Modélisation et statistique bayésienne computationnelle

Notes de cours

nicolas.bousquet@sorbonne-universite.fr

20 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

NOTATIONS ET FONCTIONS DE RÉPARTITION DE LOIS STATISTIQUES

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]

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

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 :

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

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 indifférement de la variable aléatoire

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

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 :

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

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

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

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 $\mathbf{x_n} = (x_1, \dots, x_n)$. 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 $x_1, \dots, x_n$ conditionnellement à θ.

Définition 2 (Échangeabilité.) *Les données $x_1, \dots, x_n$ sont dites échangeables si, pour toute permutation $\sigma : I\!N^n \to I\!N^n$, la loi jointe $f(x_{\sigma(1)}, \dots, x_{\sigma(n)})$ 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 -25)

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

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 :

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

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 ?

   Réponse. La vraisemblance dans les deux cas est

   $$\ell(\theta|x_1, x_2) \quad \propto \quad \exp\left\{-(\bar{x} - \theta)^2\right\}$$

   et qui devrait donc conduire à la même inférence sur θ. Mais $g(x_1, x_2|\theta)$ est très différente de la première distribution (par exemple, l'espérance de $x_1 - x_2$ n'est pas définie). Les estimateurs de θ auront donc des propriétés fréquentistes différentes s'ils ne dépendent pas que de $\bar{x}$ (**ex** : estimateur des moments). En particulier, les régions de confiance pour θ peuvent différer fortement car g possède des queues plus épaisses.

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.

Une dernière difficulté théorique posée par les estimateurs fréquentiels apparaît lorsqu'on cherche à mener une *prévision*. Considérons en effet Soit $\mathbf{X_n} = (X_1, \ldots, X_n) \stackrel{iid}{\sim} f(.|\theta)$. 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.

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 vériable 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*

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

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 ?*

Réponse. Notons $\mathbb{P}(Y_2|\theta)$ la mesure de probabilité associée à Y_2. On définit la gauche par l'événement $0 \leq Y_2 < \theta \leq 1$. Alors

$$\mathbb{P}(Y_2 < \theta|\theta) \quad = \quad \theta$$

et l'indicatrice $\mathbb{1}_{Y_2<\theta}$ suit alors une loi de Bernoulli $\mathcal{B}(\theta)$. Alors après n répétitions iid, on a

$$X \quad = \quad \sum_{i=1}^n \mathbb{1}_{Y_{2,i}<\theta} \ \sim \ \mathcal{B}(n,\theta) \quad \text{(loi binomiale).}$$

Si on connaît une réalisation x de X, la vraisemblance statistique associée est donc

$$\mathbb{P}(X=x|\theta) \quad = \quad \binom{n}{x} \theta^x(1-\theta)^{n-x}.$$

Comme $\theta \in [0,1]$ et qu'on ne dispose pas d'information *a priori* sur θ, on peut simplement supposer *a priori*

$$\pi(\theta) \quad = \mathbb{1}_{\{0\leq\theta\leq 1\}}(\theta).$$

On en déduit la loi *a posteriori* (en utilisant le symbole de proportionalité $\propto$)

$$\pi(\theta|X=x) \quad \propto \quad \theta^x(1-\theta)^{n-x}\mathbb{1}_{\{0\leq\theta\leq 1\}}(\theta)$$

qui correspond au terme général de la loi bêta $\mathcal{B}_e(1+x, 1+n-x)$ d'espérance $(1+x)/(2+n) \sim x/n$ si $(x,n) \gg 1$.

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 $I\!R$" 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 $m_\pi(x)$?*

Réponse. On a

$$m_\pi(x) \quad = \quad \sigma\sqrt{2\pi} \; < \; \infty$$

donc la mesure de Lebesgue sur $I\!R$ peut être utilisée.

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

Réponse. L'intégrale de la loi *posteriori* s'écrit

$$\begin{aligned} m_\pi(x_1, \ldots, x_n) \quad &= \quad \int_{I\!R}\int_0^\infty \exp\left(-\frac{(\bar{x_n}-\mu)^2}{2\sigma^2}\right) \exp\left(\frac{\sum_{i=1}^n (x_i - \bar{x_n})^2}{2\sigma^2}\right) \frac{d\mu d\sigma}{\sigma^{n+1}} \\ &= \quad \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\sum_{i=1}^n (x_i - \bar{x_n})^2}{2\sigma^2}\right) \frac{d\sigma}{\sigma^n}. \end{aligned}$$

Pour que l'expression converge, il faut avoir $n > 1$ et la propriété suivante vérifiée :

$$\exists (i,j) \in I\!N^*, \; i \neq j \;\text{ tel que }\; X_i \neq X_j.$$

Lorsque $n > 1$ l'ensemble des vecteurs qui ne vérifient pas cette propriété est de mesure nulle et donc n'affecte pas la finitude de l'intégrale définie ci-dessus. Il est possible de donner une interprétation intuitive du résultat ci-dessus : pour estimer la dispersion (variance), au moins deux observations non égales sont nécessaires.

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.

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 **??**). 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 $X \sim f(x|\theta)$

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

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 **??** 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 **??** 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 5.

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 $g : I\!R^{d_1} \to I\!R^{d_2}$ (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).

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

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

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

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* $\pi(\theta) \propto \exp(-2\lambda\|\theta\|_2^2)$.

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

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

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*

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

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 ?

Réponse. Calculons le risque fréquentiste pour δ_1. On obtient :

$$\begin{aligned} R(\theta, \delta_1) &= \mathbb{E}_\theta \left[L(\theta, \delta_1(X)) \right], \\ &= \int_\Omega L(\theta, \delta_1(x)) f(x|\theta) \, dx. \end{aligned}$$

La fonction $L(\theta, \delta_1(x))$ vaut 0 si $\delta = \theta$ (bonne décision) et 1 sinon. On en déduit que

$$\begin{aligned} R(\theta, \delta_1) &= \int_{\{\theta-1, \theta+1\}} (1 - \mathbb{1}_\theta((x_1 + x_2)/2) \, f(x_1, x_2|\theta) \, dx_1 dx_2, \\ &= \frac{1}{4} \left\{ (1 - \mathbb{1}_\theta((\theta - 1 + \theta - 1)/2)) + (1 - \mathbb{1}_\theta((\theta + 1 + \theta - 1)/2)) + \right. \\ &\quad \left. (1 - \mathbb{1}_\theta((\theta + 1 + \theta + 1)/2)) + (1 - \mathbb{1}_\theta((\theta - 1 + \theta + 1)/2)) \right\}, \\ &= \frac{1}{4}(1 + 1) \; = \; 1/2. \end{aligned}$$

(l'estimateur est correct la moitié du temps). On trouve alors, similairement,

$$R(\theta, \delta_1) \;=\; R(\theta, \delta_2) \;=\; R(\theta, \delta_3) \;=\; 1/2$$

ce qui signifie qu'on ne peut pas classer les estimateurs sous le coût fréquentiste 0-1.

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 $x \in \Omega$,*

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

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.

Preuve. Nous prouvons ici qu'un estimateur minimisant le risque intégré R_B est obtenu par sélection, pour chaque valeur $x \in \Omega$, de la valeur $\delta(x)$ qui minimise le coût moyen *a posteriori*. Il s'agit en pratique d'une méthode de calcul d'un estimateur bayésien. En effet, L'application du théorème de Fubini est possible par la finitude des intégrales impliquées ci-dessous : avec $L(\theta, \delta(x)) \geq 0$, il vient

$$\begin{aligned}
R_B(\delta|\pi) &= \int_\Theta \int_\Omega L(\theta, \delta(x)) f(x|\theta)\, dx \pi(\theta)\, d\theta, \\
&= \int_\Omega \int_\Theta L(\theta, \delta(x)) f(x|\theta)\pi(\theta)\, d\theta dx, \\
&= \int_\Omega \int_\Theta L(\theta, \delta(x))\pi(\theta|x)\, d\theta m_\pi(x)\, dx, \\
&= \int_\Omega R_P(\delta|\pi, x) m_\pi(x)\, dx.
\end{aligned}$$

On peut en déduire que pour tout $\delta \in \mathcal{D}$, $R_P\left(\delta^\pi(x)|\pi, x\right) \leq R_P\left(\delta|\pi, x\right)$implique $R_B\left(\delta^\pi|\pi\right) \leq R_B\left(\delta|\pi\right)$ce qui permet de conclure la démonstration du corollaire.

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.*

Preuve. Soit un estimateur bayésien δ^π de risque R fini. Pour $\delta_0 \in \mathcal{D}$ tel que $\forall\theta \in \Theta$, $R(\theta, \delta) \leq R(\theta, \delta_0)$, on note

$$\mathcal{A}_0 \quad = \quad \{\theta \in \Theta, \; R(\theta, \delta) \leq R(\theta, \delta_0)\} \, .$$

On a alors

$$\int_\Theta R(\theta, \delta_0) d\Pi(\theta) - \int_\Theta R(\theta, \delta^\pi) d\Pi(\theta) \quad = \quad \int_{\mathcal{A}_0} \left(R(\theta, \delta) - R(\theta, \delta_0)\right) d\Pi(\theta) \; \leq \; 0$$

avec égalité si et seulement si $\pi(\mathcal{A}_0) = 0$. Or, comme δ^π est bayésien et le risque fini, $R(\theta, \delta_0) \geq R(\theta, \delta^\pi)$. Donc l'intégrale ci-dessus est négative et positive, donc nulle, ce qui sous-entend qu'en effet $\pi(\mathcal{A}_0) = 0$ (on dit alors que δ^π est $\pi-$admissible). Supposons cependant que δ^π n'est pas admissible. On déduit de la démarche précédente que $\exists\delta_0$ tel que $\forall\theta$ tel que $R(\theta, \delta_0) \leq R(\theta, \delta^\pi)$ et $\theta_0 \in \Theta$ tel que $R(\theta, \delta_0) < R(\theta, \delta^\pi)$. La fonction définie sur Θ par $\theta \to R(\theta, \delta_0) - R(\theta, \delta^\pi)$ est continue par hypothèse. Donc il existe un voisinage ouvert $V_0 \subset \Theta$ de θ_0 tel que $\forall\theta \in V_0$, $R(\theta, \delta_0) < R(\theta, \delta^\pi)$. On a $\pi(\mathcal{A}_0) \geq \pi(V_0)$. Or π est supposé strictement positive sur Θ, donc $\pi(V_0) > 0$. L'ensemble $\mathcal{A}_0$ est donc de mesure non nulle, ce qui contredit la première partie de la démonstration. En conclusion, 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.*

Preuve.
Supposons que δ^π est non admissible. Alors, d'après la Définition 12, $\exists\delta_0 \in \mathcal{D}$ tel que $\forall\theta \in \Theta$, $R(\theta, \delta^\pi) \geq R(\theta, \delta_0)$, et $\exists\theta_0 \in \Theta$ tel que $R(\theta_0, \delta^\pi) > R(\theta_0, \delta_0)$. En intégrant la première inégalité, il vient :

$$\int_\Theta R(\theta, \delta_0) \, d\Pi(\theta) \quad \leq \quad \int_\Theta R(\theta, \delta^\pi) \, d\Pi(\theta) \; = \; R_B(\delta|\pi)$$

donc δ_0 est aussi un estimateur bayésien associé à L et π, et $\delta_0 \neq \delta^\pi$ d'après la seconde inégalité. Ce qui contredit à l'hypothèse du théorème. Par contraposée, on en déduit le résultat de ce thèorème. Remarquons que l'unicité de l'estimateur impliqué la finitude du risque :

$$\int_\Theta R(\theta, \delta^\pi) \, d\Pi(\theta) \quad < \quad \infty$$

sinon tout estimateur minimise le risque.

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

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

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*

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

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* $\pi(\theta|\mathbf{x_n})$

Preuve. On a $\mathcal{D} = \Theta \in I\!R^d$ (ou plus généralement un espace de Hilbert) et $L(\theta,\delta) = \|\theta-\delta\|^2$ (norme euclidienne au carré). Par simplificité travaillons sur $\Theta = I\!R$ (d=1). Alors

$$\begin{aligned} R_B\left(\delta|\pi\right) &= \int_\Theta (\theta-\delta)^2 \pi(\theta|x)\, d\theta, \\ &= \mathbb{E}[\theta^2|x] - 2\delta\mathbb{E}[\theta|x] + \delta^2. \end{aligned}$$

En dérivant en δ, on obtient

$$R'_B\left(\delta|\pi\right) \;=\; 2\delta - 2\mathbb{E}[\theta|x]$$

valant 0 en $\delta = \mathbb{E}[\theta|x]$. Or

$$R''_B\left(\delta|\pi\right) \;=\; 2 \;>\; 0$$

donc $\delta \to R_B\left(\delta|\pi\right)$ est convexe, ce qui signifie que la solution de $R'_B\left(\delta|\pi\right) = 0$ est bien un minimisateur du risque.

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*

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

ou plus généralement une fonction linéaire par morceaux

$$L_{c_1,c_2}(\theta,\delta) \;=\; \begin{cases} c_2(\theta-\delta) & \textit{si } \theta > \delta \\ c_1(\delta-\theta) & \textit{sinon} \end{cases} \tag{7}$$

Proposition 2 *L'estimateur de Bayes associé à toute loi a priori* π *et au coût (8) 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*).*

Preuve. Considérons la situation générique où

$$L_{c_1,c_2}(\theta,\delta) \quad = \quad \left\{ \begin{array}{ll} c_2(\theta-\delta) & \text{si } \theta > \delta \\ c_1(\delta-\theta) & \text{sinon} \end{array} \right. \tag{8}$$

Le risque de Bayes s'écrit

$$R_B(\delta|\pi) \quad = \quad \int_{-\infty}^{\delta} c_2(\theta-\delta)\pi(\theta)\, d\theta + \int_{\delta}^{\infty} c_1(\delta-\theta)\pi(\theta)\, d\theta.$$

En raisonnant par intégration par parties (IPP), il vient :

$$\begin{aligned} R_B(\delta|\pi) \quad &= \quad [c_2(\theta-\delta)\Pi(\theta)]_{-\infty}^{\delta} + c_2\Pi(\theta<\delta|x) + [c_1(\delta-\theta)\Pi(\theta]_{\delta}^{\infty} - c_1\Pi(\theta>\delta|x), \\ &= \quad c_2\Pi(\theta<\delta|x) + c_1\left(1-\Pi(\theta<\delta|x)\right) \end{aligned}$$

car $\lim_{\theta\to-\infty}\Pi(\theta) = \lim_{\theta\to\infty}\Pi(\theta) = 0$. Ce risque est minimum lorsque $R_B(\delta|\pi) = 0$ soit lorsque

$$\Pi(\theta<\delta^\pi|x) \quad = \quad \frac{c_1}{c_1+c_2}$$

ce qui confère donc à l'estimateur δ^π le sens du quantile de seuil $\frac{c_1}{c_1+c_2}$.

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**

$$L(\theta,\delta) \quad = \quad \left\{ \begin{array}{ll} 1-\delta & \textit{si } \theta\in\Theta_0 \\ \delta & \textit{sinon} \end{array} \right. \tag{9}$$

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*

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

Preuve. Cette fonction de perte est utilisée dans le contexte des tests statistiques. On suppose partitionner Θ en Θ_0 et Θ_1. La fonction de perte correspondante est alors

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

Le risque *a posteriori* est alors

$$R_B(\delta|\pi) \quad = \quad \mathbb{1}_{\delta=1}\Pi(\theta\in\Theta_0|x) + \mathbb{1}_{\delta=0}\Pi(\theta\in\Theta_0|x).$$

Ainsi $\delta^\pi = 1$ équivaut à $\Pi(\theta\in\Theta_0|X) \leq \Pi(\theta\in\Theta_1|X)$.

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 :

$$L(\theta, d) = \begin{cases} 0 & \text{si } d = \mathbb{1}_{\Theta_0} \\ a_0 & \text{si } \theta \in \Theta_0 \text{ et } d = 0 \\ a_1 & \text{si } \theta \notin \Theta_0 \text{ et } d = 1 \end{cases} \quad (10)$$

qui donne l'estimateur bayésien

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

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 ou divergences 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.

Rappels sur la divergence de Kullbacl-Leibler (KL)

La divergence KL a un sens issu de la théorie de l'information (qui sera rappelé plus loin dans le cours). Rappeler ses principales propriétés peut être utile. Soit $\pi_1(\theta)$ et $\pi_2(\theta)$ deux densités de probabilité définies sur un même espace Θ, π_1 étant absolument continue par rapport à $\pi_2(\theta)$. Alors

$$\begin{aligned} \mathrm{KL}(\pi_1||\pi_2) &= \mathbb{E}_{\pi_1}\left[\log\frac{\pi_1(\theta)}{\pi_2(\theta)}\right], \\ &= \int_\Theta \pi_1(\theta)\log\frac{\pi_1(\theta)}{\pi_2(\theta)}\,d\theta. \end{aligned}$$

Proposition 4 *Si $\pi_1(\theta)$ et $\pi_2(\theta)$ sont propres, alors KL$(\pi_1||\pi_2) \geq 0$ et vaut 0 si et seulement si $\pi_1 = \pi_2$ (presque partout)*

La preuve de ce résultat est fondée sur l'application de l'inégalité de Jensen à la fonction $\pi_1 \to \pi_1 \log \pi_1$, qui est différentiable deux fois sur l'espace des densités de probabilités, et dont la différentielle seconde vaut $1/\pi_1 > 0$. Elle est donc convexe. Alors $\mathrm{KL}(\pi_1||\pi_2) \geq -\log 1 = 0$.

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 ?

Réponse. Si $\tilde{L}(\theta,\ell,\delta) = (\ell - L(\theta,\delta))^2$ et $L(\theta,\delta) = (\theta-\delta)^2$, alors l'estimateur bayésien est

$$\ell^\pi = \mathbb{E}_\pi\left[L(\theta,\delta|X)\right] = \mathbb{E}_\pi\left[((\theta-\delta)^2|X\right].$$

En choisissant en toute logique $\delta = \delta^\pi = \mathbb{E}[\theta|X]$ il vient donc

$$\ell^\pi = \mathrm{Var}_\pi[\theta|X].$$

3.5 Mode *a posteriori* (MAP)

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

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

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. D'une façon générale, supposons vouloir tester deux hypothèses de modèles, M_0 et M_1, l'une contre l'autre, et pouvoir assigner à ces modèles des probabilités *a priori* non nulles

$$\Pi(M_0),\ \Pi(M_1)$$

d'expliquer des données X non encore observées. Typiquement, on peut vouloir :

- pour un même modèle de vraisemblance $f(X|\theta)$, tester *a posteriori* deux sous-ensembles différents de Θ : $\Pi(\theta \in \Theta_0|X) > 0$ et $\Pi(\theta \in \Theta_1|X) > 0$;
- plus généralement tester un couple $(f_0(x|\theta_0), \pi_0(\theta_0))$ contre un couple $(f_1(x|\theta_1), \pi_1(\theta_1))$, avec $\theta_i \in \Theta_i$.

Le facteur de Bayes est le *rapport de la vraisemblance marginale* de ces deux hypothèses concurrentes :

$$B_{01}(X) \quad = \quad \frac{P(X|M_0)}{P(X|M_1)} \tag{11}$$

où

$$\begin{aligned} P(X|M_i) \quad &= \quad \int_{\Theta_1} f_i(X|\theta_i)\pi_i(\theta_i)\, d\theta_i, \\ &= \quad \frac{\Pi(M_i|X)P(X)}{\Pi(M_i)} \end{aligned}$$

où $P(X)$ représente la loi inconnue des données. Le rapport (11) peut alors se simplifier en

$$B_{01}(X) \quad = \quad \frac{\Pi(M_0|X)}{\Pi(M_1|X)}\frac{\Pi(M_1)}{\Pi(M_0)}. \tag{12}$$

Le facteur de Bayes est donc une transformation bijective de la probabilité *a posteriori*, qui a fini par être l'outil le plus utilisé pour choisir un modèle bayésien. Lorsque les deux modèles sont également probables *a priori*, alors $\frac{\Pi(M_1)}{\Pi(M_0)} = 1$ et le rapport de Bayes est simplement le rapport de leurs probabilités *a posteriori*.

Le cas le plus fréquemment rencontré en pratique est celui où l'hypothèse de modèle M_i se réduit à $\theta \in \Theta_i$. Il amène à la définition suivante.

.

Définition 13 *Le facteur de Bayes associé au problème du choix entre $\theta \in \Theta_0$ et $\theta \in \Theta_1$ est le rapport des probabilités a posteriori des hypothèses nulle et alternative sur le rapport a priori de ces mêmes hypothèses*

$$B_{01}(X) \quad = \quad \left(\frac{\Pi(\theta \in \Theta_0|X)}{\Pi(\theta \in \Theta_1|X)}\right) \Big/ \left(\frac{\Pi(\theta \in \Theta_0)}{\Pi(\theta \in \Theta_1)}\right) \tag{13}$$

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

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

Sous le coût généralisé (10), 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

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

Dans le pratique, on utilise souvent des échelles logarithmiques pour faire une sélection de modèle (échelle de Jeffreys améliorée par Kass et Raftery) :

(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.

Exercice 8 *Soit $X \sim \mathcal{B}(\theta)$ (loi de Bernoulli) avec $\Theta = [0, 1]$. Soit M_0 un modèle défini par $\{\theta = 1/2\}$ et M_1 un modèle défini par un θ inconnu dans $[0, 1]$, avec $\pi_1(\theta) = \mathcal{U}[0, 1]$. Un échantillon de 200 tirages fournit 115 succès et 85 échecs. Au vu de ces données, quel modèle choisir ? Ce résultat diffère-t-il significativement d'un test fréquentiste ?*

Réponse. La vraisemblance des données X est binomiale :

$$f(x|\theta) = \binom{200}{115} \theta^{115}(1-\theta)^{85}$$

ce qui permet de calculer

$$P(X|M_0) \quad = \quad \binom{200}{115} (1/2)^{200} \ \simeq \ 0.006$$

alors que pour M_1 on a

$$P(X|M_1) \quad = \quad \int_0^1 \binom{200}{115} \theta^{115}(1-\theta)^{85} \, d\theta \ = \ 1/201 \ \simeq \ 0.005.$$

Le facteur de Bayes vaut alors 1.2, ce qui indique uniquement que la certitude que H_0 est fausse est faible (on pointe très légèrement vers le modèle M_0).
Un test d'hypothèse fréquentiste de M_0 indiquerait que M_0 doit être rejeté par exemple au niveau de signification 5%, car la probabilité d'obtenir 115 succès ou plus à partir d'un échantillon de 200 si $\theta = 1/2$ est de 0.02. On en conclut qu'un test classique donnerait des résultats significatifs permettant de rejeter H_0 tandis qu'un test bayésien ne pourrait considérer le résultat comme extrême.

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

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

Dans la définition (13), 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.

Considérons le cas suivant : on dispose d'un modèle de régression

$$y \quad = \quad \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

où $\beta_d \in I\!R$, et on veut mener un test de significativité sur $\theta = \beta_d$:

$$M_0 : \ \{\beta_d = 0\} \quad vs \quad M_1 : \ \{\beta_d \neq 0\}$$

On peut alors noter $\Theta_0 = \{\theta_0\} = \{0\}$ et $\Theta_1 = I\!\!R^*$. Il est clair que $\Pi(\theta \in \Theta_0) == 0$ car la mesure dominante est Lebesgue. Donc la formulation (13) n'est pas applicable. On peut alors reposer le test en redéfinissant plus généralement

$$M_0 : \ \{|\beta_d - \theta_0| = \epsilon\} \quad vs \quad M_0 : \ \{|\beta_d| \neq \epsilon\}$$

avec un ϵ tendant vers 0 par valeurs positives. Reste la difficulté majeure que ϵ est inconnu et dépend du contexte de l'étude.

Plus généralement, et pour mieux formaliser les choses, il convient alors d'introduire une masse de Dirac δ_{θ_0} en θ_0 et de considérer un poids $\rho_\epsilon \in [0,1]$ à ϵ fixé, valant

$$\rho_\epsilon = \Pi(\theta||\theta - \theta_0| < \infty) \ = \ \Pi(\theta \in \Theta_0)$$

Le mélange

$$\pi(\theta) \quad = \quad \rho_\epsilon \delta_{\theta_0} + (1-\rho_\epsilon)\pi_1(\theta),$$

désigne la loi *a priori* qui généralise $\Pi(\theta \in \Theta_0)$ et $\Pi_1(\theta) = \Pi(\theta \in \Theta_1)$, la densité de cette dernière loi étant absolument continue par rapport à la mesure de Lebesgue. Dans ce cas, l'application de la formule (13) donne directement

$$\begin{aligned} B_{01}(X) \quad &= \quad \frac{\rho_\epsilon f(X|\theta_0)}{(1-\rho_\epsilon)\int_{\theta\in\Theta_1} f(X|\theta)\pi_1(\theta)\,d\theta} \left(\frac{1-\rho_\epsilon}{\rho_\epsilon}\right), \\ &= \quad \frac{f(X|\theta_0)}{\int_{\theta\in\Theta_1} f(X|\theta)\pi_1(\theta)\,d\theta}. \end{aligned}$$

(on retrouve bien le rapport des lois marginales).

Cas d'un ou plusieurs priors impropres

Une remarque importante concerne le cas où π_0 ou π_1 n'est pas un *prior propre* (mesure non intégrable). Dans ce cas, B_{01} n'est alors pas défini de manière unique. En effet, si une loi est impropre (par exemple π_0) alors elle est définie à une constante multiplicative près. Pour $\pi_0^*(\theta) = c\pi_0(\theta)$, alors le facteur de Bayes est lui aussi multiplié par c : $B_{01}^* = cB_{01}$ et par conséquent les ordres de grandeur de B_{01} n'ont plus de sens : il n'est plus possible de comparer ses valeurs à une échelle dédiée.

EXEMPLE 10. *Soit $X \sim \mathcal{N}(\theta, 1)$ et $\pi(\theta) = c \neq 0$. On considère le test suivant : $H_0 : \theta = 0$ vs $H_1 = \theta \neq 0$. Dans ce cas, le facteur de Bayes est*

$$B_{01}(X = x) \quad = \quad \frac{\exp(-x^2/2)}{c\sqrt{2\pi}}$$

c étant inconnu, le facteur de Bayes ne peut avoir d'interprétation.

Pour pallier ce problème, des démarches ont été inventées, qui visent à séparer l'échantillon X en sous-échantillons et transformant les priors impropres en posteriors propres (mais faiblement informatifs) : les *facteurs de Bayes fractionnels* ou les *facteurs de Bayes intrinsèques*. Ils ont de bonnes propriétés asymptotiques. Toutefois, dans les dernières années, la recherche se tourne plutôt vers des approches par *mélanges de modèles bayésiens*, qui permettent de remplacer la sélection d'un sous-modèle via le facteur de Bayes par la sélection d'un sous-modèle via les poids du mélange. Ce type d'approche permet de se débarasser, sous certaines conditions, des problèmes posés par les priors impropres.

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 $\mathbf{e_n} = (\mathbf{x_n}, \boldsymbol{\theta}_\mathbf{n})$.

Questions.

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

Réponses. **1.** Dans un cadre bayésien, probabilisons l'espace $\Theta = \Theta_0 \oplus \Theta_1$ et nommons Π la mesure de probabilité associée. On note également P la mesure de probabilité associée à l'espace χ supposé probabilisé. On souhaite prévoir la probabilité que $\theta \in \Theta_0$ sachant x et e_n. *Via* une règle de Bayes, on produit le classifieur classique (dit *classifieur de Bayes*)

$$
\begin{aligned}
\Pi(\theta \in \Theta_0 | x, e_n) &= \frac{P(X = x | \theta \in \Theta_0, e_n)\Pi(\theta \in \Theta_0 | e_n)}{P(X = x | e_n)}, \\
&= \frac{P(X = x | \theta \in \Theta_0, e_n)\Pi(\theta \in \Theta_0 | e_n)}{P(X = x | \theta \in \Theta_0)\Pi(\theta \in \Theta_0 | e_n) + P(X = x | \theta \notin \Theta_0)\left[1 - \Pi(\theta \in \Theta_0 | e_n)\right]}
\end{aligned}
$$

Le terme de vraisemblance $P(X = x | \theta \in \Theta_0, e_n)$ peut être estimé de nombreuses manières (ex : par la fréquence d'observation de l' évènement $X = x$ dans les situations recensées pour lesquelles $\theta \in \Theta_0$). Le classifieur de Bayes *naïf* repose ainsi sur une simplification de cette vraisemblance, etc. (voir cours d'apprentissage statistique).

2. Le calcul de la probabilité $P(\theta \in \Theta_0 | x, e_n)$ ne suffit cependant pas pour prendre une décision opérationnelle. Intuitivement, on souhaiterait pourtant choisir de mener une intervention si

$$
\begin{aligned}
\Pi(\theta \in \Theta_0 | X = x, e_n) \ \geq \ \Pi(\theta \notin \Theta_0 | X = x, e_n) &= \Pi(\theta \in \Theta_1 | X = x, e_n), \\
&= 1 - \Pi(\theta \in \Theta_0 | X = x, e_n)
\end{aligned}
$$

et donc à choisir (ou recommander) d'intervenir si

$$
\Pi(\theta \in \Theta_0 | X = x, e_n) \geq 1/2. \tag{14}
$$

Mais cette règle est en fait simpliste, car elle n'intègre pas les risques d'erreur liés au fait qu'on utilise un échantillon de taille finie n pour mener le calcul de cette probabilité. Une bonne façon de faire est de placer le problème de classification dans un problème de décision plus vaste.

La décision qu'un possible intervenant sur le réseau routier souhaite prendre est binaire : sachant $X = x$, on intervient ou non. À partir des données e_n, il tente donc de définir un estimateur statistique $\hat{\delta}_n(x)$ d'une décision *idéale* $\delta(x)$ vivant dans un espace de décision $\mathcal{D} = \{0, 1\}$ où :

- $\delta(x) = 0 \Leftrightarrow$ pas d'intervention,

- $\delta(x) = 1 \Leftrightarrow$ intervention.

Pourquoi parle-t-on d'estimateur statistique ? Parce que la décision idéale $\delta(x)$ est inaccessible par nature – elle sous-entend que le possible intervenant est omniscient, que θ est parfaitement connu et que nécessairement $n = \infty$.

On construit tout estimateur statistique comme le minimiseur d'une *fonction de coût*

$$\delta(x) \in \mathcal{D} \mapsto L(\theta, \delta(x))$$

que l'on cherche à définir si la vérité sur θ pouvait être connue. Dans le cas qui nous intéresse, on aurait :

- $L(\theta, \delta(x)) = C_1 =$ le coût prévisionnel d'une intervention à raison, donc si $\theta \in \Theta_1$ (ou $\theta \notin \Theta_0$) et $\delta(x) = 1$;

- $L(\theta, \delta(x)) = C_2 =$ le coût prévisionnel d'une non-intervention à tort (*erreur de 1ère espèce*), si $\theta \in \Theta_1$ et $\delta(x) = 0$;

- $L(\theta, \delta(x)) = C_3 =$ le coût prévisionnel d'une intervention à tort (*erreur de 2ème espèce*), si $\theta \in \Theta_0$ et $\delta(x) = 1$;

- $L(\theta, \delta(x)) = C_4 = 0$ le coût (nul) d'une non-intervention à raison, si $\theta \in \Theta_0$ et $\delta(x) = 0$.

On peut alors écrire, de façon plus condensée :

$$L(\theta, \delta(x)) = C_1\delta(x)\mathbb{1}_{\{\theta\in\Theta_1\}} + C_2(1-\delta(x))\mathbb{1}_{\{\theta\in\Theta_1\}} + C_3\delta(x)\mathbb{1}_{\{\theta\in\Theta_0\}}. \tag{15}$$

Or la vérité sur θ ne peut être parfaitement connue. On dispose simplement de la connaissance *a posteriori* $\Pi(\theta \in \Theta|X = x, e_n)$. Si l'on souhaite prendre une décision qui prenne en compte l'incertitude épistémique sur θ, il faut définir un *risque* $R(\delta(x), \Pi, e_n)$ qui puisse intégrer la connaissance de $\Pi(\theta \in \Theta|X = x, e_n)$, la construction de $L(\theta, \delta(x))$ et qui soit minimisable en un choix unique d'estimateur de $\delta(x)$. Pour obtenir un *ordre total* sur l'espace des applications $\delta(x) \mapsto R(\delta(x), \Pi, e_n)$, il faut nécessairement définir ce risque comme le *risque de Bayes*

$$R(\delta(x), \Pi, e_n) = \int_\Theta L(\theta, \delta(x)) d\Pi(\theta \in \Theta|X = x, e_n)$$

et d'en déduire donc la *décision optimale* (et non pas *idéale*)

$$\hat{\delta}_n(x) = \arg\min_{\delta(x)\in\mathcal{D}} R(\delta(x), \Pi, e_n).$$

Après intégration,

$$\begin{aligned} R(\delta(x), \Pi, e_n) &= C_1\delta(x)\Pi(\theta \in \Theta_1|X = x, e_n) + C_2(1-\delta(x))\Pi(\theta \in \Theta_1|X = x, e_n) \\ &\quad + C_3\delta(x)\left[1 - \Pi(\theta \in \Theta_1|X = x, e_n)\right]. \end{aligned}$$

On en déduit la règle de décision (ou de recommandation) suivante : ayant observé l'évènement $X = x$, on décide d'intervenir ($\hat{\delta}_n(x) = 1$) si le risque associé à cette décision est moins élevé que le risque associé à la décision contraire ($\hat{\delta}_n(x) = 0$), soit si

$$\begin{aligned} R(0, \Pi, e_n) = C_2\Pi(\theta \in \Theta_1|X = x, e_n) \;\geq\; R(1, \Pi, e_n) &= C_1\Pi(\theta \in \Theta_1|X = x, e_n) \\ &\quad + C_3\left[1 - \Pi(\theta \in \Theta_1|X = x, e_n)\right], \end{aligned}$$

c'est-à-dire quand

$$\Pi(\theta \in \Theta_1 | X = x, e_n) \quad \geq \quad \frac{C_3}{C_2 - C_1 + C_3} \tag{16}$$

Finissons par quelques remarques importantes :

- il n'est pas utile de disposer des coûts absolus C_i pour prendre une décision, car des rapports de coûts suffisent, ce qui est en général plus simple à estimer dans une vraie démarche opérationnelle ;
- on peut légitimement supposer que $C_1 \leq C_3 < C_2$, car le coût C_1 d'une intervention utile peut être réduit par l'effet d'assurances, tandis que le coût C_3 d'une intervention inutile ne l'est pas. Enfin, le coût (prévisionnel) C_2 d'une non-intervention qui aurait pu être utile peut éventuellement intégrer celui de vies humaines ; remarquons qu'en toute rigueur, $C_2 = C_2(t)$ où t représente le temps depuis l'occurence de l'événement, et que cette fonction est très certainement croissante.
- Il faut que $C_3 = C_2$ et $C_1 = 0$ pour obtenir l'équivalent décisionnel de la règle (14). On perçoit bien qu'une décision purement intuitive est à rejeter, car elle pré-suppose une contradiction forte avec la deuxième remarque.
- Cette règle de décision prend intégralement en compte les incertitudes sur la véritable nature de l'événement θ, conditionnellement à la validité des hypothèses.
- Le choix de la fonction de coût (15) est arbitraire ; mais retenons qu'en l'absence d'arguments permettant de rationaliser ce choix, les coûts sont généralement assemblés de façon additive

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 :

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

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

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

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

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

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*

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

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*

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

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

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

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 $X_1, \ldots, X_n|\theta \sim f(X|\theta)$,*

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

Preuve. On va montrer que $\Pi(\theta|X_1, \ldots, X_n) \to 0 \; \forall\theta \neq \theta_0$. On a

$$\log \frac{\Pi(\theta|X_1, \ldots, X_n)}{\Pi(\theta_0|X_1, \ldots, X_n)} \quad = \quad \log \frac{\Pi(\theta)}{\Pi(\theta_0)} + \sum_{i=1} Y_i \qquad (15)$$

où les

$$Y_i = \log \frac{f(X_i|\theta)}{f(X_i|\theta_0)}$$

sont des v.a. iid, telles ques

$$\mathbb{E}\left[Y_i\right] \quad = \quad KL\left(\theta_0\right) - KL\left(\theta\right)$$

(rappelons que l'intégration se fait par rapport à la vraie loi inconnue des données) qui vaut 0 si $\theta = \theta_0$, et qui est négatif sinon car θ_0 est l'unique minimiseur de $KL(\theta)$. Ainsi, si $\theta \neq \theta_0$, le second terme de (15) est une somme de termes iid avec une espérance négative. Par la loi des grands nombres, on obtient que $\lim_{n\to\infty} \sum_{i=1} Y_i = -\infty$ si $\theta \neq \theta_0$. Tant que le premier terme de (15) est fini (soit tant que $\Pi(\theta = \theta_0) > 0$), l'expression totale pour (15) tend également vers $-\infty$. Nécessairement,

$$\frac{\Pi(\theta|X_1, \ldots, X_n)}{\Pi(\theta_0|X_1, \ldots, X_n)} \quad \xrightarrow[\mathbb{P}]{n\to\infty} 0$$

et donc $\Pi(\theta|X_1, \ldots, X_n) \to 0 \; \forall\theta \neq \theta_0$. Comme toutes les probabilités somment à 1, nécessairement

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

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

$$\Pi(\theta \in V_{\theta_0}|x_1, \ldots, x_n) \quad \xrightarrow[\mathbb{P}]{n\to\infty} \quad 1.$$

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,*

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

Preuve. Le théorème 6 montre qu'on peut concentrer l'étude sur un voisinage de θ_0. Obtenir la loi limite requiert deux étapes :

- montrer que le mode *a posteriori* est consistant, c'est-à-dire qu'il se situe dans le voisinage de θ_0 où se situe presque toute la masse ;
- montrer l'approximation gaussienne centrée en le mode *a posteriori*.

Pour simplifier, le schéma de preuve donné ici considère que θ est un scalaire. Notons $\tilde{\theta}_n$ le mode *a posteriori*

$$\tilde{\theta}_n = \arg\max_{\theta\in\Theta} \left\{\log f(x_1,\ldots,x_n|\theta) + \log\pi(\theta)\right\}.$$

La preuve de consistance du MLE peut être adaptée pour montrer que $\tilde{\theta}_n \to \theta_0$ quand $n\to\infty$, presque sûrement. On peut alors approximer la log-densité *a posteriori* par un développement de Taylor centré autour de $\tilde{\theta}_n$ (approximation quadratique de $\log\pi(\theta|x_1,\ldots,x_n)$) :

$$\begin{aligned}
\log\pi(\theta|x_1,\ldots,x_n) &= \log\pi(\tilde{\theta}_n|x_1,\ldots,x_n) + \frac{1}{3}(\theta-\tilde{\theta}_n)^2\frac{\partial^2}{\partial\theta^2}\left[\log\pi(\theta|x_1,\ldots,x_n)\right]_{\theta=\tilde{\theta}_n} &(15)\\
&\quad + \frac{1}{6}(\theta-\tilde{\theta}_n)^3\frac{\partial^3}{\partial\theta^3}\left[\log\pi(\theta|x_1,\ldots,x_n)\right]_{\theta=\tilde{\theta}_n} + \ldots &(16)
\end{aligned}$$

Le terme linéaire est nul car par définition du mode :

$$\frac{\partial}{\partial\theta}\left[\log\pi(\theta|x_1,\ldots,x_n)\right]_{\theta=\tilde{\theta}_n} = 0.$$

Le premier terme de (16) est constant. Le coefficient du second terme (sous l'hypothèse iid) est

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\left[\log\pi(\theta|x_1,\ldots,x_n)\right]_{\theta=\tilde{\theta}_n} &= \frac{\partial^2}{\partial\theta^2}\left[\log\pi(\theta)\right]_{\theta=\tilde{\theta}_n} + \sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\left[\log f(x_i|\theta)\right]_{\theta=\tilde{\theta}_n},\\
&= cte + \sum_{i=1}^n Y_i
\end{aligned}$$

où les Y_i sont des variables aléatoires iid d'espérance négative sous l'hypothèse $X\sim\tilde{f}(x)$. En effet, si $\tilde{\theta}_n=\theta_0$, on a

$$\mathbb{E}\left[Y_i\right] = \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(x_i|\theta_0)\right] = -I_{\theta_0}$$

et sinon

$$\mathbb{E}\left[Y_i\right] = --\frac{\partial^2}{\partial\theta^2}KL\left(\theta\right)_{\theta=\tilde{\theta}_n} < 0$$

par convexité. Ainsi, le coefficient du second terme converge vers $-\infty$ à la vitesse n. Quand $n\to\infty$, $|\tilde{\theta}_n-\theta_0|\to 0$, et les termes suivants du développement de Taylor tendent vers 0. On a donc

$$\log\pi(\theta|x_1,\ldots,x_n) \sim -\alpha(\theta-\tilde{\theta}_n)^2$$

et la loi limite de $\pi(\theta|x_1,\ldots,x_n)$ est donc une gaussienne.

4.3 Régions de crédibilité et régions HPD

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 ∞,

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

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 $1-\alpha$. 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 $\alpha-$crédibles, 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

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

$A_{\alpha,\pi}$ 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.*

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-difféomorphisme (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 $\eta = \exp(\theta)$.*

Réponse.
En général, on peut constater que $\tilde{A}_{\alpha,\pi} = g\left(A_{\alpha,\pi}\right)$. En effet,

$$\pi(\eta|X) = \pi(\theta(\eta)|X)\left|\frac{d\theta}{d\eta}\right|$$

donc

$$\left\{\theta\;;\;\pi(\eta|X) \geq \tilde{h}_\alpha\right\} \quad = \quad \left\{\theta\;;\;\pi(\theta(\eta)|X) \geq \tilde{h}_\alpha\left|\frac{d\theta}{d\eta}\right|^{-1}\right\}$$

et on a donc égalité si et seulement si

$$\tilde{h}_\alpha\left|\frac{d\theta}{d\eta}\right|^{-1} \quad = \quad h_\alpha.$$

Si l'on suppose par exemple que $X \sim \mathcal{N}(\theta,1)$ et $\pi(\theta) \propto 1$, alors

$$\theta|X \quad \sim \quad \mathcal{N}(X,1).$$

Pons alors $\eta = \exp(\theta)$. Comme $\pi(\theta|X) \geq h_\alpha \Leftrightarrow (\theta - X)^2 \leq \left(\phi^{-1}(1-\alpha)\right)^2$, alors

$$A_{\alpha,\pi} \quad = \quad \left[X - \phi^{-1}(1-\alpha)\;;\;X + \phi^{-1}(1-\alpha)\right].$$

Par ailleurs,

$$\pi(\eta|X) \quad \propto \quad \frac{1}{\eta}\exp\left(-\frac{1}{2}\left(X - \log\eta\right)^2\right).$$

Donc

$$\begin{aligned}
\pi(\eta|X) \geq \tilde{h}_\alpha \quad &\Leftrightarrow \quad \frac{1}{2}\left(X - \log\eta\right)^2 + \log\eta \leq h_\alpha,\\
&\Leftrightarrow \quad \frac{1}{2}\left(2X - \log\eta\right)^2 - \frac{1}{2}X^2 \leq h_\alpha,\\
&\Leftrightarrow \quad \left(2X - \log\eta\right)^2 \leq 2h_\alpha,\\
&\Leftrightarrow \quad \log\eta \in \left[2X - \phi^{-1}(\alpha/2)\;,\;2X + \phi^{-1}(\alpha/2)\right]
\end{aligned}$$

Ainsi, le lien entre $\tilde{A}_{\alpha,\pi}$ et $A_{\alpha,\pi}$ ne correspond pas à la transformation initiale (exponentielle).

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 :

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

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* $\theta \in I\!R$*. Alors les intervalles de quantiles empiriques de la forme* $[\theta^{(\alpha/2)}, \theta^{(1-\alpha/2)}]$ *sont tels que*

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

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.

4.4 TP : Comparatif des approches bayésienne et fréquentiste pour les régions HPD

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

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

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.
4. Déterminez la loi *a posteriori* marginale $\pi(\sigma^2 | x_1, \ldots, x_n)$; le calcul de la région HPD est-il simple ?

Remarque. "Déterminer" signifie indiquer si la loi appartient à une famille connue, par exemple largement implémentée sur machines. La connaissance des lois gamma, inverse gamma et Student est peut-être nécessaire pour répondre aux questions.

5 Méthodes de calcul bayésien

5.1 Introduction

Nous nous plaçons dans le cadre à présent bien connu :

- Soit $\{X \sim f(.|\theta), \pi(\theta)\}$ un modèle bayésien servant à prendre une décision $\delta \in \mathcal{D}$.
- Dans un cadre d'**analyse** (*a posteriori*), on a observé des données $\mathbf{x_n} = (x_1, \ldots, x_n) \sim f(.|\theta)$.
- La loi *a posteriori* $\pi(\theta|\mathbf{x_n})$ décrit l'ensemble des incertitudes sur $\theta \in \Theta$, vecteur inconnu qui "paramétrise" l'état de nature.
- Toute décision δ peut être jugée par un *coût* $L(\theta, \delta)$, c'est-à-dire son écart par rapport à une décision idéale inatteignable, affecté par la distribution de probabilité *a posteriori* $\pi(\theta|\mathbf{x_n})$.
- La *décision optimale* s'obtient en cherchant l'optimum de la fonction du coût moyen *a posteriori* (expected opportunity loss) : $\delta^\pi = \arg \min\limits_{\delta \in \mathcal{D}} R_B(\delta|\pi)$ avec

$$R_B(\delta|\pi) = \int_\Theta L(\theta, \delta)\, d\Pi(\theta|\mathbf{x_n}) = \int_\Theta L(\theta, \delta) \pi(\theta|\mathbf{x_n})\, d\theta$$

  La question fondamentale du calcul bayésien est double : peut-on obtenir une expression explicite pour δ^π ? sinon, comment peut-on l'évaluer numériquement ?

En général, l'estimateur de Bayes n'a pas de caractère explicite, car

$$\pi(\theta|\mathbf{x_n}) = \frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{\int f(\mathbf{x_n}|\theta)\pi(\theta) d\theta}$$

et le dénominateur n'est pas explicitement connu ($\pi(\theta|\mathbf{x_n})$ est définie à une constante d'intégration près). On peut essayer de l'estimer par *intégration numérique* : les techniques classiques de Newton-Cotes ou Runge-Kutta mènent fréquemment à des instabilités numériques lorsque $\dim \Theta$ augmente. Toutefois, les méthodes de *quadrature bayésienne* permettent de pallier ce problème. Elles relèvent encore du domaine de la recherche, et seront abordées plus tard dans ce cours.

Fort généralement, le calcul d'estimateur de Bayes, mais aussi la sélection de modèle, la production de régions $\alpha-$crédibles (par exemple), nécessitent de pouvoir **simuler** *a posteriori*. Par des techniques de Monte Carlo, fondée sur une capacité algorithmique de simulation pseudo-aléatoire (voir Annexe C), on peut ainsi mener les calculs suivants :

1. **Calcul d'une moyenne** *a posteriori*. On peut estimer δ^π par un estimateur de Monte Carlo

$$\hat{\delta}^\pi_M = \frac{1}{M} \sum_{i=1}^M \theta_i$$

   - Ingrédients : Loi Forte des Grands Nombres ($\hat{\delta}^*_M \xrightarrow{p.s.} \delta^*$) + Théorème Central Limite

2. **Estimation d'un quantile** *a posteriori* **d 'ordre** α. On peut estimer δ^π par l'inversion de la fonction de répartition empirique *a posteriori*

$$\hat{\Pi}_M(\theta|\mathbf{x_n}) = \frac{1}{M} \sum_{i=1}^M \mathbb{1}_{\{\theta \leq \theta^*_i\}} \qquad \text{(Théorème de Glivenko-Cantelli)}$$

   soit en prenant

$$\hat{\delta}^\pi_M = \begin{cases} \frac{1}{2}\left(\theta^*_{\alpha \cdot M} + \theta^*_{\alpha\cdot(M+1)}\right) & \text{si } \alpha \cdot M \text{ est entier} \\ \theta^*_{\lfloor \alpha \cdot M \rfloor + 1} & \text{sinon} \end{cases} \qquad \text{(Théorème de Mosteller)}$$

Un autre intérêt (et même une nécessité) apporté par la simulation *a posteriori* est le suivant : l'analyse prédictive.

Plus précisément, plaçon-nous dans ce cadre : **simuler des réalisations de** X est une nécessité lorsqu'on s'intéresse au comportement Y d'un phénomène (par exemple physique) modélisé ainsi :

$$Y = g(X, \nu) + \epsilon$$

où :

- g est une fonction (ou un code de calcul) déterministe ;
- ν est un indice ou une variable indexant typiquement des conditions environnementales ;
- ϵ est un "bruit" stochastique qui modélise l'erreur entre la réalité du phénomène Y et la sortie de g.

Dans ce problème de *propagation d'incertitudes*, on cherche à reproduire un grand nombre de configurations de Y pour calculer (par exemple) la probabilité que Y dépasse un certain seuil.

EXEMPLE 11. *Y représente une hauteur d'eau aval, g est un code hydraulique, X est un débit d'eau amont, ν caractérise le frottement de la rivière et ϵ tient compte de la méconnaissance du terrain, de la précision du code, etc.*

Comment doit être simulé X en entrée de g ? La loi *prédictive* de densité

$$f(x|\mathbf{x_n}) = \int f(x|\theta)\pi(\theta|\mathbf{x_n})\, d\theta$$

permet de simuler une prochaine observation x_{n+1} *crédible* sachant qu'on a déjà observé les $\mathbf{x_n}$. Si l'on cherche à simuler de façon crédible la succession de **deux** futures observations (x_{n+1}, x_{n+2}), on doit procéder ainsi :

$$\begin{aligned} X_{n+1} &\sim f(x|\mathbf{x_n}), \\ X_{n+2} &\sim f(x|\mathbf{x_n}, x_{n+1}), \quad \text{etc.} \end{aligned}$$

Un algorithme simple de simulation repose donc sur la simulation de la loi *a posteriori*.

Remarque 8 Liens avec le *machine /deep learning*. *Les techniques de simulation et d'échantillonnage, dans le cadre spécifique du machine / deep learning, servent notamment à :*

- *initialiser des algorithmes d'optimisation de métriques, fonctions de coût complexes... ;*
- *élaborer des modèles génératifs, tels les Generative Adversarial Networks (GAN) ;*
- *résoudre des problèmes de complétion de données manquantes ou en trop faible nombre (data augmentation) ;*
- *proposer des façons de sélectionner des mini-batchs utiles parmi un ensemble de données trop grand pour être processé intégralement (problématique de plan d'expérience) .*

5.1.1 Principe de la simulation indirecte

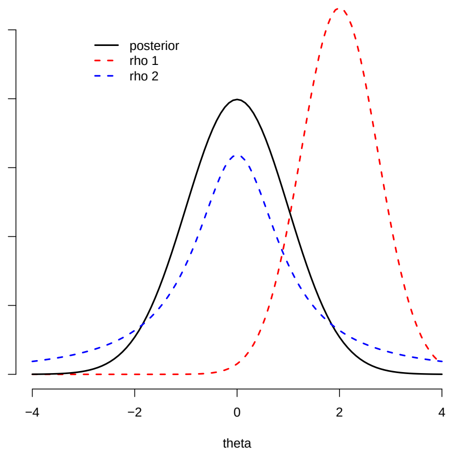
1. on simule un tirage θ_i suivant une *loi instrumentale* $\rho(\theta)$ (facile à simuler) ;
2. on utilise un *test* pour déterminer si θ_i aurait également pu être un *tirage plausible* de $\pi(\theta|\mathbf{x_n})$.

Plus $\rho(\theta)$ est "proche" de $\pi(\theta|\mathbf{x_n})$, plus ce test doit accepter les θ_i. Plus précisément :

1. La densité $\rho(\theta)$ doit être facilement simulable (ex : mélanges gaussiens si $\pi(\theta|\mathbf{x_n})$ est multimodale...).

2. Le *support*[6] de $\rho(\theta)$ contient *nécessairement* celui de $\pi(\theta|\mathbf{x_n})$.

3. Les queues de $\rho(\theta)$ devraient être plus lourdes que celles de $\pi(\theta|\mathbf{x_n})$.



4. Lorsque $\dim \Theta$ est petite (1 ou 2), on peut tracer $\pi(\theta|\mathbf{x_n})$ à un coefficient près pour sélectionner une forme intéressante pour $\rho(\theta)$.

Un candidat logique peut parfois être la loi *a priori* $\pi(\theta)$, car elle respecte automatiquement la règle d'inclusion du support. Si l'*a priori* est très informatif par rapport aux données, l'*a posteriori* en sera proche. Une quantification de cette "force" relative d'information est donc pratique pour choisir $\rho(\theta)$. Toutefois, ce choix peut être délicat : si l'*a priori* est très large (peu informatif), alors

- il peut privilégier indûment des régions où la vraisemblance (comme fonction de θ) est nulle ou quasi-nulle ;
- il faudra beaucoup de tirages pour atteindre les régions HPD (de plus haute densité) *a posteriori*, ce qui entraînera un coût algorithmique très fort

Ce choix est aussi à proscrire si l'*a priori* privilégie des régions de Θ qui sont éloignées de celles privilégiées par les données. Une indication en faible dimension est de mesurer l'éloignement du mode *a priori* de θ et du maximum de vraisemblance $\hat{\theta}_n$.

5.2 Méthodes d'échantillonnage dans la loi *a posteriori*

Dans cette partie, on cherche doncà obtenir **indirectement** des tirages qui suivent (en général approximativement) la loi *a posteriori*. Citons quelques algorithmes classiques que nous étudierons (quasiment tous) dans ce cours :

1. algorithmes d'acceptation-rejet ;
2. échantillonnage d'importance (préférentiel) ;
3. méthodes de Monte Carlo par chaînes de Markov (MCMC) ;
4. filtrage particulaire (pour les modèles à espace d'état).

6. *Le domaine de Θ où la densité est non nulle.*

Ces méthodes – et leurs hybrides – sont les outils actuels les plus puissants pour simuler des lois connues semi-explicitement (à une constante/une intégrale près). Ils connaissent plusieurs approches d'**accélération** dont nous discuterons également, et qui amènent par ailleurs à faire un lien avec les techniques classiques du *machine learning* : une technique de gradient stochastique, qui vise à estimer un mode *a posteriori* (et sous-entend en général que la loi *a posteriori* est approximativement gaussienne), peut être vue comme une forme dégénérée de MCMC adaptivement accélérée.

5.2.1 Rappel : approches par inversion et transformations simples

Rappelons avant de commencer que la méthode générique de simulation de $\theta \sim \pi(\theta|X)$ repose sur *l'inversion de la fonction de distribution* Π qui, en unidimensionnel, est la fonction de répartition.

Théorème 12 *Si $U \sim \mathcal{U}[0,1]$ et $\Pi(\theta|X)$ la fonction de répartition de $\theta|X$, alors $\Pi^{-1}(U|X)$ a la même loi que θ.*

La preuve vient du fait que par définition, $\Pi(\Pi^{-1}(U|X) \leq \theta|X) = \Pi(I \leq \Pi(\theta|X)|X) = \Pi(\theta|X)$. Si Π n'est pas parfaitement croissante, on prend $\Pi^{-1}(u|X) = \inf\{\theta\ ;\ \Pi(\theta|X) \geq u\}$.

EXEMPLE 12.

- *Loi binomiale $\mathcal{B}(n,p)$: $F_X(x) = \sum_{i\leq x} \binom{n}{i} p^i(1-p)^{n-i}$ et $F_X^{-1}(u)$ s'obtient numériquement.*
- *Loi exponentielle $\mathcal{E}(\lambda)$: $F_X(x) = 1 - \exp(\lambda x)$ et $F_X^{-1}(u) = -\log(\mathbf{u})/\lambda$.*
- *Loi de Cauchy $\mathcal{C}(0,1)$: $F_X(x) = \frac{1}{\pi}\arctan(x) + \frac{1}{2}$ et $F_X^{-1}(u) = \tan(\pi(\mathbf{u} - 1/2))$.*

.

La critique de cette approche est aisée : $\Pi^{-1}(.|X)$ est rarement disponible, et ce “théorème" (plutôt un lemme) d'inversion ne s'applique qu'en dimension 1. Pour "contrer" ces problèmes, on peut proposer quelques transformations (voir ci-dessous), mais cela ne permet de régler que des cas particuliers.

Définition 17 Transformation de Box-Müller *Pour la loi normale $\mathcal{N}(0,1)$, si $X_1, X_2 \overset{iid}{\sim} \mathcal{N}(0,1)$, alors*

$$X_1^2 + X_2^2 \sim \chi_2^2, \qquad \arctan(X_1/X_2) \sim \mathcal{U}([0,2\pi]).$$

Comme χ_2^2 est identique à $\mathcal{E}(1/2)$, il vient par inversion :

$$X_1 = \sqrt{-2\,\log(U_1)}\,\sin(2\pi U_2) \qquad X_2 = \sqrt{-2\,\log(U_1)}\,\cos(2\pi U_2).$$

.

Les lois de Student et de Fisher se déduisent naturellement de la loi normale et de la loi du chi-deux. La loi de Cauchy se déduit de la loi normale par la règle suivante : si $X_1, X_2 iid \mathcal{N}(0,1)$, alors $X_1/X_2 \sim \mathcal{C}(0,1)$. La loi Beta $\mathcal{B}_e(\alpha,\beta)$, de densité

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}\,,$$

s'obtient à partir de la loi gamma par la règle suivante : si $X_1 \sim \mathcal{G}a(\alpha,1)$, $X_2 \sim \mathcal{G}a(\beta,1)$, alors

$$\frac{X_1}{X_1+X_2} \sim \mathcal{B}_e(\alpha,\beta).$$

5.2.2 Simulation multidimensionnelle

Le cas de la simulation multidimensionnelle est réglé également en principe par la règle en cascade suivante :

Définition 18 Cascade rule. *Supposons vouloir générer dans $\mathbb{R}^p$ l'échantillon $(X_1, \ldots, X_p) \sim f(x_1, \ldots, x_p)$ dont les composantes ne sont pas nécessairement indépendantes. La densité jointe s'écrit alors*

$$f(x_1, \ldots, x_p) = f_1(x_1) \times f_{2|1}(x_2|x_1) \ldots \times f_{p|-p}(x_p|x_1, \ldots, x_{p-1}).$$

On peut donc en déduire la règle d'implémentation suivante :

Simuler pour $t = 1, \ldots, T$

1 $X_1 \sim f_1(x_1)$

2 $X_2 \sim f_{2|1}(x_2|x_1)$

$\vdots$

$X_p \sim f_{p|-p}(x_p|x_1, \ldots, x_{p-1})$

5.2.3 Algorithmes d'acception-rejet (AR)

Ce type d'algorithme permet de simuler de façon **exacte** et **indépendante** selon la loi *a posteriori*. Il repose sur l'hypothèse suivante sur $\rho(\theta)$:

$$0 \;<\; K \;=\; \sup_{\theta\in\Theta} \frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{\rho(\theta)} \;<\; \infty.$$

Algorithme AR :

---

1. simulation indirecte : soit $\theta_i \sim \rho(\cdot)$
2. test :
   - soit $U_i \sim \mathcal{U}[0,1]$
   - si $U_i \leq \dfrac{f(\mathbf{x_n}|\theta_i)\pi(\theta_i)}{K\rho(\theta_i)}$ alors θ_i suit la loi $\pi(\theta|\mathbf{x_n})$

Preuve. Soit $\tilde{\theta}$ la variable aléatoire dont les tirages sont acceptées par le test. Alors, en définissant P la mesure de probabilité usuelle sur $[0,1]$, et $\tilde{\Pi}$ la mesure de probabilité produit sur $\Theta \times [0,1]$, d'après la formule de Bayes,

$$\begin{aligned}
\Pi\left(\tilde{\theta} \leq y|U \leq \frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{K\rho(\theta)}\right) &= \frac{\tilde{\Pi}\left(\tilde{\theta} \leq y \;,\; U \leq \frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{K\rho(\theta)}\right)}{\tilde{\Pi}\left(U \leq \frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{K\rho(\theta)}\right)}, \\
&= \frac{\int_{-\infty}^{y}\int_0^1 \frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{K\rho(\theta)}\, du\rho(\theta)\, d\theta}{\int_{-\infty}^{\infty}\int_0^{\frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{K\rho(\theta)}} du\rho(\theta)\, d\theta}, \\
&= \frac{\int_{-\infty}^{y} f(\mathbf{x_n}|\theta)\pi(\theta)\, d\theta}{K\int_{-\infty}^{\infty}\frac{1}{K} f(\mathbf{x_n}|\theta)\pi(\theta)\, d\theta}, \\
&= \Pi(\theta \leq y).
\end{aligned}$$

(Preuve en cours)

Observons que la loi du nombre de tirages nécessaires selon $\rho(\theta)$ jusqu'à en accepter un suit la loi géométrique de probabilité $1/(K \cdot C)$ où C est la constante d'intégration inconnue

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

donc $K \cdot C$ est l'espérance du nombre de tirages nécessaires avant l'acceptation. *Optimiser l'algorithme* revient donc à *diminuer* K.

Exercice 10 *On suppose $X \sim \mathcal{N}(\theta, 1)$ et on suppose connaître un échantillon $\mathbf{x_n}$ composé de :*

- *quelques observations $x_1, \ldots, x_{n-1}$ supposées iid.*
- *une pseudo-observation y qui est un cas-limite masquant (censurant) une observation x_n qui aurait dû être faite : $y < x_n$*

A priori, on suppose $\theta \sim \mathcal{N}(\mu, 1)$. Pouvez-vous produire un algorithme d'AR qui génére des réalisations de la loi a posteriori de θ ?

Réponse. La vraisemblance s'écrit

$$f(\mathbf{x_n}|\theta) \quad \propto \quad \underbrace{\exp\left(-\frac{1}{2}\sum_{k=1}^{n-1}(x_k-\theta)^2\right)}_{\text{terme régulier}} \quad \underbrace{1-\Phi(y-\theta)}_{\substack{\text{terme dû à la censure}\\ = \ P(X\ >\ y)}}.$$

L'*a posteriori* sur θ s'écrit alors

$$\pi(\theta|\mathbf{x_n}) \quad \propto \quad \tilde{\pi}(\theta|\mathbf{x_n}) \ = \ \exp\left\{-\frac{n}{2}\left[\theta-\frac{1}{n}\left(\mu+\sum_{k=1}^{n-1}x_k\right)\right]^2\right\}\left\{1-\Phi(y-\theta)\right\}.$$

On remarque que si $y = x_n$, on aurait un modèle conjugué et $\pi(\theta|\mathbf{x_n})$ serait une loi normale. Il semble donc pertinent de proposer, comme choix de loi instrumentale,

$$\rho(\theta) \equiv \mathcal{N}\left(\frac{1}{n}\left(\mu+\sum_{k=1}^{n-1}x_k\right), 1/n\right).$$

Puisque $1-\Phi(y-\theta) \leq 1$, on a

$$\tilde{\pi}(\theta|\mathbf{x_n}) \quad \leq \quad \underbrace{\sqrt{\frac{2\pi}{n}}}_{K}\cdot\left\{1-\Phi(y-\theta)\right\}\cdot\rho(\theta).$$

On peut donc mettre en oeuvre l'algorithme comme suit : on accepte θ_i si $U_i \leq 1-\Phi(y-\theta_i)$. Le nombre moyen d'appels nécessaires à $\rho(\theta)$ varie proportionnellement à $1/\sqrt{n}$, donc plus l'échantillon de données grandit, plus l'algorithme est efficace. Si cependant, on fait le choix $\rho(\theta) = \pi(\theta)$, alors

$$K \quad = \quad \sqrt{2\pi}\exp\left(\frac{1}{2}\left[\frac{1}{n-1}\sum_{k=1}^{n}x_k-\mu\right](1-\sqrt{n})\right)$$

Voir la figure 1 pour une illustration de la mise en oeuvre de cet algorithme.

Remarque 9 *On peut améliorer (faire baisser) le taux de rejet en encadrant la loi a posteriori entre 2 densités instrumentales (acceptation-rejet par enveloppe).*

Le principe de l'AR est parfait en théorie, mais en pratique il est réservé aux cas simples ($\dim\Theta$ petite). De plus, cet algorithme est en général très coûteux en temps d'attente.

5.2.4 Algorithmes d'échantillonnage préférentiel ou d'importance (IS)

Ce type d'algorithme vise surtout à produire un estimateur consistant d'une quantité d'intérêt *a posteriori*, mais il peut être utilisé dans un but de produire un échantillonnage exact mais non indépendant de la loi-cible (approche SIR).

Algorithme IS :

1. Soit $(\theta_1,\ldots,\theta_M)$ un tirage i.i.d. selon une densité instrumentale $\rho(\theta)$.

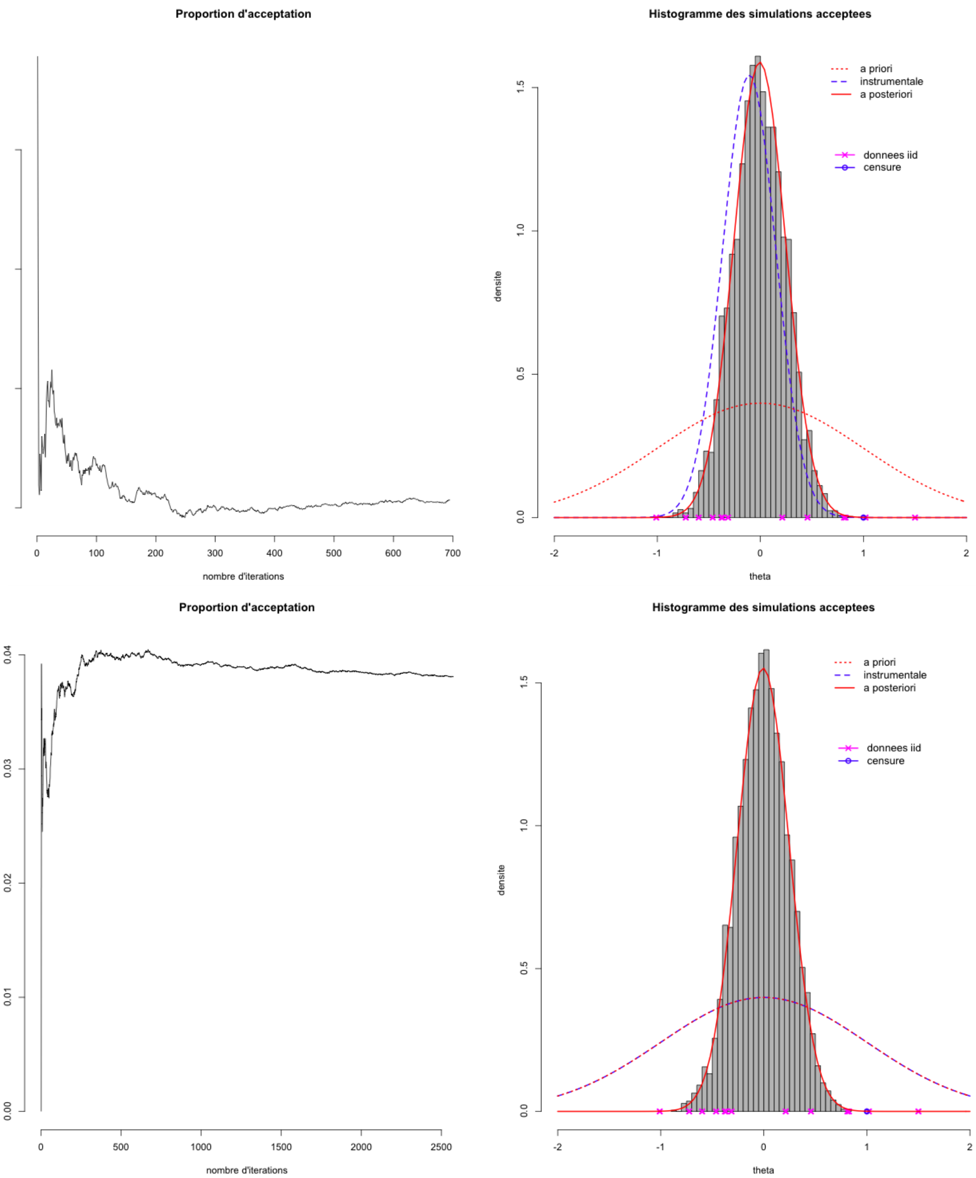


FIGURE 1 – Essai de simulation par AR en utilisant (en haut) une loi instrumentale “proche" du vrai posterior ; (en bas) le prior comme loi instrumentale.

2. Soit $(\omega_1,\ldots,\omega_M)$ les **poids d'importance** définis par

$$\omega_i \propto \frac{f(\mathbf{x_n}|\theta_i)\pi(\theta_i)}{\rho(\theta_i)}$$

et normalisés de façon à ce que leur somme fasse 1.

Théorème 13 *Geweke 1989. Toute fonction prédictive*

$$h(x|\mathbf{x_n}) = \int_\Theta h(x|\theta)\pi(\theta|\mathbf{x_n})\, d\theta$$

(ex : $h = f$) peut être estimée de façon consistante, lorsque $M \to \infty$, par

$$\hat{h}(x|\mathbf{x_n}) = \sum_{i=1}^{M} \omega_i h(x|\theta_i).$$

L'approche *Sampling-Importance Resampling* (SIR), proposée par Rubin (1988), se fonde sur le résultat suivant :

Théorème 14 *Les tirages*

$$\tilde{\theta}_1,\ldots,\tilde{\theta}_P \sim \mathcal{M}_{ultinomial}\left(\theta_1,\ldots,\theta_M|\omega_1,\ldots,\omega_M\right)$$

suivent la loi a posteriori $\pi(\theta|\mathbf{x_n})$.

L'heuristique de Rubin consiste à prendre $P < M/20$ pour diminuer la dépendance dans l'échantillon resimulé. On peut aussi ainsi estimer les caractéristiques de $\pi(\theta|\mathbf{x_n})$ (moyenne, variance, etc.).

EXEMPLE 13. *En reprenant une solution proposée pour l'exemple 10, par exemple $\rho(\theta) \equiv \mathcal{N}\left(\mu + \sum_{k=1}^{n-1} x_i, 1/n\right)$, les poids sont simplement proportionnels à*

$$\omega_i \propto 1 - \Phi(y - \theta_i)$$

qu'on normalise en divisant le membre de droite par la somme des $1 - \Phi(y - \theta_i)$, $i = 1,\ldots,M$. Les poids les plus hauts sont donc ceux pour lesquels $y \ll \theta_i$.

Remarquons que fort logiquement, plus la densité instrumentale $\rho(\theta)$ est "proche" de $\pi(\theta|\mathbf{x_n})$, plus les poids sont équilibrés (donc meilleur est le rééchantillonnage). Plutôt qu'une loi unique ρ, on peut mettre en place des algorithmes *adaptatifs* qui construisent itérativement une suite de densités $\{\rho_k(\theta)\}_k$ convergeant vers $\pi(\theta|\mathbf{x_n})$, pour améliorer encore le rééchantillonage. Une revue de tels algorithmes est proposée dans [19].

Notons le résultat suivant, important et dû à Rubinstein (1981), qui guide ces recherches d'algorithmes IS optimaux. Ce résultat n'est pas exploitable directement, car il revient à avoir déjà résolu que l'on cherche à résoudre, mais il sert à construire des lois instrumentales qui progressivement vont se rapprocher de cette optimalité.

Théorème 15 *Importance sampling optimal. Soit l'estimateur de la fonction d'intérêt $h(\theta) \in I\!R$ par IS :*

$$\hat{h}_M = \frac{1}{M}\sum_{i=1}^{M} \frac{\pi(\theta_i|x)}{\rho(\theta_i)} h(\theta_i) \to \mathbb{E}_\pi[h(\theta|X] \quad p.s.$$

où les $\theta_i \overset{iid}{\sim} \rho(\theta)$. Alors le choix de ρ qui minimise la variance de l'estimateur $\hat{h}_M$ est

$$\rho^*(\theta) = \frac{|h(\theta)|\pi(\theta|X)}{\int_\Theta |h(\theta)|\pi(\theta|X)\, d\theta}.$$

Preuve. On note d'abord que

$$\mathbb{V}_\rho\left[\frac{h(\theta)\pi(\theta|x)}{\rho(\theta)}\right] \;=\; \mathbb{E}_\rho\left[\frac{h^2(\theta)\pi^2(\theta|x)}{\rho^2(\theta)}\right] - \left(\mathbb{E}_\rho\left[\frac{h(\theta)\pi(\theta|x)}{\rho(\theta)}\right]\right)^2$$

et que le second terme ne dépend pas de ρ. Il suffit donc de minimiser le premier terme. D'après l'inégalité de Jensen, on a donc,

$$\begin{aligned}\mathbb{E}_\rho\left[\frac{h^2(\theta)\pi^2(\theta|x)}{\rho^2(\theta)}\right] &\geq \left(\mathbb{E}_\rho\left[\frac{|h(\theta)|\pi(\theta|x)}{\rho(\theta)}\right]\right)^2, \\ &\geq \left(\int_\Theta |h(\theta)|\pi(\theta|x)\, d\theta\right)^2\end{aligned}$$

qui est une borne inférieure indépendante de ρ, qui est atteinte en ρ^*.

Un outil important, dès qu'on aborde le problème de la génération de données de même loi, mais corrélées, est la **taille d'échantillon effective**, notée en général ESS (*Effective Sample Size*). L'ESS relie la variance d'un estimateur de Monte Carlo idéal (échantillonnant directement dans la loi-cible) à la variance d'un estimateur fondé sur un échantillonnage corrélé, comme celui produit par l'IS, dans le cas où les deux estimateurs utilisent le même nombre de tirages. L'ESS mesure donc l'efficacité d'un algorithme d'échantillonnage corrélé.

Définition 19 Effective Sample Size. *Soit un échantillonnage instrumental $(\theta_1,\ldots,\theta_M)$ associé à des poids d'importance normalisés $(\omega_1,\ldots,\omega_M)$. Alors on définit*

$$ESS \;=\; \left(\sum_{i=1}^M \omega_i^2\right)^{-1}.$$

Lorsque l'échantillon produit est bien indépendant, les poids normalisés sont tous égaux à $1/M$, et donc ESS $= M^2/M = M$.

Remarque 10 *Dans le cas des MCMC, la définition précédente nécessite d'être remaniée pour tenir compte de la corrélation dans l'échantillonnage obtenu.*

Une autre propriété intéressante des techniques d'IS est de permettre de mener facilement certains types d'**analyse de sensibilité au prior**. Nous l'analysons dans l'exercice suivant.

Exercice 11 *Considérons une fonction d'intérêt $h(\theta)$ que l'on cherche à résumer par un estimateur calculé sous un coût quadratique ; il s'agit donc de l'espérance a posteriori*

$$h \;=\; \mathbb{E}_\pi[h(\theta)|x_1,\ldots,x_n] \;=\; \int_\Theta h(\theta)\pi(\theta|x_1,\ldots,x_n)\, d\theta \tag{-2}$$

que l'on suppose pouvoir estimer simplement, de façon consistante, par Monte Carlo. Supposons vouloir modifier le prior : $\pi(\theta)\to\pi'(\theta)$, sans modifier le support, mais de façon à ce que la nouvelle loi a posteriori ne soit plus directement simulable. Peut-on (et sous quelles conditions) ne pas faire de calcul supplémentaire pour simuler le nouveau posterior $\pi'(\theta_1,\ldots,x_n)$?

Réponse. On considére donc une fonction d'intérêt $h(\theta)$ que l'on cherche à résumer par son espérance *a posteriori*

$$h \ = \ \mathbb{E}_{\pi}[h(\theta)|x_1,\ldots,x_n] \quad = \quad \int_{\Theta} h(\theta)\pi(\theta|x_1,\ldots,x_n)\, d\theta \tag{-1}$$

que l'on suppose pouvoir estimer simplement, de façon consistante, par Monte Carlo :

$$\hat{h}_M \quad = \quad \frac{1}{M}\sum_{k=1}^{M} h(\theta_i) \quad \text{avec } \theta_i \overset{iid}{\sim} \pi(\theta|x_1,\ldots,x_n)$$

Supposons vouloir modifier le prior : $\pi(\theta) \to \pi'(\theta)$, sans modifier le support, mais de façon à ce que la nouvelle loi *a posteriori* ne soit plus directement simulable. En supposant que $\pi(\theta|x_1,\ldots,x_n) > 0$ pour tout $\theta \in \Theta$, on peut néanmoins recalculer facilement le nouvel estimateur de Bayes :

$$\begin{aligned} h' \ = \ \mathbb{E}_{\pi'}[h(\theta)|x_1,\ldots,x_n] \quad &= \quad \int_{\Theta} h(\theta)\pi'(\theta|x_1,\ldots,x_n)\, d\theta, \\ &= \quad \int_{\Theta} \omega(\theta_i)h(\theta)\pi(\theta|x_1,\ldots,x_n)\, d\theta \end{aligned}$$

avec

$$\omega(\theta_i) \quad = \quad \frac{\pi'(\theta|x_1,\ldots,x_n)}{\pi(\theta|x_1,\ldots,x_n)} \ = \ \left(\frac{\pi'(\theta)}{\pi(\theta)}\right)\left(\frac{m_{\pi}(x_1,\ldots,x_n)}{m_{\pi'}(x_1,\ldots,x_n)}\right)$$

Le second terme entre parenthèses ne dépend pas de θ. En faisant l'hypothèse suivante :

$$\forall i \in \{1,\ldots,M\}, \ \ \pi(\theta_{i>}0,$$

on peut donc proposer l'estimateur IS suivant pour h', qui réutilise les calculs faits pour l'estimateur $\hat{h}_M$:

$$\hat{h'}_M \quad = \quad \frac{1}{M}\sum_{k=1}^{M} \tilde{\omega}_i h(\theta_i)$$

avec

$$\tilde{\omega}_i = C\tilde{\omega}_i^* \quad \text{et} \quad \tilde{\omega}_i^* = \left(\frac{\pi'(\theta_i)}{\pi(\theta_i)}\right),$$

et C la constante de proportionnalité

$$C \quad = \quad \frac{m_{\pi}(x_1,\ldots,x_n)}{m_{\pi'}(x_1,\ldots,x_n)}$$

dont le calcul nécessite en théorie uniquement des tirages des deux priors (par Monte Carlo). On remarque, par la loi forte des grands nombres, que

$$\frac{1}{M}\sum_{i=1}^{M}\tilde{\omega}_i^* \quad \xrightarrow[n\to\infty]{p.s.} \quad \mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\tilde{\omega}_i^*\right] \ = \ \frac{1}{M}\sum_{i=1}^{M}\int_{\Theta}\pi'(\theta)\, d\theta \ = \ 1.$$

Cependant, on sait que le calcul par Monte Carlo selon des tirages *a priori* du rapport C des lois marginales peut être fortement instable. Il est donc simplement conseillé d'adopter la démarche suivante, suivant le théorème 14 :

1. calculer les poids relatifs $\tilde{\omega}_i^*$;
2. resimuler avec remise dans $\theta_1,\ldots,\theta_M$ selon les poids $\tilde{\omega}_i^*$ pour obtenir de nouveaux tirages *a posteriori* de $\pi'(\theta|x_1,\ldots,x_n)$.

Ce faisant, on crée de la corrélation entre les deux estimateurs de h.

5.2.5 Méthodes de Monte Carlo par Chaînes de Markov (MCMC)

Le principe des MCMC est de partir d'un tirage d'une densité $\tilde{\pi}_0(\theta)$ arbitraire, puis de produire une *chaîne de Markov* de réalisations $\theta^{(1)}, \ldots, \theta^{(M)}$ qui a pour loi **stationnaire** $\pi(\theta|\mathbf{x_n})$.

Définition 20 Noyau de transition. *Une chaîne de Markov homogène est déterminée par un* noyau de transition*, défini sur* $\Theta \times \mathcal{B}(\Theta)$ *à l'itération* i *par*

$$\mathcal{K}(\theta|A) \quad = \quad P(\theta^{(i)} \in A|\theta^{(i-1)} = \theta) \;=\; \int_A \underbrace{\kappa(\theta, \tilde{\theta})}_{\text{densité de transition sur } \tilde{\theta}} \quad d\tilde{\theta},$$

telle que $\mathcal{K}(.|A)$ *est mesurable* $\forall A \in \mathcal{B}(\theta)$*. Cette notion de noyau généralise au cadre continu celle de matrice de transition d'un état à un autre dans un cadre discret.*

Toute la structure d'une chaîne de Markov, que l'on considèrera toujours d'ordre 1 dans ce cours, dépend seulement du choix d'un noyau de transition et de l'état initial (ou la distribution initiale) de la chaîne, comme l'exprime la définition suivante.

Définition 21 Chaîne de Markov. *Sachant un noyau de transition* $\mathcal{K}$*, une suite* $\theta_0, \ldots, \theta_n, \ldots$ *de variables aléatoires est une chaîne de Markov d'ordre 1 si,* $\forall n \geq 0$*, la distribution de* θ_n *conditionnelle à la* $\sigma-$*algèbre (filtration) générée par* $\theta_{n-1}, \theta_{n-2}, \ldots, \theta_0$ *est la même que celle de* $\theta_n|\theta_{n-1}$ *:*

$$\pi\left(\theta_n \in \mathcal{A}|\theta_{n-1}, \theta_{n-2}, \ldots, \theta_0\right) \quad = \quad \pi\left(\theta_n \in \mathcal{A}|\theta_{n-1}\right), \;=\; \mathcal{K}(\theta_{n-1}|\mathcal{A}).$$

Des éléments fondamentaux de théorie des chaînes de Markov sont rappelés en Annexe D. Nous en retenons surtout un résultat fondamental, le *théorème ergodique*. Celui-ci nous donne "le droit", sous certaines conditions, de mener des calculs de Monte Carlo à partir de chaînes de valeurs corrélées produites par une chaîne de Markov. On parlera alors de méthode de Monte Carlo par chaîne de Markov (MCMC).

Théorème 16 (Théorème ergodique). *Si la chaîne de Markov* $(\theta_n)_{n\geq 0}$ *est réccurente positive* [a]*, alors pour toute fonction* $h : \Theta \to I\!R$ *telle que* $\mathbb{E}_\pi[|h|\,|x_1, \ldots, x_n]] < \infty$*, alors*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(\theta_i) = \int_\Theta h(\theta)\, d\Pi(\theta|x_1, \ldots, x_n).$$

Si de plus la chaîne $(\theta_n)_{n\geq 0}$ *est réversible, alors*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(h(\theta_i) - \mathbb{E}_\pi[h|x_1, \ldots, x_n]\right) \quad \xrightarrow[n\to\infty]{\mathcal{L}} \quad \mathcal{N}(0, \gamma^2)$$

avec

$$0 \;<\; \gamma^2 = \mathbb{E}_\pi\left[h^2(\theta_0)|x_1, \ldots, x_n\right] + 2\sum_{k=1}^{\infty} \mathbb{E}_\pi\left[h(\theta_0)h(\theta_k)|x_1, \ldots, x_n\right] \;<\; \infty.$$

a. Plus précisément Harris-récurrente positive, cf. Annexe D.

Caractéristiques générales d'une MCMC. À l'itération i d'une MCMC, la densité de probabilité d'un θ simulé est

$$\tilde{\pi}_i(\theta) = \int_{\hat{\theta}\in\Theta} \tilde{\pi}_{i-1}(\hat{\theta})\kappa(\hat{\theta}, \theta)\, d\hat{\theta}$$

et converge en loi vers une *unique densité stationnaire* $\tilde{\pi}_\infty(\theta)$, indépendamment de $\tilde{\pi}_0$, sous des conditions très générales de convergence et d'unicité :

- tout état (ou sous-ensemble) de Θ est accessible à partir de n'importe quel autre état (*irréductibilité*) ;
- le nombre minimal d'états intermédiaires est nul (*apériodicité*) ; cela se traduit, lorsque Θ est discret, par le fait que la chaîne ne peut pas boucler sur un ensemble d'états ;
- l'espérance du temps de retour en n'importe quel état est fini (*récurrence positive*).

Les caractéristiques majeures d'une MCMC sont les suivantes (cf. figure 2) :

- le début de la chaîne (dit *temps de chauffe*) sert à explorer l'espace Θ et trouver les zones de **haute densité** *a posteriori* ;
- on ne conserve que la seconde partie de l'ensemble des $\theta^{(i)}$ produits, qui suivent la distribution stationnaire (la chaîne "oublie" son état initial) ;
- la fréquence de visite de chaque état (ou sous-ensemble) de Θ est la même pour toute trajectoire MCMC ;
- on ajoute souvent une étape de *rééchantillonage* (SIR) ou de *décorrélation* des $\theta^{(i)}$ conservés pour obtenir un échantillon approximativement indépendant de $\tilde{\pi}_\infty(\theta)$.

Application au bayésien. Si on veut appliquer le principe des MCMC au bayésien, il faut que la loi stationnaire $\tilde{\pi}_\infty(\theta)$ soit la loi *a posteriori* $\pi(\theta|\mathbf{x_n})$. Pour cela, le noyau $\mathcal{K}$ doit être construit en fonction de la vraisemblance des données $\mathbf{x_n}$ et de l'*a priori* $\pi(\theta)$. On peut prosaïquement réutiliser la structure de l'algorithme d'Acceptation-Rejet, en créant un noyau résultant du mélange de deux actions à l'itération i :

- on accepte un nouveau candidat-tirage de $\tilde{\pi}_i(\theta)$ avec une probabilité α_i ;
- on refuse et on conserve le tirage précédent dans la chaîne avec probabilité $1-\alpha_i$.

Ce type d'algorithme a été formalisé et est connu sous le nom d'**algorithme de Hastings-Metropolis** (HM). Sous certaines conditions de conditionnement explicite *a posteriori*, on peut accepter des candidats avec probabilité 1. Ceci donne une forme particulière de MCMC, connue sous le nom d'**algorithme de Gibbs**.

Algorithme HM. La structure de l'algorithme est la suivante.

Étape i:

---

1. Simuler selon une loi instrumentale $\tilde{\theta} \sim \rho(\theta|\theta^{(i-1)})$.
2. Calculer la probabilité

$$\alpha_i \quad = \quad \min\left\{1, \left(\frac{f(\mathbf{x_n}|\tilde{\theta})\pi(\tilde{\theta})}{f(\mathbf{x_n}|\theta^{(i-1)})\pi(\theta^{(i-1)})}\right)\cdot\left(\frac{\rho(\theta^{(i-1)}|\tilde{\theta})}{\rho(\tilde{\theta}|\theta^{(i-1)})}\right)\right\}.$$

3. $$\left.\begin{array}{l}\texttt{Simuler } U \sim \mathcal{U}_{\text{unif}}[0,1]. \\ \texttt{Si } U \leq \alpha_i \texttt{ choisir } \theta^{(i)} = \tilde{\theta}. \\ \texttt{Sinon choisir } \theta^{(i)} = \theta^{(i-1)}.\end{array}\right\} \texttt{ Accepter } \tilde{\theta} \texttt{ avec probabilité } \alpha_i.$$

Le noyau markovien est alors constitué d'un mélange d'un Dirac en $\theta^{(i-1)}$ et de la loi instrumentale, mélange pondéré par la probabilité de transition α_i. La partie continue du noyau de transition (de θ vers θ') s'écrit

$$p(\theta,\theta') \quad = \quad \alpha(\theta,\theta')\rho(\theta'|\theta)$$

avec $\alpha(\theta, \theta')$ la probabilité de transition

$$\alpha(\theta, \theta') = \min\left\{1, \left(\frac{f(\mathbf{x_n}|\theta')\pi(\theta')}{f(\mathbf{x_n}|\theta)\pi(\theta)}\right) \cdot \left(\frac{\rho(\theta|\theta')}{\rho(\theta'|\theta)}\right)\right\}.$$

On a

$$\pi(\theta) \times p(\theta, \theta') = \pi(\theta') \times p(\theta', \theta).$$

La chaîne MCMC produite est alors dite *réversible* et ceci suffit à montrer, si la chaîne est irréductible et apériodique, que :

- elle est ergodique ;
- la distribution des itérés $\theta^{(i)}, \ldots, \theta^{(j)}$ de la chaîne converge en loi vers une loi-limite unique ;
- celle-ci est proportionnelle à $f(\mathbf{x_n}|\theta)\pi(\theta)$: il s'agit donc de $\pi(\theta|\mathbf{x_n})$.

L'irréductibilité peut être facilement assurée par la contrainte suivante : le support de la loi instrumentale doit inclure le support de la loi-cible.

Le rapport de Metropolis fait intervenir le rapport des lois *a posteriori* (qui permet d'ôter la constante d'intégration inconnue) : à l'étape k, si $\tilde{\theta}_k$ se situe plus haut dans la zone de haute densité de $\pi(\theta|\mathbf{x_n})$ que θ_{k-1}, ce rapport est plus grand que 1. Le rapport inverse des lois instrumentales interdit d'automatiser l'acceptation de ce nouveau point, en permettant à la chaîne de Markov d'explorer exhaustivement l'espace Θ.

Exercice 12 *Soit X la variable "débit maximal de rivière". ELle est supposée suivre une loi des extrêmes (Gumbel) de densité*

$$f(x|\theta) = \lambda\mu\exp(-\lambda x)\exp(-\mu\exp(-\lambda x)).$$

avec $\theta = (\mu, \lambda)$.

Considérons n observations $\mathbf{x_n} = (x_1, \ldots, x_n)$ supposées iid selon cette distribution.

1. *Comment s'écrit la vraisemblance ?*
2. *considère l'a priori $\pi(\mu, \lambda) = \pi(\mu|\lambda)\pi(\lambda)$ avec*

$$\begin{aligned}\mu|\lambda &\sim \mathcal{G}\left(m, b_m(\lambda)\right), \\ \lambda &\sim \mathcal{G}\left(m, m/\lambda_e\right)\end{aligned}$$

et $b_m(\lambda) = \left[\alpha^{-1/m} - 1\right]^{-1} \exp(-\lambda x_{e,\alpha})$.. Ces hyperparamètres ont le sens suivant :

- $x_{e,\alpha}$ = *quantile prédictif a priori d'ordre* α *:*

$$P\left(X < x_{e,\alpha}\right) = \int P\left(X < x_{e,\alpha}|\mu,\lambda\right)\pi(\mu,\lambda)\, d\mu d\lambda \;=\; \alpha;$$

- m = *taille d'échantillon fictif, associée à la "force" de la connaissance a priori* $x_{e,\alpha}$ *;*
- $1/\lambda_e$ = *moyenne de cet échantillon fictif.*

Pouvez-vous produire un algorithme de type MCMC qui permette de générer une loi jointe a posteriori pour (μ,λ) *?*

Réponse. En posant $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{b}_{\mathbf{x_n}}(\lambda) = \sum_{i=1}^{n} \exp(-\lambda x_i)$, la vraisemblance s'écrit alors

$$f(\mathbf{x_n}) = \lambda^n \mu^n \exp(-\lambda n \bar{x}_n) \exp\{-\mu \bar{b}_{\mathbf{x_n}}(\lambda)\}.$$

En conséquence, la loi *a posteriori* s'obtient sous la forme *hiérarchisée* suivante :

$$\pi(\mu, \lambda | \mathbf{x_n}) = \pi(\mu | \lambda, \mathbf{x_n}) \, \pi(\lambda | \mathbf{x_n})$$

où

$$\mu | \lambda, \mathbf{x_n} \sim \mathcal{G}\left(m + n, b_m(\lambda) + \bar{b}_{\mathbf{x_n}}(\lambda)\right)$$

et

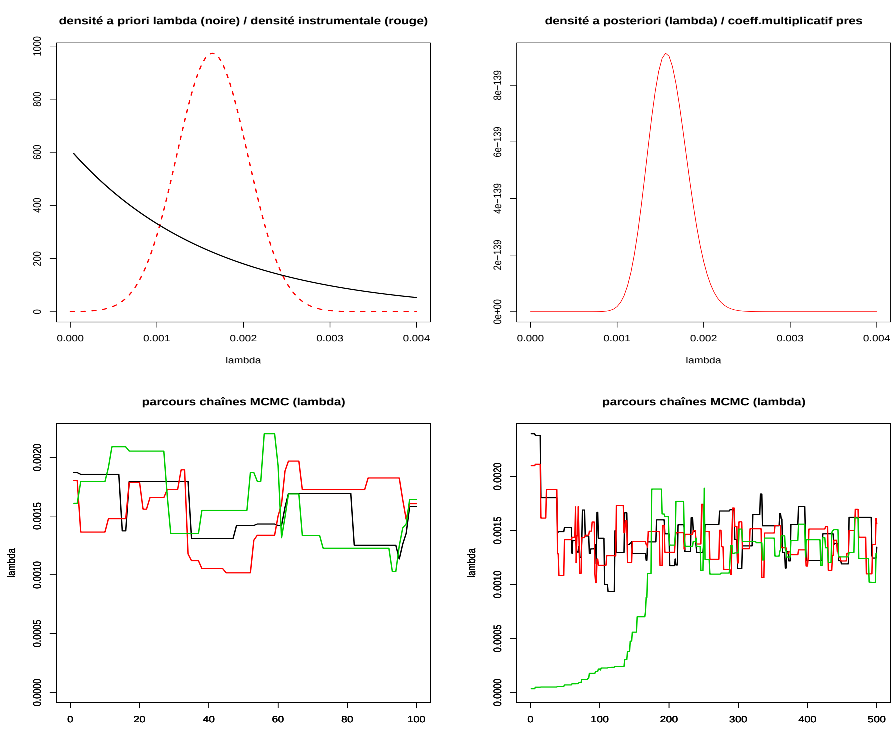
$$\pi(\lambda | \mathbf{x_n}) = \gamma(\lambda) \cdot \mathcal{G}(m + n, m/\lambda_e + n\bar{x}_n)$$

avec

$$\gamma(\lambda) \propto \frac{b_m^m(\lambda)}{\left(b_m(\lambda) + \bar{b}_{\mathbf{x_n}}(\lambda)\right)^{m+n}}$$

La loi *a priori* est donc *semi-conjuguée*, et il suffit de simuler λ *a posteriori* pour obtenir un tirage joint *a posteriori* de (μ, λ). On trace ci-dessous quelques graphiques typiquement obtenus. Plusieurs choix de lois instrumentales pour $\rho(\lambda|\lambda^{(i-1)})$ peuvent être faits. On peut notamment proposer :

- la loi *a priori* $\pi(\lambda)$;
- une loi qui “semble proche" : $\mathcal{G}_{amma}\,(m + n, m/\lambda_e + n\bar{x}_n)$;
- une loi normale de moyenne $\lambda^{(i-1)}$ et de coefficient de variation petit (5%) ou grand (25 ou 50%).



Heuristique de progression du taux d'acception moyen α. Dans une chaîne MCMC produite par HM, la *stationarité* est l'atteinte par une chaîne d'un tirage stationnaire dans la loi *a posteriori*. La rapidité de convergence vers la stationarité est induite par le taux d'acceptation α. Au début de la MCMC, on cherche à *explorer l'espace* : α grand ($\simeq 0.5$). Si α est petit, la simulation est fortement dépendante du passé de la chaîne : l'exploration de l'espace est très lente. Si α reste grand, chaque chaîne évolue solitairement et elles risque de se mélanger lentement. De différents travaux appliqués et théoriques, on a tiré une règle du pouce : un $\alpha = 0.25$ est souvent considéré, en pratique (en particulier lorsque $\dim \Theta$ est grande) comme un bon objectif de renouvellement à la stationnarité. Par ailleurs, la calibration de $\rho(\theta|\theta^{(i-1)})$ (en général, le choix de sa variance) peut être en général faite de façon **empirique** en "testant" le taux d'acceptation effectif.

Heuristique de choix d'une loi instrumentale $\rho(\theta|...)$. Dans le cas le plus simple, on choisit volontiers $\rho(\theta|\theta^{(i-1)}) = \rho(\theta)$ (*loi statique*). Mais une modélisation standard est de choisir $\rho(\theta|\theta^{(i-1)})$ centrée sur $\theta^{(i-1)}$, et donc seule la variance doit être calibrée (ou le coefficient de variation).

EXEMPLE 14. *Marche aléatoire $\theta \sim \theta^{(i-1)} + \sigma\epsilon_i$ où $\epsilon_i \sim \mathcal{N}(0,1)$.*

À la différence du noyau, le caractère markovien de ρ peut être relâché : on peut construire des ρ **adaptatives** en utilisant tout le passé de la chaîne et non pas le dernier état connu $\theta^{(i-1)}$. Il existe une très vaste littérature à ce sujet, plutôt du domaine de la recherche que de la règle du pouce ou la "boîte à outils". Là encore, on conseille de se référer à l'ouvrage [19].

Arrêt de chaîne MCMC. Une fois que le *temps de chauffe* est passée $\equiv$ la *phase ergodique* est atteinte. De nombreux diagnostics de convergence vers la stationarité ont été proposés [Cowles & Carlin 1996] et *nécessitent d'avoir lancé plusieurs chaînes parallèles*. À la stationarité, ces chaînes parallèles se sont bien mélangées et ont "oublié" le passé de chacune. Les diagnostics sont surtout *visuels* : on regarde l'évolution du comportement d'une statistique informant sur la stabilité de la distribution des θ.

Parmi ces diagnostics de convergence, les statistiques de Gelman-Rubin (1992, cas 1D) et de Brooks-Gelman (1998, cas multidimensionnel) sont très standards : elles sont ondées sur la comparaison de variances inter et intra chaînes.

Définition 22 Diagnostic de Gelman-Rubin.

- *Soit P trajectoires (chaînes) parallèles de longueur n (en pratique, $P = 3$.)*
- *Soit $\theta_k^{(i)}$ la $i^{\text{ième}}$ réalisation sur la trajectoire k.*
- *Soit B l'estimateur de la* variance de θ inter-chaînes.

$$B = \frac{n}{P-1}\sum_{k=1}^{P}\left(\bar{\theta}_k - \bar{\theta}\right)^2$$

avec

$$\bar{\theta}_k = \frac{1}{n}\sum_{i=1}^{n}\theta_k^{(i)} \quad \textit{et} \quad \bar{\theta} = \frac{1}{P}\sum_{k=1}^{P}\bar{\theta}_k$$

- *soit W l'estimateur de la* variance de θ intra-chaînes *(***ergodique***)*

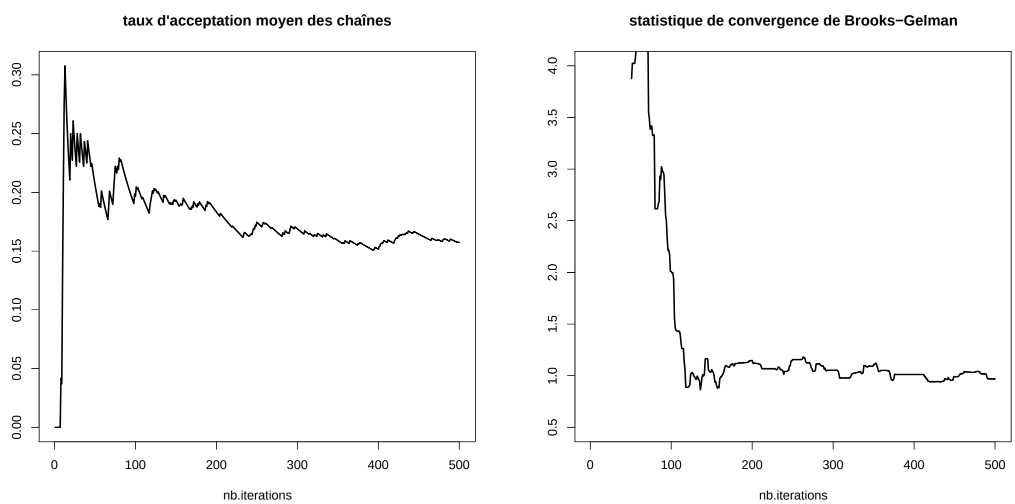
$$W = \frac{1}{P}\sum_{k=1}^{P}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\theta_k^{(i)} - \bar{\theta}_k\right)^2\right]$$

Alors, le rapport (statistique de Gelman-Rubin*)*

$$R = \frac{\frac{(n-1)}{n}W + \frac{1}{n}B}{W}$$

tends vers 1 par valeurs supérieures.

En reprenant l'exemple 12, on représente ci-dessous des exemples de trajectoires du taux d'acceptation et du diagnostic de Brooks-Gelman, qui généralise Gelman-Rubin en tenant compte de la covariance entre dimensions d'une même chaîne.



Décorrélation d'un échantillon MCMC. Soit M_c le nombre d'itérations d'une MCMC avant qu'on atteigne la stationnarité (*temps de chauffe*, cf. figure 2). En sortie de la MCMC, on obtient un échantillon de $M - M_c$ vecteurs $\theta^{(M-M_c+1)}, \ldots, \theta^M$ qui suivent la loi stationnaire $\pi(\theta|\mathbf{x_n})$.

De par le caractère markovien de la MCMC, ces valeurs peuvent être très dépendantes le long d'une chaîne. Si on dispose de beaucoup de chaînes parallèles indépendantes, il suffit de prélever une valeur dans chacune... (mais c'est peu faisable en pratique).

Pour obtenir un échantillon décorrélé (qui offre une meilleure information sur les caractéristiques de la loi $\pi(\theta|\mathbf{x_n})$), une bonne façon de faire repose sur l'usage de l'*autocorrélation*, de façon similaire au traitement des *séries temporelles*.

On peut procéder comme suit :

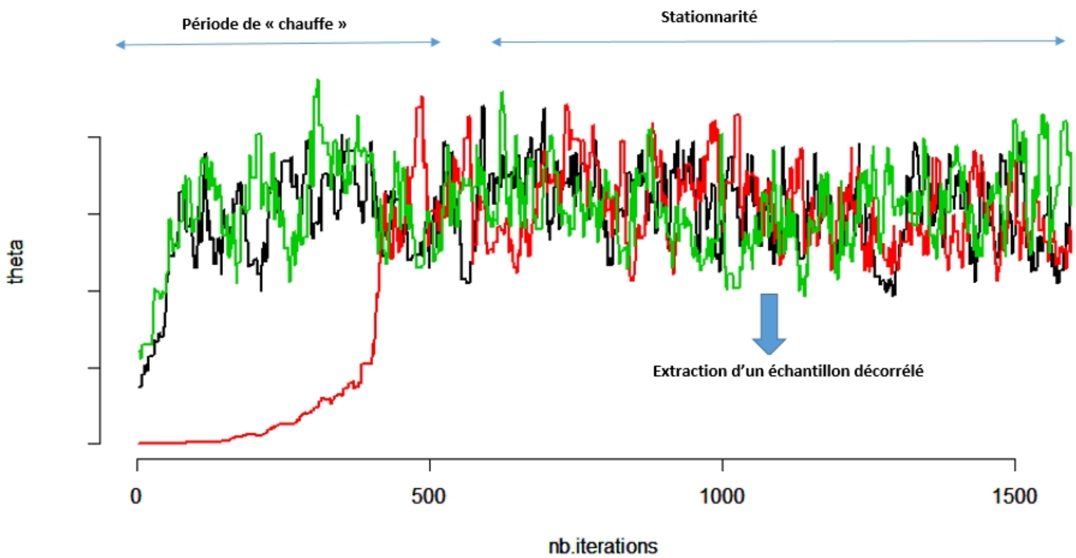


FIGURE 2 – Trois chaînes MCMC convergeant en parallèle vers la même loi-cible *a posteriori*. La période de *chauffe* correspond au nombre d'itérations nécessaire au mélange des chaînes, qui est un indicateur de l'atteinte de la stationnarité.

Procédure de décorrélation.

1. On estime l'autocorrélation des éléments d'une chaîne :

$$\text{Aut}_{i,i+j} \quad = \quad \frac{\mathbb{E}\left[\left(\theta^{(i)} - \mathbb{E}[\theta|\mathbf{x_n}]\right)\left(\theta^{(i+j)} - \mathbb{E}[\theta|\mathbf{x_n}]\right)\right]}{\text{Var}[\theta|\mathbf{x_n}]}$$

   à valeur dans $[-1, 1]$.

   - À i fixé, $\text{Aut}_{i,i+j}$ tends vers 0 lorsque j augmente $\Leftrightarrow$ $\theta^{(i+j)}$ devient de plus en plus décorrélé de $\theta^{(i)}$.
   - On considère que cette décorrélation est effective lorsque l'estimateur de $\text{Aut}_{i,i+j}$ est un **bruit blanc gaussien**.
   - On peut donc, en moyenne sur les i, estimer le nombre d'itérations nécessaire t pour obtenir 2 valeurs décorrélées de θ.

2. Sur chaque chaîne, on sélectionne le sous-échantillon (*thinning*)

$$\theta^{(M-P+1)}, \theta^{M-P+1+t}, \theta^{M-P+1+2t}, \ldots.$$

3. On baisse encore la dépendance des éléments de l'échantillon final en prélevant dans les chaînes indépendantes.

5.2.6 Échantillonneur de Gibbs et approches hybrides

Dans un cas où le paramètre $\theta = (\theta_1, \ldots, \theta_d)$ est multidimensionnel (ce qui concerne notamment pour les modèles hiérarchiques), il est recommandé d'utiliser une algorithmique d'*échantillonnage de Gibbs* [28] qui tire parti du principe de *cascade rule* (§ 5.2.2).

Étant donné un échantillon $\mathbf{x_n}$, supposons disposer des lois *a posteriori* conditionnelles

$$\begin{array}{rcl} \theta_1|\theta_2, \ldots, \theta_d, \mathbf{x_n} & \sim & \pi(\theta_1|\theta_2, \ldots, \theta_d, \mathbf{x_n}), \\ \theta_2|\theta_1, \theta_3, \ldots, \theta_d, \mathbf{x_n} & \sim & \pi(\theta_2|\theta_1, \theta_3, \ldots, \theta_d, \mathbf{x_n}), \\ \ldots & & \ldots. \end{array}$$

Alors la chaîne de Markov de vecteurs

$$\theta^{(1)} = \left(\begin{array}{c} \theta_1^{(1)} \\ \ldots \\ \theta_d^{(1)} \end{array}\right), \quad \theta^{(i)} = \left(\begin{array}{c} \theta_1^{(M)} \\ \ldots \\ \theta_d^{(M)} \end{array}\right), \ldots$$

produite par la simulation conditionnelle itérée converge également vers la loi *a posteriori*-cible $\pi(\theta|\mathbf{x_n})$, sous des conditions très générales.

Lorsque les lois *a posteriori* conditionnelles ne sont elles-mêmes pas complètement explicites, une démarche hybride, dite de *Metropolis-Hastings-within-Gibbs*, consiste à faire appel à un test de Metropolis pour chaque dimension.

Remarque 11 *La solution proposée pour l'exercice 12 est d'ailleurs un exemple d'algorithme de Gibbs en dimension 2, incluant une étape de Metropolis-Hastings pour l'estimation du paramètre λ.*

L'algorithmique de Gibbs (hybride), qui exploite au maximum la structure conditionnelle des modèles hiérarchiques, est donc très générale, et elle est notamment particulièrement intéressante lorsqu'on traite des *problèmes à données manquantes* (ex : présentant des données censurées, ou des variables latentes comme les modèles de mélange...). Dans un cadre bayésien, il permet de considérer ces données comme des paramètres inconnus à simuler *(augmentation de données)*. L'exemple suivant illustre ce procédé.

Exercice 13 (Retour à l'exercice 10).. *On suppose de nouveau connaître un échantillon $\mathbf{x_n} \sim \mathcal{N}(\theta, 1)$ composé de quelques observations $x_1, \ldots, x_{n-1}$ supposées iid de loi $\mathcal{N}(\theta, 1)$, et d'une pseudo-observation y qui est un cas-limite masquant (censurant) une observation x_n qui aurait dû être faite : $y < x_n$. On considère toujours $\theta \sim \mathcal{N}(\mu, 1)$ a priori. Pouvez-vous produire un algorithme d'échantillonage par Gibbs qui génére des réalisations de la loi a posteriori de θ ?*

ANNEXES

A Rappels : concepts et outils fondamentaux de l'aléatoire

Remarque 12 *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 $\Omega = \{1, 2, 3, \ldots, \}$) 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 ce cours 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 $I\!R^d$, 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

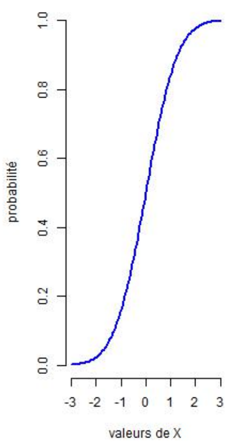
$$F(x) = \int_{-\infty}^{x} f_X(u)\, du$$

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 [7]) par sa fonction de répartition ou sa densité (figure 4).

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 4). 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.



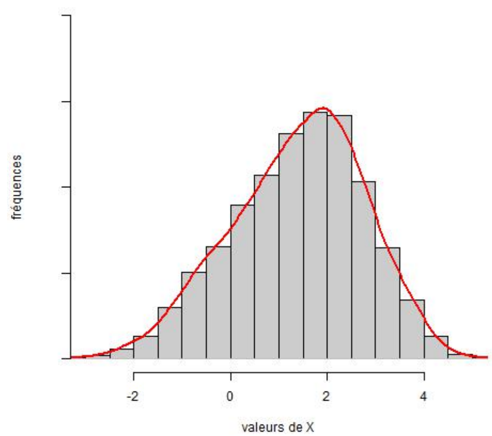


FIGURE 3 – Gauche : exemple de fonction de répartition. Droite : histogramme en fréquence de valeurs de X et densité de probabilité correspondante (courbe).

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).

7. 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

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

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 ce cours. Ces modèles seront, dans le cadre de ce cours, 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 4). 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 4). 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 4). 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*.

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

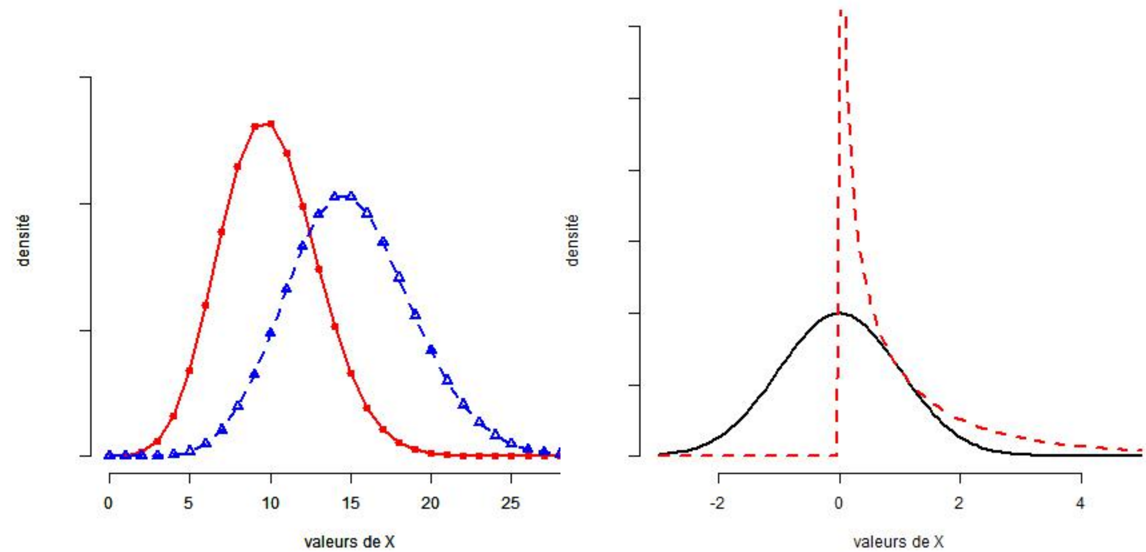


FIGURE 4 – Gauche : fonction de masse des lois discrètes binomiale $\mathcal{B}_n(100, 0.1)$ (carrés) et Poisson $\mathcal{P}(15)$ (triangles). Droite : densités de probabilité continues de la loi normale centrée réduite $\mathcal{N}(0,1)$ (courbe pleine) et χ^2_1.

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 15. **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

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

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

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

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 4 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 16. **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*

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

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

$$R_n = \sqrt{n \sum_{k=1}^M \frac{(\hat{p}_k - p_k)^2}{p_k}} \tag{-28}$$

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

Théorème 17 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*

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

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

$$R_n \quad \overset{n\to\infty}{\sim} \quad \chi^2_q. \tag{-27}$$

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 (-27) 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 (-27) doit être remplacé par

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

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

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

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

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

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

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

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

$$\rho_{i,j} = \frac{\mathbb{C}ov(X_i, X_j)}{\sqrt{\mathbb{V}[X_i]\mathbb{V}[X_j]}}. \tag{-27}$$

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 ce cours constituent un cas particulier des *processus aléatoires* (ou stochastiques) en temps discret [8], 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 23 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 24 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 5). 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 (on parle aussi de *distribution stationnaire*), 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.

8. Les processus aléatoires en temps continu ne sont pas traités dans ce cours.

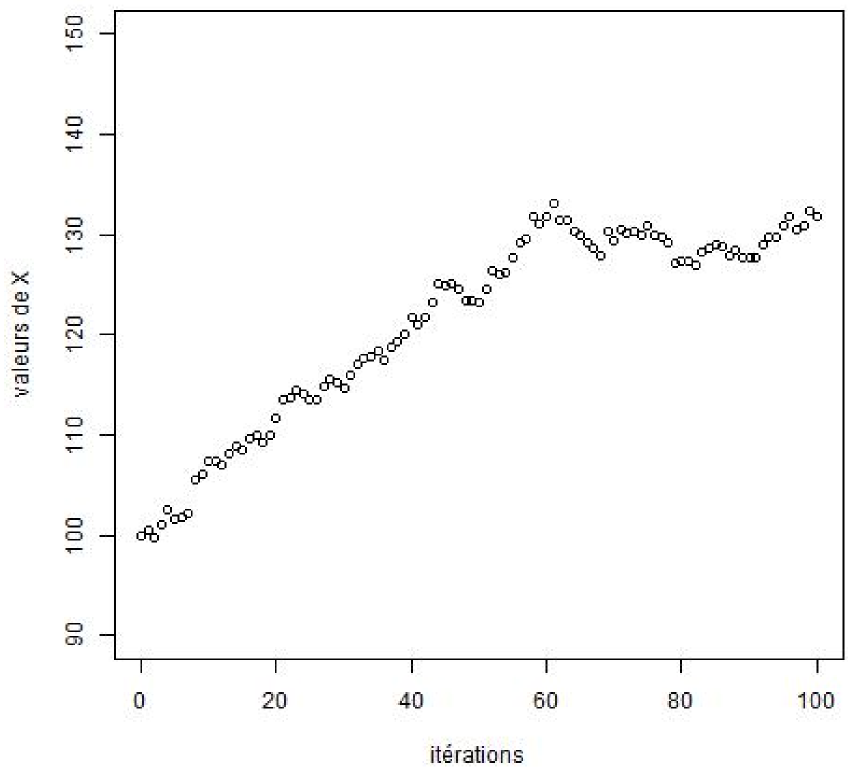


FIGURE 5 – Exemple d'une chaîne de Markov d'ordre 1 non stationnaire.

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

$$x^* \quad = \quad x + \epsilon \tag{-27}$$

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 (-27) 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 ce cours 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 18 *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 19 *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*

$$\text{EQ}(\hat{\theta}_n) = \mathbb{E}\left[(\hat{\theta}_n - \theta)^T(\hat{\theta}_n - \theta)\right], \quad (-30)$$

   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

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

   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 (-30) 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 20 Méthode Delta multivariée *[22]. Soit $\theta_1, \ldots, \theta_n$ un processus stochastique dans $I\!R^d$ et soit $g : I\!R^d \rightarrow R^q$ 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 $\Sigma \in I\!R^{2d}$. 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

$$\hat{\theta}_n \quad = \quad \arg\min_{\hat{\theta}} \widetilde{\text{EQ}}(\hat{\theta}) \quad (-29)$$

où $\widetilde{\text{EQ}}$ 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 (-29) 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 :

$$\ell(x_1, \ldots, x_n|\theta) \quad = \quad \prod_{i=1}^{n} f_X(x_i|\theta). \tag{-28}$$

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 18. **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 (-28), 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]

$$\hat{\theta}_n \quad = \quad \arg\max \ell(X_1, \ldots, X_n|\theta) \tag{-27}$$

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 19. **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 (-28) par*

$$\begin{aligned} & P\left(x_{i,\min} \leq X \leq x_{i,\max}|x_{i,\min}, x_{i,\max}\right) && \text{(-26)} \\ & \qquad = P\left(X \leq x_{i,\max}|x_{i,\max}\right) - P\left(X \leq x_{i,\min}|x_{i,\min}\right), \\ & \qquad = F(x_{i,\max}) - F(x_{i,\min}). && \text{(-26)} \end{aligned}$$

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 (-26) 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).

12. Pour des raisons de commodité, on remplace souvent ℓ par la log-vraisemblance $\log \ell$ dans la définition (-27).

Théorème 21 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),*

$$\hat{\theta}_n \quad \xrightarrow{\mathcal{L}} \quad \mathcal{N}_q\left(\theta, I_\theta^{-1}\right) \tag{-26}$$

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)

$$I_\theta^{(i,j)} \quad = \quad -\mathbb{E}_X\left[\frac{\partial \log \ell(X_1, \ldots, X_n|\theta)}{\partial\theta_i\partial\theta_j}\right]. \tag{-25}$$

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,

$$\mathbb{P}\left(\hat{\theta}_n \in C_{\hat{\theta}_n,\alpha}\right) \quad = \quad \alpha. \tag{-26}$$

En particulier, lorsqu'on s'intéresse à une dimension spécifique θ_i, le théorème 21 permet de définir l'*intervalle de confiance (asymptotique)* $1-\alpha$ associé à $\hat{\theta}_n$:

$$\mathbb{P}\left(\theta_i \in \left[\hat{\theta}_{n,i} - z_{\alpha/2}\sqrt{\sigma_{i,i}^2}\ ,\ \hat{\theta}_{n,i} + z_{\alpha/2}\sqrt{\sigma_{i,i}^2}\right]\right) \quad = \quad 1-\alpha, \tag{-25}$$

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 (-26) et (-25) 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 (-26) 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 $x \rightarrow \hat{\theta}(x)$ devrait être telle que $y=d(x) \rightarrow \hat{\theta}'(y)=h(\hat{\theta}(x))$.

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 $\mathcal{B}(n,p)$	$\mathbb{P}(X=k)=C_n^k\,p^k q^{n-k}$ $q=1-p, \quad k=0,1,\dots,n$	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	$\mathbb{P}(X=k)=C_{k+r-1}^{r-1}\,p^r q^k$ $q=1-p, \quad k=0,1,\dots$	$\dfrac{rq}{p}$	$\dfrac{rq}{p^2}$	$\left(\dfrac{p}{1-qz}\right)^r$
Pascal	$\mathbb{P}(X=k)=C_{k-1}^{r-1}\,p^r q^{k-r}$ $q=1-p, \quad k=r,r+1,\dots$	$\dfrac{r}{p}$	$\dfrac{rq}{p^2}$	$\left(\dfrac{pz}{1-qz}\right)^r$

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*	$\mathbb{E}(X)$	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 : $K_\nu(z)=\dfrac{\pi}{2}\dfrac{I_{-\nu}(z)-I_\nu(z)}{\sin\pi\nu}$ où $I_\nu(z)=\left(\dfrac{z}{2}\right)^\nu\displaystyle\sum_{n=0}^{+\infty}\frac{1}{n!\,\Gamma(n+\nu+1)}\left(\frac{z^2}{4}\right)^n$

- 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 Éléments sur la simulation pseudo-aléatoire

Les générateurs pseudo-aléatoires sont des éléments central des méthodes de simulation : elles reposent toutes sur la transformation de variables uniformes $\mathcal{U}(0, 1)$.

Définition 25 Générateur pseudo-aléatoire. *Un* générateur pseudo-aléatoire *est une transformation déterministe* Ψ *de* $]0, 1[$ *dans* $]0, 1[$ *telle que, pour toute valeur initiale* u_0 *et tout* n*, la suite*

$$\{u_0, \Psi(u_0), \Psi(\Psi(u_0)), \dots, \Psi^n(u_0)\}$$

a le même comportement statistique qu'une suite iid $\mathcal{U}(0, 1)$.

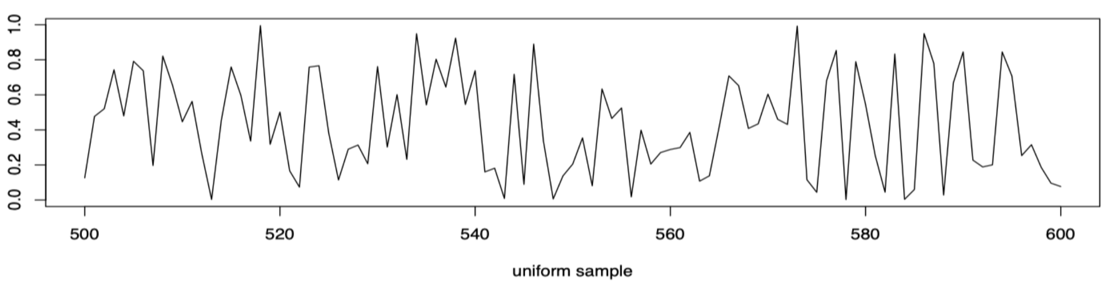
Sans appel au “hasard", la suite déterministe $(u_0, u_1 = \Psi(u_0), \dots, u_n = \Psi(u_{n-1}))$
doit ressembler à une suite aléatoire.

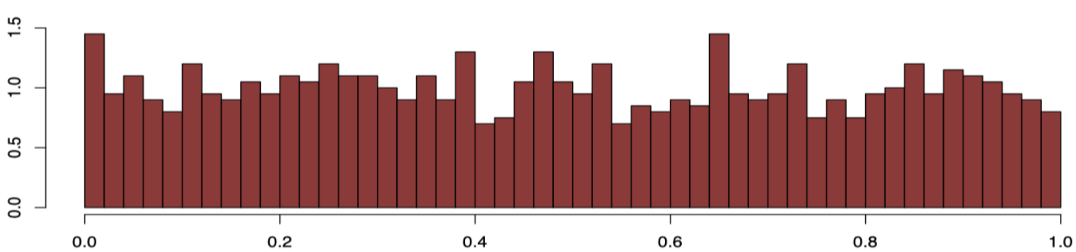
- En **Python**, il faut faire appel à la procédure `random.seed( )`

Description :
'random.seed(a=None, version=2)' generates pseudo-random values with seed 'a'.

Exemple :
u = random.seed(20)

Generally, the seed value is the previous number generated by the generator. However, When the first time you use the random generator, there is no previous value. So by-default current system time is used as a seed value.





- En **C**, il faut faire appel à la procédure `rand()` / `random()`

SYNOPSIS
\# include <stdlib.h>
long int random(void) ;

DESCRIPTION
The random() function uses a non-linear additive feedback random number generator employing a default table of size 31 long integers to return successive pseudo-random numbers in the range from 0 to RAND MAX. The period of this random generator is very large, approximately 16*((2**31)-1).

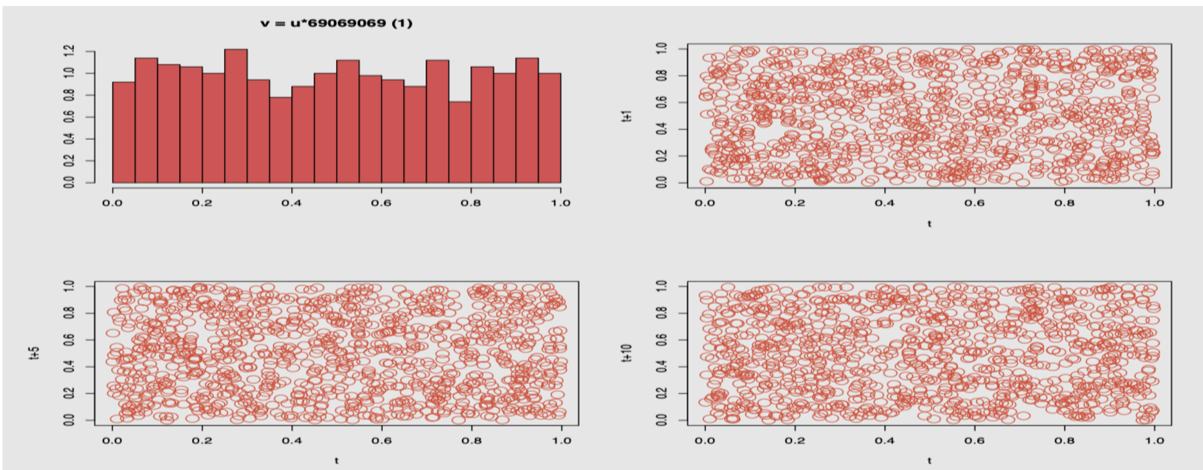
RETURN VALUE
random() returns a value between 0 and RAND MAX.

Un générateur usuel est le suivant :

Définition 26 *Générateur congruenciel Le générateur congruenciel*

$$D(x) = (ax + b) \bmod (M + 1).$$

est de période M pour les bons choix de (a, b) et se transforme en générateur sur $]0, 1[$ par division par $M + 2$.



Il faut toujours utiliser la fonction appropriée sur l'ordinateur ou le logiciel en service plutôt que de construire un générateur aléatoire de mauvaise qualité.

D Rappels sur les chaînes de Markov

Rappelons les deux définitions exprimées dans le cours.

Définition 27 Noyau de transition. *Une chaîne de Markov homogène est déterminée par un* noyau de transition*, défini sur* $\Theta \times \mathcal{B}(\Theta)$ *à l'itération* i *par*

$$\mathcal{K}(\theta|A) \quad = \quad P(\theta^{(i)} \in A|\theta^{(i-1)} = \theta) \; = \; \int_A \underbrace{\kappa(\theta, \tilde{\theta})}_{\text{densité de transition sur } \tilde{\theta}} \quad d\tilde{\theta},$$

telle que $\mathcal{K}(.|A)$ *est mesurable* $\forall A \in \mathcal{B}(\theta)$*. Cette notion de noyau généralise au cadre continu celle de matrice de transition d'un état à un autre dans un cadre discret.*

Toute la structure d'une chaîne de Markov, que l'on considèrera toujours d'ordre 1 dans ce cours, dépend seulement du choix d'un noyau de transition et de l'état initial (ou la distribution initiale) de la chaîne, comme l'exprime la définition suivante.

Définition 28 Chaîne de Markov. *Sachant un noyau de transition* $\mathcal{K}$*, une suite* $\theta_0, \ldots, \theta_n, \ldots$ *de variables aléatoires est une chaîne de Markov d'ordre 1 si,* $\forall n \geq 0$*, la distribution de* θ_n *conditionnelle à la* $\sigma-$*algèbre (filtration) générée par* $\theta_{n-1}, \theta_{n-2}, \ldots, \theta_0$ *est la même que celle de* $\theta_n|\theta_{n-1}$ *:*

$$\pi\left(\theta_n \in \mathcal{A}|\theta_{n-1}, \theta_{n-2}, \ldots, \theta_0\right) \quad = \quad \pi\left(\theta_n \in \mathcal{A}|\theta_{n-1}\right), \; = \; \mathcal{K}(\theta_{n-1}|\mathcal{A}).$$

EXEMPLE 20. **Marche aléatoire.** *Soit* $\Theta \in I\!R^p$*. La marche aléatoire (*random walk*) gaussienne est une chaîne de Markov de noyau* $\mathcal{K}(\theta|.)$ *associé à la distribution* $\mathcal{N}_p(\theta, \tau^2 I_p)$ *:*

$$\theta_{n+1} \quad = \quad \theta_n + \tau\epsilon_n \quad \textit{avec } \epsilon_n \sim \mathcal{N}(0, 1).$$

L'irréductibilité est une mesure de la sensibilité de la chaîne de Markov aux conditions initiales, qui fournit une garantie de convergence de cette chaîne : tout ensemble de Θ a une chance d'être visité par la chaîne de Markov.

Proposition 5 Irréductibilité.

- *Si* Θ *est discret, la chaîne est irréductible sur tous les états communiquent :*

$$P_\theta\left(\tau_{\theta'} < \infty\right) \quad > \quad 0 \;\; \forall(\theta, \theta') \in \Theta^2$$

  où $\tau_{\theta'}$ *est le premier temps (*> 0*) de visite de* θ'*.*

- *Si* Θ *est continu, la chaîne est irréductible pour une mesure* ψ *si,* $\forall\theta \in \Theta$ *et pour presque tout* $\mathcal{A} \in \mathcal{B}(\Theta)$ *avec* $\psi(\mathcal{A}) > 0$*,* $\forall n < \infty$

$$\mathcal{K}^n(\theta|\mathcal{A}) > 0.$$

L'irréductibilité est une condition trop faible pour être sûr que $(\theta_n)_n$ visite suffisament de fois n'importe quel sous-ensemble $\mathcal{A} \in \mathcal{B}(\Theta)$. Il faut également vérifier des conditions de stabilité pour garantir une approximation acceptable de la distribution-cible : la notion de *réccurence* formalise de telles conditions.

Pour un espace d'état Θ discret, la récurrence d'un état est équivalent à une probabilité 1 de retour certain en cet état. Elle est dite **récurrente positive** lorsque le temps moyen de retour est fini (récurrente nulle sinon). Lorsque les chaînes sont irréductibles et si Θ est fini (borné), l'irréductibilité implique la récurrence positive. De manière plus générale, on oppose la *Harris-récurrence* à la *transcience* :

Définition 29 Harris-récurrence et transcience. *Un ensemble* $\mathcal{A}$ *est* Harris-récurrent *si*

$$P_\theta\left(\eta_{\mathcal{A}} = \infty\right) \quad = \quad 1 \;\; \forall\theta \in \Theta \tag{-28}$$

où $\eta_{\mathcal{A}}$ désigne le nombre de visites dans un ensemble $\mathcal{A}$. La propriété (-28) implique que

$$\mathbb{E}[\eta_{\mathcal{A}}] \quad = \quad \infty.$$

La transcience *correspond à la propriété contraposée :*

$$\mathbb{E}[\eta_{\mathcal{A}}] \quad < \quad \infty.$$

L'étude d'une chaîne de Markov est aussi l'étude de son éventuelle *mesure invariante* π :

$$\theta_{n+1} \sim \pi \quad \text{si} \quad \theta_n \sim \pi.$$

Définition 30 Mesure invariante. π *est invariante par* $\mathcal{K}(\theta|\mathcal{A})$ *si*

$$\pi(\mathcal{A}) = \int_{\Theta} \mathcal{K}\left(\theta, \mathcal{A}\right) \ d\Pi(\theta) \quad \forall \mathcal{A} \in \mathcal{B}(\Theta).$$

On peut comprendre l'invariance de π de la façon suivante : soit $\pi_\mu(\theta_n \in \cdot)$ la loi de θ à l'étape n de la chaîne, μ désignant la loi de départ de cette chaîne. Si une mesure limite γ_μ existe telle que, $\forall \mathcal{A} \in \mathcal{B}(\Theta)$,

$$\pi_\mu(\theta_n \in \mathcal{A}) \xrightarrow[\mathcal{L}]{n\to\infty} \gamma_\mu(\mathcal{A}),$$

alors

$$\begin{aligned}
\gamma_\mu(\mathcal{A}) &= \lim_{n\to\infty} \int_{\Theta} \mu(d\theta)\mathcal{K}^n\left(\theta, \mathcal{A}\right), \\
&= \lim_{n\to\infty} \int_{\Theta} \int \mathcal{K}^{n-1}\left(\theta, d\theta\right) \mathcal{K}\left(\theta, \mathcal{A}\right), \\
&= \int_{\Theta} \gamma_\mu(d\theta)\mathcal{K}\left(\theta, \mathcal{A}\right)
\end{aligned}$$

car la convergence de $\int_\Theta \mu(d\theta)\mathcal{K}^n\left(\theta, .\right)$ implique la convergence des intégrales de fonctions mesurables bornées. Ainsi, si la chaîne de Markov converge vers une distribution limite, il s'agit d'une mesure invariante.

Cette mesure-limite, invariante (dite aussi *stationnaire*), peut présenter une propriété tout à fait intéressante : elle peut être **ergodique**, c'est-à-dire indépendante de la loi initiale μ. Cette propriété se traduit de façon générale par une **convergence en norme en variation totale** entre la mesure $\mathcal{K}^n\left(\theta, \theta\right)$ et la loi stationnaire $\pi(\theta|...)$, qui peut être raffinée dans le cas où Θ est discret. Dans le cas discret et fini, les équations de Chapman-Kolmogorov permettent d'obtenir la mesure-limite.

Définition 31 Norme en variation totale entre mesures. *Soit* (μ_1, μ_2) *deux mesures sur* $\mathcal{A}$*. Alors la norme en variation totale entre* μ_1 *et* μ_2 *est*

$$\|\mu_1 - \mu_2\|_{TV} \quad = \quad \sup_{\mathcal{A}} |\mu_1(x) - \mu_2(x)|.$$

Théorème 22 Ergodicité d'une chaînes de Markov. *Si* $(\theta_n)_n$ *est Harris récurrente positive et apériodique, alors, pour presque toute distribution initiale* μ*,*

$$\lim_{n\to\infty} \left\| \int_{\Theta} \mathcal{K}^n\left(\theta, .\right) \mu(d\theta) - \pi(.) \right\|_{TV} \quad = \quad 0$$

Cette converenge en variation totale implique que pour presque toute fonction bornée $h : \Theta \to I\!R$*,*

$$\lim_{n\to\infty} |\mathbb{E}_\mu\left[h(\theta_n)\right] - \mathbb{E}_\pi\left[h(\theta)\right]| \quad = \quad 0.$$

On déduit de ce résultat le théorème ergodique (théorème 16), qui pour permettre l'obtention d'un théorème de limite centrale nécessite que la chaîne de Markov soit *réversible* :

$$\theta_{n+1}|\theta_{n+2} = x \quad \sim \quad \theta_{n+1}|\theta_n = x.$$

E Calcul bayésien avec OpenBUGS et JAGS

E.1 Contexte de développement

WinBUGS et son successeur OpenBUGS font partie du projet BUGS (*Bayesian inference Using Gibbs Sampler*) qui vise à rendre simple la pratique des méthodes MCMC aux statisticiens. Il a été développé par l'université de Cambridge. **Seul OpenBUGS est actuellement maintenu**.

WinBUGS et OpenBUGS peuvent être utilisé de différentes manières :

- Via une interface « clique-bouton » qui permet de contrôler l'analyse,
- En utilisant des modèles définis par des interfaces graphiques, appelés `DoddleBUGS`,
- Via d'autres logiciels tels que `R` (en particulier via le package `R2WinBUGS`).

WinBUGS et OpenBUGS sont des logiciels libres et gratuits, Cependant, afin d'accéder à la version non restreinte de WinBUGS, il est nécessaire d'obtenir la clé d'utilisation. Le site internet pour WinBUGS et OpenBUGS, **The BUGS Project** présente les deux logiciels et fournit de la documentation. Le site spécialisé pour OpenBUGS est `http://www.openbugs.net/`

JAGS est une version "rapide" de BUGS développée par Martyn Plummer. Il repose sur le même language, à quelques différences subtiles près. Il a la particularité de ne pas présenter d'interface graphique pour la gestion des chaînes (mais il peut aussi être appelé par `R`). Traditionnellement, il possède moins de distributions de probabilité que sous OpenBUGS.

E.2 Un exemple “fil rouge" : le modèle bêta-binomial

Dans une parcelle, on compte le nombre d'arbres de l'espèce A. On répète l'opération en J parcelles. On veut modéliser ce processus d'échantillonage.

On se place dans une parcelle i donnée. Si on suppose que l'espèce de chaque arbre est indépendante de l'espèce de ses voisins, on peut modéliser le nombre d'arbre de l'espèce A par une loi binomiale de paramètres p et N.

$$Y_i \sim \mathcal{B}(N_i, p_i) \tag{1.1}$$

avec N_i le nombre d'arbres de la parcelle, p_i la proportion inconnue d'arbre de type A.

Si on suppose en prime que toutes les parcelles sont équivalentes et que la proportion d'arbres d'espèce A est la même dans toutes les parcelles alors on ajoute l'hypothèse

$$p_i = p \quad \text{pour tous les i} \tag{1.2}$$

Enfin si on suppose que les parcelles sont indépendantes les unes des autres, le modèle s'écrit

$$Y_i \overset{i.i.d}{\sim} \mathcal{B}(N_i, p) \tag{1.3}$$

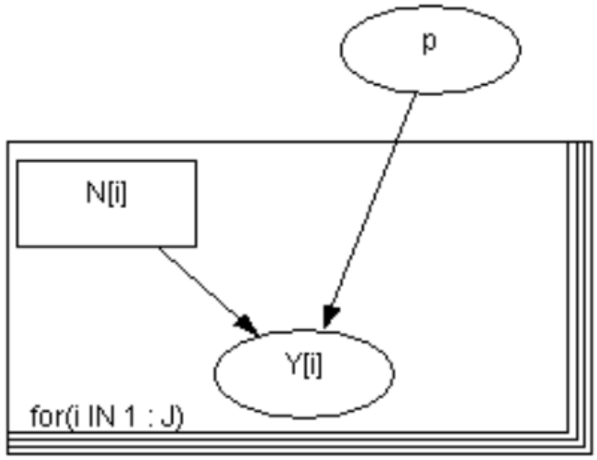


FIG. 1.1 – DAG du modèle binomial

Que sait-on sur le paramètre p qui régit la proportion des arbres de type A dans nos parcelles ?

1. A peu près rien. C'est une proportion donc ca prend des valeurs continues entre 0 et 1. on peut encoder ca au travers d'une loi de probabibilité uniforme sur $[0; 1]$.

2. on a une expérience d'autres étude set on se souvient qu'en moyenne p tourne autour de 10% mais avec une certaine variabilité. On décide d'encoder cette connaissance par une loi Beta.

$p \sim \beta(1, 9)$

3. On sait d'expérience que p ne dépasse "jamais" 0.2, et qu'en moyenne il vaut 0.1. Ca signifie que $\mathbb{E}(p) = 0.1$, or si $p \sim \beta(a, b)$ alors $\mathbb{E}(p) = a/(a + b)$, donc $b = 9a$. On "résoud" ensuite l'inégalité qui dit que p est très rarement plus grand que 0.2, i.e $\mathbb{P}(p > 0.2) < 0.001$, par exemple par dichotomie dans `R`.

On peut mener un calcul *a posteriori* par conjugaison. Avec
$p \sim \beta(a, b) \quad \text{et} \quad y|p \sim \mathcal{B}(N, p)$ alors
$p|y \sim \beta(a + y, b + N - y)$.

E.3 Fonctionnement résumé d'OpenBUGS

1. Ouverture de la fenêtre OpenBUGS
2. Création d'un fichier "modele.txt" contenant l'écriture formelle de la *vraisemblance* et de *la distribution a priori*
   - nécessité d'utiliser une boucle sur les données pour la vraisemblance

- `langage BUGS différent de R (mais plutôt compréhensible)`

3. `Création d'un fichier "data.txt" avec des données entrées en vectoriel (possibilité de tableaux)`
   - `les "données" regroupent aussi les constantes du problème : taille des données` n`, etc.`
4. `Éventuellement création d'un fichier d'initialisation pour les paramètres (chaînes MCMC)`

Testons ici l'implémentation du cas d'étude bêta-binomial. Ouvrons un fichier `model-beta-binomial.txt` (à ouvrir comme "Text") :

```
#------------------------------
# A simple binomial model
#------------------------------

model {

# Likelihood of the binomial distribution
  for (k in 1:n)
    {
      x[k] ~ dbin(p, N)
    }

# The prior of the unknown parameter
    p ~ dbeta(a,b)
}
```

1. **Vérification du modèle** via `"Model/Model Specification"` puis `"Check Model"`
   - ⇒ *model is syntactically correct*
2. **Enregistrement des données** via :
   (a) sélection du fichier `data-beta-binomial.txt`

```
list(N=10,
x=c(7, 7, 5, 8, 3, 4, 6, 4, 5, 4, 4, 4, 6, 3, 5, 1, 5, 7, 7, 3),
n=20,
a=12.5,
b=112.5
)
```

   (b) `"Load data"`
   - ⇒ *data loaded*
3. **Sélection du nombre de chaînes MCMC puis compilation du modèle** via `"compile"`
   - ⇒ *model compiled*
4. **Initialisation des chaînes MCMC** : 2 façons possibles
   (a) Sélection dans le fichier `init-beta-binomial.txt` puis `"load inits"`, chaîne par chaîne
   (b) Génération automatique via `"gen inits"` pour toutes les chaînes
      - ⇒ *model initialized*
5. **Ouverture de la fenêtre de monitoring des chaînes** via `"Inference/Sample Monitor Tool"`
   - Écrire `"p"` dans la fenêtre `"node"`, valider avec `"set"`

6. **Ouverture de la fenêtre de lancement des chaînes** via "`Model/Update`"

    - cliquer sur "`update`" pour lancer une première fois les chaînes
    - ⇒ *model is updating*

7. **Monitorer les chaînes** via la fenêtre consacrée :

    - Aller chercher "`p`" dans la fenêtre "`node`", puis cliquer sur :
    - "`trace`" pour tracer l'évolution des chaînes associées à p
    - "`trace`" pour tracer la densité *a posteriori* courante (approximative)
    - "`coda`" pour récupérer les chaînes
    - "`stats`" pour obtenir un résumé statistique de la loi *a posteriori* courante
    - etc.

E.4 Quelques détails supplémentaires concernant OpenBUGS

Menu "Infererence".

1. **Sous-menu "`Correlation`"**

    - "`Correlation Tool`"
        - `scatter` ⇒ trace un "scatterplot" entre 2 dimensions
        - `matrix` ⇒ dessine la matrice de corrélation (par niveaux de gris)
        - `print` ⇒ calcule le coefficient de corrélation linéaire

2. **Sous-menu "`Compare`"**

    - "`Comparison Tool`"
        - `boxplot` ⇒ trace une "boîte à moustaches" d'une dimension sélectionnée

Menu "Model".

- Commande *"`latex`"* : fournit le code latex du fichier sélectionné (utile pour le fichier de modèle !)

E.5 Liste des distributions de probabilités disponibles

Discrete Univariate
- Bernoulli
- Binomial
- Categorical
- Negative Binomial
- Poisson
- Geometric
- Geometric (alternative)
- Non-central Hypergeometric

Continuous Univariate
- Beta
- Chi-squared
- Double Exponential
- Exponential
- Flat
- Gamma
- Generalized Extreme Value
- Generalized F
- Generalized Gamma
- Generalized Pareto
- Generic LogLikelihood Distribution
- Log-normal
- Logistic
- Normal
- Pareto
- Student-t
- Uniform
- Weibull

Discrete Multivariate
- Multinomial

Continuous Multivariate
- Dirichlet
- Multivariate Normal
- Multivariate Student-t
- Wishart

E.6 Noeuds logiques et indexation

Les noeuds logiques sont définis par une flèche et sont **toujours indexés** :

```
mu[i] <- beta0 + beta1 * z1[i] + beta2 * z2[i] + b[i]
```

On peut utiliser une fonction de lien (log, logit, probit) :

```
logit(mu[i]) <- beta0 + beta1 * z1[i] + beta2 * z2[i] + b[i]
```

On peut définir des tableaux :

```
Y[(i + j) * k, l]
```

Par ailleurs, toute variable (noeud logique ou stochastique ~) ne peut apparaître qu'une fois dans la partie gauche d'une expression (sauf dans le cas d'une transformation de données) du type :

```
for (i in 1:N) {
   z[i] <- sqrt(y[i])
   z[i] ~ dnorm(mu, tau)
}
```

Enfin, on peut créer des noeuds multiparités : soient μ et τ deux vecteurs de taille K. On peut alors définir la boucle suivante :

```
for (i in 1 : I) {
   x[i, 1 : K] ~ dmnorm(mu[], tau[ , ])
}
```

L'aide sur les fonctions utiles est disponible ici : `http://www.openbugs.net/Manuals/ModelSpecification.html`

E.7 Pièges à éviter

Deux pièges peuvent fréquemment survenir :

- **Se tromper de paramétrisation**. Il faut toujours aller vérifier de quelle façon les distributions de probabilité sont définies ! C'est piégeant en particulier pour les lois normale et log-normale :

Log-normal

`x ~ dlnorm(mu, tau)`

$$\sqrt{\frac{\tau}{2\pi}}\frac{1}{x}\exp\left(-\tfrac{\tau}{2}(\log x-\mu)^2\right); \quad x>0$$

Normal

`x ~ dnorm(mu, tau)`

$$\sqrt{\frac{\tau}{2\pi}}\exp\left(-\tfrac{\tau}{2}(x-\mu)^2\right); \quad -\infty<x<\infty$$

- **Confondre censure et troncature**. La *censure* est possible en utilisant la notation suivante :

```
x ~ dnorm(mu, tau)C(lower, upper)
```

Il s'agit d'une censure par intervalle. On laisse un blanc à gauche (*resp.* à droite) si la donnée est censurante à droite (*resp.* à gauche). La *troncation* est possible en utilisant la notation suivante :

```
x ~ dnorm(mu, tau)T(lower, upper)
```

Notons enfin que Les priors impropres (non intégrables) ne sont pas utilisables en BUGS. Il faut les approcher avec des distributions propres mais de variance très large (ce qui est dangereux hélas). Une "règle du pouce" pour les paramètres de variance inverse est la suivante :

$$\tau \sim \text{dgamma}(0.001, 0.001)$$

mais l'usage d'OpenBUGS, JAGS, etc. nécessite donc de toujours mener des études de sensibilité *a posteriori*. Enfin, l'implémentation d'une nouvelle distribution est possible, mais il faut utiliser des subterfuges, typiquement par une transformation de variables (ex : Box-Müller pour simuler des gaussiennes).

E.8 Utilisation d'OpenBUGS avec R

Plusieurs packages sont disponibles pour appeler un code OpenBUGS depuis R : BRugs, R2WinBUGS et R2OpenBUGS. Nous recommandons l'usage du dernier, qui nécessite d'installer également le package CODA (gestion des MCMC) :

1. Installer le package R2OpenBUGS (nécessite CODA).
2. Le charger dans un programme R avec `library(R2OpenBUGS)`.
3. Définir le répertoire où se trouve les fichiers BUGS comme répertoire de travail courant, par exemple :

```
> setwd("E:/RepertoireTravail Courant/Docs Perso")
> getwd()
[1] "E:/RepertoireTravail Courant/Docs Perso"
```

`getwd()` sert à vérifier que le changement de répertoire a bien été pris en compte et doit retourner le chemin du répertoire de travail courant.

4. Spécification du modèle :

```
# Definir le nom du fichier contenant le modèle BUGS
filename <- "model-beta-binomial.txt"

# Définir les données
donnees <- list(N=10,
x=c(7, 7, 5, 8, 3, 4, 6, 4, 5, 4, 4, 4, 6, 3, 5, 1, 5, 7, 7, 3),
n=20,
a=12.5,
b=112.5
)

# Définir les paramËtres

params <- c("p")

# Initialisation des chaînes

inits <- function()
{
 list(p=0.5)
 list(p=0.1)
 list(p=0.9)
}

# Simulation des chaînes de Markov
out <- bugs(donnees,inits,params,filename,n.chains=3,debug=T,n.iter=1000,working.directory=getwd())

print(out)
plot(out)
```

5. Obtention des résultats : avec `debug=T`, fermer la fenêtre OpenBUGS qui s'est ouverte pour achever le traitement numérique. Le répertoire courant doit se présenter ainsi :

CODAchain1.txt	03/03/2016 04:07	Document texte	12 Ko
CODAchain2.txt	03/03/2016 04:07	Document texte	12 Ko
CODAchain3.txt	03/03/2016 04:07	Document texte	12 Ko
CODAindex.txt	03/03/2016 04:07	Document texte	1 Ko
data.txt	03/03/2016 04:07	Document texte	1 Ko
data-beta-binomial.txt	02/03/2016 17:30	Document texte	1 Ko
exemple-beta-binomial.r	03/03/2016 04:10	Fichier R	1 Ko
init-beta-binomial.txt	02/03/2016 17:31	Document texte	1 Ko
inits1.txt	03/03/2016 04:07	Document texte	1 Ko
inits2.txt	03/03/2016 04:07	Document texte	1 Ko
inits3.txt	03/03/2016 04:07	Document texte	1 Ko
log.odc	03/03/2016 04:07	Microsoft Office D...	15 Ko
log.txt	03/03/2016 04:07	Document texte	1 Ko
model-beta-binomial.txt	02/03/2016 17:22	Document texte	1 Ko
script.txt	03/03/2016 04:07	Document texte	1 Ko

E.9 Détails sur JAGS : exemple de script

```
% Script for the bash execution of the posterior computation

model in model.txt
data in SKJ-jags.r
compile, nchains(3)
initialize

update 10
update 10



monitor M
monitor logP0
monitor Ft
monitor Ff
monitor F
monitor selectivity.at.age
monitor recovery.rate
monitor tau2
monitor tau.selec
monitor mu.selec
monitor pR
monitor sigma

update 100

coda *
```

E.10 D'autres outils (R/Python)

`Nimble` est un outil permettant de lancer des tâches de calcul bayésien à partir de `R`. Pour `Python 3`, on pourra explorer et utiliser `Pymc`.

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.