CHAPITRE 5

Modèles bayésiens pour la moyenne de la loi normale

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

Jonathan Jalbert – Hiver 2023

Ce chapitre introduit les concepts fondamentaux de la statistique bayésienne pour le cas particulier de la loi normale avec la variance connue. 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.
- Calculer la loi prédictive pour une observation future.
- Estimer une espérance avec une méthode Monte-Carlo.
- Implémeter l'algorithme de Metropolis-Hastings pour simuler un échantillon aléatoire d'une loi de probabilité.

Contexte

Dans ce chapitre, nous réanalyserons les données d'une expérience fondamentale en physique moderne : l'expérience de Michelson-Morley effectuée en 1927 par Illingworth[1].

L'expérience de Michelson-Morley avait pour but de mesurer la vitesse de l'éther, la substance dans laquelle on supposait que la lumière se propageait. Le dispositif consistait à mesurer la différence de la différence de la vitesse de la lumière entre les directions parallèle

1. Illingworth, K. K. (1927). A Repetition of the Michelson-Morley Experiment Using Kennedy's Refinement. *Physical Review*, 30(5), 692–696.

TABLE 5.1 – 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

et perpendiculaire à l'éther. Puisque la vitesse de la lumière est trop grande pour être mesurée directement, la différence de vitesse se mesurait par interférométrie optique en reliant le déplacement des franges d'interférence de la lumière à la vitesse de l'éther. Pour le montage utilisée par Illingworth, la vitesse de l'éther par rapport à la Terre était donnée par $V = 112Y^{\frac{1}{2}}$, où Y correspond aux nombres de frange d'interférence de déplacement.

Le résultat de l'expérience est que peu importe l'orientation du montage, on obtient que la vitesse de l'éther est nulle. Cela suggère que l'éther n'existe pas. Cette expérience est fondamentale en physique moderne et d'ailleurs, Michelson a d'ailleurs reçu le prix Nobel de physique pour le développement de celle-ci. Elle a été répétée plusieurs fois toujours de façon plus précise avec les mêmes conclusions.

Dans ce chapitre, nous utiliserons le jeu de données `illingworth1927.csv` disponible sur le site web du cours. Ces données proviennent de l'expérience de Michelson-Morley effectuée par Illingworth en 1927. Les données mesurées sont illustrées à la figure 5.1. Ces données sont intéressantes pour ce chapitre car Illingworth a très bien étudié la précision de son montage avant d'effectuer les mesures. Il a estimé que l'erreur de mesure de mesure standard de son montage correspondait à 1.5 millième de frange d'interférence.

5.1 Modèle gaussien

Considérons pour le moment les 6 observations observations effectuées à 5 a.m. avec le montage orienté dans la direction N répertoriées dans le tableau 5.1. Pour chacune de ces observations, le modèle statistique suivant a été supposé :

$$Y_i = \mu + \varepsilon_i \tag{5.1}$$

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 mesures sont indépendantes et identiquement distribuées selon la loi normale de moyenne nulle et de variance σ^2, alors le modèle décrit à l'équation (5.1) peut s'exprimer sous la forme suivante :

$$Y_i \sim \mathcal{N}(\mu, \sigma^2) \tag{5.2}$$

pour $1 \leq i \leq n$.

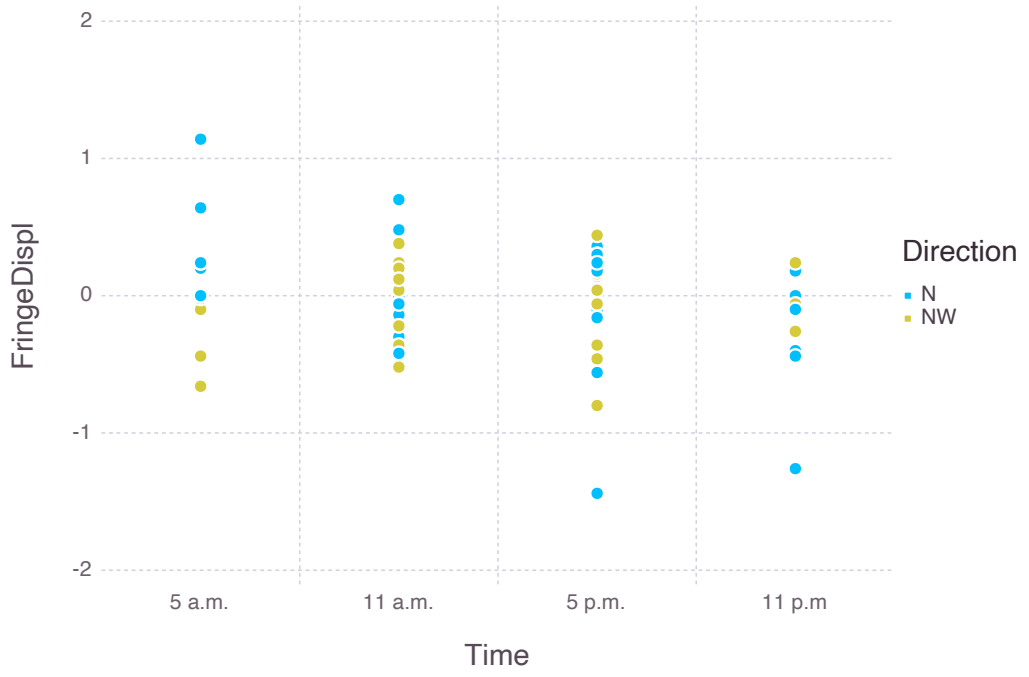

FIGURE 5.1 – Mesures du déplacement des franges d'interférence (mesurée en millième) en fonction du moment de l'expérience et de la direction du montage.

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. La loi normale, aussi appelée loi gaussienne ou loi de Laplace–Gauss, est la loi continue possédant la densité suivante sur les réels :

$$f_{(Y|\mu,\sigma^2)}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}, \quad \text{pour } y \in \mathbb{R}; \tag{5.3}$$

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

Remarque. *Dans le cadre de ce cours, la densité de la loi normale de paramètre* (μ, σ^2) *évaluée à* y *sera dénotée par*

$$\mathcal{N}(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}, \quad \textit{pour } y \in \mathbb{R};$$

afin de simplifier l'écriture mathématique.

Exercice 1: Estimation classique

Pour la loi normale lorsque σ^2 est connue, calculez l'estimateur du maximum de la vraisemblance pour μ ? Calculez également un intervalle de confiance à 95% pour μ en vous servant de la relation bien connue :

$$\frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0, 1^2).$$

Utilisez ces estimateurs sur les données obtenues du tableau 5.1. Considérez que $\sigma^2 = (3/2)^2$, la variance de l'erreur estimée par Illingworth.

5.2 Estimation bayésienne

En statistique bayésienne, le paramètre inconnu μ est considérée comme une **variable aléatoire**. Ce changement de paradigme a des répercussions majeures. Il sera désormais possible de traiter μ en utilisant la théorie des probabilités ; par exemple les densités, les densités conditionnelles, les lois marginales, etc. L'élément fondamental de l'analyse bayésienne constitue la densité conditionnelle $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu)$. Cette loi modélise l'information sur le paramètre inconnu μ après avoir considéré les données $\boldsymbol{y}$.

Remarque. *La plupart du temps, comme c'est le cas pour* μ *dans la loi normale, le paramètre inconnu est une variable continue. C'est pourquoi on utilise la notation* f *pour dénoter une densité.*

Pour obtenir cette densité conditionnelle, le théorème de Bayes dans sa version continue doit être utilisée, d'où la formulation statistique bayésienne :

$$f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu) = \frac{f_{(\boldsymbol{Y}|\mu)}(\boldsymbol{y}) \times f_{\mu}(\mu)}{\displaystyle\int_{-\infty}^{\infty} f_{(\boldsymbol{Y}|\mu)}(\boldsymbol{y}) \times f_{\mu}(\mu)\, d\mu}. \tag{5.4}$$

La densité $f_{\mu}(\mu)$ est appelée la **loi *a priori*** du paramètre μ sur l'espace paramètre $\mathbb{R}$. La loi *a priori* modélise l'information que l'on possède sur le paramètre μ avant même d'obtenir les observations. 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 $\mathbb{R}$. La loi *a posteriori* correspond à la mise à jour de cette information après avoir incorporé l'information apportée par les observations. L'information apportée par les observations est modélisée par la densité $f_{(\boldsymbol{Y}|\mu)}(\boldsymbol{y})$ appelée **vraisemblance**. Le dénominateur, parfois dénoté par $m(\boldsymbol{y})$, correspond à la loi marginale de l'échantillon aléatoire évaluée aux observations, ce qui constitue la constante de normalisation de la loi *a posteriori*.

Avant de développer davantage sur la loi *a priori*, considérons un exemple d'inférence bayésienne avec les données du tableau 5.1.

Exemple 1

Considérons le modèle de l'équation (5.2) avec le paramètre μ inconnue et la variance connue σ^2. La vraisemblance de ce modèle est donnée par

$$f_{(\boldsymbol{Y}|\mu)}(\boldsymbol{y}) = \prod_{i=1}^{n} \mathcal{N}(y_i \mid \mu, \sigma^2).$$

Considérons la loi *a priori* suivante pour μ :

$$f_{\mu}(\mu) = \mathcal{N}(\mu \mid 0, \sigma^2). \tag{5.5}$$

Nous reviendrons plus en détails sur ce choix de loi *a priori* mais on peut dire pour le moment que la moyenne 0 est choisie par les expériences précédentes et σ^2 par l'estimation de l'erreur de mesure effectuée par Illingworth.
La loi *a posteriori* de μ est calculée en utilisant le théorème de Bayes :

$$f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu) = \frac{\prod_{i=1}^{n} \mathcal{N}(y_i \mid \mu, \sigma^2) \times \mathcal{N}(\mu \mid 0, \sigma^2)}{\displaystyle\int_{-\infty}^{\infty} \prod_{i=1}^{n} \mathcal{N}(y_i \mid \mu, \sigma^2) \times \mathcal{N}(\mu \mid 0, \sigma^2)\; d\mu}.$$

Nous montrerons en classe que la loi *a posteriori* s'exprime sous la forme suivante :

$$f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu) = \mathcal{N}\left(\mu \,\middle|\, \frac{n\bar{y}}{n+1}, \frac{\sigma^2}{n+1}\right). \tag{5.6}$$

La figure 5.2 illustre la loi *a priori* et la loi *a posteriori* avec les données du tableau 5.1.

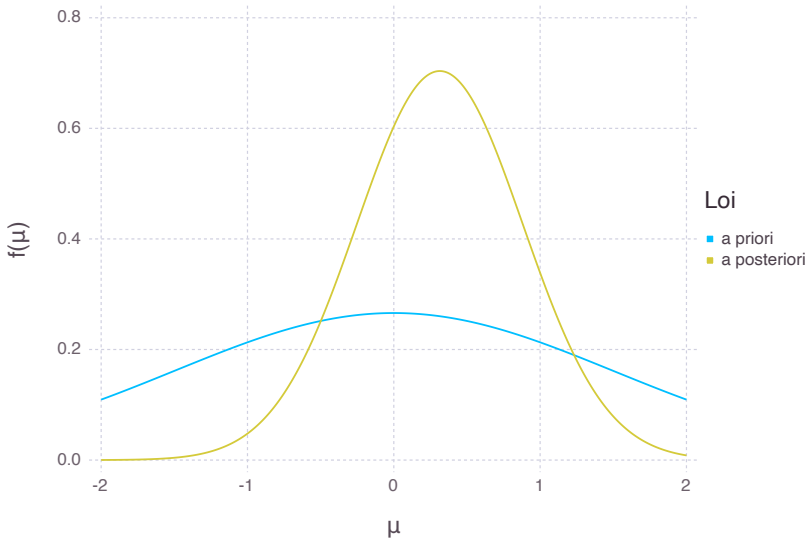

FIGURE 5.2 – Lois *a priori* et *a posteriori* de l'exemple 1

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.

On peut remarquer que la moyenne de la loi *a posteriori* (eq. 5.6) de l'exemple 1 peut s'écrire comme la moyenne pondérée des données et de la loi *a priori* :

$$\frac{n}{n+1}\,\bar{y} + \frac{1}{n+1}\,0.$$

Pour la variance de la loi *a posteriori*, le fait de considérer la loi *a priori* (5.5) est similaire à avoir une observation additionnelle qui diminue la variance de $(\mu|\boldsymbol{Y} = \boldsymbol{y})$. En effet, la variance de l'erreur est divisée par $(n+1)$.

5.3 Loi *a priori*

L'utilisation du théorème de Bayes pour l'estimation du paramètre μ requiert la définition d'une loi *a priori* pour μ. 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. Si aucune information *a priori* n'est disponible sur le paramètre avant d'effectuer l'expérience, alors une loi *a priori* non informative peut être utilisée.

Dans le cas de l'exemple 1, la loi *a priori* utilisée est informative. En effet, les valeurs de μ autour de 0 sont favorisées avant même de considérer les données.

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 l'approche pragmatique qui consiste à prendre la loi *a priori* conjuguée du modèle statistique. Les lois conjuguées font l'objet de la prochaine section.

5.3.1 Les 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'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 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

Une autre loi *a priori* informative aurait pu être utilisée dans l'exemple 1, en l'occurrence la la loi de Student à 5 degrés de liberté :

$$f_\mu(\mu) = t_5(\mu|0, \sigma),$$

où $t_\nu(y|\mu, \sigma)$ dénote la densité de la loi de Student à ν degrés de liberté, de paramètre de localisation μ et paramètre d'échelle σ. En utilisant cette loi, les valeurs de μ autour de 0 sont plus également *a priori* plus probables mais puisque la loi de Student possède des queues beaucoup plus lourdes que la loi normale, l'incertitude *a priori* sur μ est plus grande.
Si on utilise cette loi *a priori* informative, on trouve la forme suivante pour la loi *a*

posteriori en utilisant le théorème de Bayes :

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

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*.

Puisque les lois *a priori* conjuguées possède une certaine forme paramétrique, par exemple la loi normale dans le cas de ce présent chapitre, la moyenne et la variance de la loi normale *a priori* doivent être fixées avant même de considérer les données. Les paramètres de la loi *a priori* sont appelés *hyperparamètres*. Suffisamment d'information doit être disponible avant de considérer les données pour spécifier les hyperparamètres. L'information *a priori* peut être fournie par une expérience précédente ou le savoir d'un expert.

Exemple 3

Dans le cas de l'exemple 1, la loi *a priori* pour $f_\mu(\mu) = \mathcal{N}(\mu \mid 0, \sigma^2)$ a été choisie de la façon suivante.
Puisque les 9 expériences précédentes avaient obtenues un déplacement nul, on a supposé que $\mathbb{E}(\mu) = 0$. La variance sur cette valeur a été fixée en étudiant le montage expérimental : $\mathbb{V}ar(\mu) = \sigma^2$. Pour faciliter les calculs de la loi *a posteriori*, on a choisi d'encoder ces informations avec la loi normale qui est la loi conjuguée.

Dans le cas du modèle normal avec variance connue, la loi *a priori* conjuguée est la loi normale générale suivante :

$$f_\mu(\mu) = \mathcal{N}(\mu \mid \nu, \tau^2), \tag{5.7}$$

où les hyperparamètres $\nu \in \mathbb{R}$ et $\tau^2 > 0$ sont fixés avant de considérer les données. La loi *a posteriori* conjuguée correspondante est la suivante :

$$f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu) = \mathcal{N}\left\{\mu \,\middle|\, \frac{\frac{1}{\tau^2}\nu + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}, \left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)^{-1}\right\}. \tag{5.8}$$

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 [2] de l'échantillon n/σ^2 et par la précision de la loi *a priori* $1/\tau^2$.

Exercice 2

En utilisant l'équation (5.8) avec $\nu = 0$ et $\tau^2 = \sigma^2$, vérifiez que vous obtenez bien la loi *a posteriori* de l'exemple 1.

5.3.2 Les 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

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

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 4

Soit la densité *a priori* non information et impropre

$$f_\mu(\mu) \propto 1, \quad \text{pour } \mu \in \mathbb{R} \tag{5.9}$$

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 pour les données du tableau 5.1 est illustrée à la figure 5.3.

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

$$\int_{-\infty}^{\infty} f_{(\boldsymbol{Y}|\mu)}(\boldsymbol{y})\, f_\mu(\mu)\, d\mu < \infty \tag{5.10}$$

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

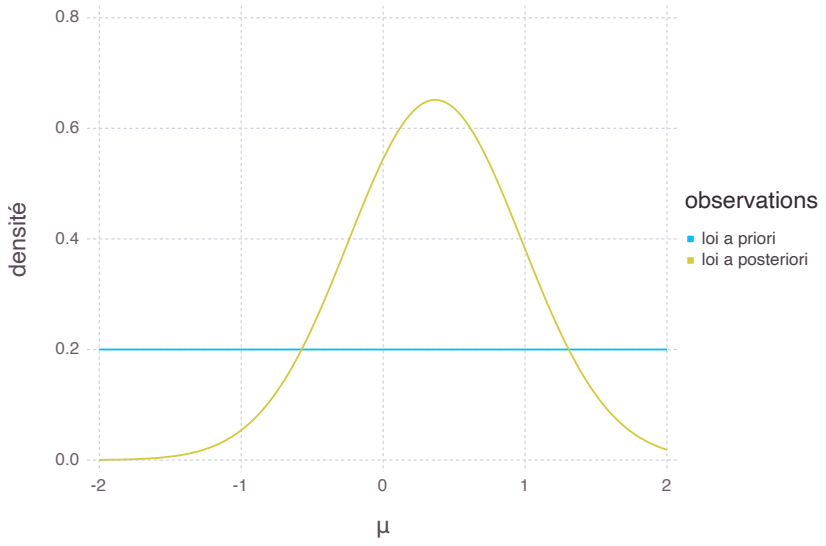

FIGURE 5.3 – Lois *a priori* et *a posteriori* de l'exemple 4.

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

Exemple 5

Pour la loi normale avec variance connue σ^2, la loi *a priori* impropre $f_\mu(\mu) \propto 1$ satisfait la condition (5.10) lorsque $n \geq 1$.

Exemple 6

La figure 5.4 illustre les lois *a posteriori* pour les données du tableau 5.1 lorsqu'une loi *a priori* conjuguée est utilisée (exemple 1) et lorsque la loi *a priori* impropre est utilisée (exemple 4). On remarque que la loi correspondante au cas informatif possède une plus petite variance et est légèrement décalée vers 0, la moyenne de la loi *a priori*.

5.3.3 La loi *a priori* : un choix subjectif

En statistique bayésienne, le choix de la loi *a priori* est une étape où la subjectivité du statisticien peut jouer un rôle, par exemple choisir cette loi *a priori* plutôt qu'une autre. Il est possible de complètement rationaliser le choix de la loi *a priori* en stipulant que l'on devrait toujours utiliser des lois non informatives. Cette approche ne permet cependant pas d'exploiter toute la puissance de la statistique bayésienne. D'une part, la loi *a priori* permet

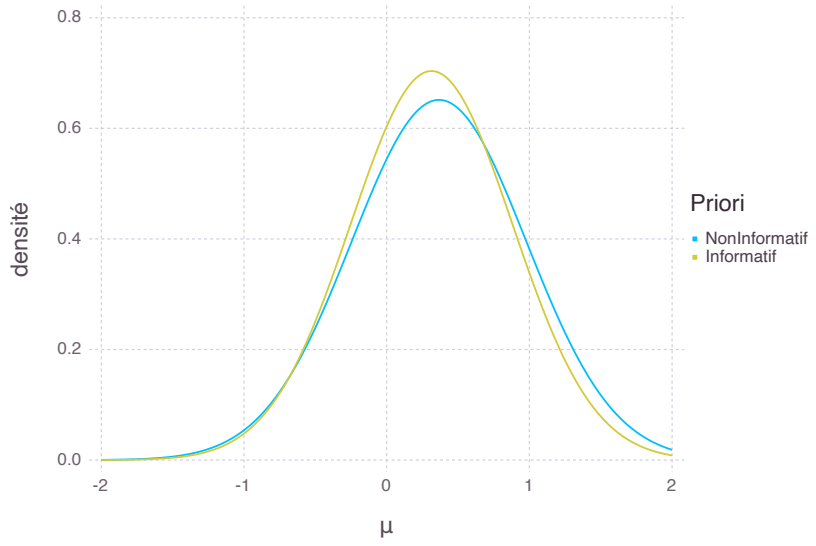

FIGURE 5.4 – Lois *a posteriori* correspondants aux exemples 1 et 4 où respectivement une loi *a priori* informative et non informative ont été utilisées.

de tirer profit de la connaissance déjà établit. D'autre part, d'un point de vue apprentissage machine, la loi *a priori* informative permet de limiter le surapprentissage des données. Nous le verrons dans les prochains chapitres. Cela étant dit, lorsqu'aucune information *a priori* n'est disponible, l'utilisation d'une loi non informative demeure pertinente.

Par ailleurs, lorsqu'une loi *a priori* informative est utilisée, son influence diminue au fur et à mesure que le nombre d'observations augmente. Par exemple, la loi *a posteriori* exprimée à l'équation (5.8) devient à toute fin pratique la densité suivante

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

lorsque n est très grand. Le caractère subjectif du choix de la loi *a priori* disparaît lorsque la taille d'échantillon augmente.

5.4 Inférence bayésienne

5.4.1 Estimations ponctuelles

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.

5.4.2 Intervalles de crédibilité

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

$$\mathbb{P}\left(\mu \in I | \boldsymbol{Y} = \boldsymbol{y}\right) = 1 - \alpha. \tag{5.11}$$

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 (5.11) est satisfaite.

Pour un niveau de confiance donnée, les deux intervalles de crédibilité concordent si la densité *a posteriori* est symétrique. Si la densité *a posteriori* est asymétrique, alors l'intervalle HPD est le plus court. La figure 5.5 illustre cette situation pour une loi asymétrique.

5.5 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

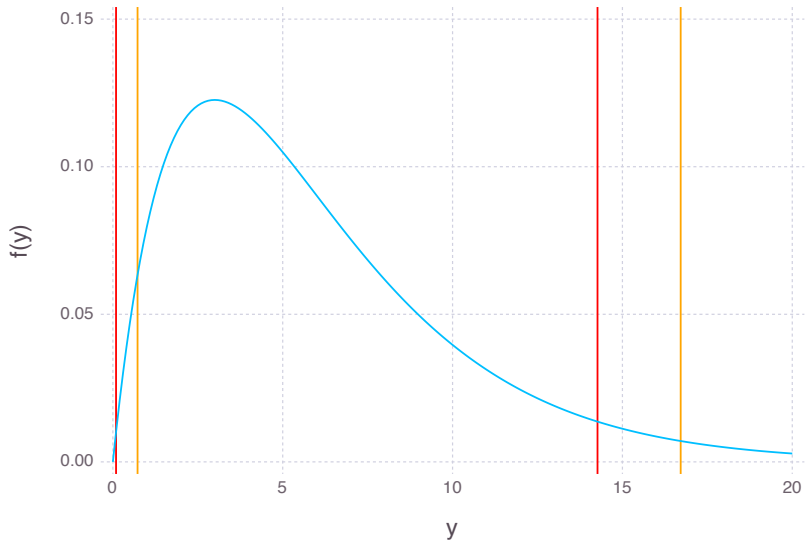

FIGURE 5.5 – Intervalles de crédibilité bayésien de niveau 95% pour une loi asymétrique. L'intervalle défini entre les lignes oranges constitue l'intervalle basé sur les quantiles tandis que l'intervalle défini entre les lignes rouges constitue l'intervalle HPD.

prédictive d'une observation future. Autrement dit, on souhaite calculer la densité d'une observation future, en sachant les observations :

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

Autrement dit, on ne conditionne pas sur les paramètres car ce sont des quantités incertaines. On intègre alors leurs incertitudes en utilisant la loi des probabilité totales :

$$f_{(Y_{n+1}|\boldsymbol{Y}=\boldsymbol{y})}(\tilde{y}) = \int_{-\infty}^{\infty} f_{(Y_{n+1}|\boldsymbol{Y}=\boldsymbol{y},\mu)}(\tilde{y}) \times f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu)\ d\mu \tag{5.12}$$

Le concept de distribution prédictive n'existe qu'en statistique bayésienne puisque l'utilisation de la loi des probabilités totales requiert que le paramètre μ soit une variable aléatoire.

Exemple 7

Quelle est la distribution prédictive d'une 7^e mesure qui serait effectuée par Illingworth à 5 a.m. dans la direction N ? En utilisant la loi *a priori* non informative exprimée à l'équation (5.9), on obtiendrait la densité prédictive suivante :

$$f_{(Y_7|\boldsymbol{Y}=\boldsymbol{y})}(\tilde{y}) = \mathcal{N}\left(\tilde{y}\,\middle|\,\bar{y}, \frac{7}{6}\sigma^2\right)$$

Nous montrerons ce résultat en classe.

5.6 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.

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

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 `illingworth1927.csv` disponible sur le site du cours. Supposez que les observations sont distribuées selon la loi $\mathcal{N}\left\{\mu, (3/2)^2\right\}$. Supposez la loi *a priori* non informative $f_\mu(\mu) \propto 1$. Si on considère les 64 observations sans les relier aux conditions expérimentales, dans quel intervalle la 65^e mesure aurait 95% de chance de se retrouver ? Vous pouvez utiliser le fait que la moyenne du déplacement des franges est de $\bar{y} = -0.0297$ pour les 64 observations.

6. 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 ?

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 $f_Y(y)$. On a que

$$I = \int_{-\infty}^{\infty} h(y) f_Y(y)\, dy. \tag{5.13}$$

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 8

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 $y^{(1)}, y^{(2)}, \ldots, y^{(m)}$ 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 9

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 $y^{(1)}, y^{(2)}, \ldots, y^{(m)}$ 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 la densité *a posteriori* de la moyenne μ de la loi normale. Toutefois, l'algorithme présenté est beaucoup plus général : il peut s'appliquer à toute densité de probabilité. Supposons que l'on ne connaît que la forme fonctionnelle $g_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu)$ de la loi *a posteriori* $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu)$ du paramètre μ :

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

avec la constante de normalisation C inconnue. Un échantillon aléatoire de la densité $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu)$ sera obtenu en n'utilisant que la forme fonctionnelle $g_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu)$. L'algorithme de Metropolis-Hastings est une procédure itérative qui fonctionne de la façon suivante. Un état initial dénoté $\mu^{(1)}$ est fixé pour μ. Un candidat est proposé pour la valeur suivante de μ. Si cette valeur est plus favorable, on accepte le candidat. Sinon, la candidat est tout de même accepté avec une certaine probabilité dépendant de $g_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu)$. 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 $f_{(\mu|\boldsymbol{Y}=\boldsymbol{y})}(\mu)$, 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 $\mu^{(t-1)}$ l'état de μ à l'itération $(t-1)$. Dénotons par $\tilde{\mu}$ le candidat proposé pour l'état à l'itération t. La marche aléatoire consiste à ajouter un pas aléatoire à l'état précédent :

$$\tilde{\mu} = \mu^{(t-1)} + \delta;$$

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 $\left\{\mu^{(t)} : t = 0, \ldots, m\right\}$ 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 $f_{(\boldsymbol{\mu}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\mu})$ 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 $\mu^{(0)}$.
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 précédent du paramètre $\mu^{(t-1)}$, générer un pas δ pour obtenir le candidat pour l'état au temps t :

$$\tilde{\mu} = \mu^{(t-1)} + \delta.$$

2. Calculer

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

3. Attribuer l'état au temps t de la façon suivante :

$$\mu^{(t)} = \begin{cases} \tilde{\mu} \text{ avec probabilité } \rho, \\ \mu^{(t-1)} \text{ avec probabilité } 1-\rho. \end{cases}$$

end for

Pour que la chaîne générée $\left\{\mu^{(t)} : t = 0, \ldots, m\right\}$ 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%.