \lim_{n\to\infty} \mathbb{P}\left(|X_n - X| > \epsilon\right) = 0$$

   joue un rôle important dans un grand nombre de démonstrations de convergences en loi, et implique également la convergence presque sûre d'une sous-suite de $(X_n)_n$; elle permet à X_n de s'écarter de X, mais de moins en moins significativement à mesure que n croît ;

2. la convergence *en loi*, qui est entraînée par la convergence en probabilité et qui constitue l'équivalent de la convergence simple [10] dans le monde probabiliste

$$X_n \xrightarrow{\mathcal{L}} X \quad \Leftrightarrow \quad \lim_{n\to\infty} F_n(x) = F(x)$$

   où (F_n, F) sont les fonctions de répartition de X_n et X, respectivement, pour tout x où F est continue. Cette notion de convergence ne caractérise pas les valeurs des processus stochastiques, mais uniquement les comportements aléatoires : celui de X_n ressemble de plus en plus à celui de X. On parle alors de *consistance faible* [1]. Cette convergence caractérise notamment les statistiques de test (§ A.2).

D'autres notions de convergence (en norme L^p en particulier) sont également utilisées. Leur emploi est en général de s'assurer des convergences *déterministes* utiles, par exemple celles des espérances (moments), comme l'expriment les deux théorèmes suivants.

Théorème 13 *Supposons que X_n converge en norme L^1 vers X dans $\Omega \in I\!R$:*

$$\lim_{n\to\infty} \mathbb{E}\left[\|X_n - X\|^1\right] \quad = \quad 0.$$

Alors $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$.

Théorème 14 *Supposons que $X_n \xrightarrow{\mathcal{L}} X$ avec $(X_n, X) \in \Omega^2$ avec $\Omega \subset I\!R$. Alors, pour toute fonction réelle, continue et bornée g (en particulier l'identité),*

$$\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)].$$

Un ensemble de résultats techniques permet de combiner ces différentes convergences et leurs transformations par des fonctions continues (*mapping theorem*) pour étudier des modèles complexes. Pour une exploration approfondie des notions évoquées dans ce paragraphe et leur généralisation dans un monde multidimensionnel, nous suggérons au lecteur l'ouvrage [35].

1. Bien que *stricto sensu*, la *consistance* est une propriété locale d'un estimateur, qui est induite par la convergence (possédant un sens global).

10. Au sens de la *fonction caractéristique* pour les spécialistes (théorème de continuité de Lévy [30]).

Estimation statistique classique

L'*inférence* est l'ensemble des méthodologies permettant de construire un ou plusieurs estimateurs de θ

$$\hat{\theta}_n = T(X_1, \ldots, X_n)$$

où T est une fonction des variables aléatoires associées aux réalisations x_i du phénomène étudié. Comme indiqué précédemment, $\hat{\theta}_n$ est donc lui-même une variable aléatoire, et sa valeur *observée* $T(x_1, \ldots, x_n)$ est appelée un *estimé*.

Principales propriétés des estimateurs statistiques Il existe une infinité d'estimateurs possibles pour un vecteur de paramètre θ, et il est donc indispensable de pouvoir opérer une sélection parmi eux. En statistique classique, les principales règles utilisées pour classer les estimateurs sont les suivantes :

1. *asymptotiquement* il doit y avoir *consistance* :

   $$\hat{\theta}_n \xrightarrow{?} X$$

   où ? représente, au mieux, la convergence presque sûre ;

2. l'*erreur quadratique*

   $$\mathrm{EQ}(\hat{\theta}_n) = \mathbb{E}\left[(\hat{\theta}_n - \theta)^T(\hat{\theta}_n - \theta)\right], \tag{-19}$$

   doit être la plus faible possible ; celle-ci peut s'écrire comme la somme du déterminant de la matrice de variance-covariance de $\hat{\theta}_n$, qui est une mesure de l'imprécision non-asymptotique de cet estimateur, et du carré du *biais* [11] de l'estimateur

   $$\mathrm{B}(\hat{\theta}_n) = \mathbb{E}\left[\hat{\theta}_n\right] - \theta,$$

   que l'on peut définir comme l'*erreur non-asymptotique* en espérance. Ces deux termes ne peuvent être minimisés simultanément, et la minimisation de de (-19) procède donc nécessairement d'un *équilibre biais-variance*. .

Remarquons que produire un estimateur $\hat{\theta}_n$ faiblement consistant pour un paramètre inconnu θ permet de produire un autre estimateur faiblement consistant sur n'importe quelle fonction $h(\theta)$ de ce paramètre, pourvu que h soit différentiable. Lorsque la loi de convergence est gaussienne, le procédé de dérivation permettant de le construire est connu sous le nom de *méthode Delta*.

Théorème 15 Méthode Delta multivariée *[22]. Soit $\theta_1, \ldots, \theta_n$ un processus stochastique dans $\mathbb{R}^d$ et soit $g : \mathbb{R}^d \to R^q$ une fonction différentiable et non nulle en θ. Notons $J_g(\theta)$ la jacobienne de g en θ. Supposons que $\sqrt{n}(\theta_n - \theta)$ converge en loi vers la loi normale multivariée $\mathcal{N}_d(\mathbf{0}_d, \Sigma)$, de moyenne le vecteur nul $\mathbf{0}_d$ en dimension d et de variance-covariance $\Sigma \in \mathbb{R}^{2d}$. Alors*

$$\sqrt{n}\left(g(\theta_n) - g(\theta)\right) \quad \xrightarrow{\mathcal{L}} \quad \mathcal{N}_q\left(0, J_g^T(\theta)\Sigma J_g(\theta)\right).$$

Estimation des moindres carrés La classe des *estimateurs des moindres carrés* (EMC), qui cherchent à réaliser un compromis entre biais et variance, est donc naturellement définie par une règle du type

$$\hat{\theta}_n \quad = \quad \arg\min_{\hat{\theta}} \widetilde{\mathrm{EQ}}(\hat{\theta}) \tag{-19}$$

où $\widetilde{\mathrm{EQ}}$ est une approximation empirique de EQ, construite comme une fonction de $X_1, \ldots, X_n$. Les estimateurs ainsi produits possèdent souvent de bonnes propriétés de consistance, mais peuvent s'avérer sensibles aux choix du modèle et de la paramétrisation θ. Ainsi, il n'est pas évident que l'espérance et/ou la variance impliquées dans le critère (-19) existent.

11. Un estimateur $\hat{\theta}_n$ dont l'espérance est égale à θ est dit *sans biais*.

Estimation par maximisation de vraisemblance

Principe de vraisemblance. Une règle plus générale est donc nécessairement fondée sur une représentation plus exhaustive, générique et toujours définie de l'information apportée par $X_1, \ldots, X_n$ sur le modèle paramétré par θ. Une telle représentation est la *vraisemblance statistique* ℓ, qui est définie (pour des X_i continus) comme la densité jointe des observations $X_i = x_i$ conditionnelle à θ. Ainsi, dans un cas où les observations x_i sont des réalisations iid :

$$\ell(x_1, \ldots, x_n|\theta) \quad = \quad \prod_{i=1}^{n} f_X(x_i|\theta). \tag{-18}$$

La vraisemblance peut prendre des formes plus compliquées lorsque les réalisations ne sont pas indépendantes, non identiquement distribuées ou sont *manquantes* et ont été remplacées par des valeurs-seuils, par exemple parce que les limites du procédé de mesure ont été atteintes.

EXEMPLE 13.**Mesure de la vitesse du vent.** *Certains vieux anémomètres ne peuvent mesurer la vitesse du vent au-delà d'une certaine valeur, et remplacent l'observation x_i qui aurait dû être faite par une vitesse maximale de vent mesurable, notée c. On parle alors d'observation statistique censurée à droite. Ce type d'observation partielle est fréquente en analyse de survie [18]. Le terme de densité $f(x_i)$ correspondant à une observation correcte est alors remplacé par la probabilité que la donnée manquante $P(X \geq c) = F(c)$ dans l'écriture de la vraisemblance (-18), où F est la fonction de répartition de X.*

L'exhaustivité de l'information portée par la vraisemblance constitue un principe fondamental de la théorie statistique classique. Alors, l'*estimateur du maximum de vraisemblance* (EMV) [12]

$$\hat{\theta}_n \quad = \quad \arg\max \ell(X_1, \ldots, X_n|\theta) \tag{-17}$$

définit la variable aléatoire dont la réalisation est la *valeur la plus probable* de θ ayant généré l'ensemble de réalisations $\{X_i = x_i\}_i$. Par le caractère générique de sa dérivation, sa signification et ses bonnes propriétés de consistance, il est l'estimateur statistique le plus courant et l'un des plus naturels.

EXEMPLE 14.**Données censurées par intervalles.** *Très fréquemment, une donnée historique unidimensionnelle, ou mal mesurée, peut être simplement décrite comme une valeur manquante x_i entre deux bornes connues $x_{i,\min} < x_{i,\max}$. Le terme de densité $f(x_i)$ doit alors être remplacé dans la vraisemblance (-18) par*

$$\begin{aligned} P\left(x_{i,\min} \leq X \leq x_{i,\max}|x_{i,\min}, x_{i,\max}\right) & \qquad (-16)\\ &= P\left(X \leq x_{i,\max}|x_{i,\max}\right) - P\left(X \leq x_{i,\min}|x_{i,\min}\right), \\ &= F(x_{i,\max}) - F(x_{i,\min}). \qquad (-16) \end{aligned}$$

Si l'on fait de plus l'hypothèse que les valeurs $(x_{i,\min}, x_{i,\max})$ sont des données elles-mêmes aléatoires (par exemple bruitées), décrites comme des réalisations de variables $(X_{i,\min}, X_{i,\max})$ de lois respectives $f_{i,\min}$, $f_{i,\max}$), le terme de vraisemblance (-16) devient

$$\iint P\left(X_{i,\min} \leq X \leq X_{i,\max}|X_{i,\min} = y_1, X_{i,\max} = y_2\right) f_{i,\min}(y_1) f_{i,\max}(y_2)\, dy_1 dy_2.$$

Lorsque la donnée est multivariée, plusieurs situations peuvent se présenter : une ou plusieurs dimensions de X peuvent être censurées par intervalles, et des traitement approfondis doivent être menés pour obtenir des spécifications statistiques utiles (voir [13] pour les analyses de survie, et [29] pour le cas spécifique des extrêmes multivariés).

Théorème 16 Limite centrale pour l'EMV. *Supposons que $X_1, \ldots, X_n$ soient indépendants et identiquement distribués. Soit q la dimension de θ. Alors, sous des conditions de régularité très générales (dites de Wald),*

$$\hat{\theta}_n \quad \xrightarrow{\mathcal{L}} \quad \mathcal{N}_q\left(\theta, I_\theta^{-1}\right) \tag{-16}$$

12. Pour des raisons de commodité, on remplace souvent ℓ par la log-vraisemblance $\log \ell$ dans la définition (-17).

où $\mathcal{N}_q$ représente la loi normale multivariée en dimension q, de variance-covariance I_θ^{-1}, et I_θ est la matrice d'information de Fisher dont le terme $(i,j) \in \{1,\ldots,q\}^2$ est défini par (sous ces mêmes conditions de régularité)

$$I_\theta^{(i,j)} \quad = \quad -\mathbb{E}_X \left[\frac{\partial \log \ell(X_1,\ldots,X_n|\theta)}{\partial\theta_i \partial\theta_j} \right]. \tag{-15}$$

Quelques informations supplémentaires sur la notion d'information et la matrice de Fisher sont indiquées au § A.6. Deux propriétés importantes de l'EMV sont d'être *asymptotiquement sans biais* et *fortement consistant et efficace asymptotiquement* : sa covariance asymptotique, fournie par l'inverse de la matrice de Fisher, est *minimale* pour tous les estimateurs sans biais de θ.

Information de Fisher La notion d'information a été proposée dans les années 1920 par le chercheur anglais Ronald A. Fisher (considéré comme le père de la statistique mathématique). La démarche de Fisher est la suivante : si l'on s'intéresse aux caractéristiques d'une population nombreuse (voire infinie, qui est le cas limite auquel on est ramené en permanence), on ne peut ni connaître ni traiter les informations trop abondantes relatives à chacun des individus qui la composent. Le problème devient donc d'être capable de décrire correctement la population au moyen d'indicateurs de synthèse pouvant être fournis par des échantillons issus de la population à étudier. Plus les données chiffrées que l'on peut extraire d'un échantillon représentent correctement la population de référence et plus l'information contenue dans cet échantillon doit être considérée comme élevée.

Partant de cette hypothèse, Fisher a défini techniquement l'information comme la valeur moyenne du carré de la dérivée du logarithme de la loi de probabilité étudiée. L'inégalité de Cramer permet alors de montrer que la valeur d'une telle information est proportionnelle à la faible variabilité – c'est-à-dire au fort degré de certitude – des conclusions qu'elle permet de tirer. Cette idée, qui est à la racine de toute la théorie de l'estimation et de l'inférence statistique, est exactement celle que l'on retrouvera vingt ans plus tard chez Shannon, exprimée cette fois en des termes non plus statistiques mais probabilistes.

Si X est un échantillon de densité de probabilité $f(x|\theta)$, on définit l'information de Fisher par

$$I_\theta \quad = \quad \mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial\theta} \right)^2 \right].$$

Dans le cas où la distribution de probabilité dépend de plusieurs paramètres, θ n'est plus un scalaire mais un vecteur. L'information de Fisher n'est plus définie comme un scalaire mais comme une matrice de covariance appelée matrice d'information de Fisher :

$$I_{\theta_i,\theta_j} \quad = \quad \mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial\theta_i} \right) \left(\frac{\partial \log f(X|\theta)}{\partial\theta_j} \right) \right]$$

Intervalles de confiance Dans la pratique, la loi asymptotique de $\hat{\theta}_n$ est à son tour estimée en remplaçant le terme inconnu I_θ par un estimateur consistant $\hat{I}_n$ (en général $\hat{I}_n = I_{\hat{\theta}_n}$), ce qui permet de définir des *zones de confiance* $C_{\hat{\theta}_n,\alpha}$ associées à l'estimateur $\hat{\theta}_n$ telles que, lorsque n croît vers l'infini,

$$\mathbb{P}\left(\hat{\theta}_n \in C_{\hat{\theta}_n,\alpha} \right) \quad = \quad \alpha. \tag{-17}$$

En particulier, lorsqu'on s'intéresse à une dimension spécifique θ_i, le théorème 16 permet de définir l'*intervalle de confiance (asymptotique)* $1-\alpha$ associé à $\hat{\theta}_n$:

$$\mathbb{P}\left(\theta_i \in \left[\hat{\theta}_{n,i} - z_{\alpha/2}\sqrt{\sigma^2_{i,i}}\ ,\ \hat{\theta}_{n,i} + z_{\alpha/2}\sqrt{\sigma^2_{i,i}} \right] \right) \quad = \quad 1-\alpha, \tag{-16}$$

où z_α est le quantile d'ordre α de la loi normale centrée réduite et $\sigma^2_{i,i}$ le terme diagonal (i,i) de l'*estimé* de l'inverse $\hat{I}_n^{-1}$.

Les équations (-17) et (-16) permettent d'évaluer la précision de l'estimation de θ à partir de l'échantillon $x_1, \ldots, x_n$. Cependant, observons que la mesure de probabilité $\mathbb{P}$ dans l'équation (-17) concerne $\hat{\theta}_n$ et non θ (qui est inconnu mais fixe) ; une zone de confiance n'est donc pas définie par la probabilité $1-\alpha$ que θ s'y situe, mais comme une zone où il y a *a priori* une très forte probabilité $1-\alpha$ d'obtenir un *estimé* de θ. En simulant un grand nombre de fois des échantillons similaires à $x_1, \ldots, x_n$, la distribution de ces estimés a $100(1-\alpha)\%$ chances en moyenne de contenir la vraie valeur θ. L'intervalle de confiance sur une dimension i de θ vise donc à encadrer la vraie valeur θ_i avec une certaine probabilité reliée à la loi asymptotique de l'estimateur $\hat{\theta}_n$, qui présuppose que le modèle statistique est correct (et non selon une sorte de probabilité absolue, indépendante de tout modèle).

Tout comme l'EMC, l'EMV n'est pas toujours explicite et doit être en général calculé par des méthodes numériques. EMC et EMV peuvent ne pas être uniques pour des modèles complexes, et l'EMV peut aussi ne pas être défini (menant à une vraisemblance infinie). Toutefois ces cas restent rares dans le cadre de la théorie des valeurs extrêmes. À la différence de l'EMC, l'EMV est toujours invariant par reparamétrisation : l'EMV de $h(\theta)$ est $h(\hat{\theta}_n)$ pourvu que h soit une fonction bijective. Cette propriété est cruciale pour éviter des paradoxes et des inconsistances : si on remplace l'observation x par une transformation bijective $y = d(x)$, le modèle paramétré par θ est remplacé par un modèle paramétré par une transformation bijective $\theta' = h(\theta)$. Or l'information apportée par x et y est la même, et donc toute règle d'estimation $x \to \hat{\theta}(x)$ devrait être telle que $y = d(x) \to \hat{\theta}'(y) = h(\hat{\theta}(x))$.

Un dernier argument plaide en faveur de l'EMV : la vraisemblance maximisée constitue l'ingrédient fondamental de la plupart des techniques de *sélection de modèle* : assortie d'un facteur de pénalisation lié au nombre de degrés de liberté du modèle [31, 1], l'*estimé* de $\ell(x_1, \ldots, x_n|\hat{\theta}_n)$ fournit un diagnostic utile, supplémentaire aux résultats de tests stastistiques, pour évaluer la pertinence d'un modèle par rapport à un autre sur un même jeu de données. Nous renvoyons le lecteur intéressé par ce sujet à l'ouvrage spécialisé [25].

B Descriptif de quelques modèles statistiques utiles

Les tableaux suivants sont extraits d'un formulaire proposé par Aimé Lachal (Univ. Lyon).

B.1 Lois discrètes

distribution	*loi de probabilité*	$\mathbb{E}(X)$	$\mathrm{var}\,(X)$	*fonction génératrice* $\mathbb{E}(z^X)$
Bernoulli	$\mathbb{P}(X=0)=q,\ \mathbb{P}(X=1)=p$ $q=1-p$	p	pq	$pz+q$
Binomiale $\mathcal{B}(n,p)$	$\mathbb{P}(X=k)=C_n^k\,p^kq^{n-k}$ $q=1-p,\quad k=0,1,\dots,n$	np	npq	$(pz+q)^n$
Poisson $\mathcal{P}(\lambda)$	$\mathbb{P}(X=k)=e^{-\lambda}\dfrac{\lambda^k}{k!}$ $k=0,1,\dots$	λ	λ	$e^{\lambda(z-1)}$
Géométrique $\mathcal{G}(p)$	$\mathbb{P}(X=k)=pq^{k-1}$ $q=1-p,\quad k=1,2,\dots$	$\dfrac{1}{p}$	$\dfrac{q}{p^2}$	$\dfrac{pz}{1-qz}$
Hypergéométrique $\mathcal{H}(N,n,p)$	$\mathbb{P}(X=k)=\dfrac{C_{Np}^k\,C_{Nq}^{n-k}}{C_N^n}$ $q=1-p$ $\max(0,n-Nq)\leqslant k\leqslant\min(Np,n)$	np	$npq\dfrac{N-n}{N-1}$	$\dfrac{C_{Nq}^n}{C_N^n}F(-n,-Np;Nq-n+1;z)$
Binomiale négative	$\mathbb{P}(X=k)=C_{k+r-1}^{r-1}\,p^rq^k$ $q=1-p,\quad k=0,1,\dots$	$\dfrac{rq}{p}$	$\dfrac{rq}{p^2}$	$\left(\dfrac{p}{1-qz}\right)^r$
Pascal	$\mathbb{P}(X=k)=C_{k-1}^{r-1}\,p^rq^{k-r}$ $q=1-p,\quad k=r,r+1,\dots$	$\dfrac{r}{p}$	$\dfrac{rq}{p^2}$	$\left(\dfrac{pz}{1-qz}\right)^r$

Fonction hypergéométrique : $F(a,b;c;z)=\displaystyle\sum_{n=0}^{+\infty}\frac{a(a+1)\dots(a+n-1)\ b(b+1)\dots(b+n-1)}{c(c+1)\dots(c+n-1)}\,\frac{z^n}{n!}$

- La somme de n v.a. indépendantes suivant la loi de Bernoulli de paramètre p suit une loi binomiale $\mathcal{B}(n,p)$.
- La somme de deux v.a. indépendantes suivant les lois binomiales $\mathcal{B}(m,p)$ et $\mathcal{B}(n,p)$ suit la loi binomiale $\mathcal{B}(m+n,p)$.
- La somme de deux v.a. indépendantes suivant les lois de Poisson $\mathcal{P}(\lambda)$ et $\mathcal{P}(\mu)$ suit la loi de Poisson $\mathcal{P}(\lambda+\mu)$.
- La somme de deux v.a. indépendantes suivant les lois binomiales négatives de paramètres (r,p) et (s,p) suit la loi binomiale négative de paramètres $(r+s,p)$.
- La somme de r v.a. indépendantes suivant la loi géométrique $\mathcal{G}(p)$ suit la loi de Pascal de paramètres (r,p).

B.2 Lois continues

distribution	*loi de probabilité*	$\mathbb{E}(X)$	$\mathrm{var}\,(X)$	*fonction caract.* $\mathbb{E}(e^{itX})$
Uniforme $\mathcal{U}(a,b)$	$\frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{ibt}-e^{iat}}{i(b-a)t}$
Exponentielle $\mathcal{E}(\lambda)$	$\lambda e^{-\lambda x}\mathbb{1}_{\mathbb{R}^+}(x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-it}$
Normale $\mathcal{N}(m,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$	m	σ^2	$e^{imt-\frac{1}{2}\sigma^2t^2}$
Weibull $\mathcal{W}(\lambda,a)$	$\lambda a x^{a-1}e^{-\lambda x^a}\mathbb{1}_{]0,+\infty[}(x)$	$\lambda^{-\frac{1}{a}}\Gamma\left(\frac{1}{a}+1\right)$	$\lambda^{-\frac{2}{a}}[\Gamma\left(\frac{2}{a}+1\right)$ $-\Gamma\left(\frac{1}{a}+1\right)^2]$	
Cauchy $\mathcal{C}(a,b)$	$\frac{a}{\pi(a^2+(x-b)^2)}$	non définie	non définie	$e^{ibt-a\lvert t\rvert}$
Gamma $\Gamma(a,\lambda)$	$\frac{\lambda^a}{\Gamma(a)}\,x^{a-1}e^{-\lambda x}\mathbb{1}_{]0,+\infty[}(x)$	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$	$\left(\frac{\lambda}{\lambda-it}\right)^a$
Bêta $B(a,b)$	$\frac{1}{B(a,b)}\,x^{a-1}(1-x)^{b-1}\mathbb{1}_{]0,1[}(x)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$M(a,a+b;it)$
Khi-Deux $\chi^2(n)$	$\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\,x^{\frac{n}{2}-1}e^{-\frac{x}{2}}\mathbb{1}_{]0,+\infty[}(x)$	n	$2n$	$(1-2it)^{-n/2}$
Student $\mathcal{T}(n)$	$\frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\,\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$	0 si $n>1$	$\frac{n}{n-2}$ si $n>2$	$\frac{2}{\Gamma(\frac{n}{2})}\left(\frac{\lvert t\rvert\sqrt{n}}{2}\right)^{\frac{n}{2}}K_{\frac{n}{2}}(\lvert t\rvert\sqrt{n})$
Fisher $\mathcal{F}(m,n)$	$\frac{m^{\frac{m}{2}}n^{\frac{n}{2}}}{B(\frac{m}{2},\frac{n}{2})}\frac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}\mathbb{1}_{]0,+\infty[}(x)$	$\frac{n}{n-2}$ si $n>2$	$\frac{2n^2(m+n-2)}{m(n-4)(n-2)^2}$ si $n>4$	$M\left(\frac{m}{2};-\frac{n}{2};-\frac{n}{m}it\right)$

Fonction Gamma : $\Gamma(a)=\displaystyle\int_0^{+\infty}x^{a-1}e^{-x}\,dx$

Fonction Bêta : $B(a,b)=\displaystyle\int_0^1 x^{a-1}(1-x)^{b-1}\,dx$

Fonction de Kummer : $M(a;b;z)=\displaystyle\sum_{n=0}^{+\infty}\frac{a(a+1)\dots(a+n-1)}{b(b+1)\dots(b+n-1)}\frac{z^n}{n!}$

Fonction de Bessel modifiée : $K_\nu(z)=\frac{\pi}{2}\frac{I_{-\nu}(z)-I_\nu(z)}{\sin\pi\nu}$ où $I_\nu(z)=\left(\frac{z}{2}\right)^\nu\displaystyle\sum_{n=0}^{+\infty}\frac{1}{n!\,\Gamma(n+\nu+1)}\left(\frac{z^2}{4}\right)^n$

- La somme de n v.a. indépendantes suivant la loi exponentielle $\mathcal{E}(\lambda)$ suit la loi Gamma $\Gamma(n,\lambda)$.
- La somme de deux v.a. indépendantes suivant les lois Gamma $\Gamma(a,\lambda)$ et $\Gamma(b,\lambda)$ suit la loi Gamma $\Gamma(a+b,\lambda)$.
- Si les v.a. indépendantes X et Y suivent les lois Gamma $\Gamma(a,\lambda)$ et $\Gamma(b,\lambda)$, alors $\frac{X}{X+Y}$ suit la loi Bêta $B(a,b)$.
- La somme de deux v.a. indépendantes suivant les lois normales $\mathcal{N}(m_1,\sigma_1^2)$ et $\mathcal{N}(m_2,\sigma_2^2)$ suit la loi normale $\mathcal{N}(m_1+m_2,\sigma_1^2+\sigma_2^2)$.
- Le quotient de deux variables indépendantes suivant la loi normale $\mathcal{N}(0,1)$ suit la loi de Cauchy $\mathcal{C}(1,0)=\mathcal{T}(1)$.
- La somme des carrés de n v.a. indépendantes suivant la loi normale $\mathcal{N}(0,1)$ suit la loi du Khi-Deux $\chi^2(n)=\Gamma(\frac{n}{2},\frac{1}{2})$.
- Si les v.a. indépendantes X et Y suivent les lois normale $\mathcal{N}(0,1)$ et du Khi-Deux $\chi^2(n)$, alors $\frac{X}{\sqrt{Y/n}}$ suit la loi de Student $\mathcal{T}(n)$.
- Si les v.a. indépendantes X et Y suivent les lois du Khi-Deux $\chi^2(m)$ et $\chi^2(n)$, alors $\frac{mX}{nY}$ suit la loi de Fisher $\mathcal{F}(m,n)$.

C Annales corrigées 1

C.1 Fonction de coût

Soit $\theta \in \Theta = I\!R$ le paramètre d'un modèle, sur lequel on dispose d'une loi *a priori* $\pi(\theta)$ et de données $x_1, \ldots, x_n$. On suppose que la loi a posteriori de densité $\pi(\theta|x_1, \ldots, x_n)$ est propre et telle que $\mathbb{E}_\pi[\exp(k\theta)|x_1, \ldots, x_n] < \infty$ pour tout $k \in I\!R$. On considère la fonction de coût pour l'estimation δ de θ définie sur $I\!R$ par

$$L_a(\theta, \delta) \quad = \quad \exp(a(\theta - \delta) - a(\theta - \delta) - 1$$

où a est un réel.

1. Montrer que $L_a(\theta, \delta) \geq 0$ pour tout $\theta \in \Theta$ et pour tout a et qu'elle est convexe en θ ; représenter cette fonction de coût comme une fonction de $(\theta - \delta)$ lorsque $a = \{0.1, 0.5, 1, 2\}$.

2. On suppose que $a > 0$. À quelles conditions cette fonction pénalise-t-elle les coûts de sous-estimation et de surestimation de θ de façon similaire ? Au contraire, à quelles conditions cette fonction pénalise-t-elle les coûts de sous-estimation et de surestimation de θ de façon très dissymétrique ?

3. On suppose que $a \neq 0$. Donner l'expression de l'estimateur de Bayes $\hat{\delta}_a$ sous cette fonction de coût.

4. Supposons que les données sont issues de $\mathcal{N}(\theta, 1)$ et que $\pi(\theta) \propto 1$; donnez l'estimateur de Bayes associé.

Réponses.

1. Avec

$$\frac{\partial L_a}{\partial \delta}(\theta, \delta) \quad = \quad a\,(1 - \exp(a(\theta - \delta))$$

qui s'annule en $\delta = \theta$, et

$$\frac{\partial^2 L_a}{\partial \delta^2}(\theta, \delta) \quad = \quad a^2 \exp(a(\theta - \delta) \;\geq\; 0,$$

on a clairement que $\delta \to L_a(\theta, \delta)$ est convexe et de minimum 0 en $\delta = \theta$. Un petit tableau de variations peut achever de nous en convaincre. Un code R minimal pour représenter le comportment de la fonction est le suivant :

```
f <- function(a) {
    curve(exp(a*x)-a*x-1, xlim=c(-10,10))
}

par(mfrow=c(2,2))
f(0.1)
f(0.5)
f(1)
f(2)
```

2. Pour $a > 0$, $L_a(\theta, \delta)$ se comporte comme une fonction linéaire pour des grandes valeurs négatives de l'écart $\theta - \delta$, soit pour des surestimations de θ. Elle se comporte comme une fonction exponentielle pour des grandes valeurs positives de l'écart $\theta - \delta$, soit pour des sous-estimation de θ. Elle pénalise donc bien plus fortement les sous-estimations de θ que les surestimations de θ. Elle se comporte similairement comme $a(\theta - \delta)^2$ pour $\delta \to \theta$ (à gauche comme à droite). On en déduit que cette fonction de coût est appropriée dans les cas où les petites erreurs de sous-estimation et de surestimation ne provoquent pas un coût très différent, mais où les grandes erreurs amènent à des coûts très différents.

3. L'estimateur de Bayes est défini par

$$\hat{\delta}_a \quad = \quad \arg\min_{\delta} \underbrace{\int_{\Theta} L_a(\theta,\delta)\pi(\theta|x_1,\ldots,x_n)\,d\theta}_{J(\delta)}.$$

Comme la fonction de coût est convexe, l'estimateur est donc défini comme la valeur de δ qui annule la dérivée du terme $J(\delta)$. Alors

$$\begin{aligned} J'(\hat{\delta}) = 0 \quad &\Leftrightarrow \quad \int_{\Theta} \frac{\partial L_a}{\partial \delta}(\theta,\hat{\delta})\pi(\theta|x_1,\ldots,x_n)\,d\theta = 0, \\ &\Leftrightarrow \quad \exp(-a\hat{\delta}) \int_{\Theta} \exp(a\theta)\pi(\theta|x_1,\ldots,x_n)\,d\theta = 1, \end{aligned}$$

le terme de droite étant bien défini car on suppose $\mathbb{E}_\pi[\exp(k\theta)|x_1,\ldots,x_n] < \infty$ pour tout $k \in \mathbb{R}$. Il vient alors (avec $a \neq 0$)

$$\hat{\delta} \quad = \quad \frac{1}{a}\log \int_{\Theta} \exp(a\theta)\pi(\theta|x_1,\ldots,x_n)\,d\theta. \tag{-21}$$

4. Avec $x_1,\ldots,x_n \sim \mathcal{N}(\theta,1)$ et $\pi(\theta) \propto 1$, il vient

$$\begin{aligned} \pi(\theta|x_1,\ldots,x_n) \quad &\propto \quad \exp\left(-\frac{n}{2}\theta^2 + \theta\sum_{i=1}^{n} x_i\right), \\ &\propto \quad \exp\left(-\frac{n}{2}\left\{\theta^2 - \frac{n}{2}2\theta\bar{x}_n\right\}\right), \\ &\propto \quad \exp\left(-\frac{n}{2}\left\{\theta - \bar{x}_n\right\}^2\right) \end{aligned}$$

et donc $\pi(\theta|x_1,\ldots,x_n)$ est la densité de la loi $\mathcal{N}(\bar{x}_n, 1/n)$. En appliquant (-21), on déduit alors

$$\begin{aligned} \hat{\delta} \quad &= \quad \frac{1}{a}\log \int_{\Theta} \frac{n}{2\pi}\exp\left(a\theta - \frac{n}{2}\left\{\theta - \bar{x}_n\right\}^2\right)\,d\theta, \\ &= \quad \frac{1}{a}\log \int_{\Theta} \frac{n}{2\pi}\exp\left(-\frac{n}{2}\theta^2 - \frac{n}{2}\bar{x}_n^2 + 2\frac{n}{2}\theta(\bar{x}_n + a/n)\right)\,d\theta, \\ &= \quad \frac{1}{a}\log\exp\left(-\frac{n}{2}\bar{x}_n^2 + \frac{n}{2}(\bar{x}_n + a/n)^2\right) \int_{\Theta} \frac{n}{2\pi}\exp\left(-\frac{n}{2}\left\{\theta - (\bar{x}_n + a/n)\right\}^2\right) \end{aligned}$$

On reconnaît dans le terme intégral la densité d'une loi $\mathcal{N}(\bar{x}_n + a/n, 1/n)$. Avec $\Theta = \mathbb{R}$, cette intégrale vaut donc 1, et

$$\begin{aligned} \hat{\delta} \quad &= \quad \frac{1}{a}\left(-\frac{n}{2}\bar{x}_n^2 + \frac{n}{2}(\bar{x}_n + a/n)^2\right), \\ &= \quad \bar{x}_n + a/2n. \end{aligned}$$

Remarque. Cette fonction de coût alternative aux fonctions classiques (coûts absolu, quadratique...) est dite LINEX (*linear-exponential*) et a été introduite par Varian en 1974 puis très utilisée par Zellner en 1986.

C.2 Élicitation d'*a priori* non informatif

On considère le problème suivant

$$x_i \quad \sim \quad \mathcal{N}(\mu_i, \sigma^2) \quad \text{pour } i = 1,\ldots,n$$

où les x_i sont indépendants.

1. Quelle est la densité jointe des données $x_1,\ldots,x_n$?
2. Calculer la matrice d'information I de Fisher pour ce jeu de données
3. En déduire la mesure *a priori* de Jeffreys $\pi^J(\theta)$ pour $\theta = (\mu_1,\ldots,\mu_n,\sigma)$
4. Que peut-on dire de $\pi^J(\sigma^2|x_1,\ldots,x_n,\mu_1,\ldots,\mu_n)$? Est-ce une loi vue en cours ?

Réponses.

1. La loi jointe des données, qui est aussi la vraisemblance, s'écrit

$$f(x_1,\ldots,x_n|\theta) \quad = \quad \frac{\sigma^{-n}}{(2\pi)^{n/2}}\exp\left(-\sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{2\sigma^2}\right). \tag{-29}$$

2. La matrice d'information de Fisher s'écrit, dans ce cas régulier, comme

$$I=-\mathbb{E}\left[\begin{array}{cccc} \frac{\partial^2}{\partial\theta_{i_1}^2}\log f(x|\theta) & \frac{\partial^2}{\partial\theta_{i_1}\theta_{i_2}}\log f(x|\theta) & \ldots & \frac{\partial^2}{\partial\theta_{i_1}\theta_{i_d}}\log f(x|\theta) \\ \frac{\partial^2}{\partial\theta_{i_1}\theta_{i_2}}\log f(x|\theta) & \frac{\partial^2}{\partial\theta_{i_2}^2}\log f(x|\theta) & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{array}\right]$$

où $x=(x_1,\ldots,x_n)$ et $d=n$. Or

$$\begin{aligned} \frac{\partial^2}{\partial\sigma^2}\log f(x|\theta) &= \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu_i)^2, \\ \frac{\partial^2}{\partial\mu_i\partial\mu_j}\log f(x|\theta) &= 0 \quad \text{si } i\neq j, \\ \frac{\partial^2}{\partial\mu_i^2}\log f(x|\theta) &= -\frac{1}{2\sigma^2} \end{aligned}$$

et

$$\frac{\partial^2}{\partial\sigma\partial\mu_i}\log f(x|\theta) \quad = \quad -\frac{1}{\sigma^4}(x_i-\mu_i).$$

Avec $\mathbb{E}[X_i-\mu_i]=0$ et $\mathbb{E}[(X_i-\mu_i)^2]=\sigma^2$, il vient donc

$$I=-\mathbb{E}\left[\begin{array}{cccc} \frac{1}{2\sigma^2} & & & \\ & \frac{1}{2\sigma^2} & \ldots & (\mathbf{0}) \\ & (\mathbf{0}) & \ldots & \\ & & & \frac{1}{\sigma^2}(n/2-1) \end{array}\right]$$

et donc

$$\pi^J(\theta) \quad \propto \quad \sigma^{-n-1}.$$

3. En utilisant (-29), la loi *a posteriori* s'écrit sous une forme condensée comme

$$\pi^J(\theta|x_1,\ldots,x_n) \quad \propto \quad \sigma^{-2n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_i)^2\right).$$

En opérant le changement de variable $\sigma\to\sigma^2$, on obtient alors

$$\begin{aligned} \pi^J(\sigma^2|x_1,\ldots,x_n,\mu_1,\ldots,\mu_n) &\propto \sigma^{-1}\pi^J(\theta|x_1,\ldots,x_n), \\ &\propto \sigma^{-2(n+1)}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_i)^2\right) \end{aligned}$$

et on reconnaît le terme général d'une loi inverse gamma $\mathcal{IG}\left(n,\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_i)^2\right)$ pour la variable aléatoire σ^2.

C.3 Élicitation et calcul bayésien pour un problème de Gumbel

La loi de Gumbel, de fonction de répartition

$$P(X < x|\theta) \quad = \quad \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} \quad \text{avec } \sigma > 0,\ \mu \in I\!R \text{ et } x \in I\!R$$

et $\theta = (\mu, \sigma)$, est souvent utilisée en météorologie pour modéliser le comportement d'un échantillon de *maxima* d'une variable environnementale. Son espérance vaut $\mathbb{E}[X|\theta] = \mu + \sigma\gamma$ où γ est la constante d'Euler. On suppose connaître un échantillon de données de pluies (en mm) $\mathbf{x_n} = (x_1, \ldots, x_n)$ suivant cette loi. Elles sont fournies dans la table 1 et correspondent aux années 1987 à 2013. Par ailleurs on dispose d'une expertise *a priori* qui s'exerce sur la loi *a priori* prédictive de X, et est spécifiée statistiquement sous la forme $P(X < 75) = 25\%$, $P(X < 100) = 50\%$, $P(X < 150) = 75\%$.

107.6	72.4	204.5	83.8	142	95.5	316.1	177.9	87.3
81.9	109.1	89.5	150.7	122.1	98.2	113.2	104.4	66.9
136.4	275.4	125	199.8	51.2	75	168.2	106	72.8

TABLE 1 – Données de pluviométrie extrême.

On considère la mesure *a priori*

$$\pi(\mu, \sigma) \quad \propto \quad \sigma^{-m} \exp\left(m\frac{(\mu - \bar{\tilde{x}}_m)}{\sigma} - \sum_{i=1}^{m} \exp\left\{-\frac{\tilde{x}_i - \mu}{\sigma}\right\}\right)$$

où les hyperparamètres $(m, \bar{\tilde{x}}_m, \tilde{x}_1, \ldots, \tilde{x}_m)$ correspondent respectivement à la taille d'un échantillon de données *a priori* (virtuelles), sa moyenne et les données elles-mêmes (supposées calibrables).

1. Ecrivez la densité de la loi *a posteriori* conditionnelle aux données réelles $\mathbf{x_n}$. La loi *a priori* est-elle conjuguée ?

2. Produisez un algorithme qui simule la loi *a priori* prédictive de X en fonction des hyperparamètres et estime les quantiles prédictifs *a priori*. En fixant $m = 3$ et $(\tilde{x}_1, \tilde{x}_2) = (81, 93)$, testez les valeurs de $\tilde{x}_3$ suivantes : 97,101,110,120. Quelle calibration vous semble la plus adéquate vis-à-vis de l'expertise *a priori* ?

3. Pour les calibrations des hyperparamètres précédentes, écrivez un algorithme qui produit un tirage de la loi *a posteriori* de θ ainsi qu'une représentation (densité empirique) de la loi *a posteriori prédictive* sur X. Comparez avec un histogramme des données $\mathbf{x_n}$.

4. **Cette question peut être traitée indépendamment du reste.** On pose à présent $\mu > 0$ et on cherche à définir une nouvelle loi *a priori* $\pi_2(\theta)$ par maximum d'entropie qui est telle que les contraintes linéaires suivantes soient respectées :

$$\begin{aligned} \mathbb{E}[X] &= 100, \\ \mathbb{E}_\pi[\log \sigma] &= 1. \end{aligned}$$

   Formalisez et résolvez numériquement (possiblement graphiquement) le problème de maximum d'entropie en supposant que la mesure de référence est la mesure de Jeffreys $\pi_0(\theta) \propto \sigma^{-2}$ (valable pour le modèle de Gumbel). Sous quelles contraintes sur les multiplicateurs de Lagrange pouvez-vous trouver une loi jointe propre ? Celle-ci appartient-elle à une classe de lois connues ?

   Rappel : *Si Y suit une loi gamma $\mathcal{G}(a, b)$, alors $\mathbb{E}[Y] = \Psi(a) - \log(b)$ où Ψ est la fonction digamma (`digamma` en R).*

5. Adaptez le code produit à la question 3 pour produire un nouveau calcul *a posteriori*, en utilisant $\pi_2(\theta)$.

Réponses.

1. Sachant des données réelles $\mathbf{x_n}$, la loi *a posteriori* s'écrit

$$\begin{aligned}\pi(\mu,\sigma|\mathbf{x_n}) &\propto \sigma^{-m-n}\exp\left(\{m+n\}\frac{\left(\mu-\frac{m\bar{\tilde{x}}_m+n\bar{\mathbf{x_n}}}{m+n}\right)}{\sigma}\right.\\ &\left.-\sum_{i=1}^{m}\exp\left\{-\frac{\tilde{x}_i-\mu}{\sigma}\right\}-\sum_{k=1}^{n}\exp\left\{-\frac{x_k-\mu}{\sigma}\right\}\right).\end{aligned}$$

   Elle est donc en effet conjuguée, car on retrouve la même forme que la loi *a priori*.

2. Pour simuler selon la loi *a priori* marginale, le plus simple est d'utiliser l'algorithme ci-dessous :

   (a) simuler μ_i, σ_i *a priori* ;

   (b) simuler X_i selon la loi de Gumbel en μ_i, σ_i.

   Pour réaliser la première simulation (la seconde peut être faite très facilement par inversion), on peut tenter de procéder de plusieurs façons : acceptation-rejet, échantillonnage d'importance, MCMC... En regardant la forme de la loi *a priori*, on privilégie l'approche par échantillonnage d'importance en utilisant (par exemple) une loi instrumentale de densité

$$g(\mu,\sigma) \equiv \mathcal{IG}_\sigma(m-1, m\bar{\tilde{x}}_m)\mathcal{E}_\mu(\lambda)$$

   avec $m > 1$, où $\mathcal{IG}$ est une loi inverse gamma. Les poids d'importance s'écrivent alors (à un coefficient près)

$$\begin{aligned}\omega_k &= \frac{\pi(\mu_k,\sigma_k)}{g(\mu_k,\sigma_k)},\\ &\propto \exp\left(\mu_k\left[m/\sigma_k+\lambda\right]-\sum_{i=1}^{m}\exp\left\{-\frac{\tilde{x}_i-\mu_k}{\sigma_k}\right\}\right)\end{aligned}$$

   où $(\mu_k, \sigma_k) \sim g(\mu, \sigma)$. Le logarithme de ces poids non normalisés est aisé à calculer, ce qui permet une approche numérique plus stable (en jouant éventuellement sur le λ). La valeur la plus adéquate était 110 (c'est elle qui permet un meilleur *matching* avec les requis de l'expertise).

3. La loi *a posteriori* étant connue explicitement, on peut utiliser le même type d'algorithme pour mener le calcul *a posteriori*.

4. On a

$$\mathbb{E}[X] = \mathbb{E}_\pi[\mathbb{E}[X|\theta]] = \mathbb{E}_\pi[\mu+\sigma\gamma]$$

   Sous cette contrainte, et sous l'autre contrainte $\mathbb{E}_\pi[\log\sigma] = 1$, la solution du problème classique de maximisation d'entropie est

$$\begin{aligned}\pi_2(\theta) &\propto \pi_0(\theta)\exp(\lambda_1(\mu+\sigma\gamma)+\lambda_2\log\sigma),\\ &\propto \sigma^{\lambda_2-2}\exp(\lambda_1\gamma\sigma)\exp(\lambda_1\mu).\end{aligned}$$

   Avec $\mu > 0$, pour avoir une loi *a posteriori* intégrable, il nous faut avoir $\lambda_1 = -\tilde{\lambda}_1 < 0$ et $\lambda_2 = \tilde{\lambda}_2 + 1$ avec $\tilde{\lambda}_2 > 0$. Dans ce cas, on reconnaît aisément un mélange de loi gamma $\mathcal{G}(\tilde{\lambda}_2, \gamma\tilde{\lambda}_1)$ pour σ, et de loi exponentielle $\mathcal{E}(\tilde{\lambda}_1)$ pour μ. Dans ce cas, on a

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}_\pi[\mu+\sigma\gamma],\\ &= \frac{1}{\tilde{\lambda}_1}+\gamma\frac{\tilde{\lambda}_2}{\gamma\tilde{\lambda}_1},\\ &= \frac{1}{\tilde{\lambda}_1}(1+\tilde{\lambda}_2),\\ &= 100. \qquad \text{(-50)}\end{aligned}$$

et

$$\mathbb{E}_\pi[\log\sigma] \quad = \quad \Psi(\tilde{\lambda}_2) - \log(\gamma\tilde{\lambda}_1) \; = \; 1.$$

Ce système de deux équations à deux inconnues peut se résoudre numériquement. D'après (-50), on a

$$\tilde{\lambda}_1 \quad = \quad (1+\tilde{\lambda}_2)/100$$

et

$$\Psi(\tilde{\lambda}_2) - \log(1+\hat{\lambda}_2) + \log 100 - 1 \quad = \quad 0.$$

Il suffit de tracer la courbe de l'équation précédente pour obtenir

$$\begin{array}{rcl} \tilde{\lambda}_2 & \simeq & 0,3155, \\ \tilde{\lambda}_1 & \simeq & 0,013155. \end{array}$$

5. On produit ici un algorithme MCMC, car on perd la propriété de conjugaison. Voici un lien vers un code R permettant de répondre à la question :

`http://www.lsta.upmc.fr/bousquet/coursM2-2018/calcul-MCMC-gumbel.r`

D Annales corrigées 2

D.1 Construction de prior

Soientt X une variable aléatoire de loi de Poisson $\mathcal{P}(\theta)$ avec $\theta \in I\!R_*^+$, et $x_1, \ldots, x_n$ un échantillon de cette loi.

1. Déterminer la mesure *a priori* de Jeffreys $\pi^J(\theta)$.
2. Évaluer, à partir de l'existence des lois *a posteriori* si cette mesure *a priori* est préférable à la mesure *a priori* invariante par transformation d'échelle $\pi_0(\theta) \propto 1/\theta$.
3. Soit $\pi_\alpha(\theta) \propto \theta^{-\alpha}$ avec $\alpha \in I\!R^+$. Donner l'expression de la fonction de masse prédictive *a posteriori* $P_\alpha(X = k|x_1, \ldots, x_n)$ ainsi que son espérance et sa variance, et leurs conditions d'existence.

Réponses.

1. La fonction de masse (densité) de $X \sim \mathcal{P}(\theta)$ est, $\forall k \in I\!N$,

$$P(X = k|\theta) = \frac{\theta^k}{k!}\exp(-\theta)$$

et donc

$$\frac{\partial^2 \log P(X=k|\theta)}{\partial\theta^2} \quad = \quad -k/\theta^2$$

Par absolue continuité l'information de Fisher est donc $I(\theta) = -\mathbb{E}_X\left[\frac{\partial^2 \log P(X=k|\theta)}{\partial\theta^2}\right]$ et donc, avec $\mathbb{E}_X[k] = \theta$, il vient $I(\theta) = 1/\theta$ et par définition

$$\pi^J(\theta) \quad \propto \quad \theta^{-1/2}.$$

2. On a alors

$$\pi^J(\theta|x_1,\ldots,x_n) \quad \propto \quad \theta^{\sum_i x_i}\exp\left(-n\theta\right)\theta^{-1/2}$$

on reconnaît le terme général d'une loi gamma $\mathcal{G}\left(\sum_i x_i + 1/2, n\right)$ qui est propre (intégrable et donc bien définie) $\forall x \in I\!N$ et tout $n \geq 1$. Si on remplace $\pi^J(\theta)$ par $\pi_0(\theta)$, alors la loi *a posteriori* devient $\mathcal{G}\left(\sum_i x_i, n\right)$ qui peut ne pas être définie si tous les x_i valent 0 (ça serait le cas, typiquement, si $\theta \ll 1$). La mesure de Jeffreys est donc préférable.

3. On a, par définition

$$P_\alpha(X=k|x_1,\ldots,x_n) \quad = \quad \int_{I\!R_*^+} P(X=k|\theta)\pi_\alpha(\theta|x_1,\ldots,x_n)\,d\theta \tag{-57}$$

et si

$$\alpha \quad < \quad \sum_i x_i + 1, \tag{-56}$$

alors

$$\pi_\alpha(\theta|x_1,\ldots,x_n) \quad \equiv \quad \mathcal{G}\left(\sum_i x_i - \alpha + 1, n\right) \tag{-55}$$

et

$$\begin{aligned}
\mathbb{E}[X|x_1,\ldots,x_n] &= \mathbb{E}_{\pi_\alpha}\left[\mathbb{E}[X|\theta]|x_1,\ldots,x_n\right], \\
&= \mathbb{E}_{\pi_\alpha}\left[\theta|x_1,\ldots,x_n\right], \\
&= \frac{\sum_i x_i - \alpha + 1}{n}.
\end{aligned}$$

De plus

$$\begin{aligned}
\mathbb{E}[X^2|x_1,\ldots,x_n] &= \mathbb{E}_{\pi_\alpha}\left[\mathbb{E}[X^2|\theta]|x_1,\ldots,x_n\right], \\
&= \mathbb{E}_{\pi_\alpha}\left[\theta+\theta^2|x_1,\ldots,x_n\right], \\
&= \frac{\sum_i x_i - \alpha + 1}{n} + \mathbb{V}_{\pi_\alpha}\left[\theta|x_1,\ldots,x_n\right] + \left(\mathbb{E}_{\pi_\alpha}\left[\theta|x_1,\ldots,x_n\right]\right)^2, \\
&= \frac{\sum_i x_i - \alpha + 1}{n} + \frac{\sum_i x_i - \alpha + 1}{n^2} + \left(\frac{\sum_i x_i - \alpha + 1}{n}\right)^2
\end{aligned}$$

Donc

$$\begin{aligned}
\mathbb{V}[X|x_1,\ldots,x_n] &= \frac{\sum_i x_i - \alpha + 1}{n} + \frac{\sum_i x_i - \alpha + 1}{n^2}, \\
&= \frac{\sum_i x_i - \alpha + 1}{n}\left(1 + 1/n\right).
\end{aligned}$$

Enfin, en reprenant (-57) et (-55), on obtient

$$\begin{aligned}
P_\alpha(X=k|x_1,\ldots,x_n) &= \int_{I\!R_*^+} \frac{\theta^k}{k!}\exp(-\theta)\frac{n^{\sum_i x_i-\alpha+1}}{\Gamma(\sum_i x_i - \alpha + 1)}\theta^{\sum_i x_i-\alpha}\exp\left(-n\theta\right)\,d\theta, \\
&= \frac{n^{\sum_i x_i-\alpha+1}}{k!\Gamma(\sum_i x_i - \alpha + 1)}\int_{I\!R_*^+} \theta^{k+\sum_i x_i-\alpha}\exp\left(-(n+1)\theta\right)\,d\theta
\end{aligned}$$

et on reconnaît sous l'intégrale le terme général d'une loi $\mathcal{G}\left(k+\sum_i x_i - \alpha + 1, n+1\right)$, bien définie sous la condition (-56). L'intégration donne la constante de normalisation de cette densité et donc

$$P_\alpha(X=k|x_1,\ldots,x_n) \quad = \quad \frac{n^{\sum_i x_i-\alpha+1}}{k!\Gamma(\sum_i x_i - \alpha + 1)}\frac{\Gamma(k+\sum_i x_i - \alpha + 1)}{(n+1)^{k+\sum_i x_i-\alpha+1}}$$

qui peut éventuellement se simplifier en utilisant (par exemple) la relation

$$\Gamma(x+k) = \frac{(x-1)!}{(x+k-1)!}\Gamma(x).$$

ou une formule de type Stirling.

D.2 Risque d'un estimateur

Considérons une variable binomiale $X \sim \mathcal{B}(n,p)$ de probabilité $p \in [0,1]$. Soit la perte quadratique $L(\delta,p)$. On appelle *risque bayésien d'un estimateur* $\delta(x)$ la quantité $\mathbb{E}_\pi[L(\delta(x),p)|x]$, et *risque fréquentiste de* $\delta(x)$ la quantité $\mathbb{E}_X[L(\delta(x),p)]$.

1. Soit $\pi(p)$ le prior de Laplace. Définissez l'estimateur MAP (*maximum a posteriori*) $\delta_1(x)$ de p.
2. En choisissant plutôt $\pi(p)$ comme le prior de Jeffreys, calculez les risques bayésien et fréquentiste $R_b(x)$ et $R_f(p)$ de $\delta_1(x)$.
3. Comparez $r_f = \sup_p R_f(p)$ à $r_b = \sup_x R_b(x)$.

Réponses.

1. L'estimateur MAP pour le prior de Laplace (loi uniforme) est le mode de la distribution *a posteriori*

$$p|x \sim \mathcal{B}_e(1+x, n-x+1)$$

c'est-à-dire

$$\delta_1(x) = x/n.$$

2. On a

$$\begin{aligned} R_b(x) &= \mathbb{E}_\pi\left[(\delta_1(x)-p)^2|x\right] = \mathbb{E}_\pi\left[(x/n-p)^2|x\right], \\ &= \left(\frac{x+1/2}{n+1} - \frac{x}{n}\right)^2 + \frac{(x+1/2)(n-x+1/2)}{(n+1)^2(n+2)}, \\ &= \frac{(x-n/2)^2}{(n+1)^2n^2} + \frac{(x+1/2)(n-x+1/2)}{(n+1)^2(n+2)} \end{aligned}$$

car $\pi(p)$ est la loi $\mathcal{B}_e(1/2,1/2)$ (loi de Jeffreys) et donc

$$p|x \sim \mathcal{B}_e(1/2+x, n-x+1/2).$$

De plus,

$$\begin{aligned} R_f(p) &= \mathbb{E}_X\left[(\delta_1(x)-p)^2\right], \\ &= \mathbb{V}\left[x/n\right], \\ &= \frac{p(1-p)}{n}. \end{aligned}$$

3. Il est aisé de voir que $r_f = (4n)^{-1}$ et

$$r_b = \left\{4(n+2)\right\}^{-1}.$$

Ainsi, $r_b < r_f$.

D.3 Maximisation d'entropie

On définit une nouvelle méthodologie de construction de prior de la façon suivante. Étant donné un modèle d'échantillonnage $X|\theta \sim p(x|\theta)$, avec $x \in S$ et $\theta \in \Theta \in I\!R^d$, et un prior de référence $\pi^J(\theta)$, on définit

$$\pi^*(\theta) \quad = \quad \arg \max_{\pi(\theta) \geq 0} G(\Theta) \tag{-73}$$

où $G(\Theta)$ est l'information moyenne apportée par la densité p relativement à celle apportée par un prior $\pi(\theta)$:

$$G(\Theta) \quad = \quad \mathbb{E}_\theta \left[H^J(\Theta) - H(X|\theta) \right],$$

où $H(X|\theta)$ et $H^J(\Theta)$ sont respectivement l'entropie (relative à une mesure de Lebesgue) du modèle d'échantillonnage et l'entropie (relative à $\pi^J(\theta)$) du prior $\pi(\theta)$.

1. Prouvez que si $Y \sim f(y)$ sur un espace normé et mesuré $\Omega \in I\!R^q$ avec $q < \infty$ et $f \in L^2(\Omega)$, alors l'entropie relative à la mesure de Lebesgue de f est bornée.

2. Prouvez que le problème (-73), en imposant la contrainte que $p(x|\theta)$ et $\pi(\theta)$ soient respectivement L^4 ($\forall \theta \in \Theta$) sur S et sur Θ, implique que $\pi(\theta)$ est solution du problème de maximum d'entropie de $\pi(\theta)$ sous une contrainte linéaire

$$\int_\Theta Z(\theta)\pi(\theta)\, d\theta \quad = \quad c \, < \, \infty \tag{-73}$$

   où $Z(\theta)$ est l'information de Shannon (ou entropie différentielle négative) de p

$$Z(\theta) \quad = \quad \int_S p(x|\theta) \log p(x|\theta)\, dx$$

   et c prend une valeur maximale (mais finie).

3. Pour $S = I\!R^+$ et $(\beta, \eta) \in I\!R_*^+ \times I\!R_*^+$, considérons maintenant la loi de fonction de répartition de Weibull

$$P(X < x|\theta) \quad = \quad 1 - \exp\left(- \left\{ \frac{x}{\eta} \right\}^\beta \right).$$

   (a) Calculez $Z(\eta, \beta)$ pour ce modèle.

   (b) En utilisant le prior de Berger-Bernardo $\pi^J(\eta, \beta) \propto (\eta, \beta)^{-1}$ comme mesure de référence, donnez la solution formelle $\pi^*(\eta, \beta)$ du problème de maximisation d'entropie relative sous les contraintes (-73) et

$$\int_S x m_\pi(x)\, dx \quad = \quad x_e \tag{-74}$$

   où $m_\pi(x)$ est la loi *a priori* prédictive.

   (c) Placez les résultats sous la forme hiérarchique

$$\pi^*(\theta) \quad = \quad \pi^*(\eta|\beta)\pi^*(\beta).$$

   et prouvez que la loi *a priori* sur β peut s'écrire

$$\pi^*(\beta) \quad \propto \quad \tilde{\pi}^*(\beta)$$

   avec

$$\tilde{\pi}^*(\beta) \quad = \quad \frac{\beta^{-\lambda_1 - 1} \exp\left(-\lambda_1 \frac{\gamma}{\beta} \right)}{\Gamma^{\lambda_1}(1 + 1/\beta)} \tag{-75}$$

   où λ_1 est un multiplicateur de Lagrange.

(d) En plaçant des contraintes sur les multiplicateurs de Lagrange issus de l'écriture générale de $\pi^*(\eta, \beta)$, reconnaissez-vous une forme spécifique (connue) pour $\pi^*(\eta|\beta)$ et $\pi^*(\beta)$? La loi $\pi^*(\eta|\beta)$ est-elle conjuguée conditionnellement à β ?

(e) Cette loi jointe $\pi^*(\theta)$ est-elle propre (intégrale) ? Sous quelle(s) condition(s) sur les multiplicateurs de Lagrange ?

(f) Reliez formellement les multiplicateurs de Lagrange à x_e en vérifiant l'équation (-74). Doit-on connaître la constante d'intégration de $\pi^*(\beta)$ pour ce faire ?

(g) Proposez, codez et validez une méthode numérique permettant de simuler des tirages de β selon $\pi^*(\theta)$ (formule (-75)), en fixant $\lambda_1 = 1$. Pour la validation, utilisez plutôt la représentation de la variable $Y = 1/beta$ en opérant un changement de variable.

Indications.

- *Il peut être utile de prouver au préalable à la question 1 que* $\log y \leq 1 + y \; \forall y \in I\!R_*^+$
- *On rappelle que l'espérance de la loi de Weibull est*

$$\mathbb{E}[X|\theta] \quad = \quad \eta\Gamma(1 + 1/\beta) \tag{-74}$$

et que lorsque $\beta > 0$

$$\Gamma(1 + 1/\beta) \quad \geq \quad \frac{\sqrt{\pi}}{3} \tag{-73}$$

- *On rappelle les formules suivantes :*

$$\int_0^\infty (\log x) \exp(-x) \, dx \quad = \quad -\gamma \tag{-72}$$

$$\int_0^\infty x \exp(-x) \, dx \quad = \quad \Gamma(2) \tag{-71}$$

où γ *est la constante d'Euler (que vous pouvez prendre égale à 0.5772157)*

Réponses.

1. L'entropie relative à la mesure de Lebesgue sur Ω est $H(f) = -\mathbb{E}_f[\log f]$ et, via l'inégalité de Jensen ($-\log$ étant convexe),

$$-\mathbb{E}_f[\log f] \quad \leq \quad -\log \mathbb{E}_f[f] \; = \; -\int_\Omega f^2(y) \, dy$$

et $\int_\Omega f^2(y) \, dy < \infty$ puisque $f \in L^2(\Omega)$. De plus, il est aisé de montrer que $\|ogy \leq 1 + y \; \forall y \in I\!R_*^+$. On a donc que

$$-\mathbb{E}_f[\log f] \quad \geq \quad -1 - \int_\Omega f^2(y) \, dy.$$

Donc $H(f)$ est bornée.

2. La définition (-73) est celle proposée par Arnold Zellner pour définir la classe des *Maximal Data Information* (MDI) *Priors*, qui constitue une alternative souvent intéressante aux *reference priors* de Berger-Bernardo. On voit facilement que (modulo l'existence des intégrales ci-dessous)

$$\pi^*(\theta) \quad = \quad \arg \max_{\pi(\theta) \geq 0} \int_\Theta Z(\theta)\pi(\theta) \, d\theta - \int_\Theta \pi(\theta) \log \frac{\pi(\theta)}{\pi^J(\theta)}.$$

Les deux problèmes de maximisation ont le même lagrangien si l'intégrale $\int_\Theta Z(\theta)\pi(\theta)\,d\theta$ est finie. Or on a

$$\mathbb{E}_\pi[Z] \;=\; \int_\Theta Z(\theta)\pi(\theta)\,d\theta \;=\; H(\Theta) - H(X,\Theta)$$

où $H(\Theta)$ est l'entropie (non relative) de $\pi(\theta)$ et $H(X,\Theta)$ est l'entropie (non relative) de la loi jointe de (X,θ). Si $\pi(\theta)$ est L^4 sur Θ, alors elle est aussi L^2 sur Θ et $H(\Theta)$ est bornée d'après la question 1. Il suffit alors de montrer que $-H(X,\Theta)$ est fini. On a (en utilisant les résultats précédents)

$$\begin{aligned}
-H(X,\Theta) &= \int_{S\times\Omega} p(x|\theta)\pi(\theta)\log\{p(x|\theta)\pi(\theta)\}\;dxd\theta \\
&\le 1 + \int_{S\times\Omega} \left(p(x|\theta)\pi(\theta)\right)^2\;dxd\theta \quad \text{par Jensen,} \\
&\le 1 + \sqrt{\int_S p^4(x|\theta)\,dx}\sqrt{\int_\Theta \pi^4(\theta)\,d\theta}
\end{aligned}$$

d'après l'inégalité de Cauchy-Schwarz. Le terme de droite étant alors fini d'après les hypothèses, on a donc que $-H(X,\Theta)$ est fini et dont $\mathbb{E}_\pi[Z]$ est fini. Donc $\exists c<\infty$ tel que

$$\int_\Theta Z(\theta)\pi(\theta)\,d\theta \;=\; c.$$

3. Rappelons que la densité correspondante de Weibull s'écrit

$$f(x|\theta) \;=\; \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1}\exp\left(-\left\{\frac{x}{\eta}\right\}^\beta\right).$$

(a) On a, après un simple développement,

$$Z(\eta,\beta) \;=\; \log\frac{\beta}{\eta^\beta} + (\beta-1)\mathbb{E}_X\left[\log X\right] - \mathbb{E}_X\left[\left(\frac{X}{\eta}\right)^\beta\right].$$

En utilisant la transformation $u=(x/\eta)^\beta$, avec $du/dx = \beta x^{\beta-1}/\eta^\beta$, il vient

$$\begin{aligned}
\mathbb{E}_X\left[\log X\right] &= \log(\eta)\int_0^\infty \exp(-u)\,du + \frac{1}{\beta}\int_0^\infty (\log u)\exp(-u)\,du, \\
&= \log(\eta) - \gamma/\beta \quad \text{en utilisant (-72),}
\end{aligned}$$

puis

$$\begin{aligned}
\mathbb{E}_X\left[\left(\frac{X}{\eta}\right)^\beta\right] &= \int_0^\infty u\exp(-u)\,du \\
&= \Gamma(2) \;=\; 1 \quad \text{en utilisant (-71).}
\end{aligned}$$

On en déduit que

$$Z(\eta,\beta) \;=\; \log\beta - \log\eta + \gamma/\beta - (1+\gamma).$$

(b) En notant que $\theta = (\eta,\beta)$, rappelons que l'on peut écrire la contrainte sous forme linéaire (par rapport à $\pi(\theta)$)

$$\int_S x m_\pi(x)\,dx \;=\; x_e \;=\; \int_\Theta \mathbb{E}[X|\theta]\pi(\theta)\,d\theta \qquad \text{(-85)}$$

avec $\mathbb{E}_\theta[X] = \eta\Gamma(1+1/\beta)$ d'après (-74).

Alors la solution du problème de maximisation d'entropie s'écrit, en introduisant $(\lambda_1, \lambda_2) \in I\!R^2$ des multiplicateurs de Lagrange,

$$\begin{aligned}\pi^*(\theta) &\propto \pi^J(\theta) \exp\left(-\lambda_1 Z(\theta) - \lambda_2 \mathbb{E}[X|\theta]\right), \\ &\propto \beta^{-\lambda_1-1}\eta^{\lambda_1-1} \exp\left(-\lambda_2 \eta \Gamma(1+1/\beta)\right) \exp\left(-\lambda_1 \frac{\gamma}{\beta}\right)\end{aligned}$$

4. Si on impose $(\lambda_1, \lambda_2) \in I\!R^+ \times I\!R^+$, on peut alors écrire

$$\pi^*(\theta) = \pi^*(\eta|\beta)\pi^*(\beta)$$

avec

$$\begin{aligned}\eta|\beta &\sim \mathcal{G}\left(\lambda_1, \lambda_2\Gamma(1+1/\beta)\right), \\ \pi^*(\beta) &\propto \underbrace{\frac{\beta^{-\lambda_1-1}\exp\left(-\lambda_1\frac{\gamma}{\beta}\right)}{\Gamma^{\lambda_1}(1+1/\beta)}}_{\tilde{\pi}^*(\beta)}\end{aligned}$$

où $\mathcal{G}$ désigne une loi gamma
(le terme en dénominateur de $\pi^*(\beta)$ correspondant à la constante d'intégration (à un coefficient près) de $\pi^*(\eta|\beta)$)

5. On remarque que la loi $\pi^*(\eta|\beta)$ n'est pas conjuguée conditionnellement à β (il faudrait que ce soit une inverse gamma, et non une gamma).

6. Pour que la loi jointe soit propre, sachant $(\lambda_0, \lambda_1) \in I\!R_*^+ \times I\!R_*^+$, il suffit donc de montrer que $\pi^*(\beta)$ est intégrable. Or, pour tout $\beta > 0$, d'après (-73) on a $\Gamma(1+1/\beta) \geq \sqrt{\pi}/3$. Donc,

$$0 \leq \tilde{\pi}^*(\beta) \leq \left(\frac{3}{\sqrt{\pi}}\right)^{\lambda_1} \beta^{-\lambda_1-1}\exp\left(-\lambda_1\frac{\gamma}{\beta}\right)$$

qui est clairement intégrable. Le prior est donc propre sous les conditions $(\lambda_1, \lambda_2) \in I\!R^+ \times I\!R^+$.

7. On note $A(\lambda_1)$ la constante d'intégration de $\pi^*(\beta)$, telle que

$$\pi^*(\beta) = A^{-1}(\lambda_1)\tilde{\pi}^*(\beta)$$

Vérifions l'équation (-74) en utilisant l'expression (-85) :

$$\begin{aligned}x_e = \int_\Theta \mathbb{E}[X|\theta]\pi(\theta)\, d\theta &= A^{-1}(\lambda_1)\int_{I\!R^+} \Gamma(1+1/\beta)\tilde{\pi}^*(\beta)\mathbb{E}[\eta|\beta]\, d\beta, \\ &= A^{-1}(\lambda_1)\int_{I\!R^+} \tilde{\pi}^*(\beta)\frac{\lambda_1}{\lambda_2\Gamma(1+1/\beta)}\, d\beta, \\ &= \frac{\lambda_1}{\lambda_2}.\end{aligned}$$

Ce résultat est indépendant de la constante d'intégration de $\pi^*(\beta)$.

8. Dans ce cas unidimensionnel, l'idée la plus simple consiste à utiliser une **méthode d'acceptation-rejet**, qui permettra en outre de calculer numériquement la constante d'intégration $A(\lambda_1)$. Pour ce faire, la forme du terme général $\tilde{\pi}^*(\beta)$, proche d'une inverse gamma, peut nous inspirer. Si on choisit comme loi instrumentale la densité

$$g(\beta) \equiv \mathcal{IG}(\lambda_1, \lambda_1/\gamma),$$

alors il vient

$$\begin{aligned}\frac{\tilde{\pi}^*(\beta)}{g(\beta)} &= \frac{\Gamma^{\lambda_1/\gamma}(\lambda_1)}{\Gamma^{\lambda_1}(1+1/\beta)}\frac{1}{(\lambda_1/\gamma)^{\lambda_1}},\\ &\leq \left(\frac{3}{\sqrt{\pi}}\right)^{\lambda_1}\frac{\Gamma^{\lambda_1/\gamma}(\lambda_1)}{(\lambda_1/\gamma)^{\lambda_1}},\end{aligned}$$

d'après (-73), borne supérieure qui ne dépend plus de β. En utilisant la valeur $\lambda_1 = \gamma$, on obtient

$$\frac{\tilde{\pi}^*(\beta)}{g(\beta)} \leq K = \left(\frac{3}{\sqrt{\pi}}\right)^{\gamma}\Gamma(\gamma) \simeq 2.092$$

Le programme attendu (exemple en R fourni sur le fichier `AR.r`) doit donc mettre en œuvre le pseudo-code suivant :

(a) Simuler $\beta \sim g(\beta)$.

(b) Simuler $U \sim \mathcal{U}[0,1]$.

(c) Accepter β si $U \leq \frac{\tilde{\pi}^*(\beta)}{Kg(\beta)}$

et en notant p la proportion d'acceptation dans cet algorithme, on peut estimer $A(\lambda_1)$ par

$$\hat{A}(\lambda_1) = 1/(Kp)$$

puis représenter la densité $\pi^*(\beta)$ estimée par $\hat{A}^{-1}(\lambda_1)\tilde{\pi}^*(\beta)$ et la comparer avec l'histogramme des simulations acceptées. Il est plus simple visuellement de représenter plutôt la densité de la variable $Y = 1/\beta$, telle que $du = -u^2 d\beta$. La formule de changement de variable donne :

$$\pi^*_Y(Y) = \pi^*(\beta^{-1}(Y)))/Y^2$$

avec $\beta^{-1}(Y) = 1/Y$.

D.4 Calcul bayésien

On reprend la loi de Weibull $\mathcal{W}(\eta, \beta)$ de l'exercice (D.3) et on impose le prior suivant

$$\begin{aligned}\eta &\sim \mathcal{G}(m, m/\eta_0)\\ \beta &\sim \tilde{\pi}^*(\beta) = \frac{\beta^{-\lambda_1-1}\exp\left(-\lambda_1\frac{\gamma}{\beta}\right)}{\Gamma^{\lambda_1}(1+1/\beta)}\end{aligned}$$

qui est le prior (-75) sur β.

Soit l'échantillon

$$\mathbf{x_n} = \{103, 157, 39, 145, 24, 22, 122, 126, 66, 97\},$$

1. Proposez et implémentez une méthode permettant de générer des tirages *a posteriori* de (η, β).
2. Estimez numériquement l'espérance de la loi *a posteriori prédictive*

$$p(x|\mathbf{x_n}) = \iint_{I\!R^+\times I\!R^+} p(x|\eta,\beta)\pi(\eta,\beta|\mathbf{x_n})\, d\eta d\beta,,$$

en prenant $m = 2$, $\eta_0 = 100$ et $\lambda_1 = 1$

Réponse.

1. Le prior n'étant pas conjugué pour aucune des deux dimensions, il est naturel de proposer un algorithme de Gibbs dont les deux simulations (de η puis β) sont menées par des échantillonnages de Metropolis-Hastings. Cela d'autant plus qu'on n'a pas besoin de connaître la constante d'intégration de $\pi(\beta)$ (et donc d'avoir résolu complètement l'exercice (D.3). Le code-solution produit sur le fichier `calcul-bayesien.r` utilise deux marches aléatoires comme lois instrumentales.

2. On obtient numériquement, en simulant par Monte Carlo un tirage $(\eta_i, \beta_i) \sim \pi(\eta, \beta|\mathbf{x_n})$ puis en calculant

$$\begin{aligned}\mathbb{E}\left[X|\mathbf{x_n}\right] &= \mathbb{E}_\pi\left[\eta\gamma(1+1/\beta)|\mathbf{x_n}\right] \\ &\simeq \frac{1}{M}\eta_i\gamma(1+1/\beta_i), \\ &\simeq 357\end{aligned}$$

D.5 Bonus

Soit $f(x|\theta) = h(x)\exp\left(\theta\cdot x - \psi(\theta)\right)$ une distribution d'une famille exponentielle, avec $\theta \in \Theta$. Pour toute loi *a priori* π, prouvez que la moyenne *a posteriori* de θ est donnée par

$$\mathbb{E}[\theta|x] = \nabla \log m_\pi(x) - \nabla \log h(x)$$

où ∇ est l'opérateur gradient et m_π est la loi marginale *a priori* associée à x.

Réponse. L'espérance *a posteriori* de θ vaut

$$\begin{aligned}\mathbb{E}[\theta|x] &= \frac{\int_\Theta \theta h(x)\exp\left(\theta\cdot x - \psi(\theta)\right)\pi(\theta)\,d\theta}{m_\pi(x)}, \\ &= \left(\frac{\partial}{\partial x}\int_\Theta h(x)\exp\left(\theta\cdot x - \psi(\theta)\right)\pi(\theta)\,d\theta\right)\frac{1}{m_\pi(x)} - \left(\frac{\partial}{\partial x}h(x)\right)\frac{1}{h(x)}, \\ &= \frac{\partial}{\partial x}\left(\log m_\pi(x) - \log h(x)\right).\end{aligned}$$

Références

[1] H. Akaike. On entropy maximization principle. *In : Krishnaiah, P.R. (Editor). Applications of Statistics, North-Holland, Amsterdam*, pages 27–41, 1977.

[2] Anonyme. *Measuring River Eischarge in High Flow (Flood) or High Sediment Concentration Conditions*. Application Note : R&E Instruments – Acoustic Eoppler Current Profilers. Communication Technology Technical Report, 1999.

[3] D.J. Benjamin, J.O. Berger, and V.E. Johnson. Redefine statistical significance. *Nature Human Behavior*, pages DOI :10.1038/s41562–017–0189–z, 2017.

[4] N. Bouleau. *Probabilités de l'ingénieur*. Hermann, 1986.

[5] E. Cœur and M. Lang. L'information historique des inondations : l'histoire ne donne-t-elle que des leçons ? *La Houille Blanche*, 2 :79–84, 2000.

[6] M. Evans. Measuring statistical evidence using relative belief. *Computational Structural Biotechnology Journal*, 14 :91–96, 2016.

[7] R.A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1926.

[8] M. Galevski. La corrélation entre les pluies torrentielles and l'intensité de l'érosion (avant-propos de P. Reneuve). *Annales de l'École Nationale des Eaux and Foêts and de la station de recherches and expériences*, 14 :379–428, 1955.

[9] C. Gourerioux and A. Monfort. *Statistique et modèles économétriques*. Economica, Paris, 1996.

[10] S. Greenland, S.J. Senn, K.J. Rothman, J.B. Carlin, C. Poole, S.N. Goodman, and D. Altman. Statistical tests, p values, confidence interval, and power : a guide to misinterpretations. *European Journal of Epidemiology*, 31 :227–350, 2016.

[11] V.E. Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Science*, 110 :19313–19317, 2013.

[12] M. Keller, A. Pasanisi, and E. Parent. Réflexions sur l'analyse d'incertitudes dans un contexte industriel : information disponible et enjeux décisionnels. *Journal de la Société Française de Statistiques*, 2012.

[13] M.Y. Kim and X. Xue. The analysis of multivariate interval-censored survival data. *Statistics in Medicine*, 21 :3715–3726, 2002.

[14] A.N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea Publishing Co., Oxford, 1950.

[15] J.F. Le Gall. *Intégration, Probabilités and Processus Aléatoires*. Cours de l'École Normale Supérieure, 2006.

[16] H. Lebesgue. *Oeuvres scientifiques (en cinq volumes)*. Institut de Mathématiques de l'Université de Genève, 1972.

[17] E.L. Lehman. *Fisher, Neyman, and the creation of classical statistics*. New York : Springer, 2011.

[18] N.R. Mann, R.E. Schafer, and N.D. Singpurwalla. *Methods for Statistical Analysis of Reliability and Life Data*. Wiley Series in Probability and Statistics, 1974.

[19] J.-M. Marin and C.P. Robert. *Bayesian Core : A Practical Approach to Computational Bayesian Statistics*. Springer, 2007.

[20] C. Neves and M. Isabel Fraga Alves. Testing extreme value conditions – an overview and recent approaches. *REVSTAT*, 6 :83–100, 2008.

[21] R. Nuzzo. Scientific method : Statistical errors. *Nature*, 506 :150–152, 2014.

[22] G.W. Oehlert. A Note on the Delta Method. *The American Statistician*, 46 :27–29, 1992.

[23] E. Parent and J. Bernier. *Le raisonnement bayésien. Modélisation and inférence*. Springer, 2007.

[24] O. Payrastre. Utilité de l'information historique pour l'étude du risque de crues. *14ième Journées Scientifiques de l'Environnement : l'Eau, la Ville, la Vie, 12-13 mai*, 2003.

[25] J. Planzalg and R. Hamböker. *Parametric statistical theory*. Walter de Gruyter, Berlin, 1994.

[26] D.S. Reis and J.R. Stedinger. Bayesian mcmc flood frequency analysis with historical information. *Journal of Hydrology*, 313 :97–116, 2005.

[27] C.P. Robert. *The Bayesian Choice : From Decision-Theoretic Foundations to Computational Implementation (2nd edition)*. Springer, 2007.

[28] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods (second edition)*. Springer, 2004.

[29] A. Sabourin. Semi-parametric modeling of excesses above high multivariate thresholds with censored data. *Journal of Multivariate Analysis*, 136 :126–146, 2015.

[30] G. Saporta. *Probabilités, analyses des données and statistiques*. Technip, 2006.

[31] Gideon E. Schwarz, H. Estimating the dimension of a model. *Annals of Statistics*, 6 :461–464, 1978.

[32] J. Sprenger. *Bayésianisme versus fréquentisme en inférence statistique*. I. Drouet (ed.). Éditions Matériologiques, Paris, 2017.

[33] M.A. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69 :730–737, 1974.

[34] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Martens, M.G. Tadesse, M. Vannuci, A. Gelman, D. Veen, J. Willemsen, and C. Yau. Bayesian statistics and modelling. *Nature Reviews. Methods Primer*, 2021.

[35] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[36] Cochran W.G. The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, 23 :315–345, 1952.

[37] W. Xie and Barton R.B. Nelson, B.L. Multivariate input uncertainty in output analysis for stochastic simulation. *soumis*, 2016.