CHAPITRE 5

Modèles bayésiens pour la loi normale

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

Jonathan Jalbert – Automne 2019

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

- Utiliser le théorème de Bayes pour calculer la loi *a porteriori* des paramètres.
- Distinguer les lois *a priori* informatives et non informatives.
- Calculer des intervalles de crédibilité bayésien.
- Sélectionner le meilleur modèle parmi un ensemble de modèles.
- Calculer la loi prédictive pour une observation future.
- Estimer une espérance avec une méthode Monte-Carlo.

Dans ce chapitre, nous utiliserons le jeu de données **normaldata** disponible sur le site web du cours. Ces données proviennent de l'expérience de Michelson-Morley effectuée par Illingworth en 1927. Cette expérience avait pour but de mesurer la différence de la vitesse de la lumière entre les directions parallèle et perpendiculaire à l'éther. La différence de vitesse a été mesurée par interférométrie en calculant le déplacement des franges d'interférence de la lumière. La première colonne du jeu de données indique le temps de la journée où les essais se sont déroulés et la deuxième colonne correspond au déplacement moyen des franges d'interférence pour 10 essais indépendants.

5.1 Modèle gaussien

La loi normale est la loi la plus utilisée en statistique. Elle tient son importance du fait qu'il est possible de démontrer que les erreurs de mesure sont distribuées selon cette loi (voir Gauss). La loi normale, aussi appelée loi gaussienne ou loi de Laplace–Gauss, est la loi continue possédant la densité suivante sur les réels :

$$f_{(Y|\mu,\sigma^2)}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}, \quad \text{pour } y \in \mathbb{R}; \tag{5.1}$$

où $\mu \in \mathbb{R}$ correspond à la moyenne et $\sigma^2 > 0$ à la variance.

Soit un échantillon aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de taille n, pour lequel les variables aléatoires Y_i sont indépendantes et identiquement distribuées selon la loi normale de moyenne μ et de variance σ^2. Dénotons la réalisation de cet échantillon aléatoire par $\boldsymbol{y} = (y_1, \ldots, y_n)$. La fonction de vraisemblance pour cet échantillon aléatoire est donnée :

$$\begin{aligned} f_{(\boldsymbol{Y}|\mu,\sigma^2)}(\boldsymbol{y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}. \end{aligned}$$

Exercice 1

En statistique classique, quel est l'estimateur du maximum de la vraisemblance pour μ lorsque σ^2 est connue ? Trouvez également l'intervalle de confiance à 95% pour μ.

Pour l'expérience de Michelson-Morley effectuée en 1927 par Illingworth, ce-dernier a supposé que l'erreur de mesure de son montage expérimental correspondait à un écart-type de 0.75 frange de déplacement. Il supposa que chacune des $n = 64$ mesures étaient distribuées selon la loi normale $Y_i \sim \mathcal{N}(\mu, 0.75^2)$ où le paramètre μ inconnu correspond au *vrai* déplacement des franges d'interférence. Le nombre moyen du déplacement des franges d'interférence pour les 64 essais est de $\bar{y} = -0.015$.

5.2 Inférence bayésienne pour la loi normale lorsque la variance est connue

En statistique bayésienne, le paramètre inconnu est considérée comme une **variable aléatoire**. Dans le cadre de cette section, le paramètre inconnu est la moyenne μ de la

loi normale de variance connue σ^2. Le paramètre μ peut prendre des valeurs dans l'espace paramètre $\mathbb{R}$.

L'inférence bayésienne est basée sur la densité conditionnelle des paramètres sachant les observations, $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu)$, obtenue par le théorème de Bayes :

$$f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu) = \frac{f_{(\boldsymbol{Y}|\mu)}(\boldsymbol{y}) \times f_{\mu}(\mu)}{\displaystyle\int_{-\infty}^{\infty} f_{(\boldsymbol{Y}|\mu)}(\boldsymbol{y}) \times f_{\mu}(\mu)\, d\mu} \tag{5.2}$$

où $f_{\mu}(\mu)$ est appelée la loi *a priori* du paramètre μ sur l'espace paramètre $\mathbb{R}$. La densité conditionnelle du paramètre sachant les observations, $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu)$, est appelée la loi *a posteriori* du paramètre μ sur l'espace paramètre $\mathbb{R}$. Le dénominateur est parfois dénoté par $m(\boldsymbol{y})$. Il correspond à la loi marginale de l'échantillon aléatoire évaluée aux observations, ce qui constitue la constante de normalisation de la loi *a posteriori*.

La loi *a priori* modélise l'information que l'on possède sur le paramètre μ avant même d'obtenir les observations. La loi *a posteriori* correspond à la mise à jour de cette information après avoir incorporé l'information apportée par les observations. La loi *a posteriori* constitue l'outil principal de toute l'inférence bayésienne. Autrement dit, toute estimation bayésienne est basée sur la loi *a posteriori*.

5.2.1 La loi *a priori*

Puisque l'on souhaite utiliser le théorème de Bayes pour mettre à jour l'information que l'on possède sur le paramètre inconnu μ, la définition d'une loi *a priori* pour μ est nécessaire. La sélection de la loi *a priori* et de ses paramètres constitue un enjeu très important dans les analyses bayésiennes. La loi *a priori* doit être déterminée avant même de voir les données afin de ne pas introduire de biais dans l'analyse.

Il existe deux grandes familles de lois *a priori*. Les lois *a priori* informatives et les lois non informatives. Les lois *a priori* informatives procurent de l'information initiale sur les valeurs les plus probables du paramètre inconnu. L'information nécessaire peut être obtenue par les résultats d'une expérience précédente ou par le savoir d'un expert.

Pour le cas de l'expérience de Michelson-Morley, une loi a priori informative pour le paramètre inconnu μ pourrait être la suivante :

$$f_{\mu}(\mu) = \mathcal{N}(\mu|0, 0.75^2).$$

Cette loi suppose que les déplacements des franges les plus probables sont autour de 0. Une certaine quantité d'information a priori est nécessaire pour pouvoir fixer les paramètres de la loi *a priori*, par exemple avec le résultat d'une expérience antérieure ou la connaissance d'un expert. Une autre loi informative aurait pu être la loi de Student à 5 degrés de liberté :

$$f_{\mu}(\mu) = t_5(\mu|0, 0.75).$$

Dans ce cas aussi les valeurs autours de 0 sont plus probables mais la loi de Student possède des queues beaucoup plus lourdes, donc l'incertitude *a priori* sur μ serait plus grande.

Lorsqu'une loi informative est utilisée, l'information *a priori* doit être convertie en loi *a priori*. Plusieurs chapitres de livre sont dédiés à la description d'approches rigoureuses pour convertir l'information initiale que l'on possède en loi *a priori*. En pratique, on utilise souvent une approche plus simple pour spécifier la loi *a priori* qui consiste à prendre la loi a priori conjuguée du modèle statistique. Les lois conjuguées font l'objet de la section 5.2.3.

Si aucune information *a priori* n'est disponible, une loi *a priori* non informative peut être utilisée. Dans le cas continue, les lois non-informatives sont le plus souvent des lois impropres. Les lois non-informatives sont l'objet de la section 5.2.4.

5.2.2 Loi *a posteriori* de μ avec la loi *a priori* normale

Nous sommes dans le contexte où nous souhaitons calculer la loi *a posteriori* de μ d'une loi normale lorsque la variance σ^2 est connue.

Exemple 1

Dans le cas de l'expérience de Michelson-Morley, la loi *a priori* informative pour μ pourrait être $f_\mu(\mu) = \mathcal{N}(\mu|0, 0.75^2)$. Par conséquent, nous montrerons en classe à l'aide du théorème de Bayes que la loi *a posteriori* est la suivante :

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

La figure 5.1 illustre les densités *a priori* et *a posteriori* de μ.

L'exemple 1 illustre la différence entre l'estimation classique des paramètres et l'estimation bayésienne. En statistique classique, la méthode du maximum de la vraisemblance donne $\hat{\mu} = \bar{y}$ comme estimation du paramètre μ : un point sur la droite des réels. En statistique bayésienne, on obtient plutôt une densité de probabilité définie sur tous les réels : la loi *a posteriori*. Cette loi reflète en fait l'incertitude résiduelle sur la vraie valeur de μ après avoir pris en compte l'information *a priori* et celle apportée par les observations.

Dans le cas un peu plus général où la loi *a priori* informative suivante :

$$f_\mu(\mu) = \mathcal{N}(\mu|\nu, \tau^2),$$

est utilisée pour le paramètre inconnu μ de la loi normale avec variance connue, on peut montrer que la loi *a posteriori* correspondante est la suivante :

$$f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu) = \mathcal{N}\left\{\mu \left| \frac{\frac{1}{\tau^2}\nu + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}, \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-1}\right.\right\}. \tag{5.3}$$

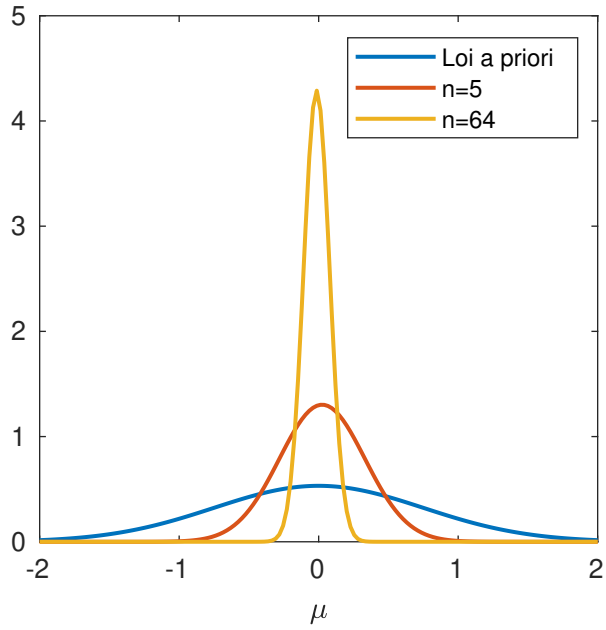

FIGURE 5.1 – Loi *a posteriori* de la moyenne θ lorsque la variance est connue pour $n = 5$ et pour $n = 64$. La densité en bleu est la loi *a priori*.

On peut remarquer de cette dernière équation que l'espérance de la loi *a posteriori* de μ s'exprime comme la moyenne pondérée de la moyenne échantillonnale $\bar{y}$ et de la moyenne *a priori*, où les poids respectifs sont donnés par la précision [1] de l'échantillon n/σ^2 et par la précision de la loi *a priori* $1/\tau^2$. On peut également noter que plus la taille de l'échantillon augmente, plus l'influence de la *loi priori* devient négligeable par rapport à l'information apportée par les données.

Remarque. *Lorsqu'une loi* a priori *conjuguée est utilisée, il est toujours possible d'exprimer l'espérance de la loi* a posteriori *comme une moyenne pondérée de l'échantillon et de la loi* a priori.

5.2.3 Lois conjuguées

Une loi est dite conjuguée si la loi *a priori* et la loi *a posteriori* partagent la même forme paramétrique. L'exemple 1 constitue un cas de loi conjuguée : la loi *a priori* ainsi que la loi *a posteriori* sont des lois gaussiennes. L'exercice 1 de la fin du chapitre vous permettra de déduire les différentes lois conjuguées en fonction de la vraisemblance du modèle statistique.

L'utilisation des lois *a priori* conjuguées est très répandue parce qu'elles permettent le calcul analytique des lois *a posteriori*. Elles possèdent donc des avantages computationnels

1. La précision est définie comme l'inverse de la variance.

très importants. L'exemple suivant illustre la complexité que peut prendre le calcul de la loi *a posteriori* si une loi *a priori* non conjuguée est utilisée.

Exemple 2

Soit la loi *a priori* $f_\mu(\mu) = t_5(\mu|0,1)$, la loi de Student centrée et réduite à 5 degrés de liberté, pour le paramètre μ de la loi normale de variance connue σ^2. En appliquant le théorème de Bayes, on trouve que

$$f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu) = \frac{\left(1+\frac{\mu^2}{5}\right)^{-3} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}}{\int_{-\infty}^{\infty}\left(1+\frac{\mu^2}{5}\right)^{-3} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}\, d\mu}.$$

Le dénominateur, qui correspond à la constante de normalisation de la loi *a posteriori* ne s'écrit pas sous une forme analytique. D'autre part, la loi *a posteriori* ne s'exprime pas sous la forme d'une densité connue. Des méthodes numériques sont alors nécessaires pour estimer cette loi *a posteriori* $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu)$. Nous verrons plusieurs techniques de la famille des méthodes Monte-Carlo permettant d'effectuer l'inférence bayésienne lorsque la densité *a posteriori* ne s'exprime pas sous une forme analytique. Pour cet exemple particulier, l'algorithme de Metropolis-Hastings (section 5.B) peut être utilisé pour obtenir un échantillon aléatoire de la loi *a posteriori*.

La loi *a priori* conjuguée possède une certaine forme paramétrique. Dans le cadre de ce chapitre, la loi *a priori* conjuguée pour le paramètre μ de la loi normale de variance connue est la loi normale. Pour utiliser cette loi *a priori* conjuguée, il faut spécifier ses paramètres avant même d'avoir vu les observations. Les paramètres de la loi *a priori* sont appelés **hyperparamètres**. Suffisamment d'information doit donc être disponible avant d'effectuer l'expérience pour bien définir la loi *a priori* conjuguée. Cette information peut être fournie par une expérience précédente ou le savoir d'un expert.

5.2.4 Lois *a priori* non informatives

Dans plusieurs cas, l'information *a priori* est soit inexistante ou très difficile à obtenir, ce qui freine l'utilisation des lois *a priori* informatives. Lorsqu'aucune information *a priori* n'est disponible, les lois non informatives sont alors très utiles. Lorsque l'espace paramètre est non borné, les densités non informatives correspondent la plupart du temps à des lois impropres, c'est-à-dire des densités non normalisées. En particulier, on dit que la loi *a priori* $f_\mu(\mu)$ pour μ est impropre si

$$\int_{-\infty}^{\infty} f_\mu(\mu)\, d\mu = \infty$$

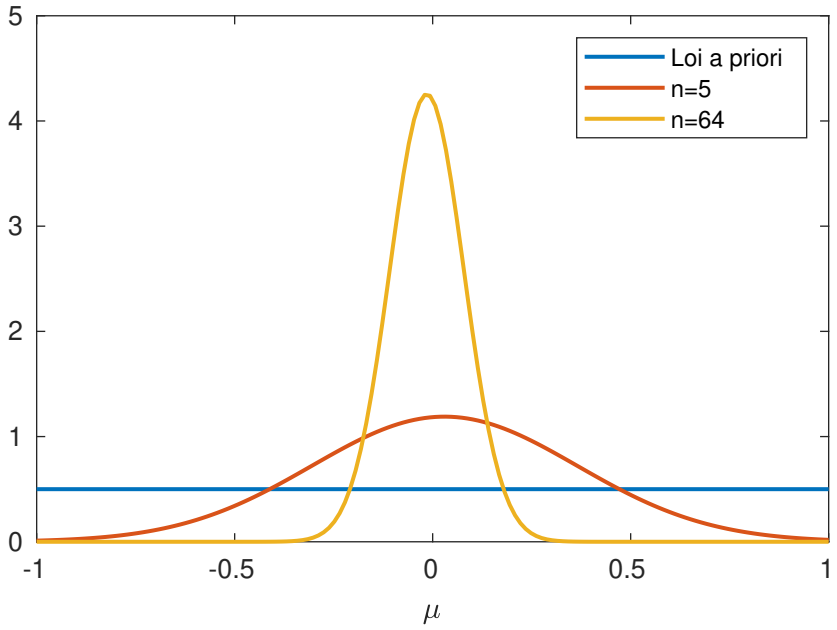

FIGURE 5.2 – Loi *a posteriori* de la moyenne μ lorsque la variance est connue pour $n = 5$ et pour $n = 64$. La densité en bleu est la loi *a priori* impropre.

Les lois impropres ne sont donc pas des densités de probabilité. Néanmoins, la densité *a posteriori* correspondante peut s'avérer valide même si une loi impropre est utilisée, tel que présenté dans l'exemple suivant.

Exemple 3

Soit la densité *a priori* impropre $f_\mu(\mu) \propto 1, \quad \text{pour } \mu \in \mathbb{R}$ pour la moyenne de la loi normale de variance σ^2 connue. La densité *a posteriori* associée un échantillon aléatoire de taille n est la suivante :

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

Cette densité est illustrée à la figure 5.2.

L'utilisation des lois *a priori* impropres n'est pas toujours possible. La condition

$$\int_{-\infty}^{\infty} f_{(\boldsymbol{Y}|\mu)}(\boldsymbol{y})\, f_\mu(\mu)\, d\mu < \infty$$

doit impérativement être vérifiée pour que la loi *a priori* impropre mène vers une loi *a posteriori* valide.

Exemple 4

Pour la loi normale avec variance connue σ^2, la densité *a priori* impropre $f_\mu(\mu) \propto 1$ impliquera toujours une densité a posteriori valide lorsque $n \geq 1$. Nous montrerons ce résultat en classe.

5.2.5 Estimations ponctuelles bayésiennes

La loi *a posteriori* résume toute l'information disponible sur les paramètres inconnus. En général, on essaie d'éviter en statistique bayésienne d'utiliser des estimations ponctuelles pour les paramètres inconnus puisque la loi *a posteriori* est beaucoup plus riche en information. De façon pragmatique, des estimateurs ponctuels bayésiens peuvent néanmoins être utilisés. Un estimateur ponctuel doit donc être défini à partir de la loi *a posteriori*. La façon rigoureuse de procéder consiste à utiliser la théorie de la décision mais c'est une approche rarement utilisée en pratique.

Dans le cadre de ce cours, nous nous limiterons à mentionner deux règles populaires pour l'obtention des estimateurs ponctuels bayésiens. La première règle consiste à prendre la moyenne de la loi *a posteriori* des paramètres :

$$\hat{\mu} = \mathbb{E}\left(\mu | \boldsymbol{Y} = \boldsymbol{y}\right) = \int_{-\infty}^{\infty} \mu \, f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu) \, d\mu.$$

La deuxième règle consiste à prendre le mode de la loi *a posteriori* des paramètres :

$$\hat{\mu} = \arg\max_{\mu \in \mathbb{R}} f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu).$$

Ce dernier estimateur ressemble davantage aux estimateurs du maximum de la vraisemblance. Il est d'ailleurs possible de démontrer que si la taille de l'échantillon n tend vers l'infini, alors cet estimateur ponctuel bayésien et l'estimateur du maximum de la vraisemblance concordent.

5.2.6 Estimations par intervalles de crédibilité bayésiens

En statistique bayésienne, une estimation de μ par intervalle est très simple à obtenir. Celle-ci est basée sur la loi *a posteriori* des paramètres $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\theta)$. Un intervalle de crédibilité I de niveau nominal $(1-\alpha)$ pour $0 \leq \alpha \leq 1$ signifie que

$$\mathbb{P}\left(\mu \in I | \boldsymbol{Y} = \boldsymbol{y}\right) = 1 - \alpha. \tag{5.4}$$

L'intervalle de crédibilité de niveau $(1-\alpha)$ le plus simple à obtenir est l'intervalle entre les quantiles d'ordre $(1-\alpha/2)$ et $\alpha/2$ de la loi *a posteriori* du paramètre μ. L'intervalle

de crédibilité de niveau $(1-\alpha)$ qui possède la plus petite longueur est appelé région HPD (pour *highest probability density*). La région HPD de niveau $(1-\alpha)$ est donnée par

$$\left\{\mu : f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu) \geq k_\alpha\right\},$$

où k_α est la constante pour laquelle l'équation (5.4) est satisfaite.

5.2.7 Distribution prédictive

En statistique bayésienne, puisque le paramètre inconnu est considéré comme une variable aléatoire, on peut utiliser la loi des probabilité totales pour définir la distribution prédictive d'une observation future. Autrement dit, on souhaite calculer la densité suivante

$$f_{(Y_{n+1}|\boldsymbol{Y}=\boldsymbol{y})}(\tilde{y})$$

en intégrant l'incertitude sur les paramètres inconnus. Le concept de distribution prédictive n'existe qu'en statistique bayésienne.

Dans le cas de l'expérience de Michelson-Morley, la densité prédictive de la 65^e observation peut s'obtenir ainsi :

$$f_{(Y_{65}|\boldsymbol{Y}=\boldsymbol{y})}(\tilde{y}) = \int_{-\infty}^{\infty} f_{(Y_{65}|\mu)}(\tilde{y}) \times f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu)\, d\mu.$$

On conditionne sur le paramètre μ puis on intègre ce paramètre en pondérant par sa densité *a posteriori*. La densité prédictive constitue donc une moyenne des densités $f_{(Y_{65}|\mu)}(\tilde{y})$ pondérée par la densité *a posteriori* du paramètre inconnu μ.

Exemple 5

Dans le cas de l'expérience de Michelson-Morley où $n = 64$ avec la loi *a priori* impropre, la densité prédictive de $(Y_{65}|\boldsymbol{Y} = \boldsymbol{y})$ est la suivante :

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

Nous montrerons ce résultat en classe.

5.3 Inférence bayésienne pour la loi normale lorsque la variance est inconnue

En pratique, la variance est rarement connue. Dans cette section, nous traiterons le cas où la variance est inconnue. Il y a donc deux paramètres inconnus : la moyenne μ et la variance σ^2 de la loi normale.

5.3.1 Loi *a priori* partiellement conjuguée

Dans le cas de la vraisemblance gaussienne lorsque la moyenne et la variance sont inconnues, il n'existe pas de loi *a priori* conjuguée pour le couple de paramètres (μ, σ^2). Il existe seulement une loi partiellement conjuguée :

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

Avec cette loi *a priori*, on obtient que

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

et

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

mais la loi conjointe $f_{\{(\mu,\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\mu, \sigma^2)$ ne s'exprime pas sous une forme analytique. La loi *a priori* est partiellement conjuguée parce que les **lois conditionnelles complètes** des paramètres s'expriment sous une forme connue. À la section 5.C, nous verrons comment ces lois conditionnelles complètes sont utiles pour mettre en place un algorithme numérique, l'échantillonnage de Gibbs, permettant de générer un échantillon aléatoire de la loi *a posteriori* $f_{\{(\mu,\sigma^2)|\boldsymbol{Y}=\boldsymbol{y}\}}(\mu, \sigma^2)$.

Pour l'expérience de Michelson-Morley avec les hyperparamètres $\nu = 0$, $\tau^2 = 1$ et $\alpha = \beta = 1$, la densité *a posteriori* est illustrée à la figure 5.3a.

Remarque. *Si on a un vecteur de paramètres inconnus $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. La loi conditionnelle complète du paramètre θ_i pour $1 \leq i \leq p$ est la densité conditionnelle de θ_i sachant tous les autres paramètres ainsi que les observations :*

$$f_{(\theta_i|\boldsymbol{Y}=\boldsymbol{y},\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_p)}(\theta_i).$$

Dans le cas de la loi normale avec la moyenne et la variance inconnues, on a deux lois conditionnelles complètes :

$$f_{(\mu|\sigma^2,\boldsymbol{Y}=\boldsymbol{y})}(\mu) \quad \textit{et} \quad f_{(\sigma^2|\mu,\boldsymbol{Y}=\boldsymbol{y})}(\sigma^2).$$

5.3.2 Loi *a priori* non informative

La variance σ^2 est un paramètre qui ne peut être que positif. Le logarithme de la variance est un réel $\phi = \ln \sigma^2 \in \mathbb{R}$. On préfère utiliser une densité impropre constante *a priori* pour le logarithme de la variance :

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

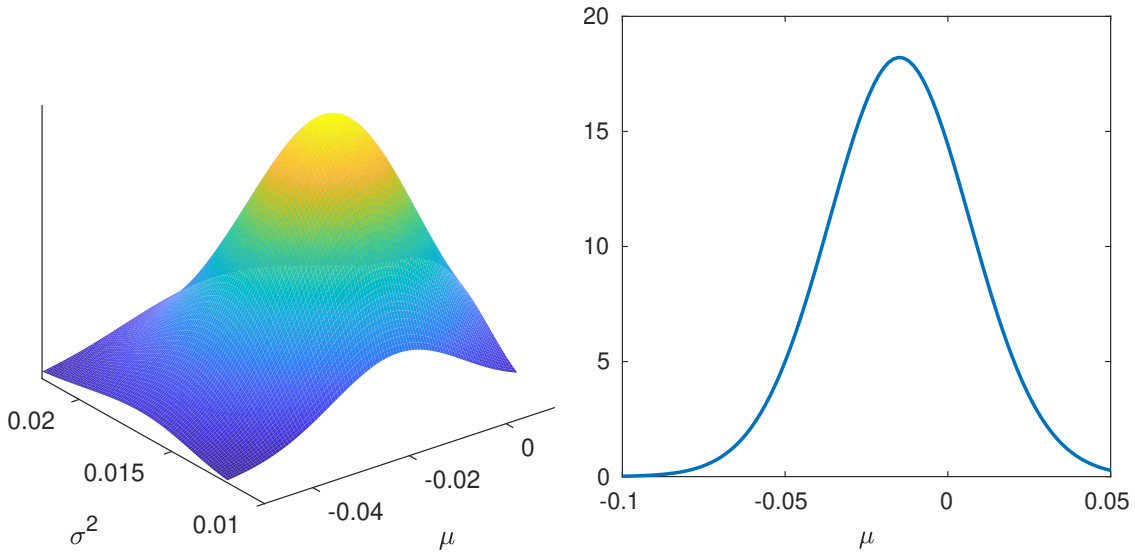

FIGURE 5.3 – (a) Densité conjointe *a posteriori* de (μ, σ^2). (b) Densité marginale *a posteriori* de μ.

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

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

Par conséquent, la densité impropre *a priori* pour la moyenne et la variance est

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

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

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

Cette densité ne s'exprime pas sous une forme analytique.

Les densités conditionnelles complètes s'expriment néanmoins sous des formes connues :

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

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 5.C permettra de générer un échantillon aléatoire de cette densité à partir des densités conditionnelles complètes.

5.3.3 Densité marginale *a posteriori* de μ.

Dans le cas de l'expérience de Michelson-Morley, le paramètre d'intérêt est μ, le déplacement moyen des franges d'interférence. La variance σ^2 constitue ce qu'on appelle un paramètre de nuisance puisqu'il ne constitue pas un paramètre d'intérêt principal. On peut donc intégrer ce paramètre de la densité *a posteriori* pour obtenir la loi marginale de μ :

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

On revient donc à un problème unidimensionnel.

Exemple 6

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 l'expérience de Michelson-Morley, la loi marginale *a posteriori* de μ est illustrée à la figure 5.2.
Pour montrer ce résultat, on intègre σ^2 de la forme fonctionnelle de la loi *a posteriori* :

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

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

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

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.*

5.4 Sélection de modèle

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 sélectionner le meilleur modèle où chacun de ceux-ci correspond à une hypothèse à tester.

5.4.1 L'indice du modèle comme variable aléatoire

La sélection de modèle consiste à identifier le modèle statistique qui est le plus cohérent avec les observations parmi un ensemble prédéfini de modèles. L'ensemble des J modèles considérés est dénoté par $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$. L'approche naturelle en statistique bayésienne consiste à attribuer à chacun des modèles considérés la probabilité qu'ils soient vrais, autrement dit la probabilité que les observations aient été générées par ces modèles.

Soit la variable aléatoire indicatrice M prenant des valeurs dans l'ensemble $\{1, \ldots, J\}$. Le cas $M = k$ correspond à l'événement *les observations ont été générées par le modèle* $\mathcal{M}_k$. L'idée est de calculer la fonction de masse conditionnelle de la variable indicatrice M sachant les observations $\boldsymbol{y} = (y_1, \ldots, y_n)$:

$$\mathbb{P}\left(M = k \mid \boldsymbol{Y} = \boldsymbol{y}\right), \quad \text{pour } k \in \{1, \ldots, J\}. \tag{5.5}$$

Dans ce contexte, le modèle le plus cohérent avec les observations correspond au mode de la fonction de masse *a posteriori* $\mathbb{P}\left(M = k \mid \boldsymbol{Y} = \boldsymbol{y}\right)$. Le calcul de cette probabilité requiert l'utilisation du théorème de Bayes :

$$\mathbb{P}\left(M = k \mid \boldsymbol{Y} = \boldsymbol{y}\right) \propto \mathbb{P}\left(M = k\right) \int_{\Theta_k} f_{(\boldsymbol{Y}|\boldsymbol{\theta}_k)}(\boldsymbol{y})\, f_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k)\, d\boldsymbol{\theta}_k,$$

où $\mathbb{P}\left(M = k\right)$ correspond à la probabilité *a priori* que le modèle $\mathcal{M}_k$ soit vrai. La plupart du temps en pratique, la probabilité *a priori* des modèles est supposée uniforme sur l'ensemble des J modèles considérés :

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

Remarque. *La loi* a priori *des paramètres du modèle* $\mathcal{M}_k$, $f_{\theta_k}(\theta_k)$ *dépend de* k, *de l'indice du modèle. En effet, différents modèles peuvent avoir des espaces paramètres* Θ_k *différents.*

Exemple 7

Dans le cas de l'expérience de Michelson-Morley effectuée par Illingworth, on peut souhaiter déterminer lequel parmi les deux modèles suivants est le meilleur :

- $\mathcal{M}_1 : Y_i \sim \mathcal{N}(0, 0.75^2)$;
- $\mathcal{M}_2 : Y_i \sim \mathcal{N}(0, \sigma^2)$;

Le modèle $\mathcal{M}_1$ suppose que la variance est connue et est égale à 0.75^2 et le modèle $\mathcal{M}_2$ suppose que la moyenne est inconnu.

Exemple 8

Dans un modèle de régression linéaire, on peut souhaiter déterminer lequel parmi les deux modèles suivants est le meilleur :

— $\mathcal{M}_1 : Y_i = \beta_0 + \beta_1 x_1 + \varepsilon$;

— $\mathcal{M}_2 : Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

5.4.2 Le facteur de Bayes

Lorsque le choix d'un modèle peut avoir des conséquences importantes, choisir naïvement le modèle qui maximise la probabilité (5.5) peut s'avérer être un choix risqué. Le facteur de Bayes permet de quantifier la certitude qu'un modèle soit faux par rapport à un autre.

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 :

$$B_{21} = \frac{\mathbb{P}(M = 2 \mid \boldsymbol{Y} = \boldsymbol{y})}{\mathbb{P}(M = 1 \mid \boldsymbol{Y} = \boldsymbol{y})} \times \frac{\mathbb{P}(M = 1)}{\mathbb{P}(M = 2)}.$$

L'interprétation du facteur de Bayes introduite par Jeffreys (1939) est résumée au tableau 5.1. Bien que cette interprétation soit arbitraire, elle est néanmoins très utile pour la sélection de modèle en l'absence d'un cadre décisionnel formel. Elle permet de choisir un modèle si celui-ci augmente *significativement* la probabilité que celui-ci soit vrai.

TABLE 5.1 – 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**

Exercice 2

Montrez que le facteur de Bayes peut s'exprimer en fonction des densités marginales des modèles :

$$B_{21} = \frac{\displaystyle\int_{\Theta_2} f_{(\boldsymbol{Y}|\theta_2)}(\boldsymbol{y})\, f_{\theta_2}(\theta_2)\, d\theta_2}{\displaystyle\int_{\Theta_1} f_{(\boldsymbol{Y}|\theta_1)}(\boldsymbol{y})\, f_{\theta_1}(\theta_1)\, d\theta_1} = \frac{m_2(\boldsymbol{y})}{m_1(\boldsymbol{y})},$$

où $m_1(\boldsymbol{y})$ et $m_2(\boldsymbol{y})$ sont respectivement les densités marginales des modèles $\mathcal{M}_1$ et

$\mathcal{M}_2$ évaluées aux observations $\boldsymbol{y}$.

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 :

Exemple 9

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

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

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

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

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

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

5.4.3 Le critère BIC

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 *Bayesian Information Criterion* (BIC) 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 [2]. En particulier, on a que le BIC approxime le logarithme de la loi marginale du modèle évaluée aux observations :

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

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

Le BIC est défini de la façon suivante :

$$\text{BIC} = \ln f_{(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}_{MV})}(\boldsymbol{y}) - \frac{k}{2}\ln n, \tag{5.6}$$

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

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

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 5.1) s'avère utile.*

5.5 Exercices

1. Loi *a priori* conjuguée. Dans le tableau suivant, montrez que la loi *a posteriori* résulte bien de la loi *a priori* et de la vraisemblance données lorsque l'on a une seule observation y. Le paramètre inconnu est dénoté θ. Les autres paramètres sont supposés connus.

$f_{(Y\|\theta)}(y)$	$f_\theta(\theta)$	$f_{(\theta\|Y=y)}(\theta)$
$\mathcal{N}(y \mid \theta, \sigma^2)$	$\mathcal{N}(\theta \mid \nu, \tau^2)$	$\mathcal{N}\left(\theta \mid \frac{\sigma^2\mu+\tau^2 y}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)$
$\mathcal{P}oisson(y \mid \theta)$	$\mathcal{G}amma(\theta \mid \alpha, \beta)$	$\mathcal{G}amma(\theta \mid \alpha + y, \beta + 1)$
$\mathcal{B}inomiale(y \mid n, \theta)$	$\mathcal{B}eta(\theta \mid \alpha, \beta)$	$\mathcal{B}eta(\theta \mid \alpha + y, \beta + n - y)$
$\mathcal{E}xponentielle(y \mid \theta)$	$\mathcal{G}amma(\theta \mid \alpha, \beta)$	$\mathcal{G}amma(\alpha + 1, y + \beta)$

2. Soit un échantillon aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de taille n de la loi $\mathcal{C}auchy(\mu, 1)$. La densité de la loi $\mathcal{C}auchy(\mu, 1)$ s'exprime

$$f_{(Y|\mu)}(y) = \frac{1}{\pi\{1 + (y-\mu)^2\}}, \text{ pour } y \in \mathbb{R},\ \mu \in \mathbb{R}.$$

   a) Existe-t-il une loi *a priori* conjuguée pour μ ?

   b) Si on utilise la loi $\mathcal{N}(0, 10)$ comme loi *a priori* de μ, quelle est la forme fonctionnelle de la loi *a posteriori* ? La forme fonctionnelle est la densité non normalisée.

3. Obtenez avec l'algorithme de Metropolis-Hastings une estimation de la moyenne et de la variance de la loi $t_5(0,1)$ en utilisant une marche aléatoire comme loi de proposition des candidats. Comparez vos résultats numériques aux valeurs théoriques.

4. Soit un échantillon aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de taille n obtenu de la densité $f_{(Y|\theta)}(y)$. Montrer que la loi obtenue en actualisant la loi *a posteriori* une observation à la fois est identique à celle obtenue si on considère l'échantillon au complet d'un seul coup.

5. Considérez le jeu de données **normaldata** disponible sur le site du cours. Supposez que les observations sont distribuées selon la loi $\mathcal{N}\left\{\mu, (0,75)^2\right\}$. Autrement dit, on suppose que la variance est connue. Supposez la loi *a priori* non informative $f_\mu(\mu) \propto 1$. La moyenne des 64 observations est de $\bar{y} = -0.0148$. Si on mesurait une 65 fois, dans quel intervalle la mesure aurait 95% de chance de se retrouver ?

6. Considérez le jeu de données **normaldata** disponible sur le site du cours. Soit le modèle $\mathcal{M}_1$ supposant que $Y_i \sim \mathcal{N}\left\{\mu, (0,75)^2\right\}$ et le modèle $\mathcal{M}_2$ supposant que $Y_i \sim \mathcal{N}\left\{\mu, \sigma^2\right\}$. Considérez les lois *a priori* suivantes :

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

   Avec les 64 observations, on obtient $\bar{y} = -0.0148$ 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 ?

7. En août 2013, le New York Times publiait un sondage effectué sur 599 personnes concernant la satisfaction à l'égard de Barack Obama. La proportion de gens satisfaits était de 52%. Considérez la loi non informative suivante $f_\theta(\theta) \propto \theta^{-1}(1-\theta)^{-1}$ pour la proportion de gens satisfaits θ.

   a) Quelle est la loi *a posteriori* de la proportion de gens satisfait ?

   b) Donnez un intervalle de crédibilité à 95% de la proportion de gens satisfait ?

   c) Si on demande à une 600e personne, quelle est la probabilité que celle-ci soit satisfaite du travail de Barack Obama ?

8. Au TD5, nous avons vu qu'une implémentation efficace et stable de l'algorithme de

Metropolis-Hastings utilisait le logarithme de ρ. Dans ce numéro, nous verrons les avantages d'utiliser le $\ln \rho$ plutôt que ρ.

a) Quelle est l'expression appropriée pour $\ln \rho$?

b) Soit la variable aléatoire $U \sim Uniforme(0, 1)$ et la variable aléatoire $V = \ln U$. Quelle est la probabilité que $V \leq \ln \rho$?

c) Si $\ln \rho$ est défini de la façon suivante en négligeant le minimum :

$$\ln \rho = \ln g_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\tilde{\theta}) - \ln g_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta^{(t)});$$

quelle est la probabilité d'accepter le candidat $\tilde{\theta}$ si celui-ci est plus probable ? Autrement dit, calculez la probabilité de l'événement $V \leq \ln \rho$ en considérant que le candidat est plus favorable.

Annexe

5.A Méthodes Monte-Carlo

L'utilisation du théorème de Bayes nécessite le calcul de la constante de normalisation de la loi *a posteriori* que l'on dénote parfois par $m(\boldsymbol{y})$. À part dans quelques cas particuliers notamment lorsque la loi *a priori* est conjuguée, le calcul analytique de la loi *a posteriori* s'avère impossible. L'utilisation des méthodes numériques d'intégration pour évaluer la constante de normalisation n'est généralement pas recommandée notamment parce que l'erreur numérique associée à ces méthodes explosent lorsque la dimension de l'espace paramètre augmente. Les méthodes dites de Monte-Carlo permettent de contourner cette difficulté numérique. Elles ont été nommées ainsi en raison du district de Monte-Carlo de la principauté de Monaco où s'agglomèrent un bon nombre de casinos. La procédure Monte-Carlo originale a été développée par le mathématicien Stanislaw Ulam (1909-1984) pour estimer la probabilité de gagner au jeu Solitaire.

Tout au long du cours, nous verrons plusieurs algorithmes de la famille des méthodes Monte-Carlo qui sont essentielles à tout statisticien bayésien. Dans le cadre du cours, nous nous limiterons à présenter ces méthodes afin de pouvoir les mettre en pratique dans les problèmes concrets. Pour traiter en profondeur ces aspects, une formation de deuxième cycle en probabilités est nécessaire.

5.A.1 Approximation d'une espérance par simulations Monte-Carlo

Supposons que l'on souhaite estimer la quantité $I = \mathbb{E}\{h(Y)\}$ où la variable aléatoire Y est distribuée selon la densité $f(y)$. On a que

$$I = \int_{-\infty}^{\infty} h(y) f(y)\, dy. \tag{5.7}$$

Si $y_1, y_2, \ldots, y_m$ constituent des réalisations indépendantes selon la densité $f_Y(y)$ et si m est suffisamment grand, alors

$$I \approx \frac{h(y_1) + h(y_2) + \ldots + h(y_m)}{m}.$$

Exemple 10

Supposons que l'on souhaite estimer l'espérance de la variable aléatoire Y distribuée selon la densité $f_Y(y)$. Si on possède un échantillon aléatoire $y_1, \ldots, y_m$ de la densité $f_Y(y)$, alors l'espérance peut être approximée par

$$\mathbb{E}(Y) \approx \frac{y_1 + \ldots + y_m}{m}.$$

Exemple 11

Supposons que l'on souhaite estimer la variance de la variable aléatoire Y distribuée selon la densité $f_Y(y)$. Si on possède un échantillon aléatoire $y_1, \ldots, y_m$ de la densité $f_Y(y)$, alors la variance peut être approximée par

$$\mathbb{V}ar(Y) = \mathbb{E}(Y^2) - \{\mathbb{E}(Y)\}^2 \approx \frac{y_1^2 + \ldots + y_m^2}{m} - \left(\frac{y_1 + \ldots + y_m}{m}\right)^2.$$

Cette méthode de simulation de base est très utile lorsque la distribution de $h(Y)$ est inconnue et qu'il est possible de générer un échantillon aléatoire de la densité $f_Y(y)$. Lorsqu'il sera impossible de générer directement un échantillon aléatoire de la de la densité $f_Y(y)$, des méthodes de simulation plus avancées seront nécessaires. L'algorithme de Metropolis-Hastings, qui constitue l'objet de la prochaine section, est un exemple de méthode avancée permettant de générer un échantillon aléatoire de la densité $f_Y(y)$.

5.B Algorithme de Metropolis-Hastings

Dans cette section, une première méthode Monte-Carlo par chaîne de Markov (MCMC) est présentée : l'algorithme de Metropolis-Hastings. Les méthodes MCMC ont gagné en popularité dans les années 1980 grâce à l'augmentation de la capacité de calcul des ordinateurs

personnels. Les méthodes MCMC ont démocratisé l'application de la statistique bayésienne et c'est pourquoi elle est aujourd'hui «largement» utilisée. Les méthodes MCMC constituent des algorithmes permettant de générer un échantillon aléatoire d'une loi de probabilité dont la constante de normalisation est inconnue. Ces méthodes sont donc particulièrement utiles dans le contexte bayésien lorsque la densité de la loi *a posteriori* ne s'exprime pas sous une forme analytique.

Dans cette section, l'algorithme de Metropolis-Hastings est illustré pour une densité *a posteriori* unidimensionnelle. Dénotons par θ le paramètre d'intérêt. Supposons que l'on ne connaît que la forme fonctionnelle $g_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta)$ de la loi *a posteriori* $f_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta)$ du paramètre θ :

$$f_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta) = \frac{1}{C} g_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta);$$

avec la constante de normalisation C inconnue. Un échantillon aléatoire de la densité $f_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta)$ sera obtenu en n'utilisant que la forme fonctionnelle $g_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta)$. L'algorithme de Metropolis-Hastings est une procédure itérative qui fonctionne de la façon suivante. Un état initial dénoté $\theta^{(1)}$ est fixé pour θ. Un candidat est proposé pour la valeur suivante de θ. Si cette valeur est plus favorable, on accepte le candidat. Sinon, la candidat est tout de même accepté avec une certaine probabilité dépendant de $g_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta)$. Ces deux dernières étapes sont répétées un très grand nombre de fois. La propriété remarquable de l'algorithme est que tôt ou tard, la procédure produira un échantillon aléatoire de $f_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta)$, peu importe l'état initial.

L'algorithme de Metropolis-Hastings nécessite une loi de proposition des candidats. La loi de proposition la plus simple et la plus utilisée en pratique correspond à une marche aléatoire autour de l'état présent. Soit $\theta^{(t)}$ l'état de θ à l'itération t. Dénotons par $\tilde{\theta}$ le candidat proposé pour l'état suivant. La marche aléatoire consiste à ajouter un pas aléatoire à l'état présent :

$$\tilde{\theta} = \theta^{(t)} + \delta;$$

où δ est une réalisation d'une variable aléatoire de densité symétrique autour de 0, par exemple $\delta \sim \mathcal{N}(0, 1^2)$.

Un échantillon obtenu par l'algorithme de Metropolis-Hastings comporte deux phases. Une phase de chauffe et une phase d'échantillonnage. La phase de chauffe est une phase transitoire où l'algorithme explore l'espace paramètre. La longueur de la phase de chauffe peut être déterminée visuellement en traçant la chaîne obtenue $\left\{\theta^{(t)} : t = 1, \ldots, m\right\}$ en fonction des itérations. La phase transitoire se termine lorsque la chaîne entre dans la partie stationnaire, appelée phase d'échantillonnage. Seulement cette dernière phase de la chaîne doit être conservée comme échantillon aléatoire de la loi cible. Le nombre d'itérations nécessaires avant d'entrer dans la phase d'échantillonnage est généralement inconnu. Il dépend notamment de l'état initial des paramètres et du modèle statistique. Ce qui est toutefois remarquable des méthodes MCMC, c'est que l'algorithme produira tôt ou tard un échantillon aléatoire de $f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta})$ et ce, peu importe les valeurs initiales.

Algorithm 1 Algorithme de Metropolis-Hastings avec une marche aléatoire dont le pas est symétrique autour de 0.

Initialiser l'état initial du paramètre $\theta^{(1)}$.
Définir la loi du pas de la marche aléatoire, par exemple $\delta \sim \mathcal{N}(0, 1^2)$.
for t = 1 à m **do**

1. Sachant l'état présent du paramètre $\theta^{(t)}$, générer un pas δ pour obtenir le candidat pour l'état au temps $(t+1)$:
$$\tilde{\theta} = \theta^{(t)} + \delta.$$

2. Calculer
$$\rho = \min\left\{ \frac{g_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\tilde{\theta})}{g_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta^{(t)})}, 1 \right\}.$$

3. Attribuer l'état au temps $(t+1)$ de la façon suivante :
$$\theta^{(t+1)} = \begin{cases} \tilde{\theta} \text{ avec probabilité } \rho, \\ \theta^{(t)} \text{ avec probabilité } 1-\rho. \end{cases}$$

end for

Pour que la chaîne générée $\left\{\theta^{(t)} : t = 1, \ldots, m\right\}$ possède des propriétés optimales, le taux d'acceptation des candidats de la phase d'échantillonnage doit être entre 40% et 70%. Un taux d'acceptation trop grand indique que l'algorithme n'explore pas suffisamment l'espace paramètres. La variance du pas de la marche aléatoire doit donc être augmentée. Un taux d'acceptation trop petit indique que l'algorithme n'est pas optimal : un nombre très important d'itérations sera nécessaire pour obtenir un échantillon de la loi cible. Dans ce cas, la variance du pas de la marche aléatoire doit être diminuée. En pratique, on utilise souvent pour le pas δ la loi normale de moyenne 0 et de variance ajustée de façon à ce que la proportion d'acceptation des candidats se situe entre 40% et 70%.

5.C 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 générique $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. 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* $g_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta})$, *i.e.*

$$f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}) = \frac{1}{C} g_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}).$$

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

Algorithm 2 Échantillonnage de Gibbs

Initialiser l'état initial des paramètres $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_p^{(0)}\right)$.
for t = 1 à N **do**
1. Tirer $\theta_1^{(t)}$ de la loi $f_{(\theta_1|\theta_2^{(t-1)},\ldots,\theta_p^{(t-1)},\boldsymbol{Y}=\boldsymbol{y})}(\theta_1)$.
2. Tirer $\theta_2^{(t)}$ de la loi $f_{(\theta_2|\theta_1^{(t)},\theta_3^{(t-1)},\ldots,\theta_p^{(t-1)},\boldsymbol{Y}=\boldsymbol{y})}(\theta_2)$.

$\vdots$

p. Tirer $\theta_p^{(t)}$ de la loi $f_{(\theta_p|\theta_1^{(t)}\ldots,\theta_{p-1}^{(t)},\boldsymbol{Y}=\boldsymbol{y})}(\theta_p)$.
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 $\left\{\theta_j^{(t)} : t = 1, \ldots, m\right\}$ pour tous les paramètres en fonction des itérations. 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.