

Modélisation et statistique bayésienne computationnelle

nicolas.bousquet@upmc.fr

Master 2, Sorbonne Université, 2019



Plan du cours

1	Introduction	6
2	Rappels de statistique	10
	• Aspects descriptifs (modélisation)	13
	• Aspects inférentiels	16
	• Aspects décisionnels liés à une estimation ponctuelle	31
	• Risque fréquentiste	36
	• Risque bayésien	39
3	Quelques autres caractéristiques du cadre bayésien	53
	• Estimateur du mode <i>a posteriori</i>	55
	• Régions de confiance et de crédibilité	56
	• Une approche rapide des tests d'hypothèse	57
	• Consistance et normalité asymptotique de la loi <i>a posteriori</i>	60
	• Interprétation subjective de la probabilité <i>a priori</i>	61
	• Une première conclusion	63
4	Rappels en statistique computationnelle	65

Plan du cours (cont.)

• Acceptation-rejet	78
• Échantillonnage d'importance (ou <i>préférentiel</i>)	84
• Méthodes de Monte Carlo par chaînes de Markov (MCMC)	88
• Algorithme de Metropolis-Hastings	
• Échantillonneur de Gibbs	
5 Récapitulatif : quand et comment faire, erreurs à éviter, principaux messages	117
6 Outils de mise en oeuvre et références	121
7 Choix du cadre bayésien	124
8 Élicitation bayésienne - présentation	148
• Élicitation par maximum d'entropie	157
• Lois <i>a priori</i> conjuguées	170
• Distributions <i>a priori</i> non-informatives	181
• Mesure de Jeffreys	
• Mesure de Berger-Bernardo	
• Mesure coïncidente	

Plan du cours (cont.)

• Convergence vers les mesures impropres	
• Exemple complet dans un cadre fiabiliste	203
• Modélisation bayésienne hiérarchique	213
• Principe	
• Exemples de dépendances statistiques apparaissant naturellement	
• Représentation des liens de causalité / conditionnement	
• Quelques soucis méthodologiques et pratiques importants	222
• Prouver l'intégrabilité de la loi <i>a posteriori</i>	
• Fusion de plusieurs <i>a priori</i>	
• Cohérence entre <i>a priori</i> et vraisemblance des données observées	
• Modélisation <i>a priori</i> informative non conjuguée	239
• Exemple d'élicitation paramétrique	
• Exemple d'élicitation par méta-analyse	
• Vers une méthodologie critique de l'élicitation	
• Analyse de sensibilité	275
• ϵ -contamination	
• <i>Exponential twisting</i>	

Plan du cours (cont.)

- Autre approche et conclusions 279

- 9 Sélection de modèle bayésien 283
 - Régions de confiance et de crédibilité 284
 - Une approche rapide des tests d'hypothèse 285
 - Facteur de Bayes 286

Contexte et objectifs du cours

- Cours tourné vers les applications réelles et l'usage pratique des statistiques bayésiennes
- Exemples théoriques et appliquées, travail computationnel avec R (Rstudio / R Markdown)
- Que signifie adopter une démarche statistique bayésienne ? Dans quel contexte ?
- Pourquoi modéliser *stochastiquement* les incertitudes caractérisant un univers risqué ?
- Comment construire une modélisation et l'enrichir ?
 - *La plupart des exemples seront issus du risque industriel et environnemental, des études de survie, des études technico-économiques, tous domaines intéressant l'assurance, la réassurance, etc.*

La statistique bayésienne

Très utile et de plus en plus utilisée en ingénierie mathématique

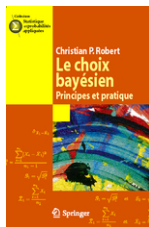
Encore relativement peu enseignée car l'une des branches les plus récemment développées des statistiques (même si historiquement la plus précoce)






Pour être bien utilisée, elle requiert d'avoir compris / assimilé

- la statistique classique (dite "fréquentiste")
- la théorie de l'information
- la théorie de la décision (ou de l'utilité)

Elle sert à modéliser et prendre une décision sous incertitude **conditionnellement** à une information transportée par des réalisations (possiblement bruitées) de lois ou processus aléatoires, et d'éventuelles autres sources d'information incertaines

De bonnes références (parmi d'autres) pour aborder le bayésien



the theory 
 that would 
 not die 
 how bayes' rule cracked
 the enigma code, 
 hunted down russian
 submarines & emerged
 triumphant from two 
 centuries of controversy
 sharon bertsch mcgraw

Rappels de statistique

(et introduction des notations)

Introduction

La Statistique peut être vue comme une théorie de la description d'un **phénomène incertain**, perçu au travers de données $\mathbf{x}_n = (x_1, \dots, x_n)$, décrites comme des **observations** d'une variable X

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 \mathbf{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

Cette prise de décision permet d'ailleurs de défendre un choix particulier de procédure inférentielle

Aspects descriptifs

La modélisation probabiliste 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 \mathbf{x}_n est représentée par une fonction de densité $f(x|\theta)$ où seul le paramètre θ (de dimension finie) est inconnu

Cette **seconde approche** est ici **privilégiée**

Deux arguments :

- un nombre fini d'observations ne peut servir à estimer qu'un nombre fini de paramètres
- l'évaluation des outils inférentiels paramétriques peut être faite avec un nombre fini d'observations

Aspects descriptifs - le cadre statistique paramétrique

On s'intéresse au comportement d'une variable aléatoire X évoluant dans un **espace mesuré et probabilisé** $(\Omega, \mathcal{A}, \mu, \mathcal{P})$ où

- ❶ Ω est l'**espace d'échantillonnage** des $X = x$, càd l'ensemble de toutes les valeurs possibles prises par X
 - on travaillera avec $\Omega = R^n$ et des échantillons *observés* $\mathbf{x}_n = (x_1, \dots, x_n)$
- ❷ la **tribu** (ou σ -algèbre) \mathcal{A} = collection des événements (sous-ensembles de Ω) mesurables par μ
 - on travaillera avec $\mathcal{A} = \mathcal{B}(R^n) = \sigma \left(\left\{ \bigotimes_{i=1}^n]a_i, b_i]; a_i < b_i \in R \right\} \right)$
- ❸ μ est une **mesure positive dominante** sur (Ω, \mathcal{A}) (Lebesgue ou Dirac dans ce cours)
- ❹ \mathcal{P} est une famille de **distributions de probabilité** dominée par μ , que suit X ,
 - supposé **paramétrique** : $\mathcal{P} = \{P_\theta; \theta \in \Theta \subset R^p\}$
 - de mesure de **densité** $f(\cdot|\theta)$, ie.

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

Aspects descriptifs - simplifions-nous la vie

Dans la suite, on parlera indifféremment 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 P_θ et $f(\cdot|\theta)$

On n'utilisera plus la notation μ qui sera induite :

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

La **vraisemblance** des données \mathbf{x}_n **conditionnelle à θ** sera notée $f(\mathbf{x}_n|\theta)$ et elle vaut

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

si tous les $x_i \stackrel{iid}{\sim} f(\cdot|\theta)$

Aspects inférentiels - l'inversion

L'inférence statistique est une **inversion** car elle cherche à déterminer

① les **causes**

- réduites au paramètre θ du mécanisme probabiliste générateur

à partir

② des **effets**

- résumés par les observations $\mathbf{x}_n = (x_1, \dots, x_n)$

alors que la modélisation caractérise le comportement des observations futures **conditionnellement** à θ

L'écriture usuelle (*fiduciaire*) de la vraisemblance témoigne de cette inversion ; on l'écrit plutôt

$$\ell(\theta|\mathbf{x}_n) = f(\mathbf{x}_n|\theta)$$

Aspects inférentiels - le théorème de Bayes

Une description générale de l'inversion des probabilités est donnée par le [théorème 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 $P(C|E)$

Une version en densité de ce théorème a été proposée par Bayes (1763)

- soit X et Y deux v.a. de lois *conditionnelle* $f(x|y)$ et *marginale* $g(y)$
- la loi conditionnelle de Y sachant $X = x$ est

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

Aspects inférentiels - modélisation bayésienne

Bayes (1763) puis Laplace (1795) ont supposé que l'**incertitude sur θ** pouvait être décrite par une distribution de probabilité de densité $\pi(\theta)$ sur Θ , appelée loi a priori

Sachant des données \mathbf{x}_n , la mise à jour de cette loi *a priori* s'opère par le conditionnement de θ à \mathbf{x}_n ; on obtient la loi a posteriori

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

Un modèle statistique bayésien

est constitué d'un **modèle statistique paramétrique $f(x|\theta)$** et d'une **distribution a priori $\pi(\theta)$** pour les paramètres

Exercice 1 : boule de billard (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 ?

Rappel (ou découverte)

Raisonner en "proportionnel" : \propto

Facilite le calcul *a posteriori*

Exercice 2 : loi gaussienne

Soit une observation $x \sim \mathcal{N}(\theta, \sigma^2)$ où σ^2 est connu

On choisit *a priori*

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

Quelle est la loi *a posteriori* de θ sachant x ?

Rappels : statistique fréquentiste (1/2)

L'idée que les paramètres θ puissent être **aléatoires** va à l'encontre du dogme généralement établi en statistiques que θ est "**inconnu, mais fixe**"

Ce dogme constitue le paradigme essentiel de la Statistique dite **fréquentiste** (ou *fréquentielle* en meilleur français)

Rappels : statistique fréquentiste (2/2)

θ est supposé **inconnu** mais **fixe**

θ est estimé par $\hat{\theta}_n = T(x_1, \dots, x_n)$ qui est une observation de la variable aléatoire appelée **estimateur** $T(X_1, \dots, X_n) \sim \mathbb{P}$

- Maximum de vraisemblance $\hat{\theta}_n = \arg \max \ell(\theta | \mathbf{x}_n)$
- Estimateur des moindres carrés
- Estimateur des moments

La validité de l'estimateur $T(X_1, \dots, X_n)$ est dépendante du caractère **reproductible** et **échangeable** de $X_1, \dots, X_n \sim P_\theta$

Elle s'exprime en termes de **région de confiance** sur θ

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

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 la Limite Centrale (TLC) :

$$\text{si } x_1, \dots, x_n \text{ sont iid} \quad \Sigma_n^{-1}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

où θ_0 est la vraie valeur inconnue du paramètre

Difficultés posés par l'inférence fréquentiste (1/4)

1 - Difficultés pratiques

- conditions non asymptotiques (n petit), données manquantes...
- d'autres formes de connaissance que les données peuvent être négligées
 - contraintes de forme, valeurs interdites pour θ ...

2 - Difficultés "philosophiques"

Sens de la probabilité : une probabilité n'est vue que comme la **limite d'une fréquence**, et une **notion de confiance** uniquement fondée sur la répétabilité des expériences peut ne pas être pertinente

Prévision : connaître parfaitement θ permet de quantifier complètement l'incertitude sur X , mais c'est une tâche impossible

1 Nature des incertitudes :

- P_θ représente la partie **aléatoire** du phénomène considéré
- l'estimation de θ souffre d'une incertitude **épistémique**, réductible si de l'information supplémentaire (données) est fournie

2 Comment incorporer les incertitudes sur θ à travers la connaissance de $\hat{\theta}_n$ pour prévoir une nouvelle valeur X_{n+1} ?

Difficultés posés par l'inférence fréquentiste (2/4)

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

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

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

L'utilisation d'un estimateur $\hat{\theta}_n$ par le statisticien fréquentiste peut **contredire le principe de vraisemblance**

Exemple (adapté de Robert 2006)

Soient x_1, x_2 deux réalisations. Nous avons deux candidats pour la loi jointe de ces observations :

- 1 la densité

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

- 2 on suppose sinon que x_1, x_2 sont de même loi gaussienne $\mathcal{N}(\theta, 1)$

Donner l'estimateur du maximum de vraisemblance. Que constate-on ?

Exemple (adapté de Robert 2006)

La vraisemblance dans les deux cas est

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

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

Difficultés posés par l'inférence fréquentiste (3/4)

Soit $\hat{\theta}$ l'estimateur du maximum de vraisemblance

Par construction, il respecte le principe de vraisemblance mais :

- peut ne pas exister (ex : modèle de Weibull à trois paramètres)
- peut ne pas être unique (modèle non-identifiable)
- peut se révéler difficile à estimer

De plus, les régions de confiance de la forme (*test du rapport de vraisemblance*)

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

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

Difficultés posés par l'inférence fréquentiste (4/4)

Une dernière difficulté apparaît lorsqu'on cherche à mener une **prévision**

Soit $\mathbf{X}_n = (X_1, \dots, X_n) \stackrel{iid}{\sim} f(\cdot|\theta)$

On cherche à prévoir le plus précisément possible ce que pourrait être le prochain tirage X_{n+1}

Dans l'approche fréquentiste, on utilise

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

et ce faisant on utilise 2 fois les données et on risque de sous-estimer les incertitudes (intervalles de confiance) en renforçant arbitrairement la connaissance

Intérêt du choix $\pi(\theta)$

La probabilisation de θ permet 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** [Aspect développé plus loin]
- 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
 - **ex** : par le biais d'experts techniques [Aspect développé plus loin]
- l'invariance $\pi(\theta|\mathbf{x}_n) = \pi(\theta)$ permet en outre d'identifier des problèmes d'**identifiabilité** du modèle

Le cadre décisionnel en statistique (1/4)

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)**

Exemples : acheter des capitaux selon leurs futurs rendement θ , déterminer si le nombre θ des SDF a augmenté depuis le dernier recensement...

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 **coût** ou **utilité** (opposé du coût)

Le cadre décisionnel en statistique (2/4)

Trois espaces dans le modèle statistique

- Ω = espace des observations x
- Θ = espace des paramètres θ
- \mathcal{D} = espace des décisions 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} = l'ensemble des fonctions de Θ dans $h(\Theta)$ où h dépend du contexte (on y reviendra)

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

Le cadre décisionnel en statistique (3/4)

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

Le cadre décisionnel en statistique (4/4)

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

On appelle **fonction de coût**

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

- Dans la pratique, le décideur construit $L(\theta, d) \geq 0$, ce qui implique qu'il n'existe pas de décision d menant à une utilité infinie
- $L(\theta, d)$ représente l'erreur due à une mauvaise évaluation de la fonction de θ d'intérêt, et on suppose donc $L(\theta, \theta) = 0$

Exemple

On considère le problème de l'estimation de la moyenne θ d'un vecteur gaussien

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

où Σ est une matrice diagonale connue avec pour éléments diagonaux σ_i^2 ($i = 1, \dots, p$)

Dans ce cas $\mathcal{D} = \Theta = R^p$ 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** $L(t) = t^2$

Ce faisant le coût est ici similaire à la log-vraisemblance négative

Le cadre décisionnel fréquentiste

Prendre une décision = **minimiser une fonction de coût**

Quand θ est inconnu, minimiser uniformément $L(\theta, d)$ en d est (souvent) impossible

Dans un contexte de gain aléatoire, l'**approche fréquentiste** propose de considérer le **coût moyen** ou **risque fréquentiste**

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

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

On appelle $\delta : \Omega \mapsto \mathcal{D}$ un estimateur et $\delta(x)$ une estimation

Difficultés

- 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

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

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

Exemple

Soient x_1 et x_2 deux observations de la loi définie par

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

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

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

appelé **coût 0-1**, qui pénalise par 1 toutes les erreurs d'estimation quelle que soit leur magnitude

Soit les estimateurs

$$\delta_1(x_1, x_2) = \frac{x_1 + x_2}{2}$$

$$\delta_2(x_1, x_2) = x_1 + 1$$

$$\delta_3(x_1, x_2) = x_2 - 1$$

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

Exemple

On trouve

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

On ne peut pas classer les estimateurs

Le cadre décisionnel bayésien

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

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

qui est l'erreur moyenne résultant de la décision d pour un x donné

On peut enfin définir le risque intégré = risque fréquentiste intégré sur les valeurs de θ selon leur distribution *a priori*

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

Associant un nombre réel à chaque estimateur δ , ce risque induit donc une relation d'ordre total sur les procédures de construction d'estimateur

Estimateur et risque de Bayes

Définition

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**

Théorème

Pour chaque $x \in \Omega$,

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

Ceci reste vrai même si $\int_{\Theta} \pi(\theta) d\theta = \infty$ (mesure *a priori* non-probabiliste) à condition que $\int_{\Theta} \pi(\theta|x) d\theta = 1$

Supériorité des estimateurs de Bayes sur les estimateurs fréquentistes (1/3)

On définit le risque minimax pour la fonction de coût L par

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

Il s'agit du **coût fréquentiste minimum dans le cas le moins favorable** (l'écart entre θ et δ , c-à-d l'erreur d'estimation, est maximal(e))

Théorème

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

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

Si elle existe, une distribution *a priori* π^* telle que $r(\pi^*) = \underline{R}$ est appelée **distribution a priori la moins favorable**

L'apport d'information *a priori* $\pi(\theta)$ ne peut qu'améliorer l'erreur d'estimation, même dans le pire des cas

Supériorité des estimateurs de Bayes sur les estimateurs fréquentistes (2/3)

Définition

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

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

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

Proposition

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

Théorème

Si un estimateur de Bayes δ^π associé à une mesure *a priori* π (probabiliste ou non) est tel que le risque de Bayes

$$r(\pi) = \int_{\Theta} R(\theta, \delta^\pi) d\theta < \infty$$

et si $\theta \mapsto R(\theta, d)$ est continu, alors δ^π est admissible

Supériorité des estimateurs de Bayes sur les estimateurs fréquentistes (3/3)

Les critères de minimaxité et d'admissibilité sont éminemment **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

Choix d'une fonction de coût $L(\theta, d)$

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 1 : fonction de coût quadratique (Legendre 1805, Gauss 1810)

Soit $\mathcal{D} = \Theta$. On pose

$$L(\theta, \delta) = (\theta - d)^2 \quad (1)$$

Critère d'évaluation le plus commun, convexe, mais pénalise très (trop) fortement les grands écarts peu vraisemblables

Justifié par sa simplicité (Gauss), le fait qu'il produit des estimateurs de Bayes intuitifs, et qu'il peut être vu comme le DL d'un coût symétrique complexe

Proposition

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

Un exemple complet (exercice de cours avec codage)

Soit x_n un échantillon de loi $\mathcal{N}(\mu, \sigma^2)$ dont la dernière valeur x_n est **censurée à droite**

On suppose *a priori* que

$$\begin{aligned}\mu | \sigma^2 &\sim \mathcal{N}(\mu_0, \sigma^2) \\ \sigma^2 &\sim \mathcal{IG}(a, b)\end{aligned}$$

Écrire une fonction qui produit, qu'on ôte x_n de l'échantillon ou non :

- des estimateurs bayésiens de (μ, σ^2) tels que on ait 80% de chance de ne pas sous-estimer ces deux paramètres *marginale*ment
- un estimateur bayésien de la prochaine valeur x_{n+1} la plus probable

Un exemple réel (formel) : création d'un système d'alerte

On s'intéresse à un évènement routier $X = x$ relevé par un système de détection (ex : Waze) vivant dans l'espace χ de dimension finie

Question : cet événement routier (bouchon, incident, accident, animal sur la voie...) est-il un bon indicateur d'un évènement $\theta \in \Theta_0$ ou $\theta \in \Theta_1$ (incidents sans gravité versus accidents nécessitant une intervention d'un opérateur routier) ?

On dispose d'un échantillon labélisé $\mathbf{e}_n = (\mathbf{x}_n, \theta_n)$

Lorsqu'une observation x apparaît, comment prévoir θ ?

Un exemple réel (formel) : création d'un système d'alerte

Le classifieur lui-même ne suffit pas à prendre une décision. Il faut se munir d'une règle de décision binaire (intervention /non intervention), opérationnelle pour l'opérateur routier

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 $\theta \in \Theta_0$ et $\delta(x) = 1$;
- $L(\theta, \delta(x)) = C_2 =$ le coût prévisionnel d'une non-intervention à tort (*erreur de 1ère espèce*), si $\theta \in \Theta_0$ et $\delta(x) = 0$;
- $L(\theta, \delta(x)) = C_3 =$ le coût prévisionnel d'une intervention à tort (*erreur de 2ème espèce*), si $\theta \notin \Theta_0$ et $\delta(x) = 1$;
- $L(\theta, \delta(x)) = C_4 = 0$ le coût (nul) d'une non-intervention à raison, si $\theta \notin \Theta_0$ et $\delta(x) = 0$.

Un exemple réel (formel) : création d'un système d'alerte

On peut alors écrire, de façon plus condensée :

$$L(\theta, \delta(x)) = [C_1\delta(x) + C_2(1 - \delta(x))] \mathbb{1}_{\{\theta \in \Theta_0\}} + C_3\delta(x) \mathbb{1}_{\{\theta \notin \Theta_0\}}.$$

Risque de Bayes :

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

Décision optimale (et non pas idéale)

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

Voir corrigé TP1

Exemple 2 : fonction de coût absolu (Laplace 1773) ou linéaire

Soit $\mathcal{D} = \Theta$ et $\dim \Theta = 1$. On pose

$$L(\theta, \delta) = |\theta - d| \quad (2)$$

ou plus généralement une fonction linéaire par morceaux

$$L_{c_1, c_2}(\theta, \delta) = \begin{cases} c_2(\theta - d) & \text{si } \theta > d \\ c_1(d - \theta) & \text{sinon} \end{cases} \quad (3)$$

Tout en étant convexes, elles croissent plus lentement que le coût quadratique et ne surpénalisent pas les erreurs grandes peu vraisemblables

Proposition

l'estimateur de Bayes associé à toute loi *a priori* π et au coût (3) est le fractile $c_2/(c_1 + c_2)$ 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)

Exemple 3 : fonction de coût 0-1

Fonction de coût non quantitative, utilisé dans l'approche classique des tests d'hypothèse

$$L(\theta, d) = \begin{cases} 1 - d & \text{si } \theta \in \Theta_0 \\ d & \text{sinon} \end{cases} \quad (4)$$

Le risque fréquentiste associé est

$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(x))] = \begin{cases} P_\theta(\delta(x) = 0) & \text{si } \theta \in \Theta_0 \\ P_\theta(\delta(x) = 1) & \text{sinon} \end{cases}$$

Proposition

l'estimateur de Bayes associé à toute loi *a priori* π et au coût (5) est

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

D'autres possibilités : les coûts **intrinsèques**

On cherche 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

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

Exemples : L^1 , Kolmogorov-Smirnov, Hellinger, Kullback-Leibler

Quelques autres caractéristiques du cadre bayésien

1 - Estimateur du mode *a posteriori*

On appelle MAP (*mode a posteriori*) l'estimateur

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

Pendant bayésien du maximum de vraisemblance

Ne dépend pas d'une fonction de coût, mais il a les **mêmes inconvénients que le maximum de vraisemblance (MV)** (non unicité, instabilité d'estimation, dépendant du choix de la mesure dominante μ sur Θ)

En outre, il ne vérifie (en général) pas la **non invariance par reparamétrisation** qui caractérise le MV

Cet estimateur, qui peut paraître intuitivement séduisant, est plutôt à éviter

2 - Régions de confiance et de crédibilité

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

Définition

Une région A de Θ est dite α -crédible si $\Pi(\theta \in A|x) \geq 1 - \alpha$

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 ∞ ,

$$P_{\theta}(\theta \in A) \geq 1 - \alpha$$

La définition bayésienne exprime la probabilité que $\theta \in A$ au vu (*conditionnellement*) des expériences déjà réalisées

- pas besoin d'avoir recours à un nombre ∞ d'expériences similaires

Une région α -crédible peut être estimée par les quantiles empiriques de la **simulation a posteriori** (voir plus loin)

3 - Une approche rapide des tests d'hypothèse

Supposons qu'on cherche à mener le test d'une *hypothèse nulle* $H_0 : \theta \in \Theta_0$

La fonction de coût $L(\theta, d)$ 0-1 est proposée dans l'approche classique de Neyman-Pearson :

$$L(\theta, d) = \begin{cases} 1 & \text{si } d \neq \mathbb{1}_{\Theta_0} \\ 0 & \text{sinon} \end{cases} \quad (5)$$

menant à l'estimateur bayésien dans $\mathcal{D} = \{0, 1\}$

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

qui fait sens intuitivement : l'*estimateur choisit l'hypothèse avec la probabilité a posteriori la plus grande*.

On peut généraliser en pénalisant différemment les erreurs suivant que H_0 est vraie ou fausse

$$L(\theta, d) = \begin{cases} 0 & \text{si } d = \mathbb{1}_{\Theta_0} \\ a_0 & \text{si } \theta \in \Theta_0 \text{ et } d = 0 \\ a_1 & \text{si } \theta \notin \Theta_0 \text{ et } d = 1 \end{cases} \Rightarrow \delta^\pi(x) = \begin{cases} 1 & \text{si } \Pi(\theta \in \Theta_0|x) > a_1/(a_0 + a_1) \\ 0 & \text{sinon} \end{cases}$$

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

Facteur de Bayes (1/2)

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

Soit $H_1 : \theta \in \Theta_1$ une hypothèse alternative

Définition

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

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

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

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

Sous le coût généralisé précédent, en posant

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

l'hypothèse H_0 est acceptée si $B_{01}(x) > (a_1\gamma_1)/(a_0\gamma_0)$

Facteur de Bayes (2/2)

En l'absence d'un cadre décisionnel véritable (qui consisterait à pouvoir fixer a_0 et a_1 , une échelle "**absolue**" a été proposée par Jeffreys (1939), remaniée depuis par Kass & Raftery (1995), pour évaluer le **degré de certitude en faveur ou au détriment de H_0 apporté par les données**

- (i) si $\Lambda = \log_{10} B_{10}(\mathbf{x}_n)$ varie entre 0 et 0.5, la certitude que H_0 est *fausse est faible*
- (ii) si $\Lambda \in [0.5, 1]$, cette certitude est *substantielle*
- (iii) si $\Lambda \in [1, 2]$, elle est *forte*
- (iv) si $\Lambda > 2$, elle est *décisive*

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

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

4 - Consistance et normalité asymptotique de la loi *a posteriori*

Théorème 1

Si $f(\cdot|\theta)$ 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

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

Théorème 2 (Bernstein-von Mises)

Soit I_θ la matrice d'information de Fisher du modèle $f(\cdot|\theta)$ 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,

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

5 - Interprétation subjective de la probabilité *a priori*

La fonction de coût et le processus décisionnel permettent de proposer une **interprétation importante de la distribution *a priori***

Elle peut être comprise comme **pari** (personnel) fait sur l'éventualité d'un évènement, et notamment un gain conditionné par l'occurrence du phénomène modélisé par $f(x|\theta)$

Cette interprétation **subjective**, proposée par de Finetti (1948), est certainement le point le plus critiqué de la démarche bayésienne

Incorporation d'information subjective *a priori*

Dans l'histoire des **théories de représentation de la connaissance**, deux grandes écoles de pensée

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

Deux axiomes en théorie bayésienne subjectiviste

- 1 La distribution *a priori* $\Pi(\theta)$ exprime un **degré de croyance** dans la proximité de θ avec θ_0 qui résume le "vrai" état caché de la nature
- 2 Un expert est considéré comme rationnel si il/elle **minimise un risque (ou coût) moyen** quand il exprime son opinion, en restant **indifférent aux effets de ce risque**

Une première conclusion

La **statistique bayésienne** est

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

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

- 1 éliciter une loi *a priori* $\pi(\theta)$ (objet de la deuxième partie de ce cours)
- 2 le coût associé aux décisions, $L(\theta, \delta)$
- 3 réaliser l'inférence *a posteriori* et produire un ou plusieurs estimateurs (objet de la troisième partie de ce cours), voire faire un choix de modèle

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

Quelques références

- ① Berger, J.O. (1985). Statistical Decision Theory and Bayesian Analysis. Springer
- ② Robert, C., P. (2006). Le choix bayésien. Principes et pratique. Springer
- ③ Parent, E., Bernier, J. (2007). Le raisonnement bayésien : modélisation et inférence. Springer
- ④ Keller, M., Pasanisi, A., Parent, E. (2012). Réflexions sur l'analyse d'incertitudes dans un contexte industriel : information disponible et enjeux décisionnels. Journal de la Société Française de Statistiques

Rappels - Méthodes numériques pour l'estimation bayésienne

Estimation bayésienne

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 $\mathbf{x}_n = (x_1, \dots, x_n) \sim f(.|\theta)$

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

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

Si, par exemple, on choisit un regret (ou un coût) **quadratique** $L(\theta, \delta) \propto (\delta - \theta)^2$, on a vu que

$$\delta^\pi = \mathbb{E}[\theta | \mathbf{x}_n] = \int_{\Theta} \theta \pi(\theta | \mathbf{x}_n) d\theta \quad (\text{moyenne } a \text{ posteriori})$$

Si, avec $\Theta = \Delta = R$, on choisit un autre regret, **impactant différemment la sur-estimation ou la sous-estimation** de θ , comme

$$L(\delta, \theta) = |\delta - \theta| \left(c_1 \cdot \mathbb{1}_{\{\delta < \theta\}} + c_2 \cdot \mathbb{1}_{\{\delta > \theta\}} \right)$$

on obtient que δ^π est le fractile *a posteriori* d'ordre $\alpha = c_1 / (c_1 + c_2)$:

$$\int_{-\infty}^{\delta^*} \pi(\theta | \mathbf{x}_n) d\theta = \frac{c_1}{c_1 + c_2}$$

Dans chaque cas, deux questions importantes :

- ❶ peut-on obtenir une expression explicite pour δ^π ?
- ❷ sinon, comment peut-on l'évaluer numériquement ?

En général...

Pas de caractère explicite, dans la très grande majorité des cas, 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** :

- Newton-Cotes, Runge-Kutta, ...
- instabilité numérique lorsque $\dim \Theta$ augmente

Une façon intéressante (voire même indispensable) de procéder est d'utiliser des **simulations** de la loi *a posteriori*

Estimation non-paramétrique : soit un échantillon *indépendent et identiquement distribué* (iid)

$$\theta_1, \dots, \theta_M \stackrel{iid}{\sim} \pi(\theta | \mathbf{x}_n)$$

Moyenne *a posteriori*. On peut estimer δ^π par un **estimateur de Monte Carlo**

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

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

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

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

soit en prenant

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

D'autres intérêts de la simulation *a posteriori*

Obtenir de façon non paramétrique des régions α —crédibles *a posteriori*

Calculer des facteurs de Bayes

Résoudre des problèmes de classification

Un dernier intérêt (une nécessité) de la simulation *a posteriori*

Plaçons-nous dans un cadre d'**analyse prédictive** : **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 : 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.

Simulation prédictive *a posteriori*

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

Attention : des valeurs simulées selon $f(x|\mathbf{x}_n)$ ne constitue JAMAIS un échantillon i.i.d. : elles sont toutes interdépendantes puisqu'elles dépendent toutes de \mathbf{x}_n

On remarque d'ailleurs que la densité jointe d'un tel échantillon n'est pas le produit des densités de chacun des tirages

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 :

$$\begin{aligned} X_{n+1} &\sim f(x|\mathbf{x}_n) \\ X_{n+2} &\sim f(x|\mathbf{x}_n, x_{n+1}) \end{aligned}$$

etc.

Simulation bayésienne indirecte

Simuler selon la loi *a posteriori* $\pi(\theta|\mathbf{x}_n)$ n'est en règle générale pas possible directement

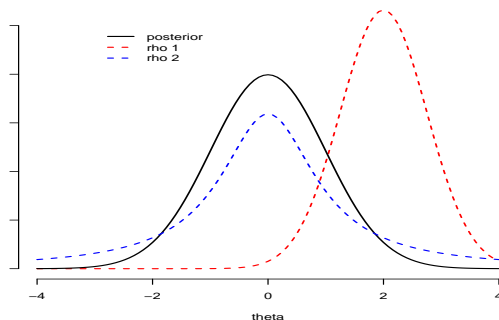
Il faut donc utiliser des techniques qui font appel à de la **simulation indirecte** :

- 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

Notion de densité instrumentale $\rho(\theta)$ (1/2)

- ❶ La densité $\rho(\theta)$ doit être facilement simulable (ex : mélanges gaussiens si $\pi(\theta|\mathbf{x}_n)$ est multimodale...)
- ❷ Le **support**¹ de $\rho(\theta)$ contient **nécessairement** celui de $\pi(\theta|\mathbf{x}_n)$
- ❸ Les queues de $\rho(\theta)$ devraient être plus lourdes que celles de $\pi(\theta|\mathbf{x}_n)$



- ❹ 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 $\rho(\theta)$

Notion de densité instrumentale $\rho(\theta)$ (2/2)

Un candidat logique peut parfois être la loi *a priori* $\pi(\theta)$, car elle respecte automatiquement la règle d'inclusion du support

- ❶ 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 $\rho(\theta)$
- ❷ Si l'*a priori* est très large (peu informatif) :
 - 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* \Rightarrow **coût algorithmique très fort**
- ❸ Ce choix est aussi à proscrire si l'*a priori* privilégie des régions de Θ qui sont éloignées de celles privilégiées par les données
 - **indication** : éloignement du mode *a priori* de θ et du maximum de vraisemblance $\hat{\theta}_n$

Méthodes d'échantillonnage dans la loi *a posteriori*

On cherche à obtenir **indirectement** des tirages qui suivent (en général approximativement) la loi *a posteriori* :

- 1 algorithmes d'acceptation-rejet
- 2 échantillonnage d'importance (préférentiel)
- 3 méthodes de Monte Carlo par chaînes de Markov (MCMC)
- 4 filtrage particulière

On présente ici les trois premières méthodes d'échantillonnage (le filtrage étant plus délicat d'utilisation, et plutôt adapté à des modèles à espace d'états)

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)

Algorithmes d'acceptation-rejet (1/2)

Permet de simuler de façon **exacte** et **indépendante** selon la loi *a posteriori*

Hypothèse supp. sur $\rho(\theta)$: $0 < K = \sup_{\theta \in \Theta} \frac{f(\mathbf{x}_n|\theta)\pi(\theta)}{\rho(\theta)} < \infty$

Algorithme :

- ❶ **simulation indirecte** : soit $\theta_i \sim \rho(\cdot)$
 - ❷ **test** :
 - soit $U_i \sim \mathcal{U}_{unif}[0,1]$
 - si $U_i \leq \frac{f(\mathbf{x}_n|\theta_i)\pi(\theta_i)}{K\rho(\theta_i)}$ alors θ_i suit la loi $\pi(\theta|\mathbf{x}_n)$
-

(Preuve en cours)

Quelques commentaires (1)

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 $K \cdot C =$ espérance du nombre de tirages nécessaires avant l'acceptation

Optimiser l'algorithme revient donc à diminuer K

Exemple : perturbation d'un modèle conjugué (1/3)

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

- quelques observations x_1, \dots, 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 : $y < x_n$

La vraisemblance s'écrit

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

A priori, on suppose $\theta \sim \mathcal{N}(\mu, 1)$

L'*a posteriori* sur θ s'écrit alors

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

Exemple : perturbation d'un modèle conjugué (2/3)

On reste proche d'une loi normale : $\rho(\theta) \equiv \mathcal{N} \left(\frac{1}{n} \left(\mu + \sum_{k=1}^{n-1} x_k \right), 1/n \right)$

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

$$\tilde{\pi}(\theta | \mathbf{x}_n) \leq \underbrace{\sqrt{\frac{2\pi}{n}}}_{K} \cdot \{1 - \Phi(y - \theta)\} \cdot \rho(\theta)$$

Mise en oeuvre : 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

Mise en oeuvre sur R [fichier "exemple-acceptation-rejet.r"]

- validation graphique
- comparaison avec le choix $\rho(\theta) = \pi(\theta)$

Exemple : perturbation d'un modèle conjugué (3/3)

Si on fait le choix $\rho(\theta) = \pi(\theta)$, alors

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

Quelques commentaires (2)

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

Principe parfait en théorie, mais en pratique réservé aux cas simples ($\dim \Theta$ petite)

Algorithme très coûteux en temps d'attente en général

Echantillonnage d'importance (ou *préférentiel*)

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

Soit $(\omega_1, \dots, \omega_M)$ 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 [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 $M \rightarrow \infty$, par

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

Sampling-Importance Resampling (SIR)

Théorème [Rubin 1988]. Les tirages

$$\tilde{\theta}_1, \dots, \tilde{\theta}_P \sim \mathcal{M}_{\text{multinomial}}(\theta_1, \dots, \theta_M | \omega_1, \dots, \omega_M)$$

suivent la loi *a posteriori* $\pi(\theta | \mathbf{x}_n)$

Il faut cependant noter qu'ils sont **fortement dépendants** ([tirage avec remise](#))

En pratique, l'heuristique de Rubin consiste à prendre $P < M/20$ pour diminuer la dépendance

On peut aussi ainsi estimer les caractéristiques de $\pi(\theta | \mathbf{x}_n)$

Exemple : perturbation d'un modèle conjugué

On reprend l'exemple précédent

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

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 $1 - \Phi(y - \theta_i)$, $i = 1, \dots, M$

Les poids les plus hauts sont donc ceux pour lesquels $y \ll \theta_i$

Mise en oeuvre sur R [fichier "exemple-IS.r"]

- validation graphique
- comparaison avec le choix $\rho(\theta) = \pi(\theta)$

Quelques commentaires

- ➊ Plus la densité instrumentale $\rho(\theta)$ est "proche" de $\pi(\theta|\mathbf{x}_n)$, plus les poids sont équilibrés (donc meilleur est le rééchantillonnage)
- ➋ 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 rééchantillonnage [Marin & Robert : The Bayesian Core]

Méthodes de Monte Carlo par chaînes de Markov (MCMC)

Principe

Partant d'un tirage d'une densité $\tilde{\pi}_0(\theta)$ arbitraire, on produit une **chaîne de Markov** de réalisations $\theta^{(1)}, \dots, \theta^{(M)}$ qui a pour loi **stationnaire** $\pi(\theta|\mathbf{x}_n)$

Soit $A \in \Theta$. Une chaîne de Markov homogène est déterminée par un **noyau de transition**, défini à l'itération i par

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

qui généralise la matrice de transition d'un état à un autre dans un cadre discret

La densité de probabilité d'un θ simulé à l'itération i 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

Convergence des MCMC

Conditions 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*)
- l'espérance du temps de retour en n'importe quel état est fini (*récence positive*)

On dit alors que la MCMC produite est **ergodique**

Caractéristiques majeures :

- 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 **rééchantillonnage** (SIR) ou de **décorrélation** des $\theta^{(i)}$ conservés pour obtenir un échantillon approximativement indépendant de $\tilde{\pi}_{\infty}(\theta)$

Application au bayésien

Pour que la loi stationnaire $\tilde{\pi}_\infty(\theta)$ soit la loi *a posteriori* $\pi(\theta|\mathbf{x}_n)$, 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 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 $1 - \alpha_i$
 - \Rightarrow Algorithme de Hastings-Metropolis

Sous certaines conditions de conditionnement explicite *a posteriori*, on peut accepter des candidats avec probabilité 1

- \Rightarrow Algorithme de Gibbs

L'algorithme de Metropolis-Hastings (cas bayésien)

Soit une densité instrumentale $\rho(\theta|\theta^{(i-1)})$

Étape i :

- 1 simuler $\tilde{\theta} \sim \rho(\theta|\theta^{(i-1)})$
- 2 calculer la probabilité

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

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

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

Propriétés particulières

La partie continue du noyau de transition (de θ vers θ') s'écrit

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

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 MCM 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 $\theta^{(i)}, \dots, \theta^{(j)}$ 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 $\pi(\theta | \mathbf{x}_n)$

Plus loin dans le cours

Comment contrôler et stopper les chaînes ?

Comment produire une ou des lois instrumentales utiles ?

Comment produire un échantillon décorrélé utile ?

Exemple général : comportement d'un débit maximal

Soit X la variable "débit maximal de rivière"

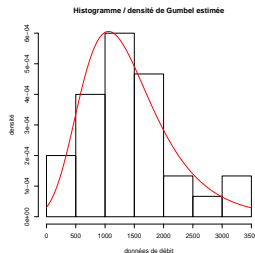
Une loi des extrêmes (Gumbel) est souvent indiquée pour modéliser X :

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

avec $\theta = (\mu, \lambda)$. L'espérance est

$$\mathbb{E}[X|\theta] = \lambda^{-1} (\log \mu + \gamma)$$

où γ est la constante d'Euler ($\simeq 0.578..$)



Vraisemblance des observations

Soient n observations $\mathbf{x}_n = (x_1, \dots, x_n)$ supposées i.i.d. selon Gumbel(μ, λ)

On pose

$$\begin{aligned}\bar{x}_n &= \frac{1}{n} \sum_{i=1}^n x_i \\ \bar{b}_{\mathbf{x}_n}(\lambda) &= \sum_{i=1}^n \exp(-\lambda x_i)\end{aligned}$$

La vraisemblance s'écrit alors

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

Choix *a priori*

On considère l'*a priori* $\pi(\mu, \lambda) = \pi(\mu|\lambda)\pi(\lambda)$ avec

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

et $b_m(\lambda) = [\alpha^{-1/m} - 1]^{-1} \exp(-\lambda x_{e,\alpha})$.

Sens des hyperparamètres :

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

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

- m = taille d'échantillon fictif, associée à la "force" de l'avis d'expert $x_{e,\alpha}$
- $1/\lambda_e$ = moyenne de cet échantillon

Loi *a posteriori*

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

$$\mu | \lambda, \mathbf{x}_n \sim \mathcal{G}_{amma}(m + n, b_m(\lambda) + \bar{b}_{\mathbf{x}_n}(\lambda))$$

et

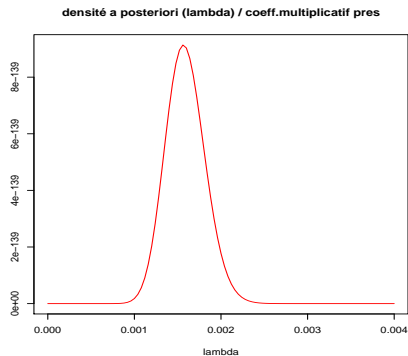
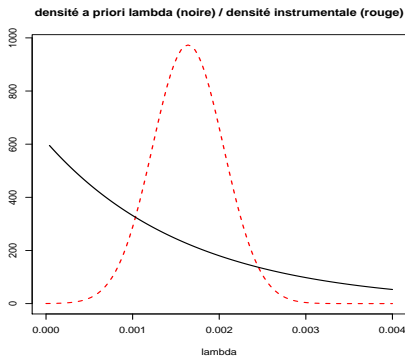
$$\pi(\lambda | \mathbf{x}_n) = \gamma(\lambda) \cdot \mathcal{G}_{amma}(m + n, m/\lambda_e + n\bar{x}_n)$$

avec

$$\gamma(\lambda) \propto \frac{b_m^m(\lambda)}{(b_m(\lambda) + \bar{b}_{\mathbf{x}_n}(\lambda))^{m+n}}$$

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)

Illustration ($m = 1$ puis $m = 10$, $x_{e,0.5} = 2000$, $\lambda_e = 1/610$)



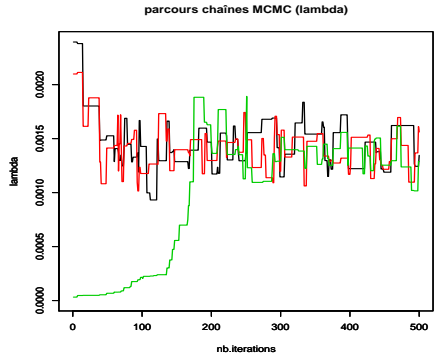
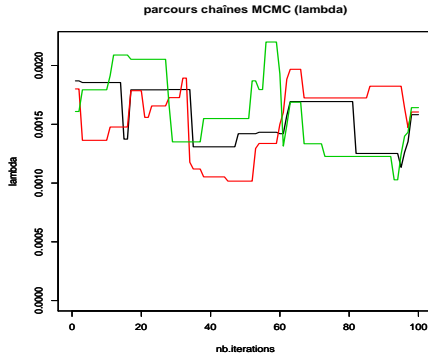
Mise en oeuvre sur R [fichier "exemple-MCMC-complet.r"]

Impact de plusieurs choix pour $\rho(\lambda|\lambda^{(i-1)})$

On peut proposer

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

Illustration ($m = 1$ puis $m = 10$, $x_{e,0.5} = 2000$, $\lambda_e = 1/610$, $\rho =$ marche aléatoire)



Heuristique de progression du taux d'acceptation moyen α

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 risquent de se mélanger lentement

Stabilisation : 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.

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

Sélection de la loi instrumentale $\rho(\theta|...)$

Dans le cas le plus simple, $\rho(\theta|\theta^{(i-1)}) = \rho(\theta)$ (**loi statique**)

Une modélisation standard est de choisir $\rho(\theta|\theta^{(i-1)})$ centrée sur $\theta^{(i-1)}$, et donc seule la variance doit être calibrée

Exemple : marche aléatoire $\theta \sim \theta^{(i-1)} + \sigma\epsilon_i$ où $\epsilon_i \sim \mathcal{N}(0, 1)$

À la différence du noyau \mathcal{K} , 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 $\theta^{(i-1)}$

Une (très) vaste littérature à ce sujet, plutôt du domaine de la recherche que de la règle du pouce ou la "boîte à outils" [Roberts & Rosenthal, Moulines et al.]

Arrêt des chaînes MCMC (1/2)

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

Statistiques de Gelman-Rubin (1992) et Brooks-Gelman (1998) :

- fondées sur la comparaison de variances inter et intra chaînes
- les plus utilisés en pratique
- Gelman-Rubin (chaînes 1D), Brooks-Gelman = généralisation

Arrêt des chaînes MCMC (2/2)

- 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**

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

avec

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

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

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

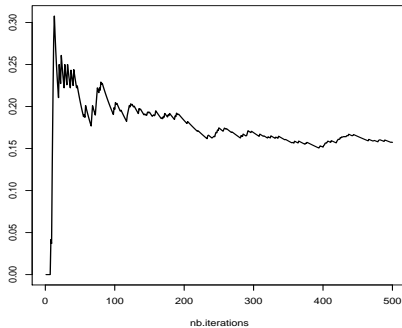
Alors, le rapport (**Statistique de Gelman-Rubin**)

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

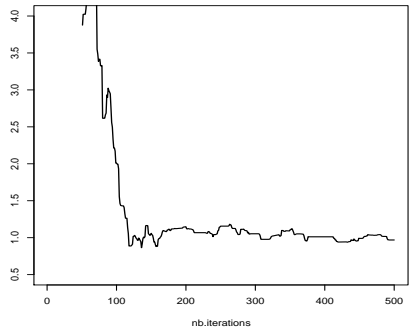
tends vers 1 par valeurs supérieures

Illustration ($m = 10$, $x_{e,0.5} = 2000$, $\lambda_e = 1/610$, $\rho =$ marche aléatoire)

taux d'acceptation moyen des chaînes



statistique de convergence de Brooks–Gelman



Comment décorréler un échantillon *a posteriori* émanant d'une MCMC ?

Soit M_c le nombre d'itérations d'une MCMC avant qu'on atteigne la stationnarité (*temps de chauffe*)

En sortie de la MCMC, on obtient un échantillon de $M - M_c$ vecteurs $\theta^{(M-M_c+1)}, \dots, \theta^M$ 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 beaucoup de chaînes parallèles indépendantes, il suffit de prélever une valeur dans chacune... (peu faisable en pratique)

Obtention d'un échantillon décorrélé

- ❶ On estime l'**autocorrélation** des éléments d'une chaîne :

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

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

- ❷ Sur chaque chaîne, on sélectionne le sous-échantillon

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

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

L'algorithme (*échantillonneur*) de Gibbs

On suppose pouvoir écrire $\theta = (\theta_1, \theta_2, \dots, \theta_d)$

On suppose pouvoir facilement simuler les lois *a posteriori conditionnelles*

$$\begin{aligned}\theta_1^{(i)} &\sim \pi(\theta_1 | \mathbf{x}_n, \theta_2^{(i-1)}, \dots, \theta_d^{(i-1)}) \\ \theta_2^{(i)} &\sim \pi(\theta_2 | \mathbf{x}_n, \theta_1^{(i)}, \theta_3^{(i-1)}, \dots, \theta_d^{(i-1)}) \\ &\vdots \\ \theta_d^{(i)} &\sim \pi(\theta_d | \mathbf{x}_n, \theta_1^{(i)}, \dots, \theta_{d-1}^{(i)})\end{aligned}$$

alors la chaîne markovienne de vecteurs

$$\theta^{(1)} = \begin{pmatrix} \theta_1^{(1)} \\ \vdots \\ \theta_d^{(1)} \end{pmatrix}, \quad \theta^{(i)} = \begin{pmatrix} \theta_1^{(i)} \\ \vdots \\ \theta_d^{(i)} \end{pmatrix}, \dots$$

est de loi stationnaire $\pi(\theta | \mathbf{x}_n)$

Quelques caractéristiques importantes

- Ne s'applique qu'aux cas multidimensionnels
- Exploite au maximum la structure conditionnelle des modèles hiérarchiques
- Très utile pour simuler des modèles conjugués en dimension au moins égale à 2

Converge en général plus vite qu'une MCMC (*temps de chauffe* moins long)

Il permet souvent de faciliter l'estimation des modèles à **données manquantes** (ex : censures)

- en considérant ces données comme des paramètres inconnus à simuler (*augmentation de données*)
- ce qui permet de retomber, parfois, dans des cas conjugués

Exemple : perturbation d'un modèle conjugué (1/2)

On reprend l'exemple $\mathbf{x}_n = (\underbrace{x_1, \dots, x_{n-1}}_{\mathcal{N}(\theta, 1)}, \underbrace{y}_{\substack{\text{censure} \\ \text{à droite}}})$ avec $\theta \sim \mathcal{N}(\mu, 1)$ *a priori*

Si on connaissait x_n , le modèle bayésien serait conjugué et

$$\theta | \mathbf{x}_n \sim \mathcal{N} \left(\frac{1}{n} \left(\mu + \sum_{i=1}^n x_i \right), (n+1)^{-1} \right)$$

On considère alors la donnée manquante x_n comme un paramètre inconnu et aléatoire.

Sachant θ et \mathbf{x}_n , on peut montrer par la règle de Bayes que la loi de x_n est la normale tronquée

$$\mathcal{N}(\theta, 1) \cdot \mathbb{1}_{\{x_n \geq y\}}$$

Éléments de preuve (dans un cas général)

Soit la variable aléatoire X_n dont x_n est une observation

La fonction de répartition de X_n est conditionnelle : $P(X_n < x | X_n > y)$

Par la règle de Bayes

$$P(X_n < x | X_n > y) = \frac{P(X_n < x \cap X_n > y)}{P(X_n > y)} = \frac{P(y < X_n < x)}{P(X_n > y)}$$

Le dénominateur est une constante (indépendante de x). Donc

$$P(X_n < x | X_n > y) \propto \int_y^x f(u) du = \int_{-\infty}^x f_X(u) \mathbb{1}_{\{y \leq u\}} du$$

où f_X est la densité d'un X non-contraint (ici gaussienne)

On en déduit que la densité de X_n est

$$f_{X_n}(x) = \frac{f(x) \mathbb{1}_{\{y \leq x\}}}{\int_{-\infty}^{\infty} f(u) \mathbb{1}_{\{y \leq u\}} du}$$

Exemple : perturbation d'un modèle conjugué (2/2)

Schéma de Gibbs :

- On part d'une valeur $\theta^{(0)}$
- Itération $i \geq 1$:
 - 1 on simule $x_n^{(i)} \sim \mathcal{N}(\theta^{(i-1)}, 1) \cdot \mathbb{1}_{\{x_n \geq y\}}$
 - 2 on simule $\theta^{(i)} \sim \mathcal{N}\left(\frac{1}{n} \left(\mu + \sum_{i=1}^{n-1} x_i + x_n^{(i)} \right), (n+1)^{-1}\right)$

Mise en oeuvre sur R [fichier "exemple-gibbs.r"]

Un problème fréquent posé par Gibbs : une vision trop "conditionnelle"

La modélisation bayésienne par conditionnement peut fréquemment entraîner le mécanisme suivant :

- 1 On construit un *a priori* hiérarchique

$$\pi(\theta) = \pi(\theta_1|\theta_2, \theta_3)\pi(\theta_2|\theta_3)\pi(\theta_3)$$

avec des *a priori* non-informatifs

- 2 Ce conditionnement est souvent choisit pour tirer parti de conjugaisons *a posteriori* : les lois conditionnelles

$$\pi(\theta_1|\mathbf{x}_n, \theta_2, \theta_3),$$

$$\pi(\theta_2|\mathbf{x}_n, \theta_1, \theta_3),$$

$$\pi(\theta_3|\mathbf{x}_n, \theta_1, \theta_2)$$

sont explicites, ce qui permet d'utiliser un algorithme de Gibbs

Problème majeur : même si ces lois *a posteriori* conditionnelles sont propres, la loi jointe peut ne pas l'être :

$$\int_{\Theta} \pi(\theta|\mathbf{x}_n) d\theta = \infty$$

TP : modèle à effets aléatoires autour d'une constante (Hobert-Casella) (1/2)

Pour $i = 1, \dots, I$ et $j = 1, \dots, J$

$$x_{ij} = \beta + u_i + \epsilon_{ij}$$

où $u_i \sim \mathcal{N}(0, \sigma^2)$ et $\epsilon_{ij} \sim \mathcal{N}(0, \tau^2)$

Application possible : β = tendance moyenne population, u_i = variation d'un groupe, ϵ_{ij} = variation au sein d'un sous-groupe

A priori de Jeffreys :

$$\pi(\beta, \sigma^2, \tau^2) \propto \frac{1}{\sigma^2 \tau^2}$$

TP : modèle à effets aléatoires autour d'une constante (Hobert-Casella) (2/2)

On note \mathbf{x}_{IJ} l'échantillon des données observées, \bar{x}_j la moyenne sur les j

On note \mathbf{u}_I l'échantillon manquant des u_1, \dots, u_I (reconstitué dans l'inférence)

Question 1. Calculer les lois conditionnelles *a posteriori* de

$$U_i | \mathbf{x}_{IJ}, \beta, \sigma^2, \tau^2$$

$$\beta | \mathbf{x}_{IJ}, \sigma^2, \tau^2, \mathbf{u}_I$$

$$\sigma^2 | \mathbf{x}_{IJ}, \beta, \tau^2, \mathbf{u}_I$$

$$\tau^2 | \mathbf{x}_{IJ}, \beta, \sigma^2, \mathbf{u}_I$$

Ces lois sont-elles bien définies ?

Question 2. Donner une formule (à un coefficient proportionnel près) pour la loi *a posteriori* jointe $\pi(\sigma^2, \tau^2 | \mathbf{x}_{IJ})$. Comment se comporte-t-elle au voisinage de $\sigma = 0$, pour $\tau \neq 0$? Que pouvez-vous en déduire ?

Question 3. Mettre en place un algorithme de Gibbs permettant d'inférer sur $(\beta, \sigma^2, \tau^2)$. Que pouvez-vous dire sur la convergence des chaînes MCMC ?

Metropolis-within Gibbs

Il arrive souvent qu'on ait, avec $\theta = (\theta_1, \theta_2)$:

- $\pi(\theta_1|\theta_2, \mathbf{x}_n)$ explicite (par conjugaison)
- $\pi(\theta_2|\theta_1, \mathbf{x}_n)$ connue à un coefficient d'intégration près

On peut alors construire un **hybride** de **Gibbs** et **Hastings-Metropolis**, qui converge encore vers la loi *a posteriori* jointe

Étape i de l'algorithme :

- 1 on simule $\theta_1^{(i)} \sim \pi(\theta_1|\theta_2^{(i-1)}, \mathbf{x}_n)$
 - on simule $\theta_2^* \sim \rho^{(i)}(\cdot|\theta_2^{(i-1)})$
 - on accepte $\theta_2^{(i)} = \theta_2^*$ avec probabilité α_i
 - sinon on conserve $\theta_2^{(i)} = \theta_2^{(i-1)}$

Tableau récapitulatif - Algorithmes de simulation *a posteriori*

	Acceptation - Rejet	Échant. d'importance	Métropolis - Hastings (MH)	Gibbs
Contexte				
Dimension de θ	1	grande	grande	grande
Nature simulation	iid	non-indep.*	non-indep. approx.*	non-indep. approx.*
Nature algo	itératif	statique	itératif	itératif
Nb. itérations typ.	quelques centaines	1	quelques dizaines de milliers	quelques milliers
Implémentation	aisée	aisée	calibration fine de $\rho(\theta)$ nécessaire	aisée
Critère d'arrêt	aucun	aucun	nécessaire	nécessaire
Risques	fort taux de rejet	poids mal équilibrés	mauvais mélange chaînes //	nécessite souvent couplage avec M-H
Temps de calcul	long	rapide	long	plutôt rapide

* : Procédures de **décorrélacion** nécessaires : rééchantillonnage, mesure d'autocorrélation

Points-clés : se poser quelques questions essentielles

- Le problème est-il proche d'un cas conjugué ? (ex : loi normale censurée)
- Si oui, que faudrait-il faire (typiquement, **simuler des données manquantes** \Rightarrow Gibbs)
- En multidimensionnel, a-t-on des propriétés de conjugaison conditionnelles (\Rightarrow Gibbs) ?
- Si aucune idée, peut-on trouver une loi $\rho(\theta)$ partageant certaines caractéristiques avec $\pi(\theta|\mathbf{x}_n)$?
 - traçage de $\pi(\theta|\mathbf{x}_n)$ (à un coefficient près) dans les cas unidimensionnels
 - calcul du mode *a posteriori* $\hat{\theta} = \text{maximum de la vraisemblance pondérée par l'a priori}$
 - une loi $\rho(\theta)$ typique est une gaussienne $\mathcal{N}(\hat{\theta}, \sigma^2 I_d)$ avec σ calibré empiriquement

D'autres approches à explorer...

- Algorithmes hybrides (MCMC-Gibbs)
- Échantillonnage d'importance adaptatif
- Méthodes de filtrage particulaire (modèles à espace d'états)
- Méthodes *Approximate Bayesian Computation* (ABC)
 - spécialement adaptées aux cas où la vraisemblance n'est pas maniable
 - très lourdes computationnellement
 - tirent cependant parti de la parallélisation croissante des moyens de calcul

WinBUGS/OpenBUGS : un outil technique pour la mise en oeuvre

Logiciel libre dédié à l'inférence *a posteriori*, voire certains calculs prédictifs

- met en oeuvre l'algorithme de Gibbs
 - très utilisé car propose une large flexibilité
 - permet notamment la définition des modèles par graphes (directs acycliques)
 - critères d'arrêt, autocorrélogrammes, etc.
- Plusieurs extensions (GeoBUGS, ...) et un concurrent sérieux : JAGS
- packages BRugs / R2WinBUGS pour être appelé directement de R et y récupérer les résultats

Cependant, un travail de fond sur les MCMC doit être fait "à la main"... en pouvant utiliser des packages R spécifiquement dédiés aux tests de convergence (CODA), à la représentation graphique des chaînes...

Quelques références

- ① Robert, C., Casella, G. (1998). Monte Carlo Statistical Methods, Springer-Verlag
- ② Marin, J.M., Robert, C. (2008). The Bayesian Core. Springer-Verlag
- ③ Geweke, J. (1989). Bayesian Inference in Econometric Models using Monte Carlo Integration, *Econometrica*, 57, pp. 1317-1339
- ④ Guillin, A., Marin, J.-M. & Robert, C.P. (2005). Estimation bayésienne approximative par échantillonnage préférentiel, *Revue de Statistique Appliquée*, 54, pp. 79-95
- ⑤ Roberts, G.O., Smith, A.F.M. (1993). Simple Conditions for the convergence of the Gibbs sampler and the Metropolis-Hastings algorithms. *Stochastic Processes and their applications*, 49, pp. 207-216
- ⑥ Rubin, D. (1988). Using the SIR Algorithm to Simulate Posterior Distributions, in *Bayesian Statistics 3*, Bernardo J., DeGroot M., Lindley D. & Smith A. (eds), Oxford University Press, pp. 395-402

Choix du cadre bayésien - quelques idées et principes utiles

Incertitudes épistémique et aléatoire : caractérisation

La statistique traite des problèmes affectées par différentes sources d'incertitude

Incertitude aléatoire.

L'incertitude aléatoire, ou naturelle est due au caractère aléatoire ou à la variabilité naturelle d'un phénomène physique (les valeurs sont précises mais différentes en raison de variations naturelles). On parle également d'*incertitude stochastique* ou de *variabilité*

Cette incertitude est généralement liée à des quantités mesurables et est considérée comme **irréductible** puisqu'inhérente à la variabilité naturelle de phénomènes physiques. L'incertitude aléatoire est généralement associée à des connaissances objectives s'appuyant sur des **données expérimentales**

Incertitudes épistémique et aléatoire : caractérisation

La statistique traite des problèmes affectées par différentes sources d'incertitude

Incertitude épistémique.

L'incertitude épistémique est due au *caractère imprécis de la connaissance ou liée à un manque de connaissance*. Elle est généralement liée à des **quantités non mesurables**, et est considérée comme **réductible** dans le sens où de nouvelles connaissances pourraient réduire voire éliminer ce type d'incertitude

Elle est principalement présente dans le cas de **données subjectives** fondées sur des croyances (avis d'expert) et pouvant être qualitatives ou quantitatives

On parle aussi d'*imprécision due au manque de données ou connaissance ou de méconnaissance*

Le jugement d'expert comme information épistémique

Au-delà du calcul mathématique, le jugement d'expert possède un rôle fondamental en prise de décision

- construction de plans d'expérience, hiérarchisation de résultats scientifiques [Cooke1991, Weinstein1993, Luntley2009]
- nourrit les études économiques [Lea1997] et actuarielles [Tredger2016] sur l'impact des risques financiers
- déterminant en arbitrage judiciaire, politique publique [Morgan2014] ou gouvernance environnementale [Miller2001, Drescher2013]

Son influence sur les choix technologiques, économiques, sociétaux ou personnels, permettant d'élaborer des stratégies de maximisation de gain, a été beaucoup étudiée par des épistémologues et des psychologues [Fischhoff1982, Luntley2009, Eagle2011]

Qu'est-ce qu'un expert ?

Un nombre infini de conceptions

Deux conceptions majeures [Weinstein1993]

- ① *ceux dont l'expertise résulte de ce qu'ils font (expertise performative)*
- ② *ceux dont l'expertise résulte de ce qu'ils ont appris (expertise épistémique) ceux dont l'expertise résulte de ce qu'ils ont appris (expertise épistémique)*

Une caractérisation usuelle, avec la capacité d'expliquer et de transmettre. Par ailleurs, pour Luntley (2009) :

I argue that what differentiates the epistemic standpoint of experts is not what or how they know [...], but their capacity for learning

Qu'est-ce qu'un expert ?

La question est, à proprement parler, "peut-on définir formellement ce qu'est un expert ?"

On parlera plutôt de **"système expert produisant une nouvelle connaissance"**

Typiquement :

- systèmes cognitifs implicites
 - humains
 - intelligences artificielles (néo-connexionnistes)
- systèmes causaux explicites
 - modèles phénoménologiques et leurs implémentation numériques (modèles de simulation)
 - intelligences artificielles symboliques

Capacité à démontrer l'expertise \Leftrightarrow capacité à prévoir/"prédire" de façon adéquate

Capacité d'apprentissage \Leftrightarrow capacité à inférer de façon cohérente quand de nouvelles

Qu'attendons-nous typiquement de la réponse d'un système expert ?

Produire de l'information épistémique sur le comportement d'une grandeur d'intérêt

$X \in \chi$

Difficultés immédiates

- biais
- impact de la subjectivité sur le processus de délivrance de l'information
- flou
- ...

résultant en **incertitude épistémique**

Information *a priori*

Information *a priori* = *information whose the value of truth is justified by considerations independent on experiment on focus* [Pegny2012]

- résultats d'essais autres que les données (par exemple sur des maquettes)
- spécifications techniques d'exploitation
- bornes physiques
- corpus référencé
- experts humains

Souvent **incomplète**, toujours **incertaine**, à cause

- de la non-existence d'un système permettant *a priori* de vérifier si l'expertise est complète ou non
- de la non-existence d'un système suffisamment précis pour spécifier que $X = x_0$ exactement (sauf dans des cas rares et pathologiques)

Des questions simples et des réponses non triviales

Que signifie "incertitude" et notamment "incertitude épistémique" ?

Question philosophique non résolue !

Pourquoi la théorie des probabilités est-elle pertinente pour aboutir une modélisation des incertitudes ?

Nombreux avantages pratiques, mais quelle pertinence théorique ?

Soulève la question de l'**auditabilité** des procédures mathématiques utilisées pour résoudre un problème = considération croissante de confiance (sociétale)

S'il y a consensus pour utiliser des probabilités, comment choisir les distributions de probabilités ?

Faire appel à quelques techniques importantes de modélisation bayésienne

Traitement de l'information *a priori*
issue de systèmes cognitifs implicites

De l'information à la connaissance (et réciproquement)

Hypothèse 1 (épistémologique) par Lakatos (1974)

- L'information sur l'état de la nature est cachée et partiellement révélée par une **théorie consensuelle** (*au sens de Popper (1972) : par décision mutuelle des protagonistes*) définissant l'**objectivité** [Gelman2015]
- La connaissance est un "filtrage" de l'information
- Ce filtrage est produit par l'intervention de symboles, ou signes, afin de la **transmettre** ou de **l'implémenter**

Hypothèse 2 (issue des neurosciences)

[Sanders2011, Salinas2011, Pouget2013, Gold2013, Dehaene2014, Chan2016]

- Face à des situations où de l'information incertaine est mobilisée, le raisonnement humain produit des inférences probabilistes
- Les difficultés apparaissent au moment de l'explicitation de cette information par **langage interprétatif** \Rightarrow **expertise utilisable**

Logique de l'information incertaine

Sous ces hypothèses, nous ne savons pas comment définir formellement la "déconvolution" retransformant la connaissance incertaine en information incertaine

Mais nous pouvons avoir des idées sur l'impact de l'ajout d'une connaissance incertaine mais utile sur la résolution du problème de détermination de X

Cet ajout se manifeste par un accroissement de l'information sur $X =$ **inférence** (mise à jour)

⇒ cette inférence doit s'appuyer sur un principe de raisonnement

⇒ ce principe de raisonnement s'établit lui-même sur une **logique** = ensemble de **règles formelles**

Propriétés souhaitées

- Pouvoir trier des assertions *atomiques* du type $X = x_0$ at chaque ajout d'information (*logique exclusive*)

- une situation initiale (**prémisse**) est moins informatif qu'une conclusion

Logique de l'information incertaine

Définition

Soit S_X un ensemble de propositions (assertions) atomiques du type $X = x_i$. L'ensemble B_X de toutes les *propositions composites* générées par

$$\begin{aligned} \neg X = x_i, \quad X = x_i \wedge X = x_j, \\ X = x_i \vee X = x_j, \quad X = x_i \Rightarrow X = x_j \\ \text{and} \quad X = x_i \Leftrightarrow X = x_j \end{aligned}$$

est appelé **état d'information**, avec $\text{Dom}(B_X) = \text{fermeture logique de } S_X$

L'état d'information B_X résume l'information existante sur un ensemble de propositions portant sur X

La même logique devrait guider la façon dont B_X évolue : il croît selon une certaine métrique quand l'information sur X croît

Plausibilité, consistance et cohérence

Définition

Considérons une proposition A on X . Sachant B_X , la **plausibilité** $[A|B_X]$ est un nombre réel, supérieurement borné par un nombre (fini ou infini) T

- **Consistence** : B_X est consistant s'il n'existe aucune proposition A pour qui $[A|B_X] = T$ et $\neg[A|B_X] = T$
- **Calcul propositionnel** : applicable à tout domaine de problème pour lequel on peut formuler des propositions utiles
 - (i) Si $A = A'$ alors $[A|B_X] \Leftrightarrow [A'|B_X]$
 - (ii) $[A|B_X, C_X, D_X] = [A|(B_X \wedge C_X), D_X]$
 - (iii) Si B_X est consistant et $\neg[A|B_X] < T$, alors $A \cup B_X$ est consistant
- **Cohérence** : il existe a fonction non croissante S_0 telle que, pour tout x et tout B_X consistant

$$\neg[A|B_X] = S_0([A|B_X])$$

- **Densité** : l'ensemble $[S_0(T), T]$ admet un sous-ensemble non vide, dense et consistant

Plausibilité, consistance et cohérence

Axiome

Considérons une proposition A on X . Sachant B_X , la **plausibilité** $[A|B_X]$ est un nombre réel, supérieurement borné par un nombre (fini ou infini) T

Cet **axiome dit de non-ambiguïté** est particulièrement important

C'est une hypothèse de *comparabilité universelle*

Conséquence : une information additionnelle (pas forcément une connaissance) peut seulement faire croître ou décroître la plausibilité d'une proposition

Plausibilité, consistance et cohérence

Axiome

Considérons une proposition A on X . Sachant B_X , la **plausibilité** $[A|B_X]$ est un nombre réel, supérieurement borné par un nombre (fini ou infini) T

Cet **axiome dit de non-ambiguïté** est particulièrement important

C'est une hypothèse de *comparabilité universelle*

Comme on le précisera après, les différences entre logique probabiliste et logique non probabiliste (ou *extra-probabiliste*) sont issues de l'accord ou du désaccord avec cette hypothèse

Jaynes (1954) justifie la validité de cette hypothèse sur des bases pragmatiques

Plausibilité, consistance et cohérence

Axiome

Considérons une proposition A on X . Sachant B_X , la **plausibilité** $[A|B_X]$ est un nombre réel, supérieurement borné par un nombre (fini ou infini) T

Cette hypothèse est notamment motivée lorsque nous parlons de quantités X possédant une **signification physique et prenant une unique valeur à chaque instant** (étant donnée, possiblement, une précision de mesure finie)

Elle peut ne pas l'être si nous parlons par exemple :

- grandeurs considérées à l'**échelle quantique** (ex : en neutronique)
- **grandeurs imaginaires** (ex : variables latentes)

Vulgarisation de la logique entière

- ❶ **Règle de reproductibilité** : deux assertions équivalentes sur X ont la même plausibilité
- ❷ **Règle de non-contradiction** : s'il existe plusieurs approches aboutissant aux mêmes conclusions sur X , celles-ci ont la même plausibilité
- ❸ **Règle de consistance** : la logique ne peut formuler une conclusion contredite par les règles élémentaires de déduction (ex : transitivité)
- ❹ **Règle d'intégrité** : la logique ne peut exclure une partie de l'information sur X pour parvenir à une conclusion sur X
- ❺ **Règle de monotonie** : la plausibilité d'une union non exclusive de deux assertions est au moins égale à la plus grande des plausibilités de chacune des assertions prises séparément
- ❻ **Règle de produit** : la plausibilité de l'intersection de deux assertions est au plus égale à la plus petite des plausibilités de chacune des assertions prises séparément

Théorème de représentation de Cox-Jaynes

Prouvé originellement (mais avec erreurs) par Cox (1946), mieux formalisé par Jaynes (1954), étendu plus rigoureusement par Paris (1994), Van Horn (2003), Dupré and Tipler (2009) (entre autres) et finalisé par Terenin and Draper (2015)

Théorème

Sous les hypothèses précédentes, il existe une fonction \mathbb{P} , croissante et continue, telle que pour toute proposition A , C et tout ensemble B_X consistant,

- (i) $\mathbb{P}([A|B_X]) = 0$ si et seulement si A est fausse étant donnée l'information sur X
- (ii) $\mathbb{P}([A|B_X]) = 1$ si et seulement si A est vraie étant donnée l'information sur X
- (iii) $0 \leq \mathbb{P}([A|B_X]) \leq 1$
- (iv) $\mathbb{P}([A \wedge C|B_X]) = \mathbb{P}([A|B_X])\mathbb{P}([C|A, B_X])$
- (v) $\mathbb{P}(\neg[A|B_X]) = 1 - \mathbb{P}([A|B_X])$

Tout système de raisonnement plausible, sous les hypothèses précédentes, est isomorphe à la théorie des probabilités

Un théorème fondamental en intelligence artificielle

Goertzel (2013) a prouvé que si la règle de consistance était affaiblie, alors les plausibilités se comportent approximativement comme des probabilités

La théorie des probabilités est pertinente pour représenter les incertitudes sur un sujet exploré par un système cognitif implicite (humain ou artificiel) qui pourrait ne pas être complétement consistant

De nombreux auteurs en intelligence artificielle [Walley1996], épistémologie [Barberousse2008] et en sciences cognitives reconnaissent la pertinence pratique de cette axiomatique pour extraire et mettre à jour de l'information, en utilisant la règle de Bayes

Critique de l'axiome de non-ambiguïté

Axiome

Considérons une proposition A on X . Sachant B_X , la **plausibilité** $[A|B_X]$ est un nombre réel, supérieurement borné par un nombre (fini ou infini) T

La "relaxation" la plus commune de cet axiome est que deux dimensions sont nécessaires pour représenter correctement la plausibilité d'une proposition

À l'origine de la **théorie des croyances** [Smets1991] et de la **théorie des possibilités** [Dubois2012]

Des expériences ont montré que cette relaxation est parfois nécessaire quand la plausibilité est interprétée comme le résumé d'une croyance personnelle, d'un *pari*

Néanmoins, cette "relaxation" reste arbitraire, et s'établit usuellement sur une interprétation de la *nature de la connaissance* (exprimée à travers un langage), et non sur la *nature de l'information* (exprimée par la réalité physique ou un modèle idéalisé de cette réalité) [Snow1998]

Quelques références (1/2)

- Pegny2012** Pegny, M. (2012). Les deux formes de la thèse de Church-Turing et l'épistémologie du calcul. *Philosophia Scientiae*
- Lakatos1974** Lakatos, I. (1974). Falsification and the methodology of scientific research programmes. In : *Criticism and the Growth of Knowledge*, I. Lakatos and A. Musgrave (eds.) Cambridge University Press
- Popper1972** Popper, K. (1972). *Objective Knowledge*. Oxford : Clarendon Pr
- Gelman2015** Gelman, A., Hennig, C. (2015). Beyond subjective and objective in statistics. *JRSS Ser. A*
- Sanders2011** Sanders, L. (2011). The Probabilistic Mind. *Science News*
- Salinas2011** Salinas, E. (2011). Prior and prejudice. *Nature Neuroscience*
- Pouget2013** Pouget, A. et al. (2013). Probabilistic brains : knowns and unknowns. *Nature Neuroscience*
- Gold2013** Gold, J.I., Heekeren, H.R. (2013). Neural Mechanisms for Perceptual Decision Making. In : *Neuroeconomics*, P.W. Glimcher and R. Fehr (eds)
- Dehaene2014** Dehaene, S. (2014). *Consciousness and the Brain : Deciphering How the Brain Codes our Thoughts*. Viking Press
- Chan2016** Chan et al. (2016). A probability distribution over latent causes, in the orbitofrontal cortex. *The Journal of Neuroscience*
- Jaynes 1954** Jaynes, E.T. (1954). *Probability Theory : The Logic of Science*. Cambridge University Press

Quelques références (2/2)

- Cox 1946 Cox, R.T. (1946). Probability, Frequency and Reasonable Expectation. *AJP*
- Paris 1994 Paris J.B. (1994). *The Uncertain Reasoner's Companion : a Mathematical Perspective*. Cambridge University Press
- Van Horn 2003 Van Horn, K.S. (2003). Constructing a logic of plausible inference : a guide to Cox's theorem. *IJAP*
- Dupré 2009 Dupré, M.J, Tipler, F.J. (2009). New axioms for rigorous Bayesian probability. *BA*
- Terenin 2015 Terenin, A., Draper, D. (2015). Rigorizing and extending the Cox-Jaynes Derivation of Probability : Implications for Statistical Practice. *arXiv :1507.06597*
- Goertzel 2013 Goertzel, B. (2013). Probability Theory Ensues from Assumptions of Approximate Consistency : A Simple Derivation and its Implications for AGI. *Proceedings of AGI-13*
- Walley1996 Walley, P. (1996). Measures of uncertainty in expert systems. *Artificial Intelligence*
- Barberousse2008 Barberousse, A. (2008). La valeur de la connaissance approchée. L'épistémologie de l'approximation d'Émile Borel. *Revue d'Histoire des Mathématiques*
- Smets1991 The transferable belief model and other interpretations of Dempster-Schafer's model. In : *Uncertainty in Artificial Intelligence, vol. 6*, Elsevier.
- Dubois2012 Dubois, D., Prade, H. (2012). *Possibility Theory*. Springer
- Snow1998 Snow, P. (1998). On the correctness and reasonableness of Cox's theorem for finite domains. *Computational Intelligence*

Introduction

Pour fixer les idées et rappeler les notations :

- des données $\mathbf{x}_n = (x_1, \dots, x_n) \in R^n$ peuvent être observées, et sont considérées comme des réalisations de la **variable aléatoire** X de densité $f(x|\theta)$ (par rapport à une mesure dominante μ)
- on décrit donc $f(x|\theta)$ comme un modèle d'**occurrence des données**
- le vecteur de paramètres $\theta \in \Theta \subset R^d$ représente **l'état de la nature**
- on suppose pouvoir munir Θ d'une structure d'espace probabilisé

L'*élicitation* de l'*a priori* est le travail d'encodage probabiliste d'une connaissance incertaine (méconnaissance) voire d'une incertitude complète sur l'état de la nature, au travers d'une distribution *a priori* $\Pi(\theta)$ de densité $\pi(\theta)$

elicere : tirer de, faire sortir, arracher, obtenir (*ex aliquo verbum elicere*)

to elicit : to get, to drawout

Qu'est-ce que $\Pi(\theta)$?

$\Pi(\theta)$ traduit la plus ou moins grande incertitude associée aux grandeurs non observables θ

Cette distribution de probabilité, dite alors **subjective**, est une mesure de l'engagement personnel en termes de pari à miser sur telle ou telle valeur de l'événement incertain

Il s'agit d'un **modèle d'expertise**, et $\Pi(\theta)$ s'interprète comme un **degré de crédibilité** des valeurs que peut prendre la grandeur incertaine θ

En conséquence, elle ne possède de sens que si elle est définie **conditionnellement** à un niveau fixé de connaissance et peut changer quand l'état de connaissance change (typiquement, quand des données \mathbf{x}_n sont apportées)

On devrait donc toujours noter

$$\pi(\theta) = \pi(\theta|H)$$

où H représente un état de connaissance *initial* de l'expérimentateur (mais cette notation ne sera pas conservée par la suite pour se simplifier la vie...)

Choisir $\pi(\theta)$

Éliciter un *a priori* $\pi(\theta)$, c'est :

- proposer une **modélisation** sous **forme hyperparamétrique** (dans ce cours)

$$\pi(\theta) = \pi(\theta|\delta)$$

où δ est appelé *vecteur des hyperparamètres*

- Ex : une forme gaussienne, exponentielle...
- proposer une **calibration** de δ en respectant le **principe de vraisemblance**, càd à partir d'une information E indépendante de données \mathbf{x}_n permettant **l'inférence *a posteriori***

L'information E peut revêtir de multiples formes :

- provenir de données antérieures à $\mathbf{x}_n \Rightarrow$ modélisation + calibration par **tests classiques**
- information "experte" sur X ou/et θ provenant de spécialistes d'un domaine (physiciens, etc.)
- contraintes physiques sur θ , etc.

Un exemple introductif en fiabilité industrielle

Soit X la durée de vie d'un matériel ou d'un composant industriel Σ

Des **modèles de durée de vie** sur X peuvent être construits à partir de considérations sur le **taux de défaillance** caractérisant Σ

$$\lambda(x|\theta) = \frac{dP_{\theta}(x < X < x + dx)}{dx}$$

En effet, on a

$$P_{\theta}(X \leq x) = 1 - \exp\left(-\int_0^x \lambda(u|\theta) du\right)$$

et

$$f(x|\theta) = \frac{dP_{\theta}(X < x)}{dx} = \lambda(x|\theta) \exp\left(-\int_0^x \lambda(u|\theta) du\right)$$

Supposons que Σ ne soit soumis qu'à des **défaillances par accident**

- son taux de défaillance est constant $\lambda(x|\theta) = \theta > 0$
- le modèle de durée de vie est alors **exponentiel** :

$$f(x|\theta) = \theta \exp(-\theta x) \mathbb{1}_{\{x \geq 0\}}$$

- le **temps moyen prédictif a priori** avant défaillance (**connaissance quantitative**), sur lequel un expert industriel peut souvent se prononcer, est alors

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\theta]] = \int_0^\infty \mathbb{E}[X|\theta] \pi(\theta) d\theta = \mathbb{E}_\pi[1/\theta]$$

Si Σ peut tomber en panne à cause du **vieillessement**

- son taux de défaillance est croissant $\lambda(x|\theta) = (x/\eta)^\beta$ avec $\theta = (\eta, \beta) \in R_+^2$
- le modèle de durée de vie est alors **Weibull** :

$$f(x|\theta) = \frac{\beta}{\eta} \left(\frac{x}{\eta}\right)^{\beta-1} \exp\left(-(x/\eta)^\beta\right) \mathbb{1}_{\{x \geq 0\}}$$

- la vitesse du vieillissement est mesurée par $\partial \lambda(x|\theta) / \partial x$
- L'expert peut savoir que **le vieillissement est sûr mais ne peut pas s'accélérer** au cours du temps (**connaissance qualitative**), donc

$$\Pi(1 \leq \beta \leq 2) = 1$$

Fragilité d'un choix de modélisation *a priori* : un exemple

(Berger 1985)

Soit $x \sim \mathcal{N}(\theta, 1)$ On suppose que la médiane *a priori* de $\theta = 0$, et que

$$\Pi(\theta < -1) = 25\% \quad \text{et} \quad \Pi(\theta < 1) = 75\%$$

- Si on fait le choix de forme *a priori* $\theta \sim \mathcal{N}(\theta_0, \sigma^2)$

- 1 $\theta_0 = 0$ et $\sigma = 1.48$

- 2 la moyenne *a posteriori* de θ sachant x (estimateur de Bayes sous coût quadratique) est

$$\theta_1^* = x \frac{\sigma^2}{1 + \sigma^2}$$

- Si on fait le choix de forme *a priori* $\theta \sim \mathcal{C}(\theta'_0, a)$: loi de Cauchy de densité

$$\pi(\theta|a, b) = \frac{1}{\pi} \left[\frac{a}{(\theta - \theta'_0)^2 + a^2} \right]$$

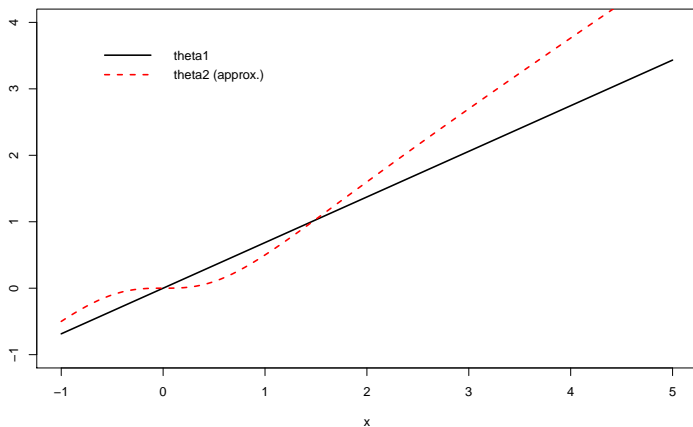
- 1 $\theta'_0 = 0$ et $a = 1$

- 2 la moyenne *a posteriori* de θ sachant x est

$$\theta_2^* \simeq \frac{x^3}{1 + x^2} \quad \text{lorsque } |x| \geq 4$$

Fragilité d'un choix de modélisation *a priori* : un exemple

(Berger 1985)



Objectif du cours

La modélisation en fonction du type de connaissance disponible est un domaine très vaste

- méthodes d'interrogation/de recueil/construction d'histogrammes
- modélisation paramétrique / non-paramétrique
- études de sensibilité sur la forme et la calibration

Dans ce cours, on va surtout s'intéresser aux **démarches formelles**, c'est-à-dire **automatiques**, visant à **choisir une forme paramétrique** pour $\pi(\theta)$ et la **calibrer** en fonction d'un **minimum ou même d'une absence d'information** sur θ

⇒ "minimum de base" pour se débrouiller dans la jungle bayésienne

Ces méthodes formelles feront notamment appel à des aspects importants de la **théorie de l'information**

1 - Élicitation par maximum d'entropie

Élicitation par maximum d'entropie

Procédure formelle de construction d'un *a priori* $\pi(\theta)$ sous un certain type de contraintes exprimant des connaissances quantitatives

Principe : on recherche un $\pi(\theta)$ dans la classe de mesures de probabilités la plus large possible respectant ces contraintes

L'**entropie** est définie comme un **indice de désordre (ou d'ignorance) associé à une distribution de probabilité**

Elle est un élément important de la **théorie de l'information**

La mise en oeuvre de ce principe aboutit à une classe de distributions importantes : **la famille exponentielle**

Le concept d'entropie (1/4)

À l'origine du concept d'entropie, un problème de recherche (tri) d'une **information discrétisée** :

Soit N sites numérotés de 1 à N où une information recherchée peut être présente

On suppose ne pouvoir poser que des questions à réponse binaire (oui ou non)

Stratégie 1 : on peut visiter chaque site et poser la question de la présence de l'information : N **questions**

Stratégie 2 (*dichotomique*) : si $\exists Q_2 \in \mathbb{N}$ tel que $N = 2^{Q_2}$, on range les sites en 2 parties et on pose la question d'appartenance au premier ou au second groupe. En itérant ce procédé, on peut trier en $Q_2 = \log_2 N$ **questions**

Lien avec la théorie de l'information : Q_2 s'interprète comme le nombre de bits (*binary digits*) nécessaire à l'écriture de N en base 2, càd la longueur du mot à utiliser pour coder N dans un alphabet de 2 caractères

Le concept d'entropie (2/4)

Si on imagine un autre alphabet de c caractères, le nombre minimal de questions sera

$$Q_c = \log_c N = \frac{\log_2 N}{\log_2 c} = \frac{\log N}{\log c}$$

Si on suppose inconnue la taille de l'alphabet (donc la nature des questions posées), le nombre de questions à poser pour identifier un site parmi N est, à une constante multiplicative près,

$$Q = \log N \quad (\text{logarithme naturel})$$

Une généralisation : on suppose qu'il existe une partition de k aires géographiques et que chaque aire contienne $N_i = N \times p_i$ sites, avec $i = 1, \dots, k$

- il suffit de poser $Q'_i = \log N_i = \log p_i + \log N$ questions pour trier l'aire i
- en moyenne sur l'ensemble des aires, on trie avec $Q' = \sum_{i=1}^k p_i Q'_i$ questions

Le concept d'entropie (3/4)

Le fait de savoir en probabilité dans quelle aire est l'information réduit le nombre de questions à poser en moyenne de la quantité

$$\Delta Q = Q - Q' = - \sum_{i=1}^k p_i \log p_i$$

quantité positive et maximale quand $p_i = 1/k$

Moins la distribution de probabilité $\Pi = (p_1, \dots, p_k)$ est *informative*, plus cette quantité est grande

Définition

L'**entropie** d'une variable aléatoire finie de distribution $\Pi = (\pi(\theta_1), \dots, \pi(\theta_k))$ est

$$\mathcal{H} = - \sum_{i=1}^k \pi(\theta_i) \log \pi(\theta_i) \quad (\text{entropie de Shannon})$$

Le concept d'entropie (4/4)

Généralisation au continu :

- le cas discret correspond à une partition fine de Θ en k intervalles dont l'étendue individuelle tend vers 0
- le résultat dépend de la mesure de partitionnement sur Θ
- l'entropie doit être invariante par tout changement de variable $\theta \mapsto \nu(\theta)$

Définition

L'**entropie** d'une variable aléatoire (continue) décrite par sa distribution de probabilité $\pi(\theta)$ est

$$\mathcal{H}(\pi) = - \int_{\Theta} \pi(\theta) \log \frac{\pi(\theta)}{\pi_0(\theta)} d\theta \quad (\text{entropie de Kullback})$$

où $\pi_0(\theta)$ est une mesure (positive) de référence sur Θ . S'il s'agit d'une mesure de probabilité, elle représente l'**ignorance complète** de la valeur θ sur Θ

Très généralement $\pi_0(\theta)$ est choisie comme la densité uniforme sur Θ

Remarque : elle n'est plus forcément positive, mais elle est maximale en $\pi(\theta) = \pi_0(\theta)$

Maximisation de l'entropie sous contraintes linéaires (1/2)

Objectif : éliciter $\pi(\theta)$ aussi vague que possible, soit

$$\pi^*(\theta) = \arg \max_{\pi \in \mathcal{P}} - \int_{\Theta} \pi(\theta) \log \frac{\pi(\theta)}{\pi_0(\theta)} d\theta \quad (6)$$

dans l'ensemble \mathcal{P} des mesures positives, sous M contraintes de forme de type **linéaire**

$$\int_{\Theta} g_i(\theta) \pi(\theta) d\theta = c_i, \quad i = 1, \dots, M$$

Le problème (1) est similaire à la **minimisation de la divergence de Kullback-Leibler** entre $\pi(\theta)$ et $\pi_0(\theta)$

La première contrainte est toujours celle de **normalisation** :

$$\int_{\Theta} \pi(\theta) = 1.$$

Maximisation de l'entropie sous contraintes linéaires (2/2)

Solution : si toutes les intégrales précédemment définies existent, la solution du problème (1) est une mesure de forme structurelle

$$\pi^*(\theta) \propto \pi_0(\theta) \exp \left(\sum_{i=1}^M \lambda_i g_i(\theta) \right)$$

(Démonstration en cours)

Cette forme caractérise les lois de la famille exponentielle

Les $(\lambda_1, \dots, \lambda_M)$ ont le sens de multiplicateurs de Lagrange et doivent être calibrés en résolvant les équations

$$\int_{\Theta} g_i(\theta) \pi^*(\theta) = c_i, \quad i = 1, \dots, M$$

Lorsque seule la contrainte de normalisation est supposée, alors

$$\pi^*(\theta) = \pi_0(\theta)$$

et correspond à une mesure (densité) de probabilité si et seulement si Θ est borné (quand π_0 est uniforme)

TP - Quelques exemples

Soit θ un paramètre réel tel que $\mathbb{E}[\theta] = \mu$ et soit π_0 la mesure de Lebesgue sur R . Quelles sont les conditions d'existence du maximum d'entropie ? Quelle est la loi correspondante ?

Supposons que $\text{Var}[\theta] = \sigma^2$. Mêmes questions

Divergence de Kullback-Leibler, calcul variationnel et exponential "twisting"

Un exemple de [calcul variationnel](#) (voir plus loin dans le cours)

Prélude à l'analyse de sensibilité *a priori*

Référence 1 : cours de M1 de A. Guyader (en ligne ; lien avec information de Fisher)

Référence 2 : *Elements of Information Theory* (Cover and Thomas, 1991)

La famille exponentielle : une modélisation de X importante

Le principe de maximisation d'entropie peut aussi être appliqué à X conditionnellement à θ , et elle permet de déboucher sur la famille paramétrée suivante

Définition

Soient $(C, h) : \Theta \times \Omega \mapsto \mathbb{R}_+^2$, et $(R, T) : \Theta \times \Omega \mapsto \mathbb{R}^k \times \mathbb{R}^k$. La famille des distributions de densité

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

est dite *famille exponentielle* de dimension k . Si R est linéaire, et lorsque $\Theta \subset \mathbb{R}^k$ et $\Omega \subset \mathbb{R}^k$, on peut écrire plus simplement (à une reparamétrisation près)

$$f(x|\theta) = h(x) \exp \{\theta \cdot x - \psi(\theta)\}$$

avec

$$\begin{aligned} \mathbb{E}_\theta[X] &= \nabla \psi(\theta) \quad (\text{gradient}) \\ \text{cov}(X_i, X_j) &= \frac{\partial^2 \psi}{\partial \theta_i \partial \theta_j}(\theta) \end{aligned}$$

On parle alors de *famille exponentielle naturelle*

Une propriété importante de la famille exponentielle

$T(x)$ est une **statistique exhaustive** (vectorielle) de x

- **Définition 1.** Si $x \sim f(x|\theta)$, une statistique T de x est exhaustive si la distribution de x conditionnellement à $T(x)$ ne dépend pas de θ

Définition 2 (Berger-Wolpert). Si $x \sim f(x|\theta)$ et si $z = t(x)$, alors z est une statistique exhaustive si et seulement si pour tout *a priori* π sur θ , $\pi(\theta|x) = \pi(\theta|z)$

Rappel sur le critère de factorisation de Neyman-Fisher (tableau)

L'exhaustivité permet de caractériser complètement la famille exponentielle

Lemme de Pitman-Koopman

Si une famille $f(\cdot|\theta)$ de **à support constant** admet une statistique exhaustive de taille fixe à partir d'une certaine taille d'échantillon, alors $f(\cdot|\theta)$ est exponentielle

Exemples

Loi de Dirichlet. (extension de la loi bêta)

$$f(x|\theta) = \frac{\Gamma(\sum_{i=1}^k \theta_i)}{\prod_{i=1}^k \Gamma(\theta_i)} \prod_{i=1}^k x_i^{\theta_i-1} \mathbb{1}_{\{S_k(x)\}}$$

définie sur le simplexe $S_k(x) = \left\{ x = (x_1, \dots, x_k); \sum_{i=1}^k x_i = 1, x_i > 0 \right\}$

Vecteur gaussien. Si $\mathbf{x}_n = (x_1, \dots, x_n) \sim \mathcal{N}_p(\mu, \sigma^2 I_p)$, alors la distribution jointe satisfait (*résultat à retrouver*)

$$f(\mathbf{x}_n|\theta) = C(\theta) h(\mathbf{x}_n) \exp \left(n\bar{x} \cdot (\mu/\sigma^2) + \sum_{i=1}^n \|x_i - \bar{x}\|^2 (-1/2\sigma^2) \right)$$

avec $\theta = (\mu, \sigma)$, et la statistique $(\bar{x}, \sum_{i=1}^n \|x_i - \bar{x}\|^2)$ est exhaustive pour tout $n \geq 2$

2 - Élicitation d'*a priori* conjugués

Une propriété majeure de la famille exponentielle : la conjugaison

Soit $X|\theta$ une loi construite par maximum d'entropie, de densité de forme :

$$f(x|\theta) = \exp \left(\sum_{j=1}^L T_j(x) d_j(\theta) \right)$$

Si, de plus, la loi *a priori* $\pi(\theta)$ est également construite par maximum d'entropie :

$$\pi(\theta) \propto \nu(\theta) \exp \left(\sum_{i=1}^M \lambda_i g_i(\theta) \right)$$

Alors, sachant l'échantillon $\mathbf{x}_n = (x_1, \dots, x_n)$, la loi *a posteriori* est ... de la **même forme structurelle** que $\pi(\theta)$:

$$\pi(\theta|\mathbf{x}_n) \propto \nu(\theta) \exp \left(\sum_{i=1}^M \lambda_i g_i(\theta) + \sum_{j=1}^L \left[\sum_{k=1}^n T_j(x_k) \right] d_j(\theta) \right)$$

L'*a priori* est alors dit **conjugué**

Dans l'écriture plus conventionnelle de la famille exponentielle naturelle

Soit

$$f(x|\theta) = h(x) \exp(\theta \cdot x - \psi(\theta))$$

alors la mesure *a priori* générée automatiquement par

$$\pi(\theta|a, b) = K(a, b) \exp(\theta \cdot a - b\psi(\theta))$$

lui est **conjuguée** (naturelle), et la mesure *a posteriori* sachant une donnée x est

$$\pi(\theta|a + x, b + 1)$$

$K(a, b)$ est la constante de normalisation

$$K(a, b) = \left[\int_{\Theta} \exp(\theta \cdot a - b\psi(\theta)) \right]^{-1}$$

qui est finie si $b > 0$ et $a/b \in \mathring{N}$

Quelques lois *a priori* conjuguées pour quelques familles exponentielles usuelles

$f(x \theta)$	$\pi(\theta)$	$\pi(\theta x)$
Normale $\mathcal{N}(\theta, \sigma^2)$	Normale $\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\varrho(\sigma^2\mu + \tau^2x), \varrho\sigma^2\tau^2)$ $\varrho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + \nu, \beta + x)$
Binomiale $\mathcal{B}(n, \theta)$	Bêta $\mathcal{B}e(\alpha, \beta)$	$\mathcal{B}e(\alpha + x, \beta + n - x)$
Binomiale Négative $\mathcal{N}eg(m, \theta)$	Bêta $\mathcal{B}e(\alpha, \beta)$	$\mathcal{B}e(\alpha + m, \beta + x)$
Multinomiale $\mathcal{M}_k(\theta_1, \dots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \dots, \alpha_k)$	$\mathcal{D}(\alpha_1 + x_1, \dots, \alpha_k + x_k)$
Normale $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}a(\alpha, \beta)$	$\mathcal{G}(\alpha + 0.5, \beta + (\mu - x)^2/2)$

tiré de C.P. Robert (2006)

Justification de l'intérêt de la conjugaison

Raisonnement d'**invariance** :

- l'information $x \sim f(x|\theta)$ transformant $\pi(\theta)$ en $\pi(\theta|x)$ est limitée
- donc elle ne devrait pas entraîner une modification de *toute* la **structure** de $\pi(\theta)$, mais simplement de ses **hyperparamètres** :

$$\pi(\theta) = \pi(\theta|\delta) \quad \Rightarrow \quad \pi(\theta|x) = \pi(\theta|\delta + s(x))$$

- cette modification devrait être de dimension finie, et un changement plus radical de $\pi(\theta)$ est peu acceptable

Une autre justification est le recours aux **données virtuelles** (cf. transparent suivant)

En pratique, l'intérêt de la conjugaison est la **commodité de traitement**

Interprétation des hyperparamètres des lois conjuguées naturelles

Soit l'*a priori* conjugué

$$\pi(\theta|x_0, m) \propto \exp\{\theta \cdot x_0 - m\psi(\theta)\} \quad (7)$$

alors l'espérance *a priori prédictive* est

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\theta]] = \mathbb{E}[\nabla\psi(\theta)] = \frac{x_0}{m}$$

et l'espérance *a posteriori prédictive*, sachant un échantillon $\mathbf{x}_n = (x_1, \dots, x_n)$, est

$$\mathbb{E}[X|\mathbf{x}_n] = \frac{x_0 + n\bar{x}}{m + n} \quad (8)$$

Autrement dit, m a le sens d'une **taille d'échantillon virtuelle**, offrant une indication de la "**force**" **informative** de l'*a priori* (d'un expert, etc.)

Théorème (Diaconis & Ylvisaker, 1979)

Si la mesure de référence est continue par rapport à la mesure de Lebesgue, alors (3) \Rightarrow (2)

Exemple : vecteur de probabilité d'un processus répétitif multivarié

On considère $X|\theta$ suivant une loi multinomiale avec $X = (X_1, \dots, X_d)$ et $\theta = (\theta_1, \dots, \theta_d)$ tel que $0 \leq \theta_i \leq 1$ et $\sum_{i=1}^d \theta_i = 1$:

$$P(X_1 = k_1, \dots, X_d = k_d | \theta) = \frac{n!}{k_1! \dots k_d!} \theta_1^{k_1} \dots \theta_d^{k_d}$$

Prior de Dirichlet. (extension de la loi bêta)

$$f(\theta | \delta) = \frac{\Gamma\left(\sum_{i=1}^d \delta_i\right)}{\prod_{i=1}^d \Gamma(\delta_i)} \prod_{i=1}^d \theta_i^{\delta_i - 1} \mathbb{1}_{\{S_d(\theta)\}}$$

définie sur le simplexe $S_d(x) = \left\{ x = (x_1, \dots, x_k); \sum_{i=1}^d x_i = 1, x_i > 0 \right\}$

Exemple : matrice de covariance d'une loi gaussienne

Soient des observations $x = (x_1, \dots, x_n)$ indépendamment issues d'une gaussienne d -multivariée $\mathcal{N}(0, \theta)$ de covariance θ .

On suppose prendre *a priori*

$$\theta \sim \mathcal{IW}(\Lambda, \nu)$$

la loi de Wishart inverse définie par la densité

$$f(\theta) = \frac{|\Lambda|^{\nu/2}}{2^{\nu d/2} \Gamma_d(\nu/2)} |\theta|^{-\frac{\nu+d+1}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\Lambda \theta^{-1}) \right\}$$

où Γ_d est la fonction gamma multivariée

Exercice : calculer la loi *a posteriori*. La loi de Wishart est-elle conjuguée ? En dimension 1, à quelle loi se réduit-elle ?

Extensions des *a priori* conjugués pour les familles exponentielles (1/2)

Proposition

Soit $\mathcal{F} = \{\pi(\theta|a, b) = K(a, b) \exp(\theta \cdot a - b\psi(\theta))\}$ la famille conjuguée naturelle de la famille exponentielle

$$f(x|\theta) = C(\theta)h(x) \exp(\theta \cdot x)$$

Alors l'ensemble des mélanges de N lois conjuguées

$$\mathcal{F}_N = \left\{ \sum_{i=1}^N \omega_i \pi(\theta|a_i, b_i); \sum_{i=1}^N \omega_i = 1, \omega_i > 0 \right\}$$

est aussi une famille conjuguée. *A posteriori*, on a

$$\pi(\theta|x) = \sum_{i=1}^N \omega'_i(x) \pi(\theta|a_i + 1, b_i + x)$$

avec

$$\omega'_i(x) = \frac{\omega_i K(a_i, b_i) / K(a_i + 1, b_i + x)}{\sum_{j=1}^N \omega_j K(a_j, b_j) / K(a_j + 1, b_j + x)}$$

Extensions des *a priori* conjugués pour les familles exponentielles (2/2)

Définition

La distance de **Prohorov** entre deux mesures π et $\tilde{\pi}$ est définie par

$$D^P(\pi, \tilde{\pi}) = \inf_A \{ \epsilon; \pi(A) \leq \tilde{\pi}(A^\epsilon) + \epsilon \}$$

où l'infimum est pris sur les ensembles boréliens de Θ et où A^ϵ indique l'ensemble des points distants de A d'au plus ϵ

Les **mélanges d'*a priori* conjugués** peuvent alors être utilisés comme **base pour approcher une loi *a priori* quelconque**, au sens où la distance de Prohorov entre une loi et sa représentation par un mélange dans \mathcal{F}_N peut être rendue arbitrairement petite

Théorème

Pour toute loi *a priori* π sur Θ , $\forall \epsilon > 0$, on peut trouver N et $\tilde{\pi} \in \mathcal{F}_N$ tel que

$$D^P(\pi, \tilde{\pi}) < \epsilon$$

Argument très fort en faveur des lois conjuguées

Au-delà de la famille exponentielle : d'autres conjugaisons possibles

Certains modèles permettant des *a priori* conjugués n'appartiennent pas à une famille exponentielle

Exemple 1 : loi de Pareto avec $\alpha > 0$ connu, et $\theta > 0$

$$f(x|\theta) = \alpha \frac{\theta^\alpha}{x^{\alpha+1}} \mathbb{1}_{] \theta, \infty[}(x)$$

admet un *a priori* conjugué, qui est Pareto sur $1/\theta$

Exemple 2 : lois uniformes

$$f(x|\theta) = \frac{\mathbb{1}_{[-\theta, \theta]}(x)}{2\theta}$$

$$f(x|\theta) = \frac{\mathbb{1}_{[0, \theta]}(x)}{\theta}$$

3 - Distributions *a priori* non-informatives

Distributions *a priori* non-informatives

Modéliser formellement une absence d'information *a priori* sur les valeurs de $\theta \in \Theta$

- 1 θ peut être artificiel : il s'agit d'un paramètre sans sens physique, biologique, etc.
- 2 Construire un *a priori* informatif comme *a posteriori* $\pi(\theta) = \pi_0(\theta|\mathbf{y}_m)$ où les \mathbf{y}_m sont des données anciennes ou "virtuelles"
- 3 Souci d'une formalisation *a priori* **objective** : on cherche à minimiser les apports subjectifs
- 4 Une loi *a priori* "vague" (ex : $\mathcal{N}(0, 100^2)$) peut donner une fausse impression de sécurité

Crûment, on peut voir $\pi_0(\theta)$ comme la "limite" d'un *a priori* subjectif (suivant une certaine topologie)

Exemple : loi exponentielle $X|\lambda \sim \mathcal{E}(\lambda)$ avec *a priori* gamma conjugué

$$\pi(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda) \mathbb{1}_{\{\lambda \geq 0\}}$$

et on fait tendre a et b vers 0 en un certain sens

Impropriété (fréquente) des distributions *a priori* non-informatives

En général, les règles formalisant l'absence d'information (cf. plus loin) aboutissent à des mesures qui sont σ -finies sur Θ mais qui ne sont pas des mesures de probabilité, soit telles que

$$\int_{\Theta} \pi_0(\theta) d\theta = \infty$$

On parle alors abusivement de **distributions impropres** (plutôt que de mesures non-intégrables)

Elles n'ont d'intérêt que si la distribution *a posteriori* est **propre** : $\int_{\Theta} \pi_0(\theta|\mathbf{x}_n) d\theta < \infty$

Un argument fort en faveur des distributions impropres est la bonne performance des estimateurs bayésiens *a posteriori*

- les estimateurs de Bayes restent **admissibles (en général)**
- on retombe souvent sur les estimateurs du maximum de vraisemblance (EMV)
- on peut voir comme des **régularisateurs probabilistes de la vraisemblance**
- ces outils réconcilient les cadres inférentiels fréquentiste et bayésien

Quelques remarques préliminaires

Le terme “non-informatif” peut être **trompeur** :

Les lois (ou distributions) non-informatives ne représentent pas une ignorance *totale* sur le problème considéré

En effet, on connaît au moins **la structure paramétrique** $X|\theta \sim f(x|\theta)$

Elles doivent être comprises comme des **lois de référence** ou **choisies par défaut**, auxquelles on peut avoir recours quand toute information *a priori* est absente

Certaines sont donc plus utiles que d'autres...

Exercice

Soit $\theta \in [1, 2]$ le paramètre d'un modèle $X \sim f(\cdot|\theta)$. On suppose ne connaître rien d'autre sur θ

Que peut-on dire de $1/\theta$?

Quelle(s) loi(s) peu ou non-informative peut-on placer de façon cohérente sur θ et $1/\theta$? Faire des tests sur machine avec la loi uniforme (par exemple)

Principes de construction d'*a priori* non informatifs

Exemple de Laplace (1773). Une urne contient un nombre n de cartes noires et blanches. Si la première sortie est blanche, *quelle est la probabilité que la proportion θ de cartes blanches soit θ_0 ?*

Laplace suppose que tous les nombres de $2/n$ à $(n-1)/n$ sont équiprobables comme valeurs de θ , donc que θ soit *uniformément distribué* sur $2/n, \dots, (n-1)/n$

Principe de raison insuffisante

En l'absence d'information, tous les événements élémentaires sont *équiprobables*, et le même poids doit être donnée à chaque valeur du paramètre, ce qui débouche automatiquement sur une distribution *a priori uniforme* $\pi(\theta) \propto 1$

Difficulté 1 : Θ doit être fini pour que $\pi(\theta)$ soit propre

Difficulté 2 : incohérence du principe des événements équiprobables en termes de partitionnement :

- si $\theta = \{\theta_1, \theta_2\}$ alors $\pi(\theta_1) = \pi(\theta_2) = 1/2$
- si on détaille plus, avec $\theta = \{\theta_1, \omega_1, \omega_2\}$, alors $\pi(\theta_1) = 1/3$

Le problème de l'invariance par reparamétrisation

Principe d'invariance par reparamétrisation

Si on passe de θ à $\eta = g(\theta)$ par une bijection g , l'information *a priori* reste inexistante et ne devrait pas être modifiée

on a

$$\pi^*(\eta) = |Jac(g^{-1}(\eta))| \pi(g^{-1}(\eta)) = \left| \det \frac{\partial \eta}{\partial \theta} \right| \pi(g^{-1}(\eta))$$

qui (en général) n'est pas constante si $\pi(\theta) = 1$

Exemple : $\eta = -\log(1 - \theta) \sim \mathcal{E}(1)$ si $\pi(\theta) = \mathbb{1}_{[0,1]}$

Principes d'invariance par reparamétrisation

On cherche à mieux préciser le sens de la "non-information" en tirant profit des caractéristiques d'invariance du problème

Exemple 1 : paramètre de position. Si on peut écrire $f(x|\theta) = f(x - \theta)$

- la famille f est *invariante par translation* : si $x \sim f$, alors $y = x - x_0 \sim f \quad \forall x_0$
- une exigence d'invariance est que $\pi(\theta)$ soit invariante par translation elle aussi :

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

Cette règle aboutit à une *loi uniforme* sur Θ

Exemple 2 : paramètre d'échelle. Si on peut écrire $f(x|\theta) = \frac{1}{\theta} f(x/\theta)$ avec $\theta > 0$

- la famille f est *invariante par changement d'échelle* : $y = x/\theta_0 \sim f \quad \forall \theta_0 > 0$
- la loi *a priori* invariante par changement d'échelle satisfait $\pi(A) = \pi(A/c)$ pour tout ensemble mesurable $A \in]0, +\infty[$ et $c > 0$

$$\pi(\theta) = \frac{1}{c} \pi\left(\frac{\theta}{c}\right)$$

et implique

$$\pi(\theta) \propto 1/\theta$$

Dans ce deuxième cas, *la mesure invariante n'est plus constante*

Principe d'invariance intrinsèque et *a priori* de Jeffreys

Ces approches impliquent la référence à une structure d'invariance, qui peut être choisie de plusieurs manières (voire ne pas exister)

Pour éviter ce choix, Jeffreys (1946) s'est intéressé à la **matrice d'information de Fisher** $I(\theta)$:

- soit $\theta \in \Theta \subset \mathbb{R}^d$; l'élément $(i, j) \in \{1, \dots, k\}^2$ de I_θ est

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

(sous des conditions de régularité suffisantes pour l'existence)

Loi *a priori* de Jeffreys

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

Pour tout changement de variable bijectif $\eta = g(\theta)$, on remarque que

$$\pi(\eta) \propto \sqrt{\det I(\eta)}$$

L'*a priori* de Jeffreys vérifie donc un principe d'invariance (*intrinsèque*) par n'importe quelle reparamétrisation

Exercice 1

Redémonstration de la preuve en cours pour la dimension 1

Exercice 2 (TP)

Calculer le prior de Jeffreys pour le paramètre $\theta \in [0, 1]$ de la loi binomiale $X \sim \mathcal{B}(n, \theta)$

Quel lien avec la minimaxité ?

Une justification supplémentaire

- $I(\theta)$ est largement accepté comme un indicateur de la quantité d'information apportée par le modèle (ou l'observation) sur θ (Fisher, 1956)
- $I(\theta)$ mesure la capacité du modèle à discriminer entre θ et $\theta + / - d\theta$ via la pente moyenne de $\log f(x|\theta)$
- Favoriser les valeurs de θ pour lesquelles $I(\theta)$ est grande équivaut à **minimiser l'influence de la loi *a priori***

L'*a priori* de Jeffreys est l'une des meilleures techniques automatiques pour obtenir des lois non-informatives

Il est le plus souvent **impropre**, sauf pour des modèles pour lesquels Θ est **borné ou/et discret** (**Exemple binomial**)

Il est généralement utilisé en **dimension 1**, où il permet d'obtenir des estimateurs bayésiens similaires au maximum de vraisemblance

En **multidimensionnel**, le prior de Jeffreys peut mener à des **incohérences** ou des **paradoxes**.

Exercice 2 (TP ; suite)

Soit une variable négative binomiale $N \sim \mathcal{NB}(x, \theta)$ de densité (fonction de masse) définie par

$$P(N = k|\theta) = \frac{\Gamma(x + k)}{k! \Gamma(x)} \theta^x (1 - \theta)^k$$

(nombre d'échecs avant un nombre x de succès), d'espérance xk/θ

Quel lien peut-on faire entre les vraisemblances d'une binomiale $x \sim \mathcal{B}_n(n, \theta)$ et celle d'une loi négative binomiale $n \sim \mathcal{NB}(x, \theta)$?

Calculer le prior de Jeffreys pour le paramètre $\theta \in [0, 1]$ de la loi négative binomiale

Peut-on conclure que le prior de Jeffreys respecte la règle de vraisemblance ?

A priori de référence de Berger-Bernardo (1979, 1992)

Une autre construction mieux adaptée au cadre multidimensionnel

Principe

- La "distance" (divergence) de Kullback-Leibler entre *a posteriori* et *a priori*

$$KL(\pi, \mathbf{x}_n) = \int_{\Theta} \pi(\theta | \mathbf{x}_n) \log \frac{\pi(\theta | \mathbf{x}_n)}{\pi(\theta)} d\mathbf{x}_n$$

mesure l'information apportée par les données observées \mathbf{x}_n sur le modèle, indépendamment de la paramétrisation θ

- L'idée est de maximiser $KL(\mathbf{x}_n)$ en π pour des données \mathbf{x}_n pouvant être typiquement observées : elles sont générées par la loi *a priori* prédictive (marginale)

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

et pour s'affranchir du choix de la taille n , on fait tendre celle-ci vers ∞

$$\text{soit } \pi^* = \arg \max_{\pi} \lim_{n \rightarrow \infty} \mathbb{E}_{f(\mathbf{x}_n)} [KL(\pi, \mathbf{x}_n)]$$

Caractéristiques principales

En dimension 1, on retombe sur la mesure de Jeffreys

Permet de résoudre des problèmes d'inconsistance *a posteriori* en dimensions supérieures

Le résultat dépend de l'**ordonnement désiré** des paramètres

- paramètres d'**intérêt**
- paramètres de **nuisance**

L'invariance par reparamétrisation est maintenue à l'intérieur des groupes (intérêt et nuisance), sous des conditions de bijectivité

Méthodologie d'élicitation cependant moins automatique que Jeffreys, plus délicate à mettre en pratique

Loi *a priori* coïncidante d'ordre i (*coverage matching prior*)

- Soit $\theta_n(\alpha)$ le quantile *a posteriori* d'ordre α
- $\forall \alpha \in [0, 1]$, on peut construire une mesure *a priori* telle que

$$\underbrace{P_\theta(\theta \leq \theta_n(\alpha))}_{\text{probabilité fréquentiste}} = \underbrace{P(\theta \leq \theta_n(\alpha) | \mathbf{X}_n)}_{\text{probabilité bayésienne}} + \mathcal{O}(n^{-i/2}).$$

On peut créer des *coverage matching priors* comme solution d'une équation différentielle stochastique

L'observation des propriétés de **recouvrement fréquentiste** permet de discriminer entre plusieurs *a priori* dits de référence

Convergence des mesures *a priori* vers des mesures impropres (1)

Proposition (Wallace 1959) : convergence en tout point

Si π est une densité *a priori* impropre, alors il existe une suite de densités *a priori* propres $\{\pi_n\}_n$ engendrant une suite d'*a posteriori* $\{\pi_n(\cdot|x)\}_n$ telle que pour tout $\theta \in \Theta$ et pour tout x ,

$$\lim_{n \rightarrow \infty} \pi_n(\theta|x) = \pi(\theta|x)$$

Ce résultat reste vrai si $\{\pi_n\}_n$ est une suite de densités telle qu'il existe une constante K et une suite $\{a_n\}_n$ telle que, pour tout θ ,

$$\lim_{n \rightarrow \infty} a_n \pi_n(\theta) = \pi(\theta) \quad \text{et} \quad a_n \pi_n(\theta) \leq K \pi(\theta)$$

Approche *rétrospective* (Stone 1965) : le jeu de données est fixé avant tout

Convergence des mesures *a priori* vers des mesures impropres (2)

D'autres notions de convergence étudiées

- ① **Convergence en probabilité** (Stone 1965) vers des mesures impropres *relativement invariantes* : continues et telles que $\pi(\theta_1, \theta_2) = \pi(\theta_1)\pi(\theta_2)$
 - Nécessite d'introduire des suites de mesures *a priori* obtenues par troncature (suite croissante de compacts sur Θ)
- ② **Convergence en variation totale** (Head et Sudderth 1989), définie par

$$\|\pi_n(\theta) - \pi(\theta)\| = \sup_{\mathcal{F}} |\pi_n(\theta) - \pi(\theta)|$$

où \mathcal{F} est une σ -algèbre de sous-ensembles de Θ

- ③ **Convergence en entropie relative** (Berger et al. 2009)

Convergence des mesures *a priori* vers des mesures impropres (3)

Une vision *prospective* ou intrinsèque (= préalable à l'occurrence de données) développée fort récemment par Bioche (2015)

Convergence vague

Soit $\{\mu_n\}_n$ et μ des mesures (de Radon). La suite $\{\mu_n\}_n$ converge vaguement vers μ si, pour toute fonction h continue à support compact,

$$\lim_{n \rightarrow \infty} \int h d\mu_n = \int h d\mu$$

Mode de convergence équivalent à la **convergence étroite** pour les mesures de probabilité

Convergence étroite

Soit $\{\mu_n\}_n$ et μ des mesures bornées. La suite $\{\mu_n\}_n$ converge étroitement vers μ si, pour toute fonction h continue bornée,

$$\lim_{n \rightarrow \infty} \int h d\mu_n = \int h d\mu$$

Convergence des mesures *a priori* vers des mesures impropres (4)

q–convergence (Bioche et al. 2015)

Une suite de mesures positives $\{\mu_n\}_n$ converge *q*–vaguement vers une mesure positive μ s'il existe une suite de réels positifs $\{a_n\}_n$ telle que $\{a_n\mu_n\}_n$ converge vaguement vers μ

Cas discrets

Si μ et μ_n sont définies sur $\Theta = \{\theta_i\}_{i \in I}$, la convergence *q*–vague est équivalente à : $\forall i \in I$,

$$\lim_{n \rightarrow \infty} a_n \mu_n(\theta_i) = \mu(\theta_i)$$

Convergence des mesures *a priori* vers des mesures impropres (5)

Cas continus

Soient μ et μ_n des mesures *a priori* sur Θ . Supposons que :

- 1 il existe une suite de réels positifs $\{a_n\}_n$ tel que la suite $\{a_n\mu_n\}_n$ converge ponctuellement vers μ
- 2 pour tout ensemble compact K , il existe un scalaire M et $N \in \mathbb{N}$ tels que, pour tout $n > N$,

$$\sup_{\theta \in K} a_n \mu_n(\theta) < M.$$

Alors $\{\mu_n\}_n$ converge *q*-vaguement vers μ

Quelques exemples

Soit $\Theta = N$ et $\mu_n = \mathcal{U}(\{0, 1, \dots, n\})$ la distribution uniforme discrète sur le compact discret $\{0, \dots, n\}$. Alors $\{\mu_n\}_n$ converge q -vaguement vers la mesure de comptage

Soit $\Theta = R$ et $\mu_n = \mathcal{U}([-n, n])$. Alors $\{\mu_n\}_n$ converge q -vaguement vers la mesure de Lebesgue

Soit $\Theta = R$ et $\mu_n = \mathcal{N}(0, n)$. Alors $\{\mu_n\}_n$ converge q -vaguement vers la mesure de Lebesgue

4 - Un exemple complet dans un cadre de fiabilité industrielle

avec incorporation d'information subjective

Exemple dans un cadre de fiabilité industrielle (1/5)

X = durée de vie d'un composant Σ , supposé tomber uniquement en panne par hasard

Le taux de défaillance λ de Σ est donc constant, ce qui implique $X \sim \mathcal{E}(\lambda)$

Un expert industriel est familier de λ

Dialogue avec l'expert :

- ❶ Considérons une décision de management (remplacement) établie sur une valeur donnée $\bar{\lambda}$ (différente de la vraie valeur inconnue λ)
- ❷ Pour un coût similaire $|\bar{\lambda} - \lambda|$, il y a 2 conséquences possibles au remplacement :
 - soit C_1 le coût positif moyen d'être trop optimiste (d'avoir $\bar{\lambda} \leq \lambda$)
 - soit C_2 le coût positif moyen d'être trop pessimiste (d'avoir $\bar{\lambda} > \lambda$)
- ❸ Pouvez-vous donner un estimé $\hat{\delta}$ du rapport des coûts moyens $\delta = C_2/C_1$?

L'axiome de rationalité dit que si l'expert n'est pas averse au risque, alors

$$\bar{\lambda} = \arg \min_x \underbrace{\int_0^{\infty} |x - \lambda| \left(C_1 \mathbb{1}_{\{x \leq \lambda\}} + C_2 \mathbb{1}_{\{x > \lambda\}} \right) \pi(\lambda) d\lambda}_{\text{fonction de coût intégrée sur toutes les valeurs possibles a priori du vrai } \lambda}$$

Exemple dans un cadre de fiabilité industrielle (2/5)

Il s'ensuit que
$$\int_0^{\bar{\lambda}} d\Pi(\lambda) = \Pi(\lambda < \bar{\lambda}) = \frac{C_1}{C_1 + C_2}$$

L'interprétation de la réponse de l'expert est que $1/(1 + \hat{\delta})$ est un estimé du quantile *a priori* d'ordre $\alpha = C_1/(C_1 + C_2)$

Avec $P(\lambda < \bar{\lambda}) = \frac{C_1}{C_1 + C_2} = \alpha$, on a :

- tant que les coûts sont équilibrés, un expert de plus en plus optimiste fournira un $\bar{\lambda}$ de plus en plus petit, car la durée moyenne avant la prochaine défaillance est

$$\mathbb{E}[X|\lambda] = \frac{1}{\lambda}$$

- cependant l'expert s'exprime plutôt sur les coûts lorsqu'on lui fournit une valeur représentative de $\bar{\lambda}$
 - plus l'expert est optimiste, plus le coût C_2 d'être optimiste (selon lui) est petit, donc α grandit vers 1 et

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\lambda]] = \mathbb{E}[1/\lambda] \text{ augmente}$$

- plus l'expert est pessimiste, plus le coût C_2 d'être optimiste augmente, donc α tombe vers 0 et

Exemple dans un cadre de fiabilité industrielle (3/5)

Attachons-nous maintenant à définir une *forme a priori* pertinente pour le décideur

On raisonne par *données virtuelles*

- L'*a priori* non-informatif de Jeffreys est utilisé pour modéliser l'absence d'expertise et la simple connaissance du modèle

$$\pi_0(\lambda) \propto \lambda^{-1} \quad (\text{paramètre d'échelle})$$

- L'information apportée par l'expert est assimilée à celle apportée par un échantillon i.i.d.

$$\mathbf{x}_m \sim \mathcal{E}(\lambda)$$

- Un "bon" prior informatif pour λ est donc $\pi(\lambda) = \pi_0(\lambda|\mathbf{x}_m)$, soit

$$\lambda \sim \mathcal{G}(m, m\bar{x}_m)$$

Exemple dans un cadre de fiabilité industrielle (4/5)

Donc $2m\bar{x}_m\lambda \sim \mathcal{G}(m, 1/2) \equiv \chi_{2m}^2$, d'où $\bar{x}_m = \frac{\chi_{2m}^2(\alpha)}{2m\bar{\lambda}}$

Le décideur peut fixer arbitrairement m selon la confiance qu'il a en l'expert (ou mettre un *a priori* hiérarchique dessus)

De plus, l'*a priori* est **conjugué** : sachant des durées de vie observées $\mathbf{x}_n = (x_1, \dots, x_n)$, la loi *a posteriori* de λ est

$$\lambda|\mathbf{x}_n \sim \mathcal{G}\left(m+n, \frac{\chi_{2m}^2(\alpha)}{2\bar{\lambda}} + n \sum_{i=1}^n x_i\right)$$

L'ingénieur s'intéresse alors à la probabilité **prédictive** que Σ tombe en panne avant la prochaine visite au temps x_0

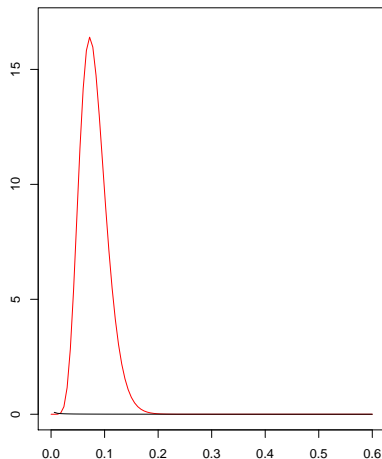
$$P(X < x_0) = \int_0^\infty P(X < x_0|\lambda) \pi(\lambda|\mathbf{x}_n) d\lambda = 1 - 1/\left(1 + \frac{x_0}{\frac{\chi_{2m}^2(\alpha)}{2\bar{\lambda}} + n \sum_{i=1}^n x_i}\right)^{m+n}$$

puis il peut introduire une fonction de coût, etc. pour prendre une décision

Illustration : cas non-informatif

10 données simulées selon $\mathcal{E}(\lambda_0)$ avec $\lambda_0 = 1/10$

densités prior / posterior



probabilité de panne

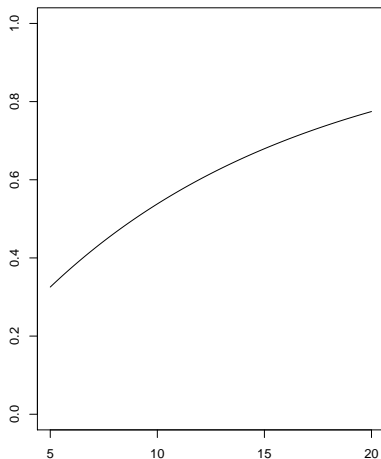
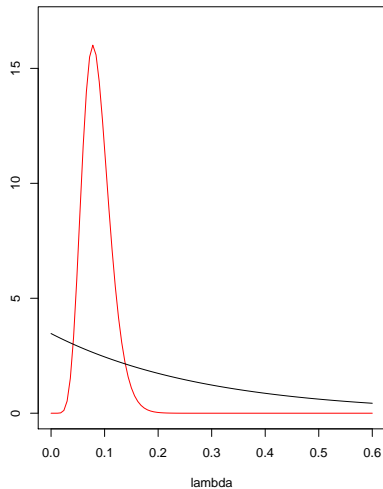


Illustration : $m = 1$, $\bar{\lambda} = 1/5$, $\alpha = 50\%$ (expert peu informatif et pessimiste)

densités prior / posterior



probabilité de panne

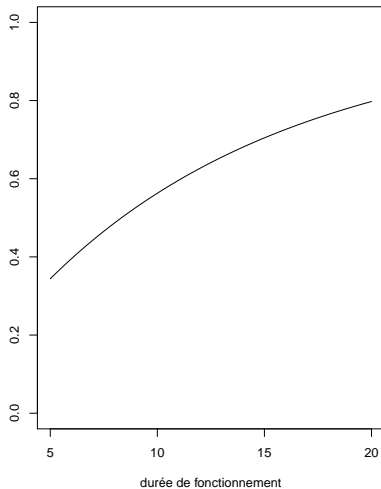
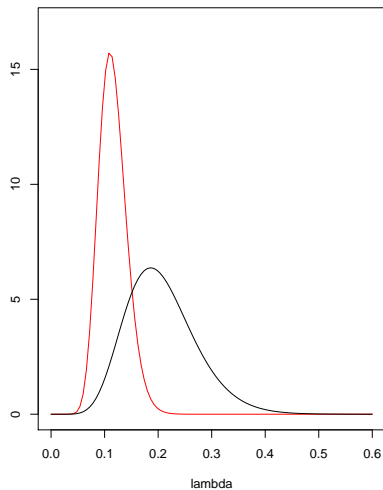


Illustration : $m = 10$, $\bar{\lambda} = 1/5$, $\alpha = 50\%$ (expert très informatif et pessimiste)

densités prior / posterior



probabilité de panne

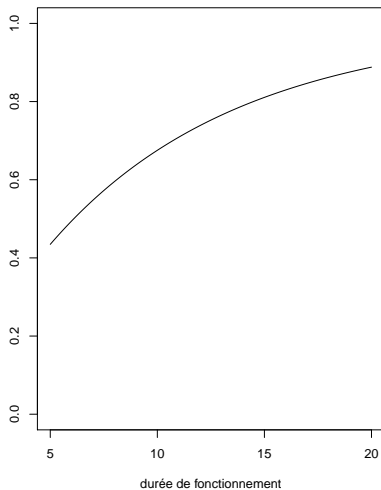
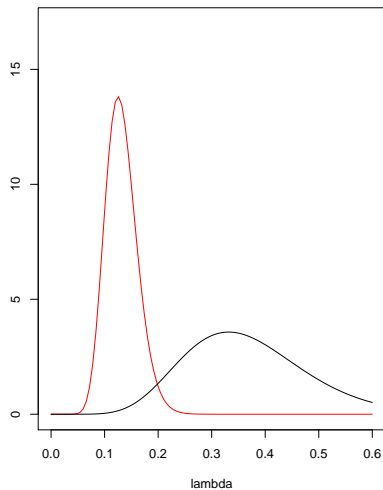
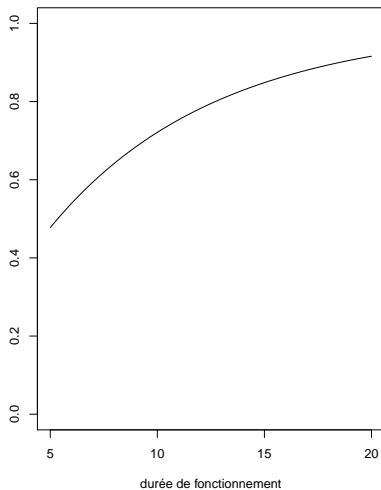


Illustration : $m = 10$, $\bar{\lambda} = 1/5$, $\alpha = 5\%$ (expert très informatif et très pessimiste)

densités prior / posterior



probabilité de panne



Exemple dans un cadre de fiabilité industrielle (5/5)

En fait, plus habituellement, l'expert préfère exprimer son opinion **quantitative** sur la durée de vie X (= **variable d'ancrage**) que sur λ , car X est **observable**

Dans ce cas, il est assimilé à un fournisseur d'estimé du **quantile prédictif *a priori*** \bar{x} :

$$\int_0^{\bar{x}} f(x) dx = \int_0^{\bar{x}} \int_{\Theta} f(x|\theta) \pi(\theta) d\theta = \alpha$$

Cette interprétation est la plus acceptée en général dans la communauté statistique bayésienne, c'est pourquoi les statisticiens fiabilistes préfèrent poser des questions comme

- Sachant les temps x_0 et $x_1 > x_0$, Σ a $1 - \alpha$ fois plus de chance de ne pas tomber en panne après x_0 qu'après x_1 . Quelle est votre évaluation de $1 - \alpha$?

5 - Modélisation bayésienne hiérarchique

Modélisation bayésienne hiérarchique

Pour des raisons liées à la modélisation des observations ou à la décomposition de l'information *a priori*, le modèle bayésien $(f(x|\theta), \pi(\theta))$ peut être défini comme **hiérarchique** : $\pi(\theta)$ est décomposé en plusieurs *lois conditionnelles*

$$\begin{aligned}\pi(\theta|\theta_1, \dots, \theta_k) &= \pi_1(\theta|\theta_1) \cdots \pi_2(\theta_1|\theta_2) \cdots \pi_k(\theta_{k-1}|\theta_k) \cdot \pi_{k+1}(\theta_k) \\ \text{et } \pi(\theta) &= \int_{\Theta_1 \times \dots \times \Theta_k} \pi(\theta|\theta_1, \dots, \theta_k) d\theta_1 \dots d\theta_k\end{aligned}$$

Exemple d'un modèle statistique classique : **modèle linéaire à effets aléatoires**

$$\begin{aligned}y|\theta &\sim \mathcal{N}_p(\theta, \Sigma_1), \\ \theta|\beta &\sim \mathcal{N}_p(X\beta, \Sigma_2)\end{aligned}$$

souvent utilisé en génétique animale pour différencier l'influence d'éléments fixes (ex : lignée, race, année) de celle de facteurs aléatoires (ex : nb de femelles dans une lignée)

Quelques caractéristiques

Le conditionnement peut s'expliquer par

- des **dépendances statistiques naturelles**
- l'appel à des **variables latentes** décrivant un **mécanisme caché**
- des grandeurs stochastiques jouant un rôle de **forçage**

En général, on ne va guère plus loin que deux ou trois niveaux de hiérarchie

En incluant l'information *a priori* aux niveaux les plus élevés, l'approche bayésienne hiérarchique permet en général de gagner en **robustesse**

Éliciter un *a priori* à partir de résultats statistiques théoriques

Un exemple : la **courbe de von Bertalanffy**

$$L(t|\theta) = L_{\infty}(1 - \exp(-g(t, \delta)))$$

est fréquemment utilisée pour produire une **clé âge-longueur**, en modélisant l'accroissement en longueur d'un organisme vivant (ex : arbre, poisson...)

On note $\theta = (L_{\infty}, \delta)$ le vecteur des paramètres inconnus **Données de capture-recapture** :

supposons avoir des couples d'observation $\{l^*(t_i), l^*(t_{i+\Delta_i})\}$ tel que

$$\begin{aligned} l^*(t_i) &= L(t_i|\theta) \exp(\epsilon_1), \\ l^*(t_{i+\Delta_i}) &= L(t_i + \Delta_i|\theta) \exp(\epsilon_2) \end{aligned}$$

où (ϵ_1, ϵ_2) sont des **bruits de mesure** (générant donc une vraisemblance)

Les estimations par maximum de vraisemblance de L_{∞} sont très sensibles à la taille des données

Comment placer un *a priori* sur L_{∞} ?

Tirer parti de propriétés asymptotiques : loi des dépassements

Le sens de L_∞ est celui d'une longueur maximale qu'un être vivant peut atteindre, en moyenne sur toutes les observations possibles

Posons alors $L_\infty^* = L_\infty \exp(\epsilon)$ la longueur maximale *observée*

Soit \bar{L} la longueur moyenne

Théorème (Pickands, 1975)

Quand \bar{L} grandit, la distribution de $L_\infty^* | \bar{L} = l$ est une Pareto généralisée :

$$P(L_\infty^* < x | L_\infty^* > \bar{L}, \sigma, \mu) = 1 - \left(1 + \mu \left(\frac{x - \bar{L}}{\sigma} \right) \right)^{-1/\mu}$$

On obtient alors une justification pour :

- ① établir une forme *a priori* pour $\pi(L_\infty)$ (la forme de ϵ étant fixée)
- ② conditionner ce prior par rapport à $\bar{L} \Leftrightarrow$ utiliser une **approche bayésienne hiérarchique**

Un autre exemple : modélisation *a priori* d'une probabilité de survie dans une population

Soit X_t un nombre d'individus dans une population

Soit $\theta = \theta_{t,t+1}$ la probabilité de survie entre t et $t + 1$

La vraisemblance peut être définie par

$$X_{t+1} | X_t, \theta_{t,t+1} \sim \mathcal{B}(X_t, \theta_{t,t+1}) \quad (\text{loi binomiale})$$

On peut alors écrire

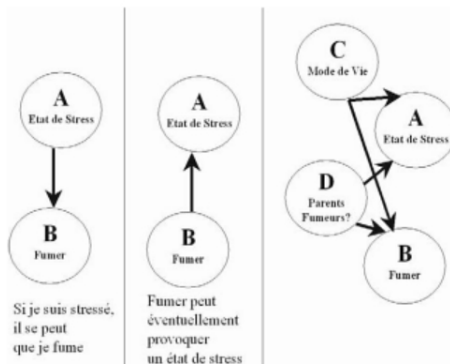
$$\theta_{t,t+1} = \prod_{i=0}^{M+1} \theta_{t+i/M, t+(i+1)/M}$$

donc, de par le Théorème de la Limite Centrale, quand $1 \ll M$,

$$\log(\theta_{t,t+1}) \sim \mathcal{N}(\mu_t, \sigma_t^2)$$

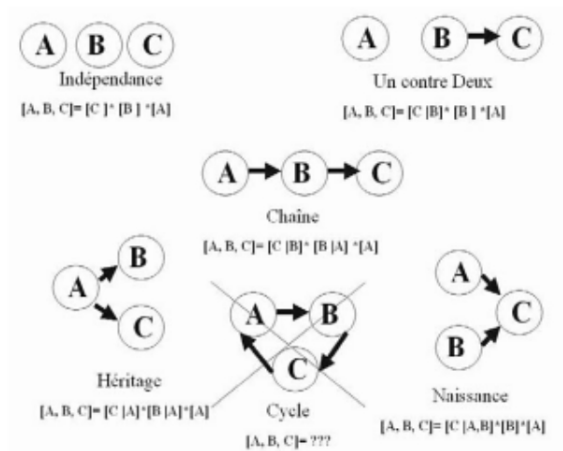
avec $\mu_t < -\sigma_t^2/2$ tel que $\mathbb{E}[\theta_{t,t+1}] \in [0, 1] \Rightarrow$ contrainte de forme sur le niveau hiérarchique $\pi(\mu_t, \sigma_t)$

Graphes acycliques orientés (1/3)

Causalité et dépendance probabiliste (*tiré de Parent et Bernier, 2007*)

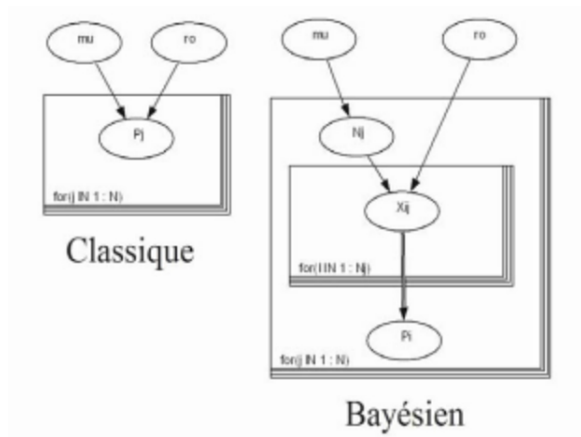
Graphes acycliques orientés (2/3)

Les relations de dépendance conditionnelles possibles entre trois variables aléatoires (tiré de Parent et Bernier, 2007)



Graphes acycliques orientés (3/3)

Un DAG incluant des variables latentes (*tiré de Parent et Bernier, 2007*)



6 - Quelques soucis méthodologiques et pratiques importants

- 1 Prouver l'intégrabilité de la loi *a posteriori*
- 2 Fusionner des *a priori*
- 3 Vérifier la cohérence entre sources d'information (données et *a priori*)
- 4 ...

1 - Prouver l'intégrabilité de la loi *a posteriori*

La modélisation bayésienne par **conditionnement** peut fréquemment entraîner le mécanisme suivant :

- 1 On construit un *a priori* **hiérarchique**

$$\pi(\theta) = \pi(\theta_1|\theta_2, \theta_3)\pi(\theta_2|\theta_3)\pi(\theta_3)$$

avec des *a priori* non-informatifs

- 2 Ce conditionnement est souvent choisit pour tirer parti de **conjugaisons** *a posteriori* : les lois conditionnelles

$$\pi(\theta_1|\mathbf{x}_n, \theta_2, \theta_3),$$

$$\pi(\theta_2|\mathbf{x}_n, \theta_1, \theta_3),$$

$$\pi(\theta_3|\mathbf{x}_n, \theta_1, \theta_2)$$

sont explicites, ce qui permet d'utiliser un algorithme de Gibbs pour le calcul *a posteriori* (cf. plus loin)

Problème majeur : même si ces lois *a posteriori* **conditionnelles** sont propres, la loi **jointe** peut ne pas l'être :

$$\int_{\Theta} \pi(\theta|\mathbf{x}_n) d\theta = \infty$$

Exemple : modèle à effets aléatoires autour d'une constante (Hobert-Casella) (1/3)

Pour $i = 1, \dots, I$ et $j = 1, \dots, J$

$$x_{ij} = \beta + u_i + \epsilon_{ij}$$

où $u_i \sim \mathcal{N}(0, \sigma^2)$ et $\epsilon_{ij} \sim \mathcal{N}(0, \tau^2)$

Application possible : β = tendance moyenne population, u_i = variation personnelle, ϵ_{ij} = variation au sein d'un sous-groupe

A priori de Jeffreys :

$$\pi(\beta, \sigma^2, \tau^2) \propto \frac{1}{\sigma^2 \tau^2}$$

Exemple : modèle à effets aléatoires autour d'une constante (Hobert-Casella) (2/3)

On note \mathbf{x}_{IJ} l'échantillon des données observées, \bar{x}_i la moyenne sur les j

On note \mathbf{u}_I l'échantillon manquant des u_1, \dots, u_I (reconstitué dans l'inférence)

Lois conditionnelles *a posteriori*

$$\begin{aligned}
 U_i | \mathbf{x}_{IJ}, \beta, \sigma^2, \tau^2 &\sim \mathcal{N} \left(\frac{J(\bar{x}_i - \beta)}{J + \tau^2 \sigma^{-2}}, (J\tau^{-2} + \sigma^{-2})^{-1} \right) \\
 \beta | \mathbf{x}_{IJ}, \sigma^2, \tau^2, \mathbf{u}_I &\sim \mathcal{N} (\bar{x} - \bar{u}, \tau^2 / IJ) \\
 \sigma^2 | \mathbf{x}_{IJ}, \beta, \tau^2, \mathbf{u}_I &\sim \mathcal{IG} \left(I/2, (1/2) \sum_{i=1}^I u_i^2 \right) \quad (\text{loi inverse gamma}) \\
 \tau^2 | \mathbf{x}_{IJ}, \beta, \sigma^2, \mathbf{u}_I &\sim \mathcal{IG} \left(IJ/2, (1/2) \sum_{i=1}^I \sum_{j=1}^J (x_{ij} - u_i - \beta)^2 \right)
 \end{aligned}$$

sont bien définies

Exemple : modèle à effets aléatoires autour d'une constante (Hobert-Casella) (3/3)

Cependant, la loi *a posteriori* jointe

$$\begin{aligned}\pi(\sigma^2, \tau^2 | \mathbf{x}_{IJ}) &= \int \pi(\beta, \sigma^2, \tau^2 | \mathbf{x}_{IJ}) d\beta \\ &= \int \left[\int_1 \dots \int_i \dots \int_I \pi(\beta, \sigma^2, \tau^2 | \mathbf{x}_{IJ}) du_i \right] d\beta\end{aligned}$$

est proportionnelle à

$$\frac{\sigma^{-2-I}\tau^{-2-IJ}}{(J\tau^{-2} + \sigma^{-2})^{I/2}} \sqrt{\tau^2 + J\sigma^2} \exp \left\{ -\frac{1}{2\tau^2} \sum_{i,j} (y_{ij} - \bar{y}_i)^2 - \frac{J}{2'\tau^2 + J\sigma^2} \sum_i (\bar{y}_i - \bar{y})^2 \right\}$$

qui se comporte comme σ^{-2} au voisinage de $\sigma = 0$, pour $\tau \neq 0$

Cette loi jointe n'est donc pas intégrable (*propre*)

2 - Fusionner plusieurs *a priori*

Dans de nombreux cas pratiques, on peut disposer de plusieurs *a priori* possibles $\pi_1(\theta), \dots, \pi_M(\theta)$ que l'on suppose ici **indépendants**

Exemple : réunions d'experts à la fin d'études pharmacologiques

Une première idée : la **fusion linéaire pondérée** (ou moyenne arithmétique)

$$\pi(\theta) = \sum_{i=1}^M \omega_i \pi_i(\theta)$$

avec $\sum_{i=1}^M \omega_i = 1$

Problèmes :

- le résultat peut être multi-modal
- l'approche n'est pas *externalement bayésienne* :

$$\pi(\theta|\mathbf{x}_n) \neq \sum_{i=1}^M \omega_i \pi_i(\theta|\mathbf{x}_n)$$

pour une ou plusieurs données \mathbf{x}_n

Une seconde idée : **la fusion logarithmique pondérée** (ou moyenne géométrique)

$$\pi(\theta) = \frac{\prod_{i=1}^M \pi_i^{\omega_i}(\theta)}{\int_{\Theta} \prod_{i=1}^M \pi_i^{\omega_i}(\theta) d\theta}$$

avec $\sum_{i=1}^M \omega_i = 1$

Elle est bien extérieurement bayésienne

Problème : l'approche n'est pas *cohérente par marginalisation*

- Soit A et B tels que $A \cap B = \emptyset$ et $C = A \cup B \Rightarrow P(C) = P(A) + P(B)$
- Soient des experts indiquant leurs opinions sur les événements A et B
- Pour chaque expert, on peut directement calculer $P(C)$ ou calculer séparément $P(A)$ puis $P(B)$
- seule la fusion linéaire permet l'égalité des résultats

En réalité, la fusion logarithmique est séduisante car elle peut s'expliquer en faisant appel à la théorie de l'information

La divergence de Kullback-Leibler

$$KL(\pi, \pi_i) = \int_{\Theta} \pi(\theta) \log \frac{\pi(\theta)}{\pi_i(\theta)}$$

exprime une **perte en terme d'information** lorsque le meilleur choix *a priori* π est remplacé par π_i

Le minimiseur de la perte pondérée

$$\pi^*(\theta) = \arg \min_{\pi} \sum_{i=1}^M \omega_i KL(\pi, \pi_i)$$

est l'*a priori* opérant la fusion logarithmique

La calibration des poids ω_i est un problème qui reste ouvert, malgré quelques réponses déjà proposées

Exemple

Considérons M priors exponentiels

$$\theta \sim \pi_i(\theta) = \lambda_i \exp(-\lambda_i \theta) \mathbb{1}_{\{\theta \geq 0\}}$$

Quelle est la loi-fusion logarithmique ?

\Rightarrow stabilité de la famille exponentielle naturelle par fusion logarithmique \Rightarrow similarité avec une

inférentielle bayésienne croissante, indépendante de l'ordre d'arrivée des informations (ex : échantillons virtuels)

3 - Vérifier la cohérence entre sources d'information (données et *a priori*)

Motivation par l'exemple (industrie nucléaire)

Durée de vie (mois) de parois de circuit secondaire

temps de défaillance $\in [72.8, 152.1]$

temps de censure $\in [66.8, 159.5]$

Opinions d'expert

	Temps médian	[5%-95%]
Expert 1	250	[200-300]
Expert 2	250	[100-500]

Les experts sont **optimistes** par rapport aux données t_n . Quelques raisons possibles : **évolution technique** (décalage temporel), **hétérogénéité** des avis, **degré de précision** de l'interrogation...

Difficulté:

les opinions subjectives et la connaissance objective des données peuvent être **conflictuelles** \Rightarrow ceci doit être remarqué avant toute inférence !

Exemple : modèle gaussien à variance connue

Soit un échantillon $\mathbf{x}_n \sim \mathcal{N}(\mu, \sigma^2)$

On suppose connaître σ , mais μ est inconnu

On place l'*a priori* conjugué $\mu \sim \mathcal{N}(m, \rho\sigma^2)$

Peut-on émettre une règle simple de cohérence entre $\pi(\mu)$ et la vraisemblance des données ?

Une idée simple

$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ est une **statistique exhaustive** de l'échantillon

Sous une hypothèse iid. des X_i , la loi de la variable aléatoire associée \bar{X}_n est, conditionnellement à (μ, σ^2)

$$\bar{X}_n | \mu, \sigma^2 \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Intégrée sur $\pi(\mu)$, la loi *a priori prédictive* de \bar{X}_n est

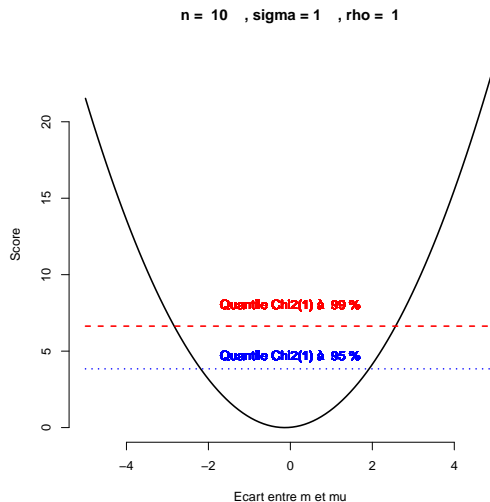
$$\bar{X}_n | \sigma^2 \sim \mathcal{N}\left(m, \sigma^2\left(\frac{1}{n} + \rho\right)\right)$$

Alors, *a priori* et *prédictivement*,

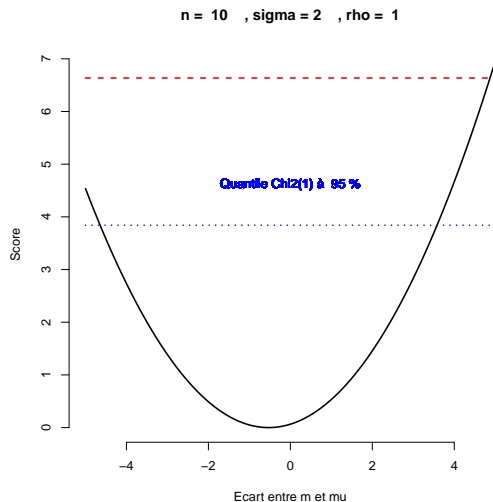
$$Z_n = \frac{(\bar{X}_n - m)^2}{\sigma^2\left(\frac{1}{n} + \rho\right)} \sim \chi_1^2$$

Il y a donc incohérence entre *a priori* et vraisemblance des données \mathbf{x}_n si la valeur *observée* de Z_n , à partir de \bar{x}_n , est une **valeur extrême** de la distribution du χ_1^2

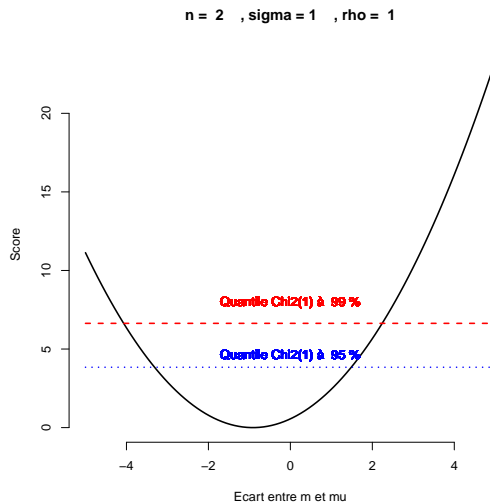
Test numériques



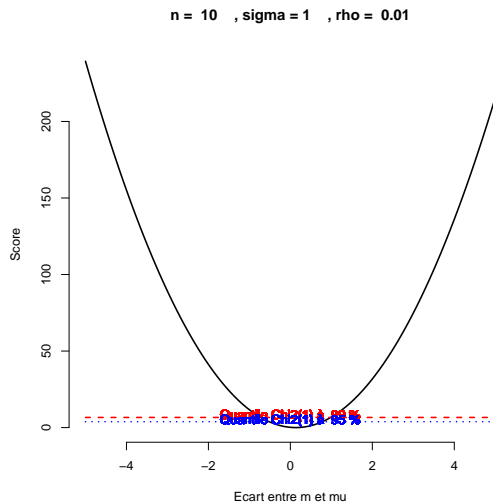
Influence d'un σ plus large



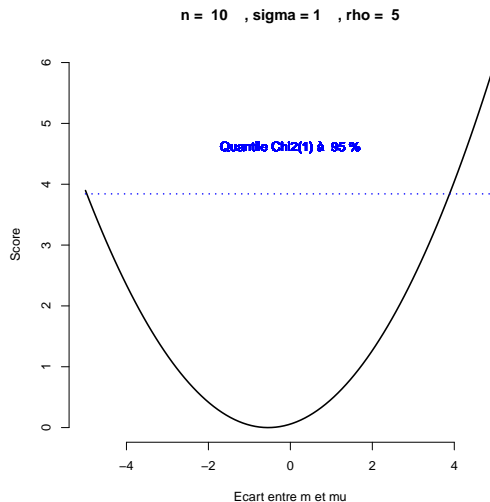
Influence d'une très faible taille d'échantillon



Influence d'une expertise extrêmement précise



Influence d'une expertise extrêmement vague



Modélisation *a priori* informative non conjuguée

La plupart des cas rencontrés en pratique sont non conjugués

De nombreux auteurs font des choix arbitraires (lois normale, gamma...) principalement en fonction des caractéristiques géométriques, de support ou d'échantillonnage de θ

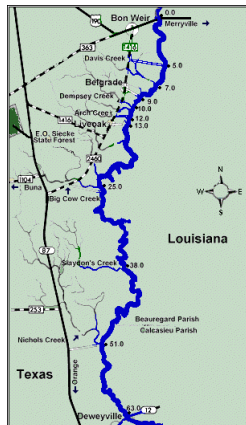
De plus, le sens *marginal* de l'information *a priori* n'est pas toujours bien compris

Exemple d'une étude typique en [statistique des extrêmes](#) (transparents suivants)

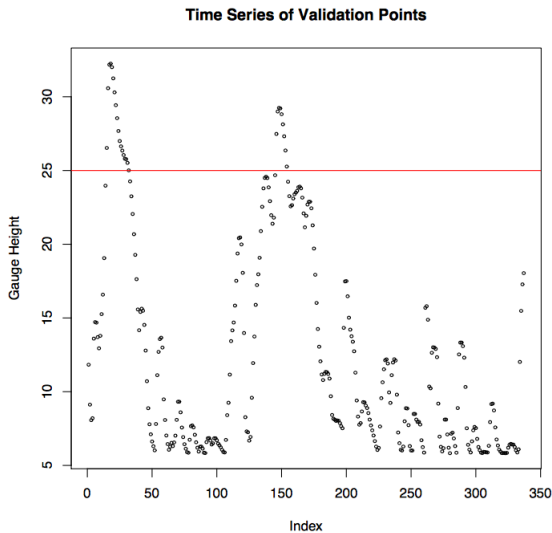
Un exemple de modélisation *a priori* subjective en risque extrême (1)

Issu de Gaioni *et al.* (2010). *Bayesian modeling of flash floods using generalized extreme value distribution with prior elicitation* (CJS).

Étude des crues éclairs de la rivière Sabine (Louisiane - Texas)



Données disponibles



Un exemple de modélisation *a priori* subjective en risque extrême (2)

Loi GEV sur les débits, de fonction de répartition

$$F(x|\theta) = \exp \left\{ - \left[1 + \lambda \left(\frac{x - \mu}{\sigma} \right) \right]_+^{-1/\lambda} \right\}$$

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

Un niveau de retour $x_q(\theta)$ correspondant à une période inter-événements $\simeq 1/q$ peut être estimé en résolvant

$$q = F(x_q|\theta)$$

Information *a priori*. Un expert connaissant la rivière est capable de produire trois couples d'estimateurs (q_i, x_{q_i}) , tels que $q_1 < q_2 < q_3$

Un exemple de modélisation *a priori* subjective en risque extrême (3)

Interprétation *paramétrique* de l'information *a priori* par les statisticiens \Rightarrow production d'un système d'équations (en supposant $\lambda \neq 0$)

$$\begin{aligned}\sigma &= \frac{\lambda(x - q_i - \mu)}{[-\log q_i]^{-\lambda} - 1} \\ \mu &= \frac{x_{q_i} K_j(\lambda) - x_{q_j} K_i(\lambda)}{K_j(\lambda) - K_i(\lambda)} \quad \text{où} \quad K_i(\lambda) = \frac{\lambda}{\sigma}(x_{q_i} - \mu) \\ 0 &= \sum_{i \neq j=1}^3 (x_{q_i} - x_{q_j}) \exp(-\log(-\log q_i))\end{aligned}$$

ce qui permet de produire un estimateur *a priori* $\tilde{\theta}$

En interrogeant plusieurs experts, ou/et un expert pouvant fournir plus de détails, on peut produire un échantillon d'estimateurs $\tilde{\theta}_1, \dots, \tilde{\theta}_k$

Les caractéristiques de cet échantillon (moyenne, variance, covariance...) peuvent (en théorie) être utilisées pour calibrer les hyperparamètres δ de $\pi(\theta|\delta)$

Posant $\delta = (\theta_0, \Sigma)$, les auteurs proposent un *a priori* log-multinormal : $\log \theta \sim \mathcal{N}_3(\log \theta_0, \Sigma)$

Une exemple de méta-analyse pour construire une modélisation *a priori* hiérarchique

Principe :

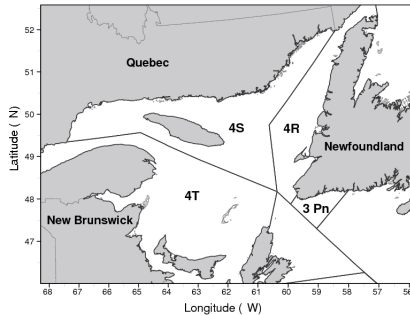
- Supposons disposer d'une observation représentative \mathbf{Y}^* sur $Y = g(\theta, c)$ où g et une fonction et c un ensemble de paramètres fixés
- Construire une vraisemblance liant \mathbf{Y}^* et θ
- Choisir un *a priori* non informatif $\pi^J(\theta)$ en fonction de cette vraisemblance
- Sélectionner π comme le posterior $\pi^J(\theta | \mathbf{Y}^*)$

Exemple : modèle à espace d'état pour une population (cohorte)

B., Chassot, Hammill, Duplisea (2008-2011)

Modélisation de l'abondance de la morue (*Gadus morhua*) dans le Golfe du Saint Laurent (Canada)

NAFO division 3Pn4RS



Dynamique

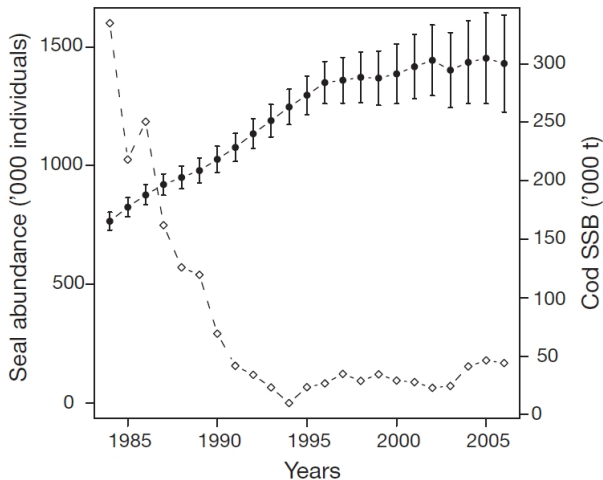
Une morue peut vivre 15 ans

Principales sources de mortalité :

- prédation par les phoques (*Phoca groenlandica*)
- pêche (en particulier durant les années 1990)
- naturelle (résiduelle, due aux variations de température de l'eau, etc.)



Observations : accroissement de la population de phoques corrélée au déclin des morues



Dynamique à espace d'états cachés de l'abondance $N_{a,t}$ de la morue

Prédation	$P_{a,t}$	$= p_{a,t}^c \cdot N_{a,t}$
Mortalité résiduelle 1	$N'_{a,t}$	$= p_{a,t}^m (N_{a,t} - P_{a,t},)$
Pêche commerciale	$C_{a,t}$	$= (1 - p_{a,t}^f) N'_{a,t}$
Abondance à mi-année	$N''_{a,t}$	$= N'_{a,t} - C_{a,t}/2$
Mortalité résiduelle 2	$N_{a+1,t+1}$	$= p_{a,t}^m (N''_{a,t} - C_{a,t}/2)$
Production d'oeufs totale	TEP_t	$= \sum_{a=1}^A N_{a,t} \xi_{a,t} \phi_{a,t} f_{a,t}$
Recrutement à l'âge 0	R_{t+1}	$= p_{t+1}^r \cdot TEP_t$
Recrutement à l'âge 0	$N_{1,t+2}$	$= (p_{0,t+1}^m)^2 R_{t+1}$

Sex ratio	ξ
Proportion de femelles matures	ϕ
Fécondité (nombre d'oeufs morue ⁻¹)	f

Observations

$$I_{a,t} = q_{\zeta_{a,s}} N_{a,t}''$$

$$\text{with } \zeta_{a,s} = \frac{1}{1 + \exp(-\gamma_s (a - \delta_s))}$$

and q

$$C_t = \sum_{a=1}^A C_{a,t}$$

$$p_{a,t,c} = C_{a,t} / \sum_{a=1}^A C_{a,t}$$

$$p_{a,t,s} = I_{a,t} / \sum_{a=1}^A I_{a,t}$$

Indice d'abondance

Sélectivité

Capturabilité

Prise totale

Probabilité observée de prise par âge

Probabilité observée d'abondance par âge

$$J_t^* = \sum_{a=1}^A J_{a,t}^* \stackrel{iid}{\sim} \mathcal{N} \left(\sum_{a=1}^A \log I_{a,t}(\theta), \psi^2 \right)$$

where $J_{a,t}^* = \log(I_{a,t}^*) = \log(I_{a,t}) + \epsilon_{a,t} + \eta_t$

$$\psi^2 = A\sigma^2 + A^2\tau^2$$

$$\log C_t^* \stackrel{iid}{\sim} \mathcal{N} \left(\log C_t(\theta) - \frac{\sigma_c^2}{2}, \sigma_c^2 \right)$$

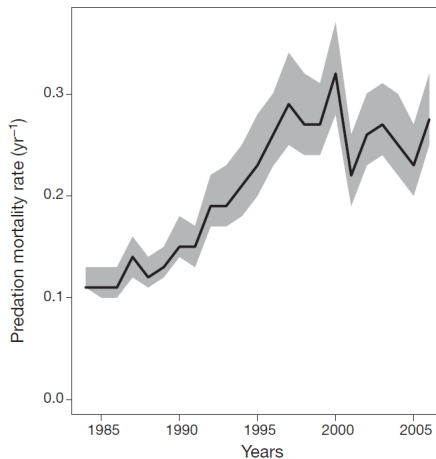
$$\sigma_c^2$$

Liste des paramètres inconnus

ζ_a	Baseline attack rate for age a (nb. attacks seal ⁻¹)
π	Normalization coefficient of attack rates
m	Shape parameter of the Holling response type
α	Intercept of the natural mortality curve (yr ⁻¹)
β	Slope of the natural mortality curve
F	Fishing mortality rate of cod (yr ⁻¹)
R_{\max}	Maximum nb. of cod recruits (NoI)
r	TEP needed to produce recruitment = $R_{\max}/2$ (NoE)
$\zeta_{a, c}$	Commercial selectivity-at-age
γ_c^1	Shape parameter of the commercial selectivity (1984-1993)
δ_c^1	Age of half-vulnerability (1984-1993)
γ_c^2	Shape parameter of the commercial selectivity (1994-2006)
δ_c^2	Age of half-vulnerability (1994-2006)
$\zeta_{a, s}$	Survey selectivity-at-age
q	Survey catchability
γ_s	Shape parameter of the survey selectivity
δ_s	Age of half-vulnerability

Résultats fréquentistes (Chassot et al. 2009)

Intervalles de confiance trop petits (bootstrap)



Élicitation *a priori* pour les paramètres de sélectivité

$$s_a = \frac{1}{1 + \exp(-\gamma\{a - \delta\})}.$$

- δ = age pour lequel 50% de la population est sensible à l'engin de pêche
- γ = paramètre de forme

Méta-analyse des estimés de la sélectivité obtenus à partir des mesures de surveillance / pêche commerciale pour des engins similaires à ceux étudiés

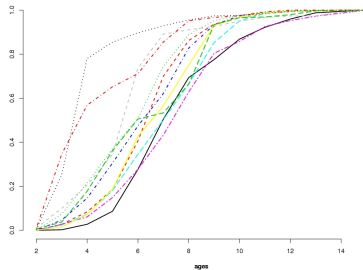
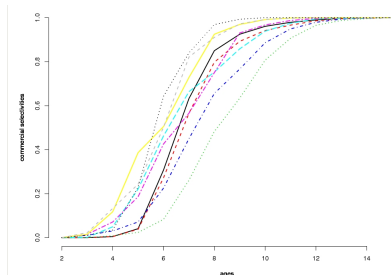
Fondé sur une idée de Harley and Myers (2001)

$M = 153$ jeux de données

Soit c_1, \dots, c_{A^+} un échantillon de prise par âge

Estimateur de Kaplan-Meier = fréquence cumulée par âge

$$\varsigma_a^* = \sum_{i=1}^A c_i \mathbb{1}_{\{i \leq a\}} / \sum_{j=1}^A c_j.$$



Notre objectif est

- de définir une vraisemblance $\ell(\varsigma_{1,i}^*, \dots, \varsigma_{A,i}^*, i = 1, \dots, M | \gamma, \delta)$
- de définir des *a priori* non informatifs pour (γ, δ) en fonction de ℓ

Ôter la dépendance à l'âge

Considérons la reparamétrisation

$$\mathbf{s}_a = -\log(\zeta_a^{-1} - 1) = \gamma(a - \delta) \quad (9)$$

et soit \mathbf{s}_a^* le vecteur correspondant des estimés non paramétriques

Estimer et tester l'hypothèse de modèle

$$\mathbf{s}_a^* = \mathbf{s}_a + \mathcal{N}(0, \sigma_a^2)$$

Des tests classiques (Shapiro-Wilks, etc.) ne nient pas l'hypothèse gaussienne (p -values $\in [0.35, 0.86]$)

Notons $\mathbf{s}_I^* = (s_1^*(i_1), \dots, s_A^*(i_A))$, avec $i_j \neq i_k$, le i_j étant choisi dans $I \subset \{1, \dots, M\}$, et

$$\bar{\mathbf{s}}^* = \frac{1}{A} \sum_{j=1}^A \mathbf{s}_j^*(i_j) = \alpha\gamma - \delta + \mathcal{N}(0, \sigma^2)$$

avec $\sigma^2 = \sum_{a=1}^A \sigma_a^2 / A$ et $\alpha = (A + 1)/2$.

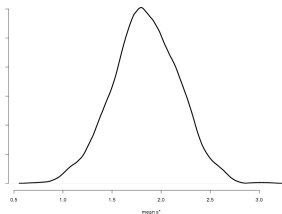
Minimiser les corrélations pour obtenir une vraisemblance simple

Il y a $(M!)/(M - A^+!)$ valeurs possibles \bar{s}^*

Elles ne sont pas indépendantes (puisqu'elles sont construites à partir de données venant des mêmes sélectivités empiriques)

Sélectionner des âges éloignés (a_1, a_2), $s_{a_1}^*(i_{a_1})$ and $s_{a_2}^*(j_{a_2})$ permet de diminuer la corrélation

Un mélange de distributions de \bar{s}^* peut être produit par simulation



La structure formelle suivante apparaît pertinente

$$\bar{s}^* = \bar{s} + \mathcal{N}(0, \nu^2)$$

Construire un *a priori* informatif sur (γ, δ)

- 1 Soit $\pi^J(\gamma, \delta)$ un *a priori* non informatif pour la vraisemblance
 - *reference prior* de Berger & Bernardo (1992)
 - des experts biologistes s'accordent sur le plus large intervalle possible pour δ : $[a_l, a_r] = [1, 6] \subset [1, A]$

$$\pi^J(\gamma, \delta) \propto \mathbb{1}_{\gamma \geq 0} \mathbb{1}_{\{a_l \leq \delta \leq a_r\}}$$

- 2 Considérons la vraisemblance d'"une donnée représentative" émanant de

$$\bar{s} = \alpha\gamma - \delta + \mathcal{N}(0, \nu^2 + \sigma^2)$$

- 3 Construisons $\pi(\gamma, \delta) = \pi^J(\gamma, \delta | \bar{s}^*)$, soit

$$\pi(\gamma, \delta) \propto \exp \left\{ -\frac{1}{2(\sigma^2 + \nu^2)} (\alpha\gamma - \delta - \bar{s})^2 \right\} \mathbb{1}_{\{\gamma \geq 0\}} \mathbb{1}_{\{a_l \leq \delta \leq a_r\}}$$

	scientifique	commercial		scientifique	commercial
σ^2	1.221	1.510	\bar{s}	1.891	1.493
ν^2	0.1146	0.1051	α	6.5	6.5

Paramètres de nuisance

Une fois que la loi *a posteriori* du vecteur θ est obtenue, des **études de projection** doivent être faites

Souvent $\theta = (\theta_I, \theta_N)$ où

- $\theta_I = \{\text{paramètres d'intérêt}\}$ (ex : paramètres de selectivité, recrutement...)
- $\theta_N = \{\text{paramètres de nuisance}\}$ (variances d'observation, capturabilité)
 - purement relatives à l'obtention des données

$$J_t^* = \sum_{a=1}^A J_{a,t}^* \stackrel{iid}{\sim} \mathcal{N} \left(\sum_{a=1}^A \log I_{a,t}(\theta), \psi^2 \right)$$

$$\log C_t^* \stackrel{iid}{\sim} \mathcal{N} \left(\log C_t(\theta) - \frac{\sigma_c^2}{2}, \sigma_c^2 \right)$$

Mesure *a priori* conditionnelle de Berger-Bernardo pour les paramètres de nuisance

Pas d'information *a priori* usuellement connue sur $\theta_N = (q, \sigma_c^2, \psi^2)$

Le choix d'un prior non informatif $\pi(\theta_N)$ doit être indépendant du choix de tout prior informatif sur θ_I

Divergence de Kullback-Leibler entre posterior et prior

$$\text{KL}(\pi|\mathbf{x}) = \int_{\Theta_N} \pi(\theta_N|\mathbf{x}) \log \frac{\pi(\theta_N|\mathbf{x})}{\pi(\theta_N)} d\theta_N$$

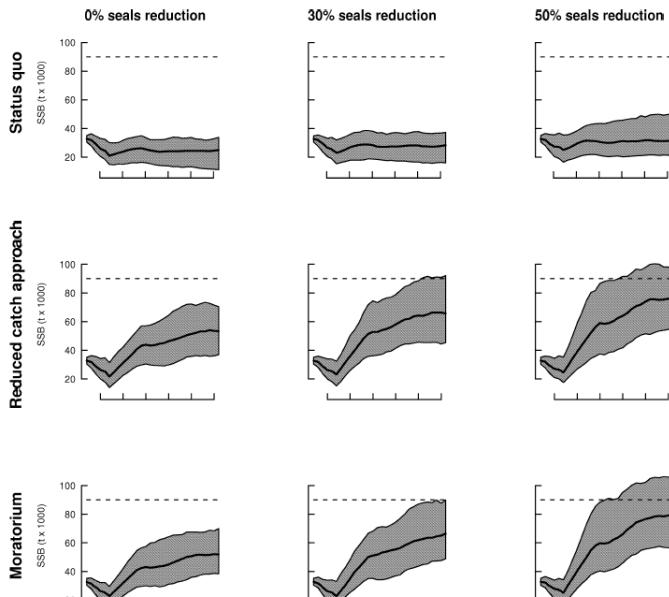
with $\pi(\theta_N) = \int \pi(\theta) d\theta_I$

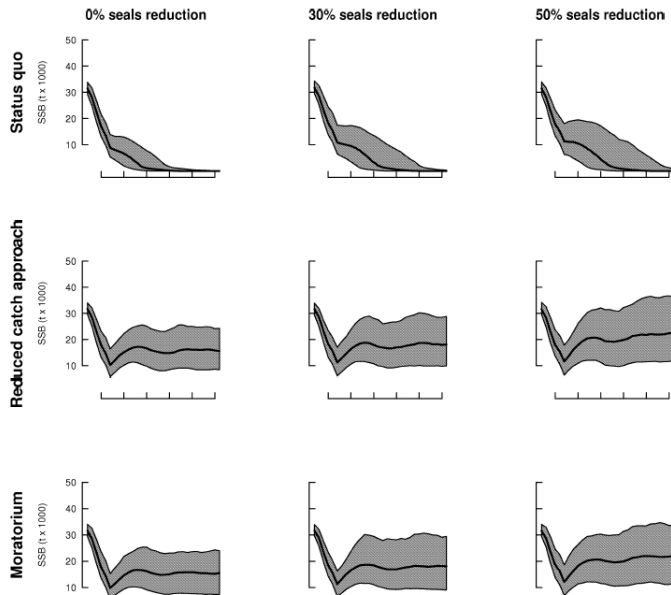
On élicite

$$\pi^* = \arg \max_{\pi} \left\{ \lim_{\text{card}(\mathbf{x}) \rightarrow \infty} \mathbb{E}_m [\text{KL}(\pi|\mathbf{X})] \right\}$$

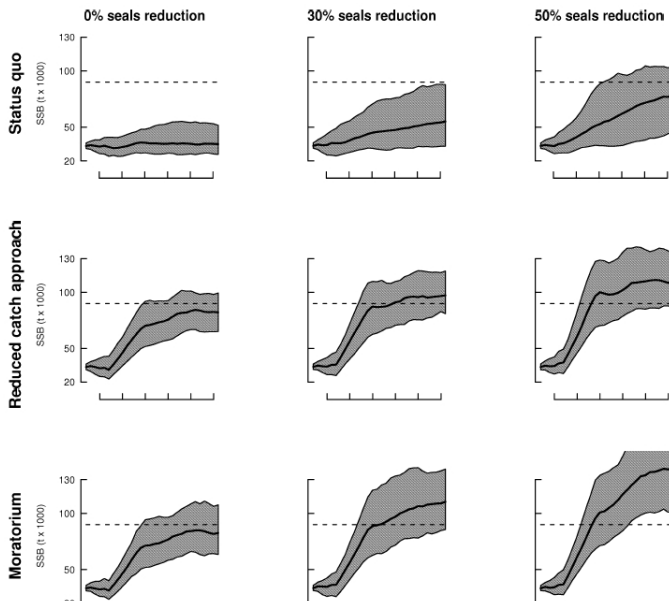
et on trouve (après des calculs complexes)

$$\pi^*(\psi^2, \sigma_c^2, q) \propto \psi^{-3} \sigma_c^{-3} q^{-1} \mathbb{1}_{\{(\psi, \phi, q) \in \mathbf{R}_{+,*}^3\}}.$$

Quelques projections *a posteriori* prédictives (B., Chassot et al. 2011)

Quelques projections *a posteriori* prédictives (2)

Quelques projections *a posteriori* prédictives (3)



Une vision critique

Interprétation paramétrique \neq compréhension marginale : l'expertise s'applique sur la **variable d'ancrage** X de loi

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

et non sur $f(x|\theta)$

Échantillon $\tilde{\theta}_1, \dots, \tilde{\theta}_k$ de taille extrêmement faible en pratique \Rightarrow statistiques peu fiables

Pourquoi un choix lognormal ?

Quelques préoccupations majeures pour faire mieux (plus défendable)

- Se rapprocher de règles produisant des *a priori* conjugués dans les cas le permettant
- Vérifier la cohérence de $\pi(\theta)$ vis-à-vis de la structure paramétrique de $f(x|\theta)$
- Donner le maximum de signification aux hyperparamètres δ (notamment des paramètres mesurant la "force" de l'information *a priori* par rapport à celle apportée par les données utilisées dans la vraisemblance classique)

Une possibilité privilégiée : construire la loi *a priori* comme une *loi a posteriori approximative*, sachant des *données virtuelles* $\tilde{\mathbf{x}}_m$ et une *mesure a priori non informative* $\pi^J(\theta)$:

$$\pi(\theta) \simeq \pi^J(\theta|\tilde{\mathbf{x}}_m)$$

m reflète la force de l'information *a priori* et peut être modulée (dans une phase de calibration ou d'analyse de sensibilité)

Permet de corrélérer automatiquement les dimensions de θ

⇒ *virtual data pseudoposterior priors*

Méthodologie (1)

Hypothèse 1 (Kadane-Berger) : il est possible d'exprimer de l'information *a priori* sous la forme de la spécification d'un ou plusieurs quantiles prédictifs *a priori* (ou *marginiaux*) $(x_{\alpha_i}, \alpha_i)_{1 \leq i \leq N}$, tels que $\alpha_i < \alpha_{i+1}$ et $x_{\alpha_i} < x_{\alpha_{i+1}}$, définis par

$$P(X \leq x_{\alpha_i}) = \int_{\Theta} P(X \leq x_{\alpha_i} | \theta) \pi(\theta) d\theta$$

ou, pour des niveaux observés x_{α_i} , de fournir des estimateurs des α_i

Arguments

- Ces informations sont corrélées (c'est en s'exprimant sur une valeur de X qu'une autre valeur de X peut être positionnée)
- La distribution jointe des informations *a priori* reste invariante par toute permutation de ces informations (*l'ordre de proposition des quantiles ne devrait pas importer*)

Remarque : ne pas oublier que des fonctions de coût se cachent derrière l'interprétation sous forme de quantiles !

Un aparté sur le théorème de De Finetti (1931)

Soit X_1, \dots, X_n, \dots 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

$$P(X_1 = x_1, \dots, X_n = x_n, \dots) = \int_{\Theta} f(x_1, \dots, x_n, \dots | \theta) \pi(\theta) d\theta$$

où $f(x_1, \dots, x_n | \theta)$ est la vraisemblance d'observations **iid** de Bernoulli

Généralisé par Hewitt, Savage (1955), Diaconis, Freedman (1980) pour l'ensemble des distributions discrétisées puis continues

Conséquences :

- La modélisation bayésienne apparaît comme une modélisation statistique naturelle de variables corrélées mais échangeables
- L'existence formelle d'un prior $\pi(\theta)$ est assurée en fonction du mécanisme d'échantillonnage
 $= \{ \text{information incertaine à propos } \theta \}$

Méthodologie (2)

Pour une densité connue $f(x|\theta)$

- ❶ sélectionner une mesure *a priori* non informative $\pi^J(\theta)$
- ❷ supposer qu'il existe un échantillon virtuel \mathbf{x}_m de taille m portant l'information *a priori*
- ❸ construire une approximation de la loi *a priori* informative $\pi(\theta) \equiv \pi^J(\theta|\mathbf{x}_m)$ comme

$$\pi(\theta) = \pi(\theta|\Delta_m)$$

où Δ_m est un ensemble de statistiques virtuelles

- ❹ estimer Δ_m par $\hat{\Delta}_m = \arg \min_{\delta_m} \mathcal{D}(\Lambda_e, \Lambda(\delta_m))$
 - Λ_e sont des caractéristiques souhaitées *a priori*
 - $\Lambda(\delta_m)$ sont des caractéristiques effectives de la loi *a priori*
 - \mathcal{D} est une distance ou divergence

sous des contraintes d'homogénéité

Méthodologie (3)

Ce qui empêche (actuellement) d'aboutir à une méthodologie complètement formalisée est l'**approximation à produire**

On souhaite retrouver des lois simples

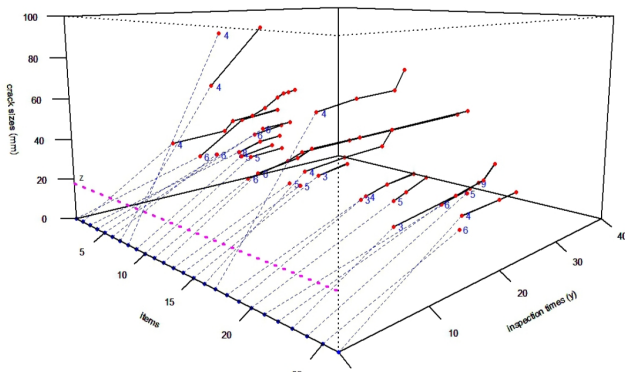
Pas de démarche unique

Exemple : processus gamma pour modéliser des accroissements de fissure

Une taille de fissure $Z_{k,t}$ sur un composant k est monotone croissante au cours du temps t

Les incréments (supposés indépendants) $X_{k,i} = Z_{k,t_i} - Z_{k,t_{i-1}}$ sont supposés obéir à des lois gamma

$$f_{\alpha(t-s),\beta}(x) = \frac{1}{\Gamma(\alpha_i(t-s))} \cdot \frac{x^{\alpha(t-s)-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha(t-s)}} \mathbb{1}_{\{x \geq 0\}}$$



Comment construire un *a priori* informatif sur (α, β) ?

Prenons une mesure non informative (Jeffreys) $\pi^J(\alpha, \beta) \propto \frac{1}{\beta} \sqrt{\alpha \Psi_1(\alpha) - 1}$

Loi *a posteriori* d'un échantillon virtuel d'incrément de fissure $\mathbf{x}_m = (\tilde{x}_1, \dots, \tilde{x}_m)$ observés aux temps $\mathbf{t}_m = (\tilde{t}_1, \dots, \tilde{t}_m)$

$$\beta | \alpha \sim \mathcal{IG}(\alpha m \tilde{t}_{e,1}, m \tilde{x}_e)$$

$$\alpha \sim \mathcal{G}(m/2, m \tilde{t}_{e,2})$$

dont la signification est fournies par

$$\tilde{t}_{e,1} = \frac{1}{m} \sum_{i=1}^m \tilde{t}_i \quad (\text{temps moyen d'observation})$$

$$\tilde{x}_e = \frac{1}{m} \sum_{i=1}^m \tilde{x}_i \quad (\text{accroissement moyen})$$

$$\tilde{t}_{e,2} = \frac{1}{m} \sum_{i=1}^m \tilde{t}_i \log \frac{\sum_{j=1}^m \tilde{x}_j / \tilde{x}_i}{\sum_{j=1}^m \tilde{t}_j / \tilde{t}_i} \quad (\text{hyperparamètre de calage})$$

Comment construire un *a priori* informatif sur (α, β) ?

En effet, on a

$$\pi(\alpha | \mathbf{x}_m, \mathbf{t}_m) \propto \exp \left(-\alpha \left\{ \sum_{i=1}^m \tilde{t}_i \log \frac{\sum_{j=1}^m \tilde{x}_j}{\tilde{z}_i} \right\} \right) \frac{\Gamma(m\alpha \tilde{t}_{e,1})}{\prod_{i=1}^m \Gamma(\alpha \tilde{t}_i)} \sqrt{\alpha \Psi_1(\alpha) - 1}$$

et le développement peut être trouvé en utilisant les approximations suivantes :

- **Formule exacte de Stirling**

$$\Gamma(x) = \sqrt{2\pi x} x^{-1/2} \exp(-x + \nu(x)) \quad \text{où } \nu(x) = \gamma/(12x) \text{ et } \gamma \in [0, 1]$$

- **Développement de Laurent**

$$\sqrt{\alpha \Psi_1(\alpha) - 1} = \frac{1}{\sqrt{2\alpha}} \left(1 + \sum_{k=1}^{\infty} \frac{B_{2k}}{\alpha^{2k-1}} \right)$$

où les B_{2k} sont les nombres de Bernoulli-Faulhaber de seconde nature

Calibrer à partir d'opinion d'expert

La valeur la plus probable *a priori* de l'accroissement de fissure moyen durant l'intervalle de temps Δ_i est

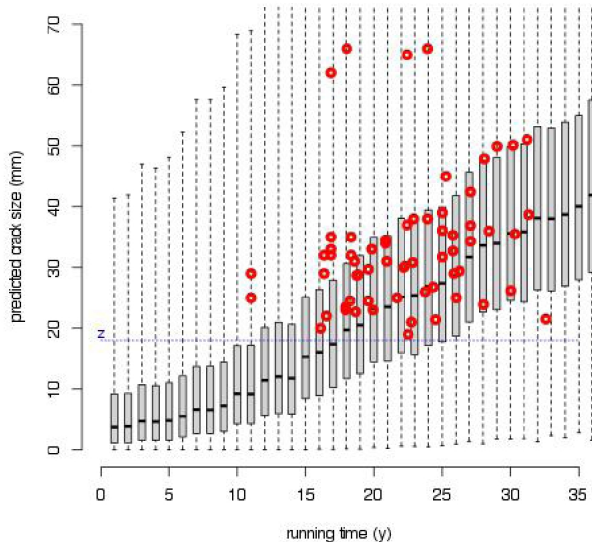
$$\hat{r}(\Delta_i) = \frac{\tilde{x}_e \Delta_i}{\tilde{t}_{e,1}}.$$

Questionner un expert. Durant les prochaines 15 puis 30 années (= valeur de $m\tilde{t}_{e,1}$), quelles sont les chances $(1 - \delta_1, 1 - \delta_2)$ qu'une quelconque fissure apparaissant sur le composant soit plus grande que $(z_1, z_2) = (5, 10)$ mm ? soit, pour $i = \{1, 2\}$,

$$P\left(Z_{m\tilde{t}_{e,1}} < z_i\right) = \delta_i = \int_0^{z_i} \int_0^\infty \frac{x^{\alpha m\tilde{t}_{e,1}-1} (m\tilde{x}_e)^{\alpha m\tilde{t}_{e,1}} \Gamma(2\alpha m\tilde{t}_{e,1})}{(m\tilde{x}_e + x)^{2\alpha m\tilde{t}_{e,1}} \Gamma^2(\alpha m\tilde{t}_{e,1})} \pi(\alpha) d\alpha dx$$

Calibrer les hyperparamètres de calage en minimisant en $(m, \tilde{t}_{e,2})$ la distance L_2 relative

$$\sum_{i=1}^2 \left\{ 1 - \delta_i^{-1} P\left(Z_{m\tilde{t}_{e,1}} < z_i\right) \right\}^2$$

Accord entre données et distribution *a priori* prédictive

Une alternative courante au critère des moindres carrés

L'adéquation entre la représentation de l'information *a priori* fournie par des couples (x_{α_i}, α_i) et ce qu'il est possible de modéliser – les couples $(x_{\alpha_i}, \tilde{\alpha}_i(\delta))$ – est définie par une fonction de perte minimisée en δ

Cooke (1991) a proposé une fonction de perte issue de la discrétisation de la divergence de Kullback-Leibler entre la loi inconnue sur X fournissant les quantiles x_{α_i} (qu'on pourrait nommer "loi d'expertise") et la loi prédictive *a priori* $f(x|\delta)$

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

Critère de Cooke

$$\delta^* = \arg \min_{\delta} \sum_{i=0}^M (\alpha_{i+1} - \alpha_i) \log \frac{(\alpha_{i+1} - \alpha_i)}{(\tilde{\alpha}_{i+1}(\delta) - \tilde{\alpha}_i(\delta))},$$

avec $\alpha_0 = \tilde{\alpha}_0 = 0$ et $\alpha_{M+1} = \tilde{\alpha}_{M+1} = 1$

Pondération possible, convexité globale non assurée

Diminuer au maximum la dimension de δ (par exemple en fixant les autres hyperparamètres, comme une taille virtuelle) permet d'accroître cette possibilité

Exemple (TP) d'élicitation *a priori* pour un modèle de Weibull

Voici un jeu de données (réel) x_n de durées de vie de tubes protecteurs de chaudière (en mois).

71.4	166.3	93.2	59.6	181.6	144.8	87.3	100.3	
90.0	173.9	95.4	44.1	149.4	73.7	86.3	145.1	167.7

Vous bénéficiez de deux experts qui vous fournissent chacun, après un processus d'interrogation minutieux, les renseignements suivants :

	Durée de vie médiane (m)	Percentile 33%	Percentile 90%
Expert 1	100*	80	200
Expert 2	130	100	200*

Proposez une modélisation *a priori* de ces informations

Analyse de sensibilité

Le choix d'un prior informatif π (conjugué ou non) est généralement subjectif

Indispensable de tester l'impact de variations de π sur le résultat *a posteriori*

Comment faire varier π dans une classe \mathcal{C} ?

Deux grandes approches (mais pas exhaustives)

- 1 Classes d' ϵ —contamination
- 2 *Exponential twisting*

Classes d' ϵ —contamination

COURS SI TEMPS RESTANT

Exponential twisting

COURS SI TEMPS RESTANT

Problèmes conceptuels et pratiques de l'interrogation d'expert

Les experts sont soumis à de nombreux biais qui limitent parfois la **pratique bayésienne subjective**, fondée sur une interprétation décisionnelle de leurs opinions sous des contraintes d'**indifférence au risque**

- ❶ les biais de situation, dus au filtre mental de l'expert vis-à-vis de la réalité, incluant :
 - les **biais cognitifs** liés aux limites intellectuelles, et à la difficulté de réviser son jugement lorsque de nouvelles informations arrivent
 - les **biais motivationnels** liés au processus d'élicitation et à la pression de l'environnement
- ❷ les biais de confiance excessive : une valeur vraie affirmée par un expert avec 90% de chance se situe en réalité autour de 30 à 60%, la moyenne de l'expert correspond à la médiane

Une grande littérature de recherche, établissant des ponts avec la psychologie des individus et des groupes, est consacrée à la vérification des contraintes (cf. Tversky et Kahneman 1973)

D'autres théories de la représentation de la connaissance ont émergé depuis les années 1960 (ex : logique floue)

Il n'en reste pas moins que la statistique bayésienne offre un **cadre décisionnel cohérent** (de par le respect des axiomes des probabilités), **pratique**, et que **les experts sont souvent la seule source d'information disponible, qui puisse permettre d'effectuer des prévisions dans un processus décisionnel**

7 - Autre approche et conclusions

Une autre approche de la construction *a priori* : le principe "bayésien" empirique

On suppose n'avoir pas d'information *a priori*

Soient x_1, \dots, x_{n+1} des observations indépendantes de densités $f(x_i|\theta_i)$

On veut inférer sur θ_{n+1} en supposant que tous les θ_i ont été produit par le même $\pi(\theta)$

On cherche donc à reconstruire $\pi(\theta)$ tel que

$$x_1, \dots, x_n \sim f(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta$$

- 1 soit de façon **non-paramétrique** en produisant un *estimateur* $\hat{\pi}_n(\theta)$
- 2 soit de façon **paramétrique** en fixant une forme hyperparamétrée $\pi(\theta|\delta)$ et en produisant un estimateur $\hat{\delta}_n$

L'ensemble de cette démarche n'est en fait pas pas bayésienne car elle utilise deux fois les données

Points-clés à retenir en élicitation *a priori*

Décider d'un *a priori de référence* (non-informatif) pour l'inférence considérée

Décider de la *forme* d'une loi *a priori* informative en se basant sur des théorèmes de convergence (ex : TLC) ou un raisonnement entièrement bayésien (ex : données virtuelles)

Décider du *sens de l'information apportée a priori* dans un cadre de théorie de la décision, lorsqu'elle est *subjective*

- **Ex** : ne pas oublier que l'interprétation en termes de quantile dans l'exemple fiabiliste provient du **choix d'une fonction de coût**

Ne pas oublier de prouver l'*intégrabilité* de la densité *a posteriori*

$$\int_{\Theta} \ell(\mathbf{x}_n|\theta)\pi(\theta) d\theta < \infty$$

Quelques références

- ❶ Clemen, R. T., Winkler, R. L. (2007). Aggregating probability distributions. In : *Advances in Decision Analysis*. Cambridge University Press. .
- ❷ O'Hagan, A., Buck, C.E., Daneshkhah, A. Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T. (2006). Uncertain Judgements : Eliciting Experts' Probabilities. Statistics in practice. Wiley
- ❸ Parent, E., Bernier, J. (2007). Le raisonnement bayésien : modélisation et inférence. Springer
- ❹ Robert, C.P. (2006). Le choix bayésien. Principes et pratique. Springer
- ❺ Kass, R., Wasserman, L. (1996). Formal rules of selecting prior distributions : a review and annotated bibliography. *Journal of the American Statistical Association*.
- ❻ Evans, M., Moshonov. H. (2006). Checking for prior-data conflict. *Bayesian Analysis*
- ❼ Bousquet, N. (2008). Diagnostics of prior-data agreement in applied Bayesian analysis. *Journal of Applied Statistics*
- ❽ Walter, G. , Augustin, T. (2009). Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice*

Régions de confiance et de crédibilité (rappel)

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

Définition

Une région A de Θ est dite α -crédible si $\Pi(\theta \in A|x) \geq 1 - \alpha$

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 ∞ ,

$$P_{\theta}(\theta \in A) \geq 1 - \alpha$$

La définition bayésienne exprime la probabilité que $\theta \in A$ au vu (*conditionnellement*) des expériences déjà réalisées

- pas besoin d'avoir recours à un nombre ∞ d'expériences similaires

Une région α -crédible peut être estimée par les quantiles empiriques de la **simulation *a posteriori***

3 - Une approche rapide des tests d'hypothèse

Supposons qu'on cherche à mener le test d'une *hypothèse nulle* $H_0 : \theta \in \Theta_0$

La fonction de coût $L(\theta, d)$ 0-1 est proposée dans l'approche classique de Neyman-Pearson :

$$L(\theta, d) = \begin{cases} 1 & \text{si } d \neq \mathbb{1}_{\Theta_0} \\ 0 & \text{sinon} \end{cases} \quad (10)$$

menant à l'estimateur bayésien dans $\mathcal{D} = \{0, 1\}$

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

qui fait sens intuitivement : l'estimateur choisit l'hypothèse avec la probabilité *a posteriori* la plus grande.

On peut généraliser en pénalisant différemment les erreurs suivant que H_0 est vraie ou fausse

$$L(\theta, d) = \begin{cases} 0 & \text{si } d = \mathbb{1}_{\Theta_0} \\ a_0 & \text{si } \theta \in \Theta_0 \text{ et } d = 0 \\ a_1 & \text{si } \theta \notin \Theta_0 \text{ et } d = 1 \end{cases} \Rightarrow \delta^\pi(x) = \begin{cases} 1 & \text{si } \Pi(\theta \in \Theta_0|x) > a_1/(a_0 + a_1) \\ 0 & \text{sinon} \end{cases}$$

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

Facteur de Bayes (1/2)

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

Soit $H_1 : \theta \in \Theta_1$ une hypothèse alternative

Définition

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

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

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

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

Sous le coût généralisé précédent, en posant

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

l'hypothèse H_0 est acceptée si $B_{01}(x) > (a_1\gamma_1)/(a_0\gamma_0)$

Facteur de Bayes (2/2)

En l'absence d'un cadre décisionnel véritable (qui consisterait à pouvoir fixer a_0 et a_1), une **échelle "absolue"** a été proposée par Jeffreys (1939), remaniée depuis par Kass & Raftery (1995), pour évaluer le **degré de certitude en faveur ou au détriment de H_0 apporté par les données**

- (i) si $\Lambda = \log_{10} B_{10}(\mathbf{x}_n)$ varie entre 0 et 0.5, la certitude que H_0 est *faus*se est *faible*
- (ii) si $\Lambda \in [0.5, 1]$, cette certitude est *substantielle*
- (iii) si $\Lambda \in [1, 2]$, elle est *forte*
- (iv) si $\Lambda > 2$, elle est *décisive*

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

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

Exemple

Pour des données discrètes x_1, \dots, x_n , on considère un modèle de Poisson $\mathcal{P}(\lambda)$ ou une loi binomiale négative $\mathcal{NB}(m, p)$ avec les *a priori*

$$\begin{aligned}\pi_1(\lambda) &\propto 1/\lambda \\ \pi_2(m, p) &= \frac{1}{M} \mathbb{1}_{\{1, \dots, M\}}(m) \mathbb{1}_{[0, 1]}(p)\end{aligned}$$

Peut-on sélectionner l'un des deux modèles ?

Difficultés posées par les *a priori* impropres

La *distribution prédictive a posteriori*

$$f(\mathbf{x}_n) = \int_{\Theta} f(\mathbf{x}_n|\theta)\pi_0(\theta) d\theta$$

est définie à une constante inconnue près C .

Cette constante affecte multiplicativement tout *rapport de Bayes*

$$B_{12} = \frac{f_1(\mathbf{x}_n)}{f_2(\mathbf{x}_n)}$$

qui interdit théoriquement une **sélection** entre les modèles $\mathcal{M}_1 = \{f_1(x|\theta_1) + \pi_1(\theta_1)\}$ et $\mathcal{M}_2 = \{f_2(x|\theta_2) + \pi_0(\theta_2)\}$,

Nécessité d'utiliser des heuristiques fondées sur la notion d'*échantillon d'entraînement* (*a priori intrinsèques* et *fractionnaires* de Berger-Perrichi-O'Hagan)