

Cours accéléré Statistiques

Partie 3: Introduction à la statistique bayésienne

Master 2 Mathématiques et Applications

Christine Keribin

¹Laboratoire de Mathématiques d'Orsay
Université Paris-Saclay

2022-2023

université
PARIS-SACLAY

FACULTÉ
DES SCIENCES
D'ORSAY



Cours accéléré
Statistiques
Partie 3:
Introduction à la
statistique
bayésienne

Christine Keribin

Introduction

Estimation

Loi a priori

Intervalle de
confiance

Test d'hypothèses



C. P. Robert.

Le choix bayésien

Springer, 2006.



J-M. Marin et C. P. Robert.

Bayesian Essentials with R.

Springer, 2014.

- Résoudre un problème inverse : déterminer le paramètre θ du mécanisme probabiliste générateur
- en proba, on est conditionnel à θ : **densité**

$$x \mapsto f_{\theta}(x) = f(x|\theta)$$

- en stat, on est conditionnel aux observations x :
vraisemblance

$$\theta \mapsto L(\theta; x) = L(\theta|x) = f_{\theta}(x)$$

Vision fréquentiste - vision bayésienne

► vision fréquentiste :

- θ est estimé par $\hat{\theta}$ avec une certaine incertitude car l'échantillon X est fini
- on évalue en moyenne

► vision bayésienne :

- modéliser l'incertitude sur θ par une loi de probabilité **a priori** $\pi(\theta)$
- l'inférence bayésienne consiste à déterminer la loi **a posteriori** $\pi(\theta|X = x)$

Théorème (Bayes)

Si A et E sont deux événements tels que $\mathbb{P}(E) \neq 0$, alors

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

Version continue du théorème de Bayes

Théorème (Bayes (1763))

Soient X et Y deux variables de loi jointe de densité $\varphi(x, y)$,
de densité conditionnelle $f(x|y)$ et de densité marginale
 $g(y) = \int \varphi(x, y) dx$,

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

Application :

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta) d\theta} = \frac{L(\theta; x)\pi(\theta)}{m(x)}$$

où $m(x)$ est appelé **vraisemblance intégrée** : constante de
normalisation de la loi a posteriori

Définition

Un *modèle statistique bayésien* est constitué d'un modèle statistique paramétrique $f(x|\theta)$ et d'une distribution a priori pour le paramètre $\pi(\theta)$.

- Le théorème de Bayes actualise l'information sur la loi de θ en "extrayant" celle contenue dans l'observation x
- novateur : paramètre inconnu \rightarrow paramètre aléatoire

Exemple $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \sim \mathcal{N}(\mu_0, \tau^2)$

Intérêts

- petits échantillons
- remplace une maximisation par une intégration
 \hookrightarrow rend les estimateurs plus stables
- connaissance experte du domaine

Choix de la loi a priori $\pi(\theta)$

l'utilisation d'une loi sur θ permet de résumer l'information disponible sur θ et l'incorporation de cette information inexacte dans le processus de décision

Choix en fonction

- ▶ de ce qu'on en connaît avant l'observation ou pas (prior non informatif)
- ▶ de considérations pratiques (faisabilité de calculs,)

Décision fréquentiste

Soit $\ell(\theta, \delta)$ une fonction de perte, et $\delta(x)$ un estimateur de $\theta \in \mathbb{R}^p$

Le **risque fréquentiste** est défini par $R(\theta, \delta) = \mathbb{E}_\theta[\ell(\theta, \delta(X))]$

- ▶ en fq, les estimateurs sont évalués suivant leur performance à long terme (en moyenne) pour toutes les valeurs possibles du paramètre θ
- ▶ mais ce n'est pas forcément optimal pour les valeurs observées sur l'échantillon
- ▶ difficultés à avoir des performances uniformément meilleures
- ▶ dépend de θ inconnu

↪ : idée bayésienne : intégrer sur l'espace des paramètres pour pallier les difficultés du risque fréquentiste

Décision bayésienne

Soit $\ell(\theta, \delta)$ une fonction de perte, et π une loi a priori,

Définition

Le *coût moyen a posteriori* ou *risque a posteriori* est défini par

$$\rho(\pi, \delta|X) = \mathbb{E}^{\pi}[\ell(\theta, \delta(X))|X]$$

↪ Le problème change suivant les données (qui sont connues)...

Définition

Le *risque intégré* est défini par

$$r(\pi, \delta) = \mathbb{E}^{\pi}[R(\theta, \delta)]$$

où $R(\theta, \delta) = \mathbb{E}_{\theta}[\ell(\theta, \delta(X))]$ est le risque (fréquentiste)

↪ on associe un nombre réel à chaque estimateur : **ordre total** sur les estimateurs

Définition

Un *estimateur de Bayes* associé à une loi a priori π et une fonction de coût ℓ est un estimateur δ^π minimisant le risque intégré $r(\pi, \delta)$
 $r(\pi) = r(\pi, \delta^\pi)$ est appelé *risque de Bayes*

Théorème (Méthode de calcul de l'estimateur de Bayes)

S'il existe une décision δ de risque intégré fini $r(\pi, \delta) < \infty$ et si

$$\forall x \in \mathcal{X}, \quad \delta^\pi(x) = \operatorname{Arg} \min_{\delta} \rho(\pi, \delta|x)$$

alors $\delta^\pi(X)$ est un estimateur bayésien.

Introduction

Estimation

Loi a priori

Intervalle de
confiance

Test d'hypothèses

Théorème

*L'estimateur de Bayes associé à la loi a priori π et au coût **quadratique** : $\ell(\theta, \delta) = \|\theta - \delta\|^2$ est la moyenne a posteriori*

Théorème

*L'estimateur de Bayes associé à la loi a priori π et à la **perte** L^1 : $\ell(\theta, \delta) = |\theta - \delta|$ est la médiane a posteriori*

Autre cas : décision binaire et **perte 0 – 1**

Théorème

Soit δ^π un estimateur bayésien associé à la loi a priori π

- S'il est unique, alors il est admissible
- S'il est de risque de Bayes fini, alors il est admissible

Exemple : Soit $S = \sum X_i$ le nombre de pièces non conformes après n tirages (avec remise). La proportion θ de pièces non conformes est inconnue. Étant donné S , que peut-on dire sur θ ?

$$X_i \sim \mathcal{B}(1, \theta), \theta \sim \mathcal{U}[0, 1]$$

[rappel Loi Beta : $Y \sim \mathcal{Be}(\alpha, \beta)$]

$$f(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)} \mathbb{1}_{[0,1]}(y) \quad \text{avec } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

$$\mathbb{E}(Y) = \alpha/(\alpha + \beta), \text{Var}(Y) = \alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)],$$

$$\text{Mode} = (\alpha - 1)/(\alpha + \beta - 2)]$$

- On utilise parfois l'estimateur du maximum de la loi a posteriori

$$\hat{\theta}_{MAP} = \text{Arg max}_{\theta} \pi(\theta|x) = \text{Arg max}_{\theta} \pi(\theta)L(\theta; x)$$

mais l'optimisation est souvent non triviale

- on peut voir la similarité avec l'EMV quand n tend vers l'infini, où le MAP en retrouve les propriétés

Choix fondamental, car peut avoir de l'influence si on a peu de données ; critiquable et critiqué !

- ▶ détermination à partir de l'expérience passée
- ▶ famille de lois conjuguées

Définition

\mathcal{F} est une *famille conjuguée* par une fonction de vraisemblance $f(x|\theta)$, si pour toute loi a priori $\pi \in \mathcal{F}$, la loi a posteriori $\pi(.|x) \in \mathcal{F}$

\hookrightarrow l'information apportée par l'échantillon se traduit uniquement par un changement de paramètre

Lois conjuguées

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

Comment choisir une loi a priori quand on n'a pas d'information ?

- ▶ approche de **Laplace** : loi uniforme
 - ▶ peut être impropre
Rem : une **loi impropre** est définie comme une mesure positive, ce qui est licite tant que la vraisemblance intégrée est finie presque sûrement
 - ▶ n'est pas invariante par reparamétrisation
- ▶ approche de **Jeffreys** : $\pi^*(\theta) \propto [\det(I(\theta))]^{1/2}$ où I est l'information de Fisher
 - ▶ est invariante par reparamétrisation

Définition

Un *intervalle de crédibilité* est un ensemble $C(x)$ de probabilité a posteriori $1 - \alpha$ pour θ

$$\pi(\theta \in C(X) | X = x) = 1 - \alpha$$

Ici, θ est la variable aléatoire (alors que c'est l'IC qui est la variable aléatoire en fréquentiste).

Définition (Intervalle High Posterior Density)

Un intervalle de crédibilité *HPD* est un ensemble $C(x)$ de probabilité a posteriori $1 - \alpha$ pour θ contenant les θ de plus haute densité. Il est de la forme

$$C(X) = \{\theta; \pi(\theta|x) \geq k_\alpha\}$$

- ▶ $(H_0) : \theta \in \Theta_0$ contre $(H_1) : \theta \notin \Theta_0$
- ▶ procédure de test : $\delta(x) = 1$ si on conserve (H_0) ,
 $\delta(x) = 0$ si on rejette (H_0)

Théorème

*Si la décision est binaire : $\delta = 1$ si $\theta \in \Theta_0$; $\delta = 0$ si $\theta \in \Theta_1$
et la fonction de **perte en 0-1 pondérée***

$$\ell(\theta, \delta) = a_0 \mathbb{1}_{\delta=0} \mathbb{1}_{\theta \in \Theta_0} + a_1 \mathbb{1}_{\delta=1} \mathbb{1}_{\theta \in \Theta_1}$$

l'estimateur de Bayes est défini par

$$\delta(X) = 1 \text{ ssi } \mathbb{P}^\pi(\Theta_0|X) \geq \frac{a_1}{a_0} \mathbb{P}^\pi(\Theta_1|X)$$

- ▶ $(H_0) : \theta \in \Theta_0$ contre $(H_1) : \theta \notin \Theta_0$
- ▶ procédure de test : $\delta(x) = 1$ si on conserve (H_0) ,
 $\delta(x) = 0$ si on rejette (H_0)
- ▶ $\ell(\theta, \delta) = a_0 \mathbb{I}_{\delta=0} \mathbb{I}_{\theta \in \Theta_0} + a_1 \mathbb{I}_{\delta=1} \mathbb{I}_{\theta \in \Theta_1}$

$$\begin{aligned} \rho(\pi, \delta(x)|x) &= \int_{\Theta} \ell(\theta, \delta(x)) \pi(\theta|x) d\theta \\ &= a_0 \mathbb{I}_{\delta(x)=0} \underbrace{\int_{\Theta_0} \pi(\theta|x) d\theta}_{\mathbb{P}^{\pi}(\theta \in \Theta_0|x)} + a_1 \mathbb{I}_{\delta(x)=1} \underbrace{\int_{\Theta_1} \pi(\theta|x) d\theta}_{\mathbb{P}^{\pi}(\theta \notin \Theta_0|x)} \end{aligned}$$

à minimiser en $\delta(x)$:

- ▶ si $a_0 \mathbb{P}^{\pi}(\theta \in \Theta_0|x) > a_1 \mathbb{P}^{\pi}(\theta \notin \Theta_0|x)$ alors $\delta(x) = 1$,
on accepte (H_0)
- ▶ si $a_0 \mathbb{P}^{\pi}(\theta \in \Theta_0|x) < a_1 \mathbb{P}^{\pi}(\theta \notin \Theta_0|x)$ alors $\delta(x) = 0$,
on rejette (H_0)

Exemple

Tester $\theta < 0$ dans le modèle $X_1 \sim \mathcal{N}(\theta, \sigma^2)$, $\theta \sim \mathcal{N}(\xi, \tau^2)$,
échantillon avec une seule observation.

Le facteur de Bayes

Pour s'affranchir du poids de l'a priori :

Définition

Le *facteur de Bayes* est le rapport des probabilités a posteriori sur les probabilités a priori

$$\begin{aligned} B_{21}^{\pi} &= \frac{\mathbb{P}^{\pi}(\theta \in \Theta_0 | X)}{\mathbb{P}^{\pi}(\theta \in \Theta_1 | X)} \div \frac{\mathbb{P}^{\pi}(\theta \in \Theta_0)}{\mathbb{P}^{\pi}(\theta \in \Theta_1)} \\ &= \frac{\int_{\Theta_0} L(\theta|x) \pi_0(\theta) d\theta}{\int_{\Theta_1} L(\theta|x) \pi_1(\theta) d\theta} = \frac{m_0(x)}{m_1(x)} \end{aligned}$$

π_0 et π_1 sont les lois a priori sous (H_0) et (H_1)

C'est un rapport de *vraisemblance intégrée* sur chaque espace des paramètres

Le facteur de Bayes pour le choix de modèle

Une autre vision : \hookrightarrow un problème de **sélection de modèle** :

► $\mathcal{M} = 1 : \{\theta; \theta < 0\}$

► $\mathcal{M} = 2 : \{\theta; \theta > 0\}$

Une procédure de Bayes choisit le modèle k qui maximise la loi a posteriori $\mathbb{P}^\pi(\mathcal{M} = k|X)$

- l'a priori π est donc défini sur une collection d'indices $\{1, \dots, K\}$, et conditionnellement à chaque modèle k , sur l'espace des paramètres correspondant Θ_k

$$\pi = p_1\pi_1(\theta) + \dots + p_k\pi_k(\theta)$$

- l'a posteriori

$$\mathbb{P}^\pi(\mathcal{M} = k|X) = \frac{p_k \int L(\theta_k|x)\pi_k(\theta_k) d\theta_k}{\sum_j p_j \int L(\theta_j|x)\pi_j(\theta_j) d\theta_j}$$

Définition

Le *facteur de Bayes* est le rapport des probabilités a posteriori sur les probabilités a priori

$$\begin{aligned} B_{21}^{\pi} &= \frac{\mathbb{P}^{\pi}(\mathcal{M} = 2|X) / \mathbb{P}^{\pi}(\mathcal{M} = 1|X)}{p_2 / p_1} \\ &= \frac{\int_{\Theta_2} L(\theta_2|x) \pi_2(\theta_2) d\theta_2}{\int_{\Theta_1} L(\theta_1|x) \pi_1(\theta_1) d\theta_1} = \frac{m_2(x)}{m_1(x)} \end{aligned}$$

C'est un rapport de *vraisemblance intégrée* sur l'espace des paramètres

A mettre en parallèle du rapport de vraisemblance fréquentiste

$$\frac{\max_{\theta \in \Theta_2} L(\theta; x)}{\max_{\theta \in \Theta_1} L(\theta; x)}$$

Le facteur de Bayes est un rapport de vraisemblance "bayésien". L'évidence apportée par les données est calibrée par l'échelle de Jeffreys

- ▶ si $0 < \log_{10}(B_{21}^{\pi}) \leq 0.5$, l'évidence contre $\mathcal{M} = 1$ est faible
- ▶ si $0.5 < \log_{10}(B_{21}^{\pi}) \leq 1$, l'évidence contre $\mathcal{M} = 1$ est substantielle
- ▶ si $1 < \log_{10}(B_{21}^{\pi}) \leq 2$, l'évidence contre $\mathcal{M} = 1$ est forte
- ▶ si $\log_{10}(B_{21}^{\pi}) > 2$, l'évidence contre $\mathcal{M} = 1$ est décisive

Remarque : attention à l'utilisation de lois impropres

Calcul du facteur de Bayes par simulation

- **Objectif** : Proposer une méthode générique pour calculer

$$I = \mathbb{E}_f[h(X)] = \int_{\mathcal{X}} h(x)f(x)dx$$

- **Solution** : simulation par Monte Carlo
Utiliser un échantillon iid (X_1, \dots, X_n) de loi de densité f pour approximer l'intégrale I par la moyenne empirique

$$\hat{h}_n = \frac{1}{n} \sum_{i=1}^n h(x_i)$$

qui converge par la Loi des Grands Nombres

$$\hat{h}_n \rightarrow \mathbb{E}_f[h(X)]$$

- La variance de l'estimation est $\text{Var}(\hat{h}_n) = \text{Var}(h(X_1))/n$ estimée par

$$\hat{v}_n := \widehat{\text{Var}(\hat{h}_n)} = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^n (h(x_i) - \hat{h}_n)^2$$

- Asymptotiquement (n grand) et si $\mathbb{E}(h(x)^2) < \infty$

$$\frac{\hat{h}_n - \mathbb{E}_f[h(X)]}{\sqrt{\hat{v}_n}} \underset{\text{appr}}{\sim} \mathcal{N}(0, 1)$$

d'où indication de la précision de l'approximation obtenue

- La simulation à partir de f (densité cible) n'est pas nécessairement de plus faible variance
- Une alternative à l'échantillonnage **direct** est l'échantillonnage **préférentiel** ou **pondéré**

$$\mathbb{E}_f[h(X)] = \int_{\mathcal{X}} h(x) \frac{f(x)}{g(x)} g(x) dx$$

qui permet d'utiliser une loi instrumentale g

↪ tirer là où les observations ont de l'importance

Introduction

Estimation

Loi a priori

Intervalle de
confiance

Test d'hypothèses

- Convergence : si $X_i \sim g$ iid, alors

$$\hat{h}_n = \frac{1}{n} \sum_{i=1}^n \frac{f(X_i)}{g(X_i)} h(X_i) \rightarrow \int_{\mathcal{X}} h(x) f(x) dx = \mathbb{E}_f[h(X)]$$

pour n'importe quel choix de loi instrumentale g de support contenant celui de f

- choisir g facile à simuler
- l'estimateur \hat{h}_n doit avoir une variance finie :

$$\int_{\mathcal{X}} h(x)^2 f(x)^2 / g(x) dx < \infty$$

- donc g doit être à queues plus épaisses que f