CHAPITRE 6

Modèles bayésiens pour la loi normale

MTH3302 - Méthodes probabilistes et statistiques pour l'I.A.

Jonathan Jalbert – Automne 2022

Ce chapitre introduit les concepts fondamentaux de la statistique bayésienne pour le cas particulier de la loi normale (avec variance inconnue). À la fin du chapitre, vous devriez être en mesure de

- Utiliser le théorème de Bayes pour calculer les lois conditionnelles complètes des paramètres.
- Générer un échantillon de la loi *a posteriori* bidimensionnelle avec l'échantillonnage de Gibbs.
- Sélectionner le meilleur modèle parmi un ensemble de modèles.

Dans ce chapitre, nous utiliserons à nouveau le jeu de données `illingworth.csv` concernant l'expérience de Michelson-Morley effectuée par Illingworth en 1927. Dans ce chapitre, nous supposerons que la variance de la loi normale est inconnue. Elle devra donc être estimée comme la moyenne en utilisant le théorème de Bayes. Nous verrons dans le TD qu'Illingworth a largement surestimé la variance de l'erreur expérimental de son montage. Autrement dit, son montage était beaucoup plus précis qu'il l'avait estimé.

TABLE 6.1 – Déplacements des franges d'interférence (mesurés en millième) pour les observations effectuées par Illingworth à 5 a.m. avec le montage orienté dans la direction N.

Observation	1	2	3	4	5	6	moyenne
Déplacement	0.24	1.14	0.00	0.20	0.64	-0.02	0.37

6.1 Modèle gaussien

Considérons pour le moment les 6 observations effectuées à 5 a.m. avec le montage orienté dans la direction N, observations répertoriées dans le tableau 6.1. À l'instar du chapitre précédent, le modèle statistique suivant est supposé pour chacune des conditions expérimentales :

$$Y_i = \mu + \varepsilon_i \tag{6.1}$$

pour $1 \leq i \leq n$, où Y_i est la mesure du déplacement des franges d'interférence de la i^e observation, μ est le vrai déplacement inconnu et ε_i est l'erreur de mesure associée à la i^e observation. Si on suppose que les erreurs de mesure sont indépendantes et identiquement distribuées selon la loi normale de moyenne nulle et de variance inconnue σ^2, alors le modèle décrit à l'équation (6.1) peut s'exprimer sous la forme suivante :

$$Y_i \sim \mathcal{N}(\mu, \sigma^2) \tag{6.2}$$

pour $1 \leq i \leq n$.

Remarque. *En pratique, la variance est rarement connue. Elle doit être estimée même si elle ne constitue pas un paramètre d'intérêt mais plutôt un paramètre de nuisance.*

6.2 Loi *a priori* informative

Dans le cas de la loi normale avec la moyenne et la variance inconnues, il n'existe pas de loi *a priori* informative conjuguée. Il existe cependant une loi *a priori* informative permettant de simplifier partiellement les calculs. Il s'agit de la loi suivante :

$$f_{(\mu,\sigma^2)}(\mu, \sigma^2) = \mathcal{N}(\mu|\nu, \sigma^2) \times \mathcal{I}nv\mathcal{G}amma(\sigma^2|\alpha, \beta). \tag{6.3}$$

Cette loi est constitué par le produit de la loi conditionnelle $f_{(\mu|\nu,\sigma^2)}(\mu)$ et de la loi marginale $f_{(\sigma^2|\alpha,\beta)}(\sigma^2)$. La loi normale pour $f_{(\mu|\nu,\sigma^2)}(\mu)$ permet d'encoder l'information *a priori* sur μ et la loi inverse-gamma informative pour $f_{\sigma^2|\alpha,\beta}(\sigma^2)$ permet d'encoder l'information *a priori* sur σ^2.

La variable aléatoire X est distribuée selon la loi inverse-gamma de paramètres $\alpha > 0$ et $\beta > 0$ si sa densité est égale à

$$f_{(X|\alpha,\beta)}(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)}(1/x)^{\alpha+1} \exp(-\beta/x) & \text{si } x > 0 \\ 0 & \text{sinon.} \end{cases}$$

La loi *a posteriori* correspondante à la loi *a priori* exprimée à l'équation (6.3) est proportionnelle à la forme suivante :

$$f_{\{(\mu,\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\mu,\sigma^2) \propto \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(\mu-\nu)^2\right\} \times \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{\beta}{\sigma^2}\right). \tag{6.4}$$

Il n'existe pas de densité de probabilités bidimensionnelle ayant la forme fonctionnelle exprimée à l'équation précédente. L'algorithme de Metropolis-Hastings pourrait être utilisée dans sa version bidimensionnelle pour générer un échantillon de cette loi. Cependant, un autre algorithme de la famille des méthodes Monte-Carlo par chaîne de Markov est plus facile à mettre en oeuvre dans les cas multidimensionnels. Il s'agit de l'échantillonnage de Gibbs (Section 6.A).

Pour implémenter l'échantillonnage de Gibbs, les **lois conditionelles complètes** sont nécessaires. Les lois conditionnelles complètes sont un ensemble de lois unidimensionnelles pour chacun des paramètres conditionnellement à tout le reste, d'où l'appellation *conditionnelles complètes*. Pour générer un échantillon de la densité exprimée à l'équation (6.4), les lois conditionnelles complètes $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y},\sigma^2)}(\mu)$ et $f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y},\mu)}(\sigma^2)$ sont nécessaires.

Pour calculer la loi conditionnelle complète $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y},\sigma^2)}(\mu)$, il suffit de traiter de σ^2 comme une constante dans la loi *a posteriori* exprimée à l'équation (6.4). On peut montrer que cette loi conditionnelle complète s'exprime sous la forme analytique suivante :

$$f_{(\mu|\boldsymbol{Y}=\boldsymbol{y},\sigma^2)}(\mu) = \mathcal{N}\left\{\mu \,\middle|\, \frac{n\bar{y}+\nu}{n+1}, \frac{\sigma^2}{n+1}\right\} \tag{6.5}$$

Pour calculer la loi conditionnelle complète $f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y},\mu)}(\sigma^2)$, il suffit de traiter de μ comme une constante dans la loi *a posteriori* exprimée à l'équation (6.4). On peut montrer que cette loi conditionnelle complète s'exprime sous la forme analytique suivante :

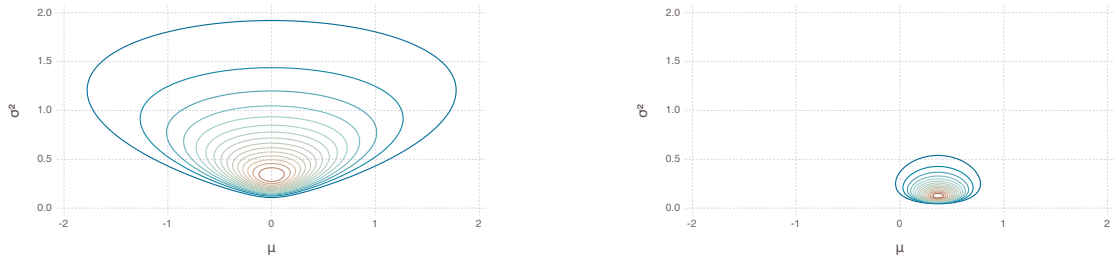
$$f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y},\mu)}(\sigma^2) = \mathcal{I}nv\mathcal{G}amma\left(\sigma^2 \,\middle|\, \alpha+\frac{(n+1)}{2}, \beta+\frac{(\mu-\nu)^2}{2}+\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}\right) \tag{6.6}$$

Exemple 1

Pour analyser les observations effectuées par Illingworth à 5 a.m dans la direction N, données répertoriées dans le tableau 6.1, supposons la loi *a priori* informative suivante :

$$f_{(\mu,\sigma^2)}(\mu,\sigma^2) = \mathcal{N}(\mu \mid 0,\sigma^2) \times \mathcal{I}nv\mathcal{G}amma(\sigma^2 \mid 1,1).$$

La loi *a priori* bidimensionnelle est illustréé à la figure 6.1. Les équations (6.5) et



(a) (b)

FIGURE 6.1 – Loi *a priori* (a) et loi *a posteriori* (b) de l'exemple 1.

(6.6) permettent de calculer les conditionnelles complètes données par les expressions suivantes :

$$f_{(\mu|\boldsymbol{Y}=\boldsymbol{y},\sigma^2)}(\mu) = \mathcal{N}\left\{\mu \,\middle|\, \frac{n\bar{y}}{n+1}, \frac{\sigma^2}{n+1}\right\}$$

$$f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y},\mu)}(\sigma^2) = \mathcal{I}nv\mathcal{G}amma\left\{\sigma^2 \,\middle|\, \frac{n+3}{2}, 1 + \frac{(\mu-\nu)^2}{2} + \frac{\sum_{i=1}^n (y_i-\mu)^2}{2}\right\}$$

La forme fonctionnelle de la densité *a posteriori* est illustrée à la figure 6.1. On peut constater que les 6 observations ont raffiner la connaissance *a priori* que nous avions sur μ et σ^2.

La loi *a priori* exprimée à l'équation (6.3) est appelée *partiellement* conjuguée parce que les lois conditionnelles complètes s'expriment sous une forme analytique. Elle n'est pas conjuguée parce que la loi *a posteriori* exprimée à l'équation (6.4) ne s'exprime pas sous une forme analytique. Le fait que les lois conditionnelles complètes s'énoncent sous une forme connue facilite l'implémentation de l'échantillonnage de Gibbs (section 6.A) pour générer un échantillon de la loi *a posteriori*.

Remarque. *Dans le cas où une loi conditionnelle complète ne correspond pas à une loi connue, l'algorithme de Metropolis-Hastings peut être utilisé pour générer une réalisation de cette loi à l'intérieur de l'échantillonnage de Gibbs.*

6.3 Loi *a priori* non informative

Pour définir la loi *a priori* non informative, elle est décomposée de la façon suivante :

$$f_{(\mu,\sigma^2)}(\mu,\sigma^2) = f_\mu(\mu) \times f_{\sigma^2}(\sigma^2).$$

Au chapitre précédent, nous avons vu que la loi non informative pour μ était la loi impropre :

$$f_\mu(\mu) \propto 1.$$

Il faut maintenant définir une loi non informative pour la variance σ^2 qui ne prend que des valeurs positives. Pour définir la loi non informative sur σ^2, on procède par la transformation de variable $\phi = \ln \sigma^2$. Dans ce cas, le paramètre ϕ prend des valeurs dans l'ensemble des réels. On peut donc utiliser la loi impropre suivante pour ϕ :

$$f_\phi(\phi) \propto 1.$$

Lorsqu'on fait la transformation inverse pour retrouver la densité pour σ^2, la loi *a priori* non informative correspond à la densité impropre

$$f_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2} \text{ pour } \sigma^2 > 0.$$

Par conséquent, la loi *a priori* non informative conjointe pour la moyenne et la variance est la densité impropre suivante :

$$f_{(\mu,\sigma^2)}(\mu, \sigma^2) \propto \frac{1}{\sigma^2}.$$

En utilisant cette densité *a priori* impropre, la loi *a posteriori* correspondante est

$$f_{\{(\mu,\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\mu, \sigma^2) \propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\} \times \frac{1}{\sigma^2}. \tag{6.7}$$

Cette densité ne s'exprime pas sous une forme analytique. Les densités conditionnelles complètes s'expriment néanmoins sous les formes connues suivantes :

$$f_{(\mu|\boldsymbol{Y}=\boldsymbol{y},\sigma^2)}(\mu) = \mathcal{N}\left(\mu \middle| \bar{y}, \frac{\sigma^2}{n}\right)$$

et

$$f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y},\mu)}(\sigma^2) = \mathcal{I}nv\mathcal{G}amma\left(\sigma^2 \middle| \frac{n}{2}, \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2}\right)$$

L'échantillonnage de Gibbs présenté à la section 6.A permettra de générer un échantillon aléatoire de cette densité à partir des densités conditionnelles complètes.

Remarque. *La loi* a posteriori *correspondante à la loi* a priori *impropre exprimée à l'équation* (6.7) *est valide lorsque* $n \geq 1$.

6.4 Densité marginale *a posteriori.*

Dans le cas de l'expérience de Michelson-Morley, le paramètre d'intérêt est μ, le déplacement réel des franges d'interférence. La variance σ^2 constitue ce qu'on appelle un paramètre de nuisance ; on n'a pas le choix de l'estimer puisqu'il fait partie du modèle statistique mais il ne constitue pas le paramètre d'intérêt principal. Il alors possible d'intégrer ce paramètre de la loi *a posteriori* pour obtenir la loi *a posteriori* marginale de μ :

$$f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu) = \int_0^\infty f_{\{(\mu,\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\mu,\sigma^2)\, d\sigma^2.$$

Cette loi incorpore dans la connaissance de μ l'incertitude que l'on possède sur σ^2. Il est assez difficile d'obtenir une la forme de cette loi. Le cas le plus simple est illustré dans l'exemple suivant. La plupart du temps en pratique, la chaîne des $\mu^{(t)}$ générée par l'échantillonnage de Gibbs est considérée comme une réalisation de la loi *a posteriori* marginale de μ.

Exemple 2

Dans le cas de la vraisemblance gaussienne avec la loi *a priori* impropre, la densité *a posteriori* marginale de μ est donnée par

$$f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu) = t_{n-1}\left(\mu \,\middle|\, \bar{y}, \frac{s}{\sqrt{n}}\right),$$

où

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

et où $t_\nu(y|\mu,\sigma)$ représente la densité évaluée à y de la loi de Student à ν degrés de liberté avec le paramètre localisation μ et le paramètre d'échelle σ.
Pour montrer ce résultat, on intègre σ^2 de la forme fonctionnelle de la loi *a posteriori* :

$$\begin{aligned} f_{(u|\boldsymbol{Y}=\boldsymbol{y})}(\mu) &\propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \times \frac{1}{\sigma^2}\, d\sigma^2 \\ &\propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} d\sigma^2. \end{aligned}$$

La forme fonctionnelle de la densité $\mathcal{I}nverse\mathcal{G}amma\left\{\sigma^2 \mid \frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}$ peut être reconnue. L'intégrale est donc égale à l'inverse de la contante de normalisation de cette densité. Alors,

$$f_{(u|\boldsymbol{Y}=\boldsymbol{y})}(\mu) \propto \frac{\Gamma(n/2)}{\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}^{n/2}}$$

$$
\begin{aligned}
&\propto \left\{ \sum_{i=1}^{n} (y_i - \mu)^2 \right\}^{-\frac{n}{2}} \\
&\propto \left\{ \sum_{i=1}^{n} (y_i - \bar{y} + \bar{y} - \mu)^2 \right\}^{-\frac{n}{2}} \\
&\propto \left\{ \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\bar{y} - \mu)^2 + 2 \sum_{i=1}^{n} (y_i - \bar{y})(\bar{y} - \mu) \right\}^{-\frac{n}{2}} \\
&\propto \left\{ \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\mu - \bar{y})^2 \right\}^{-\frac{n}{2}} \\
&\propto \left\{ (n-1)s^2 + n(\mu - \bar{y})^2 \right\}^{-\frac{n}{2}} \\
&\propto \left\{ \frac{(n-1)}{n} s^2 + (\mu - \bar{y})^2 \right\}^{-\frac{n}{2}} \\
&\propto \left\{ (n-1) + \left(\frac{\mu - \bar{y}}{s/\sqrt{n}} \right)^2 \right\}^{-\frac{(n-1)+1}{2}} \\
&\propto \left\{ 1 + \frac{1}{(n-1)} \left(\frac{\mu - \bar{y}}{s/\sqrt{n}} \right)^2 \right\}^{-\frac{(n-1)+1}{2}}
\end{aligned}
$$

La forme fonctionnelle de la loi de Student à $(n-1)$ degrés de liberté avec paramètres de localisation $\bar{y}$ et d'échelle $s/\sqrt{n}$ peut être reconnue. Il s'agit de la loi *a posteriori* marginale de μ.

Remarque. *Il n'est pas toujours possible d'obtenir une expression analytique pour la loi* a posteriori *marginale.*

6.5 Comparaison de modèles

Bien qu'il soit possible de reproduire les procédures de tests d'hypothèses en statistique bayésienne, ce n'est généralement pas l'approche privilégiée. La façon usuelle de procéder consiste à comparer les modèles statistiques où chacun de ceux-ci correspond à une hypothèse à tester.

Exemple 3

Dans le cas de l'expérience de Michelson-Morley, les modèles suivants pourraient être comparés :

— $\mathcal{M}_1 : Y_i \sim \mathcal{N}(0, \sigma^2)$
— $\mathcal{M}_2 : Y_i \sim \mathcal{N}(\mu, \sigma^2)$

pour $\mu \neq 0$. Le modèle $\mathcal{M}_1$ suppose que le vrai déplacement des franges d'interférence est nulle tandis que le modèle $\mathcal{M}_1$ suppose qu'il est différent de 0. La validité de chacun de ces modèles en fonction des données est l'objet de la prochaine section.

Remarque. *La comparaison de modèles sera aussi utile lors de la régression bayésienne pour choisir le meilleur sous-ensemble des variables explicatives.*

6.5.1 La validité du modèle

Soit une ensemble de J modèles statistiques $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$. Dénotons par $\mathbb{P}(\mathcal{M}_j)$ la probabilité *a priori* que le le modèle $\mathcal{M}_j$ soit vrai. La mise à jour de cette probabilité en fonction des observations est dénotée par

$$\mathbb{P}(\mathcal{M}_j \mid \boldsymbol{Y} = \boldsymbol{y}). \tag{6.8}$$

Elle s'interprète comme la probabilité que les observations $\boldsymbol{y}$ aient été générées par le modèle $\mathcal{M}_j$. C'est une mesure de validité du modèle ; elle mesure la cohérence entre le modèle et les données.

Le calcul de la probabilité exprimée à l'équation (6.8) requiert l'utilisation du théorème de Bayes :

$$\mathbb{P}(\mathcal{M}_j \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \mathcal{M}_j) \times \mathbb{P}(\mathcal{M}_j)}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})}$$

La probabilité $\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \mathcal{M}_j)$ dans le numérateur de l'équation précédente ne dépend pas des paramètres du modèles, disons $\boldsymbol{\theta}_j$. L'incertitude des paramètres est intégrée de la façon suivante :

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \mathcal{M}_j) = \int_{\boldsymbol{\theta}_j} f_{(\boldsymbol{Y}|\boldsymbol{\theta}_j)}(\boldsymbol{y})\, f_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_j)\, d\boldsymbol{\theta}_j.$$

Cette probabilité correspond à la constante de normalisation de la loi *a posteriori* de $\boldsymbol{\theta}_j$. Cette quantité est également appelée l'*évidence du modèle* $\mathcal{M}_j$ est est dénotée par $m_j(\boldsymbol{y})$.

Remarque. *La loi* a priori *des paramètres du modèle* $\mathcal{M}_j$, $f_{\theta_j}(\theta_j)$, *et la vraisemblance,* $f_{(\boldsymbol{Y}|\theta_j)}(\boldsymbol{y})$ *dépendent de* j, *de l'indice du modèle. En effet, différents modèles peuvent avoir des vecteurs de paramètres différents.*

La validité du modèle $\mathcal{M}_j$ s'exprime alors de la façon suivante :

$$\mathbb{P}(\mathcal{M}_j \mid \boldsymbol{Y} = \boldsymbol{y}) \propto m_j(\boldsymbol{y}) \times \mathbb{P}(\mathcal{M}_j). \tag{6.9}$$

Dans le cas où la c'est la comparaison des modèles de l'ensemble $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$, la constante de normalisation peut être définie comme suit :

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{j=1}^{J} m_j(\boldsymbol{y}) \times \mathbb{P}(\mathcal{M}_j).$$

Dans ce contexte, le modèle le plus cohérent avec les observations correspond au mode de la fonction de masse $\mathbb{P}(M = k \mid \boldsymbol{Y} = \boldsymbol{y})$.

Remarque. *Il est impossible de définir de façon absolue une mesure de validité de modèle.*

Remarque. *La plupart du temps en pratique, la probabilité* a priori *des modèles est supposée uniforme sur l'ensemble des J modèles considérés :*

$$\mathbb{P}(\mathcal{M}_j) = \frac{1}{J}.$$

6.5.2 Le facteur de Bayes

La plupart du temps, l'équation (6.9) est plutôt utilisée pour comparée deux à deux les modèles avec le **facteur de Bayes**. Le facteur de Bayes entre le modèle $\mathcal{M}_2$ et le modèle $\mathcal{M}_1$ compare les probabilités que les observations aient été générées par ces modèles. Il est défini par l'équation suivante :

$$B_{21} = \frac{\mathbb{P}(\mathcal{M}_2 \mid \boldsymbol{Y} = \boldsymbol{y})}{\mathbb{P}(\mathcal{M}_1 \mid \boldsymbol{Y} = \boldsymbol{y})} \times \frac{\mathbb{P}(\mathcal{M}_1)}{\mathbb{P}(\mathcal{M}_2)} = \frac{m_2(\boldsymbol{y})}{m_1(\boldsymbol{y})}.$$

Le modèle 2 est plus probable lorsque B_{21} est supérieur à 1.

Lorsque le choix d'un modèle peut avoir des conséquences importantes, choisir naïvement le modèle qui maximise la probabilité (6.8) peut s'avérer être un choix risqué. L'interprétation du facteur de Bayes introduite par Jeffreys (1939) est résumée au tableau 6.2. Bien que cette interprétation soit arbitraire, elle est néanmoins très utile pour la sélection de modèle en l'absence d'un cadre décisionnel formel (sans utiliser la théorie de la décision). Elle permet de choisir un modèle si celui-ci augmente *significativement* la probabilité que celui-ci soit vrai.

De façon générale, la loi marginale de l'échantillon évaluée aux observation $\boldsymbol{y}$ est assez difficile à calculer. Une alternative populaire consiste à approximer cette quantité par le *Bayesian Information Criterion* (BIC) présenté à la section suivante. Le BIC est aussi utilisé lorsque la loi *a priori* est impropre. En effet, le facteur de Bayes est incompatible avec les lois *a priori* impropres comme l'illustre l'exemple suivant.

TABLE 6.2 – Interprétation du facteur de Bayes.

Facteur de Bayes	Certitude que $\mathcal{M}_1$ est faux par rapport à $\mathcal{M}_2$
$0 < \ln(B_{21}) \leq 1/2$	**faible**
$1/2 < \ln(B_{21}) \leq 1$	**substantielle**
$1 < \ln(B_{21}) \leq 2$	**forte**
$\ln(B_{21}) > 2$	**décisive**

Exemple 4

Supposons que nous voulons comparer les modèles
- $\mathcal{M}_1 : Y \sim \mathcal{N}(0, 1^2)$
- $\mathcal{M}_2 : Y \sim \mathcal{N}(\mu, 1^2)$

à l'aide d'une observation y. Si la loi *a priori* impropre $f_\mu(\mu) \propto 1$ est utilisée pour le modèle $\mathcal{M}_2$ alors le facteur de Bayes devient

$$B_{21} = \frac{\displaystyle\int_{-\infty}^{\infty} e^{-\frac{1}{2}(y-\mu)^2}\, 1\, d\mu}{e^{-y^2/2}} = \frac{\sqrt{2\pi}}{e^{-y^2/2}}.$$

Si la loi *a priori* impropre $f_\mu(\mu) \propto 100$ est utilisée pour le modèle $\mathcal{M}_2$ alors le facteur de Bayes devient

$$B_{21} = \frac{\displaystyle\int_{-\infty}^{\infty} e^{-\frac{1}{2}(y-\mu)^2}\, 100\, d\mu}{e^{-y^2/2}} = \frac{100\sqrt{2\pi}}{e^{-y^2/2}}.$$

Si on utilise la loi *a priori* $f_\mu(\mu) \propto 100$, le modèle $\mathcal{M}_2$ est 100 fois plus probables par rapport à $\mathcal{M}_1$ que si on utilise la loi impropre $f_\mu(\mu) \propto 1$, ce qui ne fait aucun sens.

6.5.3 Le critère d'information bayésien

À part dans les situations où des lois conjuguées sont utilisées, le facteur de Bayes est très difficile à calculer. De plus, le calcul du facteur de Bayes présenté à la section précédente est incompatible avec les lois *a priori* impropres. Pour ces raisons, plusieurs alternatives au facteur de Bayes ont été proposées dans la littérature. Peut-être celle qui récolte le plus de popularité est le critère d'information bayésien (BIC ; pour *Bayesian Information Criterion*) développé par Schwarz (1978). Le BIC constitue une approximation de la loi marginale du modèle évaluée aux observations $m(\boldsymbol{y})$ lorsque la loi *a priori* est peu informative[1]. En particulier, le BIC approxime le logarithme de la loi marginale du modèle évaluée aux

1. Kass & Raftery (1995) donne un sens mathématique précis à cet énoncé.

observations :

$$\text{BIC} \approx \ln m\left(\boldsymbol{y}\right).$$

Le BIC est défini de la façon suivante :

$$\text{BIC} = \ln f_{(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}_{MV})}(\boldsymbol{y}) - \frac{k}{2}\ln n, \tag{6.10}$$

où $\hat{\boldsymbol{\theta}}_{MV}$ dénote l'estimateur du maximum de la vraisemblance de $\boldsymbol{\theta}$ et k le nombre de paramètres du modèle. L'estimation utilisée peut aussi être le mode de la loi *a posteriori* comme l'indique Fraley & Raftery (2007). Le logarithme du facteur de Bayes entre les modèles $\mathcal{M}_2$ et $\mathcal{M}_1$ peut alors être approximé par

$$\ln\left(B_{21}\right) \approx \text{BIC}_2 - \text{BIC}_1,$$

où BIC_j correspond au BIC du modèle $\mathcal{M}_j$.

Remarque. *La procédure usuelle consiste à choisir le modèle qui possède le plus grand* BIC. *Comme le critère* BIC *pénalise les modèles complexes comportant de nombreux paramètres avec le terme $-k/2\ln n$, il aura tendance à favoriser les modèles plus simples. Si un modèle complexe est sélectionné, c'est qu'il est* vraiment *meilleur par rapport aux modèles plus simples. Dans un contexte où le choix du modèle entraîne d'importantes conséquences, la table d'interprétation du facteur de Bayes proposée par Jeffreys (tableau 6.2) s'avère utile.*

6.6 Exercices

1. Soit la loi *a priori* suivante pour les paramètres μ et σ^2 de la loi normale :

$$f_{(\mu,\sigma^2)}(\mu,\sigma^2) \propto \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right) \times \left(\frac{1}{\sigma^2}\right)^2 \exp\left(-\frac{1}{\sigma^2}\right).$$

   (a) S'agit-il d'une loi informative ou non informative ? Justifiez.

   (b) Quelles sont les lois conditionnelles complètes de μ et σ^2 ?

2. Considérez le jeu de données `illingworth1927.csv` disponible sur le site du cours. Soit le modèle $\mathcal{M}_1$ supposant que $Y_i \sim \mathcal{N}\left\{\mu, (3/2)^2\right\}$ et le modèle $\mathcal{M}_2$ supposant la variance inconnue, *i.e.* $Y_i \sim \mathcal{N}\left\{\mu, \sigma^2\right\}$. Considérez les lois *a priori* suivantes :

$$f_{\mu}(\mu) \propto 1 \text{ pour le modèle } \mathcal{M}_1$$
$$f_{(\mu,\sigma^2)}(\mu,\sigma^2) \propto \frac{1}{\sigma^2} \text{ pour le modèle } \mathcal{M}_2.$$

   On considère les 64 observations pour estimer μ et σ^2. On a que $\bar{y} = -0.0148$ et $s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 0.0421$.

a) Calculez les BIC des modèles $\mathcal{M}_1$ et $\mathcal{M}_2$.

b) Selon le résultat précédent, quel est le meilleur modèle ?

c) Selon votre résultat en (b), que pouvez-vous conclure sur l'estimation *a priori* de σ^2 par Illingworth ?

3. En statistique bayésienne, la loi normale est souvent paramétrisée avec le paramètre de précision $\kappa > 0$ plutôt que la variance σ^2. La précision est définie comme l'inverse de la variance $\kappa = 1/\sigma^2$. La densité de la variable aléatoire Y distribuée selon la loi normale de moyenne μ et de précision κ peut s'écrire de la façon suivante :

$$f_{(Y|\mu,\kappa)}(y) = \sqrt{\frac{\kappa}{2\pi}} \exp\left\{-\frac{\kappa}{2}(y-\mu)^2\right\}.$$

Soit un échantillon aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de taille où les Y_i sont indépendant est distribués selon la densité précédente. Pour estimer les paramètres inconnus μ et κ, la loi *a priori* impropre suivante sera utilisée :

$$f_{(\mu,\kappa)}(\mu,\kappa) \propto \frac{1}{\kappa} \quad \text{pour } \kappa > 0.$$

a) Quelle est la vraisemblance des paramètres (μ, κ) ? Autrement dit, calculez la densité $f_{(\boldsymbol{Y}|\mu,\kappa)}(\boldsymbol{y})$.

b) Quelle est la loi conditionnelle complète de μ ? Autrement dit, calculez la densité $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y},\kappa)}(\mu)$.

c) Lorsque l'on utilise une loi *a priori* impropre, il faut toujours s'assurer que la loi *a posteriori* est propre. Dans ce cas, est-ce que la loi conditionnelle complète de μ obtenue est toujours valide ?

d) Quelle est la loi conditionnelle complète de κ ? Autrement dit, calculez la densité $f_{(\kappa|\boldsymbol{Y}=\boldsymbol{y},\mu)}(\kappa)$.

e) Est-ce que la loi conditionnelle complète de κ obtenue est toujours valide ?

4. On enregistre pendant n jours le nombre de jours où un serveur informatique tombe en panne. Soit l'échantillon aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ avec :

$$Y_i = \begin{cases} 0 & \text{si le serveur ne tombe pas en panne durant la journée } i\,; \\ 1 & \text{si le serveur tombe en panne durant la journée } i. \end{cases}$$

Posons $0 \leq \theta \leq 1$ la probabilité que le serveur tombe en panne durant la journée pour chacune des journées.

a) Quelle est la fonction de masse de Y_i ? Autrement dit, quelle est la loi $p_{(Y_i|\theta)}(y_i)$?

b) Calculez la vraisemblance associée au vecteur des observations $\boldsymbol{Y}$. Autrement dit, calculez la loi $p_{(\boldsymbol{Y}|\theta)}(\boldsymbol{y})$.

c) En utilisant la loi *a priori* suivante

$$f_\theta(\theta) = \mathcal{U}niforme(\theta \mid 0\,, 1),$$

quelle est la loi *a posteriori* du paramètre θ ?

d) Si on utilise la moyenne de la loi *a posteriori* comme estimation ponctuelle bayésienne, quelle est l'estimation ponctuelle bayésienne de θ ?

e) Calculez la probabilité qu'il y ait une panne au prochain jour, *i.e* au jour Y_{n+1}. Justifiez votre démarche.

Annexe

6.A Échantillonnage de Gibbs

L'échantillonnage de Gibbs est la deuxième et dernière méthode Monte-Carlo par chaîne de Markov qui sera présentée dans le cadre du cours. L'échantillonnage de Gibbs s'applique lorsque le nombre de paramètres est supérieur ou égal à 2. Supposons que le modèle statistique possède $p \geq 2$ paramètres inconnus dénotés par le vecteur $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)$. L'échantillonnage de Gibbs est illustré ici pour la moyenne et la variance de la loi normale mais l'algorithme s'applique à n'importe quel modèle, pourvu que $p \geq 2$. Supposons que la constante de normalisation C de la loi *a posteriori* est inconnue, on connaît seulement la forme fonctionnelle de la loi *a posteriori* $g_{\{(\mu,\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\mu, \sigma^2)$, *i.e.*

$$f_{\{(\mu,\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\mu, \sigma^2) = \frac{1}{C} g_{\{(\mu,\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\mu, \sigma^2).$$

Nous souhaitons obtenir un échantillon aléatoire de la loi $f_{\{(\mu,\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\mu, \sigma^2)$ lorsque la constante de normalisation C est inconnue. L'algorithme 1, appelé *échantillonnage de Gibbs*, permet de générer un tel échantillon aléatoire à l'aide des lois conditionnelles complètes $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y},\sigma^2)}(\mu)$ et $f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y},\mu^2)}(\sigma^2)$.

Algorithm 1 Échantillonnage de Gibbs

Définir l'état initial des paramètres $\boldsymbol{\theta}^{(0)} = \left(\mu^{(0)}, \sigma^{2\,(0)}\right)$.
for t = 1 à m **do**
1. Tirer $\mu^{(t)}$ de la loi $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y},\sigma^{2\,(t-1)})}(\mu)$.
2. Tirer $\sigma^{2\,(t)}$ de la loi $f_{(\sigma^2|\boldsymbol{Y}=\boldsymbol{y},\mu^{(t)})}(\sigma^2)$.

end for

À l'instar de l'algortihme de Metropolis-Hastings, un échantillon obtenu par l'échantillonnage de Gibbs comporte une phase de chauffe et une phase d'échantillonnage. La longueur de la phase de chauffe peut être déterminée visuellement en traçant les chaînes obtenues $\left\{\mu^{(t)} : t = 0, \ldots, m\right\}$ et $\left\{\sigma^{2\,(t)} : t = 0, \ldots, m\right\}$. La phase transitoire se termine lorsque toutes les chaînes entrent dans leur phase stationnaire. Seulement cette dernière phase de la chaîne doit être conservée pour l'inférence.

6.B Loi de Student avec un paramètre d'échelle et de localisation

On dit que la variable aléatoire Y est distribuée selon la loi de Student à $\nu > 0$ degrés de liberté, avec paramètre de localisation $\mu \in \mathbb{R}$ et paramètre d'échelle $\sigma > 0$ si sa densité s'exprime sous la forme suivante :

$$f_Y(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left\{ \frac{\nu + \left(\frac{y-\mu}{\sigma}\right)^2}{\nu} \right\}^{-\left(\frac{\nu+1}{2}\right)}.$$

On dénote ceci par $Y \sim t_\nu(\mu, \sigma)$.

Si $Y \sim t_\nu(\mu, \sigma)$, alors

— $\mathbb{E}(Y) = \mu$ si $\nu > 1$;

— $\mathbb{V}ar(Y) = \frac{\nu}{\nu-2}\,\sigma^2$ si $\nu > 2$.