Statistiques computationnelles

Master Mathématiques et Applications – 1ère année

Spécialité *Ingénierie Statistique et Numérique, Data Sciences*

Université de Lille

Charlotte Baey

Université de Lille

Table des matières

Première partie

Méthodes de ré-échantillonnage

Introduction

On appelle *méthodes de ré-échantillonnage* des méthodes basées sur la constitution de plusieurs nouveaux échantillons à partir de l'échantillon initial. Ces méthodes peuvent avoir plusieurs objectifs : estimer la loi d'une statistique (par exemple si on ne dispose que de résultats asymptotiques et que l'échantillon dont on dispose n'est pas suffisamment grand), effectuer des tests d'hypothèse, ou encore valider les résultats d'un modèle. Dans ce cours, nous aborderons trois techniques de ré-échantillonnage : le jackniffe (section 1), le bootstrap (section 2) et les méthodes de validation croisée (section 3).

Soit $\mathcal{X} = (X_1, \dots, X_n)$ un échantillon i.i.d. dont la loi commune a pour fonction de répartition F. Considérons maintenant un paramètre d'intérêt de la loi F, que l'on note $\theta(F)$. On va alors chercher à construire un estimateur de $\theta(F)$, que l'on notera $\hat{\theta}$, à partir d'une statistique T construite sur l'échantillon $\mathcal{X}$. Autrement dit, on a $\hat{\theta} = T(\mathcal{X})$. On peut notamment utiliser l'approche dite de *plug-in*, qui consiste à remplacer F par la fonction de répartition empirique F_n, ce qui donne $T(\mathcal{X}) = \theta(F_n)$.

Exemple 1. *Si $\theta(F) = F^{-1}(1/2)$ est la médiane de la loi F, on peut proposer l'estimateur de la médiane $\hat{\theta} = \theta(F_n) = F_n^{-1}(1/2)$, où F_n est la fonction de répartition empirique.*

Une fois que l'on a défini un estimateur pour notre paramètre d'intérêt, on va s'intéresser à la précision entourant cette estimation. C'est ainsi que l'on construit par exemple des intervalles de confiance, exacts ou asymptotiques. Or, dans bien des situations, il est difficile d'identifier la loi exacte ou asymptotique de l'estimateur que l'on a construit, ce qui rend délicate la construction d'intervalles de confiance. De même, si l'on cherche à faire des tests d'hypothèses, il peut s'avérer compliqué d'établir la loi de la statistique de test sous l'une ou l'autre des hypothèses, rendant difficile la construction d'une région de rejet ou le calcul de la puissance.

Exemple 2. *En reprenant l'exemple précédent de la médiane, sous l'hypothèse que F admet une densité de probabilité f, on a le théorème central limite suivant :*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow[n\to\infty]{loi} \mathcal{N}\left(0, \frac{1}{4f^2(\theta)}\right).$$

La variance asymptotique dépend du paramètre inconnu, et il est difficile d'obtenir une quantité pivotale à partir de cette expression.

Si on disposait de plusieurs échantillons $\mathcal{X}_1, \ldots, \mathcal{X}_m$, on pourrait estimer par exemple la variance de l'estimateur $\hat{\theta}$ en évaluant la statistique T sur chacun des échantillons : $\hat{\theta}_1 = T(\mathcal{X}_1), \ldots, \hat{\theta}_m = T(\mathcal{X}_m)$ et en approchant la variance de $\hat{\theta}$ par la variance empirique obtenue sur les différents échantillons :

$$\hat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \left(T(\mathcal{X}_i) - \frac{1}{m} \sum_{j=1}^{m} T(\mathcal{X}_j) \right)^2 .$$

Seulement, en pratique on ne dispose en général que d'un seul échantillon … et même si l'expérience peut parfois être répétée, pour que l'estimateur ci-dessus soit un bon estimateur de la variance empirique, il faut en particulier que m soit suffisamment grand, et donc il faudrait pouvoir répéter l'expérience un grand nombre de fois.

Les méthodes de ré-échantillonnage que l'on présente dans ce chapitre permettent "d'imiter" en un certain sens la situation où l'on aurait plusieurs échantillons à disposition, en construisant de tels échantillons *à partir de l'échantillon initial*.

Chapitre 1

Le jackknife

Le jackknife, qui signifie “couteau suisse” en anglais, est l’ancêtre du bootstrap. La méthode a tout d’abord été été introduite en 1949 par Quenouille pour estimer le biais d’un estimateur, puis quelques années plus tard Tukey (1958) l’a utilisée pour estimer la variance d’un estimateur. L’idée générale est très simple : à partir d’un échantillon initial $\mathcal{X} = (X_1, \ldots, X_n)$ de taille n, on construit n sous-échantillons, obtenus en écartant à chaque fois l’une des observations (voir Figure 1.1). Le i-ème échantillon est construit en supprimant la i-ème observation. On obtient alors n échantillons de taille $n-1$. Les échantillons ainsi obtenus sont appelés *échantillons jackknife*.

On note maintenant $\hat{\theta}_{(i)} = T(\mathcal{X}_{(i)})$ l’estimation de θ obtenue sur le i-ème échantillon jackknife. On parle aussi de la i-ème réplication jackknife de $\hat{\theta}$.

1.1 Estimation du biais

On s’intéresse tout d’abord à l’estimation du biais de l’estimateur $\hat{\theta}$.

Définition 1. *L’estimateur jackknife du biais est défini par :*

$$\hat{b}_{Jack} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}), \tag{1.1}$$

avec $\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)}$.

À partir des réplications jackknife de $\hat{\theta}$, on peut définir un nouvel estimateur pour θ, dans lequel on va corriger le biais.

Définition 2. *L’estimateur du jackknife (corrigé pour le biais) est défini par :*

$$\hat{\theta}_{Jack} = \hat{\theta} - \hat{b}_{Jack} = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)}. \tag{1.2}$$

En corrigeant l’estimateur initial par le biais estimé à l’aide de la méthode du jackknife, on obtient

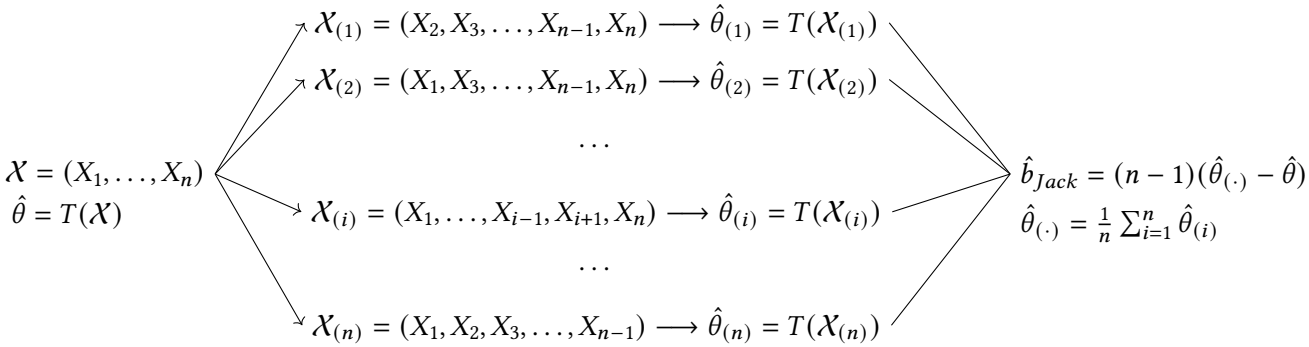


Figure 1.1 – Estimation du biais d'un estimateur $\hat{\theta}$ par la méthode du jackknife.

un estimateur avec un biais plus faible. **Attention, cela ne signifie pas que le nouvel estimateur est meilleur** ... en effet, le biais est peut-être plus faible, mais rien n'indique que l'écart-type sera aussi plus faible.

L'estimateur du jackknife peut aussi s'exprimer comme une moyenne de n termes appelés *pseudo-valeurs* du jackknife. En effet, si on définit la pseudo-valeur d'ordre i par :

$$\tilde{\theta}_{(i)} = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}, \tag{1.3}$$

on a bien $\hat{\theta}_{Jack} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\theta}_{(i)}$.

1.2 Estimation de la variance

Quelques années après Quenouille, Tukey introduit l'estimation jackknife de la variance, définie ci-dessous.

Définition 3. *L'estimateur jackknife de la variance est défini par :*

$$\hat{s}^2_{Jack} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2 \tag{1.4}$$

Comme pour l'estimateur jackknife corrigé pour le biais, on peut exprimer cet estimateur à l'aide des pseudo-valeurs. Notons $\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\theta}_{(i)}$ la moyenne empirique des pseudo-valeurs (notons que $\tilde{\theta} = \hat{\theta}_{Jack}$). On a alors :

$$\hat{s}^2_{Jack} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2. \tag{1.5}$$

Cette expression évoque la variance de la moyenne d'un échantillon i.i.d. qui serait constitué des observations $\tilde{\theta}_{(1)}, \ldots, \tilde{\theta}_{(n)}$. C'est justement l'approche de Tukey : considérer les pseudo-valeurs comme des observations i.i.d.

1.3 Conditions d'application

Une question légitime se pose après avoir défini les deux estimateurs jackknife du biais et de la variance : sous quelle(s) condition(s) les estimateurs proposés sont-ils de bons estimateurs ? Pour répondre à cette question, on peut étudier la convergence des estimateurs. Les résultats présentés ont été établis par Shao et Tu (1995). Cependant, les preuves dépassent le cadre de ce cours, et les résultats seront donc admis.

1.3.1 Estimateur jackknife du biais

On s'intéresse tout d'abord au cas où la statistique T est une fonction de la moyenne empirique, i.e. $T(\mathcal{X}) = g(\bar{X}_n)$. Pour simplifier, on suppose que $X_i \in \mathbb{R}$, mais les résultats se généralisent au cas multidimensionnel. On a le théorème suivant, dû à Shao et Tu (1995) :

THÉORÈME 1. *Soit $T(\mathcal{X}) = g(\bar{X}_n)$, avec $\mathrm{Var}(X_1) < +\infty$, g'' continue en $\mu = \mathbb{E}(X_1)$, g''' existe et est bornée au voisinage de μ, et $\mathbb{E}(|X_1^3|) < +\infty$. Alors*

$$n(\hat{b}_{Jack} - b(\hat{\theta})) \xrightarrow[n\to\infty]{p.s.} 0. \tag{1.6}$$

Autrement dit, l'estimateur jackknife du biais est un *estimateur fortement consistant* du biais de l'estimateur $\hat{\theta}$. Dans le cas plus général où T ne s'écrit pas comme une fonction de la moyenne empirique, on conserve la forte consistance du biais, mais sous des conditions plus fortes de régularité sur la fonction g.

1.3.2 Estimateur jackknife de la variance

On va également s'intéresser d'abord au cas où $T(\mathcal{X}) = g(\bar{X}_n)$. Par le théorème central limite, on sait que :

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1), \tag{1.7}$$

où $\mu = \mathbb{E}(X_1)$ et $\sigma^2 = \mathrm{Var}(X_1)$. En appliquant la méthode Delta, si g est dérivable et que sa dérivée ne s'annule pas en μ, avec $\theta = g(\mu)$, on a :

$$\sqrt{n}\,\frac{g(\bar{X}_n) - g(\mu)}{\sigma g'(\mu)} = \frac{g(\bar{X}_n) - \theta}{\sqrt{\frac{\sigma^2 (g'(\mu))^2}{n}}} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1), \tag{1.8}$$

Notons $\sigma_n^2 = \frac{\sigma^2 (g'(\mu))^2}{n}$, la variance de $T(\mathcal{X}) = g(\bar{X}_n)$. Cette variance est inconnue, et on va donc en chercher un estimateur. On peut alors légitimement se demander si l'estimateur jackknife de la variance est un bon candidat, et en particulier, si c'est un estimateur consistant de σ_n^2. La réponse est donnée par le théorème suivant, dû à Shao et Tu (1995) :

Théorème 2. *Soit $T(\mathcal{X}) = g(\bar{X}_n)$. Notons $\mu = \mathbb{E}(X_1)$ et $\sigma^2 = \mathrm{Var}(X_1)$, avec $\sigma < +\infty$. On suppose que g' est bien définie au voisinage de μ, qu'elle est continue en μ et que $g'(\mu) \neq 0$. Alors l'estimateur jackknife de la variance est fortement consistant :*

$$\frac{\hat{s}^2_{Jack}}{\sigma^2_n} \xrightarrow[n\to\infty]{p.s.} 1 \tag{1.9}$$

Que peut-on dire dans le cas plus général où la statistique T n'est pas nécessairement une fonction de la moyenne empirique ? Supposons que l'on ait $T(\mathcal{X}) = g(F_n)$ où F_n est la fonction de répartition empirique de l'échantillon. Shao et Tu (1995) montrent alors que, sous certaines conditions de régularité sur g, l'estimateur jackknife de la variance est un estimateur fortement consistant de la variance de l'estimateur. Autrement dit, le résultat du théorème 2 reste valable pour une classe plus large d'estimateurs, sous réserve que l'estimateur en question s'écrive comme une fonction *suffisamment régulière* des observations.

Que veut dire ici *suffisamment régulière* ? il s'agit principalement de notions de différentiabilité, mais celles-ci dépassent également le cadre de ce cours. On notera cependant que dans le cas où le paramètre à estimer est la médiane, et l'estimateur choisi la médiane empirique, i.e. $T(\mathcal{X}) = F_n^{-1}(1/2)$, les conditions de régularité nécessaires ne sont pas vérifiées. Dans ce cas, l'estimateur jackknife de la variance n'est pas consistant.

Chapitre 2

Le bootstrap

Le bootstrap [1] a été introduit par Efron (1979) comme une généralisation de la méthode du jackknife. Il repose sur la constitution d'*échantillons bootstrap* par ré-échantillonnage de l'échantillon initial. Les deux différences principales avec le jackknife sont les suivantes : on effectue des tirages *avec remise* de l'échantillon initial, et on constitue des échantillons de *même taille* que l'échantillon initial. Le bootstrap a connu un essor important grâce au développement des outils informatiques au début des années 1980.

2.1 Principe général

Rappelons que l'on dispose d'un échantillon initial $\mathcal{X} = (X_1, \ldots, X_n)$, i.i.d. de fonction de répartition F, et que l'on cherche à estimer un paramètre noté $\theta(F)$. On va construire B échantillons bootstrap en tirant aléatoirement et avec remise parmi les observations de $\mathcal{X}$, chaque observation ayant la même probabilité $1/n$ d'être sélectionnée (voir Figure 1.2). Pour cela, on va construire une suite d'entiers compris entre 1 et n, que l'on note $\mathbf{m}_b = \{m_{b,j}, j = 1, \ldots, n\}$, i.i.d. et indépendants de l'échantillon $\mathcal{X}$, tels que $\mathbb{P}(m_{b,j} = i) = \frac{1}{n}, \forall i = 1, \ldots, n, \forall j = 1, \ldots, n$. Cette suite d'entiers permet ensuite de construire le b-ème échantillon bootstrap, noté $\mathcal{X}_b^*$:

$$\mathcal{X}_b^* = (X_{b,1}^* = X_{m_{b,1}}, \ldots, X_{b,n}^* = X_{m_{b,n}}).$$

Exemple 3. *Supposons $n = 5$, on a donc $\mathcal{X} = (X_1, X_2, X_3, X_4, X_5)$. Si on tire la suite d'entiers suivante $\mathbf{m}_b = (5, 2, 2, 4, 1)$, alors le b-ème échantillon bootstrap sera constitué des observations suivantes : $\mathcal{X}_b^* = (X_5, X_2, X_2, X_4, X_1)$.*

1. Le terme "bootstrap" vient de l'expression américaine "to pull oneself up by one's bootstrap", difficile à traduire littéralement en français... le bootstrap est la petite boucle ou lanière que l'on trouve sur certaines bottes : l'expression fait donc référence au fait de se soulever de terre en tirant sur les boucles de ses chaussures (ce qui est impossible !). Le sens de cette expression du milieu du 19ème siècle a évolué au cours du temps, et fait plutôt référence, de nos jours, au fait de se sortir d'une situation qui semble inextricable par soi-même.

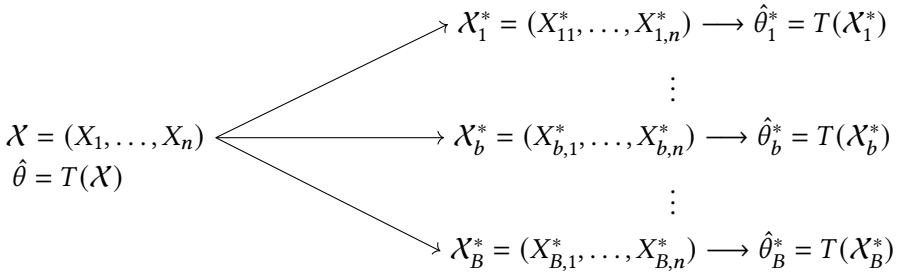


FIGURE 1.2 – Construction des échantillons bootstrap.

Par construction, certaines observations seront tirées au sort plusieurs fois pour un même échantillon bootstrap, et certaines ne le seront pas du tout pour ce même échantillon bootstrap.

Sur chacun des échantillons bootstrap, on va construire la statistique bootstrapée $\hat{\theta}_b^* = T(\mathcal{X}_b^*)$. La loi de $\hat{\theta}$, l'estimateur construit sur l'échantillon initial, dépend de F et est donc inconnue. De plus nous ne disposons que d'un échantillon et donc d'une seule observation de $\hat{\theta}$. L'objectif du bootstrap est de s'affranchir de ces difficultés. En effet, si la loi F des X_i de l'échantillon initial est *inconnue*, celle des X_{bi}^* des échantillons bootstrap est connue conditionnellement à $\mathcal{X}$: il s'agit de la loi empirique F_n. En effet, rappelons que la fonction de répartition empirique s'écrit :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq x}.$$

Il s'agit d'une fonction en escaliers, où chaque saut correspond à une observation (voir Figure 1.3). Pour le voir, considérons tout d'abord l'échantillon $\mathcal{X}$ trié par ordre croissant. On note $X_{(i)}$ la i-ème *statistique d'ordre*, c'est-à-dire la i-ème plus petite valeur. On a alors $F_n(x) = 0$ pour tout $x < X_{(1)}$, puis $F_n(x) = \frac{1}{n}$ pour $X_{(1)} \leq x < X_{(2)}$, $F_n(x) = \frac{2}{n}$ pour $X_{(2)} \leq x < X_{(3)}$, etc. La fonction F_n fait donc des sauts de taille $1/n$ à chaque nouvelle observation.

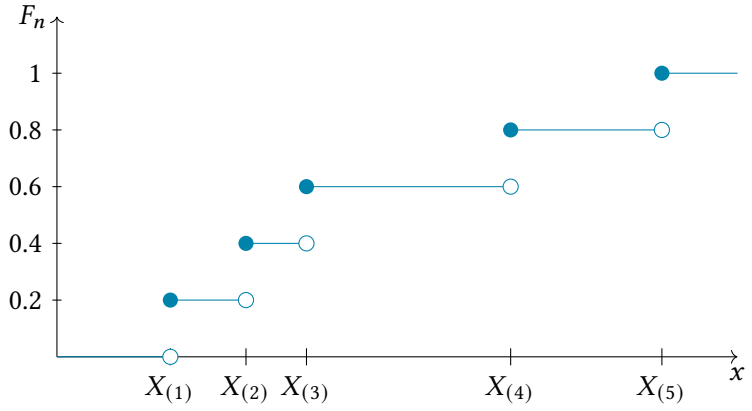


FIGURE 1.3 – Fonction de répartition empirique d'un échantillon de taille $n = 5$. $X_{(i)}$ est la i-ème statistique d'ordre de l'échantillon.

Une fonction de répartition en escaliers, dont la somme des sauts vaut 1, n'est rien d'autre que la fonction de répartition d'une loi discrète. Ici, il s'agit de la fonction de répartition de la loi uniforme

sur l'ensemble $\{X_1, \ldots, X_n\}$. Or, tirer selon la loi uniforme sur l'ensemble constitué des observations de l'échantillon initial, c'est précisément ce que l'on a fait en construisant nos échantillons bootstrap. Mais que gagne t-on en faisant cela ? L'intérêt principal c'est que l'on connaît la loi de nos échantillons bootstrap. La loi des statistiques bootstrapées $\hat{\theta}_b^*$ dépend maintenant de F_n, que l'on connaît. Cette loi est donc connue, ce qui n'était pas le cas de la loi de $\hat{\theta}$. Même si cette loi est connue en théorie, elle peut être difficile à identifier ou à calculer. Cependant, l'autre avantage du bootstrap, c'est que l'on dispose également de B observations de $\hat{\theta}^*$: on va donc pouvoir estimer sa loi, à défaut de pouvoir la calculer.

2.2 Monde réel et monde bootstrap

Dans son introduction au bootstrap, Efron propose un "monde bootstrap", qu'il conçoit comme le miroir du "monde réel" dans lequel vivent les observations de notre échantillon initial. Dans ce "monde bootstrap", toutes les lois sont connues. Le tableau 2.1 résume ces deux visions.

Table 2.1 – Comparaison des mondes "réel" et "bootstrap". Dans le monde bootstrap, on raisonne ***conditionnellement à l'échantillon initial****, et donc toutes les lois considérées sont des* ***lois sachant*** $\mathcal{X}$.

Monde réel	Monde bootstrap (**sachant** $\mathcal{X}$)
échantillon initial $\mathcal{X} = (X_1, \ldots, X_n)$	échantillon bootstrap $\mathcal{X}_b^* = (X_{b,1}^*, \ldots, X_{b,n}^*)$
X_i i.i.d. de loi inconnue F	$X_{b,i}^*$ de loi connue F_n
paramètre $\theta(F)$	paramètre $\theta(F_n) = \hat{\theta}$
estimateur $\hat{\theta} = T(\mathcal{X})$	statistique bootstrapée $\hat{\theta}_b^* = T(\mathcal{X}_b^*)$
loi de $\hat{\theta}$: G_n inconnue	loi de $\hat{\theta}^*$: G_n^* connue

2.3 Éléments de justification théorique

Rappelons que la loi de notre estimateur $\hat{\theta}$, construit sur l'échantillon initial, dépend de F et est donc inconnue. Notons G_n la fonction de répartition de $\hat{\theta}$. Dans le monde bootstrap, c'est-à-dire conditionnellement à $\mathcal{X}$, on a maintenant accès à $\hat{\theta}^*$, dont la loi conditionnellement à $\mathcal{X}$ dépend de F_n et est donc connue. Notons G_n^* la fonction de répartition de cette loi. On a :

$$G_n^*(x) = \mathbb{P}(\hat{\theta}^* \leq x \mid F_n). \tag{2.1}$$

Même si cette loi est connue, elle peut être difficile à calculer. On peut alors lui associer sa version empirique basée sur les B échantillons bootstrap :

$$\hat{G}_{n,B}^*(x) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\hat{\theta}_b^* \leq x}. \tag{2.2}$$

Ce qui nous intéresse principalement, c'est obtenir de l'information sur G, pour construire des intervalles de confiance, calculer la variance ou le biais d'un estimateur, ... la question est donc la suivante : à

quel point G_n^* est-il un bon estimateur de G_n ? et en pratique, comme on se basera souvent sur la version empirique : à quel point $\hat{G}_{n,B}^*$ est-il un bon estimateur de G_n ?

Approximation de G_n^* par $\hat{G}_{n,B}^*$.
L'approximation de G_n^* par $\hat{G}_{n,B}^*$ est contrôlée : on peut rendre ces deux fonctions aussi proches l'une de l'autre qu'on le souhaite en augmentant la valeur de B. Ce résultat est donné par le théorème de Glivenko-Cantelli, rappelé ci-dessous :

Théorème 3. *Conditionnellement à $\mathcal{X}$ (ou de façon équivalente, à F_n), on a :*

$$\sup_{x\in\mathbb{R}} \left| \hat{G}_{n,B}^*(x) - G_n^*(x) \right| \xrightarrow[B\to\infty]{p.s.} 0$$

Puis par la borne de Dvoretsky-Kiefer-Wolfowitz :

Théorème 4.

$$\forall B \geq 1, \forall y \geq 0, \mathbb{P}\left(\sqrt{B} \sup_x \left|\hat{G}_{n,B}^*(x) - G_n^*(x)\right| > y\right) \leq 2e^{-2y^2}.$$

Approximation de G_n par G_n^*.
L'approximation de la loi inconnue de l'estimateur $\hat{\theta}$ par la loi bootstrap est donnée par ce que l'on appelle les développements de Edgeworth (Hall, 2013). Dans le cas le plus simple où l'estimateur est asymptotiquement normal, i.e. s'il existe une constante σ telle que :

$$\sqrt{n}\,\frac{\hat{\theta}-\theta}{\sigma} \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0,1),$$

on a le développement d'Edgeworth suivant :

$$\mathbb{P}\left(\sqrt{n}\,\frac{\hat{\theta}-\theta}{\sigma} \leq x\right) = \Phi(x) + n^{-1/2} p_1(x) f(x) + O(n^{-1}),$$

où Φ est la fonction de répartition de la loi normale centrée réduite, p_1 est un polynôme de degré 2 et f est la densité de probabilité de la loi normale centrée réduite. Quand on approche G_n par la loi normale, on fait donc une erreur de l'ordre de $1/\sqrt{n}$.

Dans le "monde bootstrap", il existe également un développement d'Edgeworth pour la statistique bootstrapée. Si celle-ci est asymptotiquement normale, on a :

$$\mathbb{P}\left(\sqrt{n}\,\frac{\hat{\theta}^*-\hat{\theta}}{\sigma} \leq x \mid F_n\right) = \Phi(x) + n^{-1/2} \hat{p}_1(x) f(x) + O_P(n^{-1}),$$

avec $\hat{p}_1(x) - p_1(x) = O_P(n^{-1/2})$. On obtient donc :

$$\mathbb{P}\left(\sqrt{n}\,\frac{\hat{\theta}-\theta}{\sigma} \leq x\right) - \mathbb{P}\left(\sqrt{n}\,\frac{\hat{\theta}^*-\hat{\theta}}{\sigma} \leq x \mid F_n\right) = n^{-1/2}(p_1(x) - \hat{p}_1(x)) f(x) + O_P(n^{-1})$$

$$= n^{-1/2} O(n^{-1/2}) f(x) + O_P(n^{-1})$$
$$= O_P(n^{-1}).$$

Quand on approche G_n par G_n^*, on fait donc une erreur de l'ordre de $1/n$: c'est mieux que l'approximation normale !

2.4 Construction d'intervalles de confiance

2.4.1 Méthode du bootstrap classique

Pour construire un intervalle de confiance pour un paramètre $\theta(F)$, basé sur un estimateur $\hat{\theta}$, aussi noté $T(\mathcal{X})$, on peut utiliser par exemple la loi de la quantité $\hat{\theta} - \theta$. Par exemple, si on note H la fonction de répartition de $\hat{\theta} - \theta$, on peut proposer l'intervalle de confiance symétrique exact de niveau $1 - \alpha$ suivant :

$$IC(1-\alpha) = \left[\hat{\theta} - H^{-1}\left(1 - \frac{\alpha}{2}\right); \hat{\theta} - H^{-1}\left(\frac{\alpha}{2}\right)\right].$$

En effet, on a bien :

$$\begin{aligned}
\mathbb{P}\left(\hat{\theta} - H^{-1}\left(1-\frac{\alpha}{2}\right) \leq \theta \leq \hat{\theta} - H^{-1}\left(\frac{\alpha}{2}\right)\right) &= \mathbb{P}\left(-H^{-1}\left(1-\frac{\alpha}{2}\right) \leq \theta - \hat{\theta} \leq -H^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&= \mathbb{P}\left(H^{-1}\left(\frac{\alpha}{2}\right) \leq \hat{\theta} - \theta \leq H^{-1}\left(1-\frac{\alpha}{2}\right)\right) \\
&= H\left(H^{-1}\left(1-\frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1 - \alpha
\end{aligned}$$

Malheureusement, H est en général inconnue, sauf dans quelques cas particuliers (par exemple moyenne ou variance d'un échantillon gaussien, moyenne d'un échantillon de Bernoulli). On a alors recours, si elles existent, à des approximations permettant de construire des intervalles de confiance asymptotiques. C'est le cas notamment lorsqu'on utilise le théorème central limite. Avec le bootstrap, on dispose d'un nouvel outil pour construire des intervalles de confiance *non asymptotiques*. Pour cela, on va chercher à construire un estimateur de H à l'aide du bootstrap, que l'on utilisera ensuite pour construire des intervalles de confiance bootstrap.

Dans le monde bootstrap, l'équivalent de la fonction H est la fonction H^*, fonction de répartition de $\hat{\theta}^* - \hat{\theta}$ (ou $T(\mathcal{X}^*) - T(\mathcal{X})$) sachant $\mathcal{X}$. Comme dans la section précédente avec la fonction G^*, la fonction H^* peut être difficile à calculer en pratique. On va alors lui associer sa version empirique basée sur les B échantillons bootstrap. Plus précisément, à partir des B échantillons bootstrap $\mathcal{X}_1^*, \ldots, \mathcal{X}_B^*$ on va construire l'estimateur bootstrap de H^* :

$$\hat{H}_B^*(x) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{T(\mathcal{X}_b^*) - T(\mathcal{X}) \leq x} \tag{2.3}$$

$$= \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\hat{\theta}_b^* - \hat{\theta} \leq x} \tag{2.4}$$

La taille d'échantillonnage B permet de contrôler l'erreur dans l'approximation de H^* par $\hat{H}_B^*$, et on dispose des mêmes garanties que celles énoncées dans la section précédente pour l'approximation de H par H^*.

On peut alors proposer l'intervalle de confiance du bootstrap pour θ :

$$\hat{IC}(1-\alpha)^* = \left[\hat{\theta} - {H^*}^{-1}\left(1-\frac{\alpha}{2}\right); \hat{\theta} - {H^*}^{-1}\left(\frac{\alpha}{2}\right)\right], \tag{2.5}$$

que l'on peut approcher par la version empirique :

$$\hat{IC}(1-\alpha)^* = \left[\hat{\theta} - {\hat{H}_B^*}^{-1}\left(1-\frac{\alpha}{2}\right); \hat{\theta} - {\hat{H}_B^*}^{-1}\left(\frac{\alpha}{2}\right)\right] \tag{2.6}$$

Calcul des quantiles de $\hat{H}_B^*$

En pratique, il n'est pas nécessaire de calculer $\hat{H}_B^*$ et on peut ré-écrire l'intervalle de confiance bootstrap ci-dessus uniquement à l'aide des quantiles empiriques des statistiques bootstrapées. Pour le voir, remarquons que l'on a :

$$\begin{aligned} H^*(x) &= \mathbb{P}(\hat{\theta}^* - \hat{\theta} \leq x \mid F_n) \\ &= \mathbb{P}(\hat{\theta}^* \leq x + \hat{\theta} \mid F_n) \\ &= G^*(x + \hat{\theta}) \end{aligned}$$

Cherchons maintenant à exprimer ${H^*}^{-1}$ en fonction de ${G^*}^{-1}$. Soient x, y tels que $y = H^*(x) = G^*(x+\hat{\theta})$. Alors ${H^*}^{-1}(y) = x$ et ${G^*}^{-1}(y) = x + \hat{\theta}$. Donc ${H^*}^{-1}(y) = {G^*}^{-1}(y) - \hat{\theta}$.
L'intervalle de confiance bootstrap peut donc se ré-écrire :

$$\hat{IC}(1-\alpha)^* = \left[\hat{\theta} - {H^*}^{-1}\left(1-\frac{\alpha}{2}\right); \hat{\theta} - {H^*}^{-1}\left(\frac{\alpha}{2}\right)\right] \tag{2.7}$$

$$= \left[\hat{\theta} - {G^*}^{-1}\left(1-\frac{\alpha}{2}\right) + \hat{\theta}; \hat{\theta} - {G^*}^{-1}\left(\frac{\alpha}{2}\right) + \hat{\theta}\right] \tag{2.8}$$

$$= \left[2\hat{\theta} - {G^*}^{-1}\left(1-\frac{\alpha}{2}\right); 2\hat{\theta} - {G^*}^{-1}\left(\frac{\alpha}{2}\right)\right] \tag{2.9}$$

On peut alors remplacer G^* par $\hat{G}_B^*$:

$$\hat{IC}(1-\alpha)^* = \left[2\hat{\theta} - {\hat{G}_B^*}^{-1}\left(1-\frac{\alpha}{2}\right); 2\hat{\theta} - {\hat{G}_B^*}^{-1}\left(\frac{\alpha}{2}\right)\right].$$

Il s'agit alors de calculer ${\hat{G}_B^*}^{-1}$, autrement dit les quantiles de $\hat{G}_B^*$. La fonction $\hat{G}_B^*$ étant une fonction en escalier (puisqu'il s'agit d'une fonction de répartition empirique basée sur les B échantillons bootstrap), parler d'inverse n'a pas de sens. On considère alors l'*inverse généralisée* de $\hat{G}_B^*$ (que l'on notera aussi ${\hat{G}_B^*}^{-1}$ par souci de simplification des notations) :

$${\hat{G}_B^*}^{-1}(y) = \inf\{x \mid \hat{G}_B^*(x) \geq y\}$$

On a alors ${\hat{G}_B^*}^{-1}(y) = \hat{\theta}^*_{(\lceil By \rceil)}$, où $\lceil \cdot \rceil$ représente la partie entière supérieure.

Définition 4. *L'intervalle de confiance du bootstrap classique est donné par :*

$$\hat{IC}_{boot}(1-\alpha)^* = \left[2\hat{\theta} - \hat{\theta}^*_{(\lceil B(1-\frac{\alpha}{2})\rceil)}; 2\hat{\theta} - \hat{\theta}^*_{(\lceil B\frac{\alpha}{2}\rceil)}\right]. \tag{2.10}$$

En pratique, on ordonne les $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$ par ordre croissant, et on définit $\hat{\theta}^*_{\lceil B\frac{\alpha}{2}\rceil}$ comme étant la $\lceil B\frac{\alpha}{2}\rceil$-ème valeur par ordre croissant, et de même pour $\hat{\theta}^*_{\lceil B(1-\frac{\alpha}{2})\rceil}$ définie comme la $\lceil B(1-\frac{\alpha}{2}\rceil)$-ème valeur par ordre croissant.

Remarque. *On a raisonné dans cette section en partant de la loi de la quantité $\hat{\theta} - \theta$, mais on pourrait suivre le même raisonnement à partir de la loi de $\frac{\hat{\theta}}{\theta}$, si celle-ci est plus facile à obtenir (penser à l'exemple de la loi de l'estimateur de la variance dans un modèle gaussien).*

2.4.2 Méthode percentile

Supposons qu'il existe une transformation h **croissante** telle que la loi de $h(\hat{\theta})$ soit symétrique autour d'une valeur $\eta = h(\theta)$. On note $U = h(\hat{\theta})$, et $H_U(x) := \mathbb{P}(U - \eta \leq x)$. On peut alors construire l'intervalle de confiance suivant pour η :

$$IC_\eta(1-\alpha) = \left[U - H_U^{-1}\left(1-\frac{\alpha}{2}\right); U - H_U^{-1}\left(\frac{\alpha}{2}\right)\right].$$

Or, on ne connaît pas h donc on ne connaît pas non plus H_U ... on va alors, comme dans la section précédente, approcher cette fonction de répartition à l'aide des échantillons bootstrap.

Raisonnons en considérant dans un premier temps h connue. On peut alors suivre la démarche suivante :

1. construire les échantillons bootstrap $\mathcal{X}^*_1, \ldots, \mathcal{X}^*_B$
2. construire les statistiques bootstrapées $\hat{\theta}^*_1 = T(\mathcal{X}^*_1), \ldots, \hat{\theta}^*_B = T(\mathcal{X}^*_B)$ et leurs transformations par l'application h : $U^*_1 = h(T(\mathcal{X}^*_1)), \ldots, U^*_B = h(T(\mathcal{X}^*_B))$
3. définir la fonction de répartition de la statistique bootstrap :

$$H^*_U(x) = \mathbb{P}(U^*_b - U \leq x \mid F_n)$$

4. proposer l'intervalle de confiance bootstrap pour $\eta = h(\theta)$:

$$IC^*_\eta(1-\alpha) = \left[U - {H^*_U}^{-1}\left(1-\frac{\alpha}{2}\right); U - {H^*_U}^{-1}\left(\frac{\alpha}{2}\right)\right]$$

Or, on a supposé que la loi de U était symétrique par rapport à η, donc la fonction de répartition H^*_U vérifie ${H^*_U}^{-1}\left(\frac{\alpha}{2}\right) = -{H^*_U}^{-1}\left(1-\frac{\alpha}{2}\right)$ (voir Figure 1.4). On peut donc ré-écrire l'intervalle de confiance ci-dessus de la façon suivante :

$$IC^*_\eta(1-\alpha) = \left[U + {H^*_U}^{-1}\left(\frac{\alpha}{2}\right); U + {H^*_U}^{-1}\left(1-\frac{\alpha}{2}\right)\right]$$

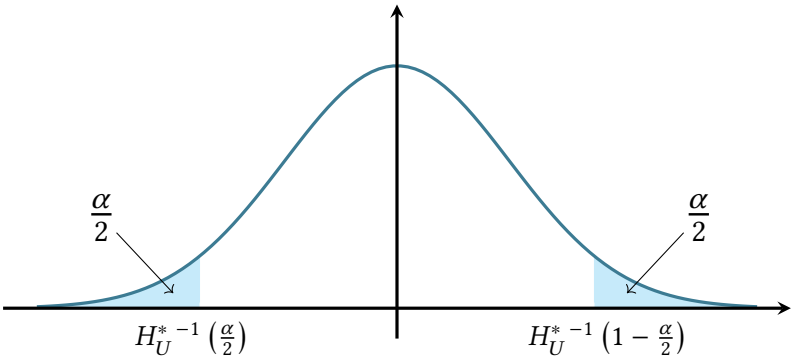


FIGURE 1.4 – Illustration de la symétrie de la loi de $U-\eta$ dans le cas où celle-ci admet une densité de probabilité. La loi de U étant symétrique par rapport à η, celle de $U-\eta$ est symétrique par rapport à 0.

TABLE 2.2 – Résumé des notations utilisées dans la section 2.4.2

Nom	Définition	Interprétation
θ		paramètre d'intérêt
η	$h(\theta)$	transformation de θ
$\hat{\theta}$	$T(\mathcal{X})$	estimation de θ sur l'échantillon $\mathcal{X}$
U	$h(\hat{\theta})$	transformation de $\hat{\theta}$
$\hat{\theta}_b^*$	$T(\mathcal{X}_b^*)$	statistique bootstrapée
U_b^*	$h(\hat{\theta}_b^*)$	transformation de la statistique bootstrapée
H_U	$H_U(x)=\mathbb{P}(U-\eta\leq x)$	f.d.r. de la transformation de l'estimateur recentrée
H_U^*	$H_U^*(x)=\mathbb{P}(U^*-U\leq x \mid F_n)$	f.d.r. de la transformée de la statistique bootstrapée recentrée
G	$G(x)=\mathbb{P}(\hat{\theta}\leq x)$	f.d.r. de l'estimateur
G^*	$G^*(x)=\mathbb{P}(\hat{\theta}^*\leq x \mid F_n)$	f.d.r. de la statistique bootstrapée
G_U^*	$G_U^*(x)=\mathbb{P}(U^*\leq x \mid F_n)$	f.d.r. de la transformée de la statistique bootstrapée
$\hat{G}_B^*$	$\hat{G}_B^*(x)=\frac{1}{B}\sum_{b=1}^{B}\mathbf{1}_{\hat{\theta}_b^*\leq x}$	estimateur bootstrap de la f.d.r. de la statistique bootstrapée

Si on veut exprimer cet intervalle de confiance à l'aide de la fonction de répartition de U^*, notée G_U^*, et non pas à l'aide de celle de U^*-U, on remarque que, comme dans la section précédente, on a $G_U^*(x)=H_U^*(x-U)$, et donc $H_U^{*\,-1}(y)=G_U^{*\,-1}(y)-U$. D'où l'intervalle de confiance bootstrap suivant pour η :

$$IC_\eta^*(1-\alpha)=\left[G_U^{*\,-1}\left(\frac{\alpha}{2}\right);G_U^{*\,-1}\left(1-\frac{\alpha}{2}\right)\right]$$

Pour obtenir un intervalle de confiance pour θ, on peut remarquer que :

$$\begin{aligned}
h(G^{*\,-1}(y)) &= h(\inf\{x \mid G^*(x)\geq y\}\\
&= \inf\{h(x) \mid G^*(x)\geq y\} \quad \text{car } h \text{ est croissante}\\
&= \inf\{h(x) \mid \mathbb{P}(\hat{\theta}^*\leq x \mid F_n)\geq y\}\\
&= \inf\{h(x) \mid \mathbb{P}(h(\hat{\theta}^*)\leq h(x) \mid F_n)\geq y\} \quad \text{car } h \text{ est croissante}
\end{aligned}$$

$$= \inf\{t \mid \mathbb{P}(U^* \leq t \mid F_n) \geq y\}$$
$$= {G_U^*}^{-1}(y)$$

On en déduit alors par croissance de la fonction h :

$$\mathbb{P}\left({G^*}^{-1}\left(\frac{\alpha}{2}\right) \leq \theta \leq {G^*}^{-1}\left(1-\frac{\alpha}{2}\right)\right) \Leftrightarrow \mathbb{P}\left({G_U^*}^{-1}\left(\frac{\alpha}{2}\right) \leq h(\theta) \leq {G_U^*}^{-1}\left(1-\frac{\alpha}{2}\right)\right)$$

Autrement dit, pour passer d'un intervalle de confiance sur η à un intervalle de confiance sur θ, il suffit de reprendre l'intervalle de confiance précédent et de remplacer ${G_U^*}^{-1}$ par ${G^*}^{-1}$:

$$IC_\theta^*(1-\alpha) = \left[{G^*}^{-1}\left(\frac{\alpha}{2}\right); {G^*}^{-1}\left(1-\frac{\alpha}{2}\right)\right].$$

Puis, comme dans les sections précédentes, on remplace ${G^*}^{-1}$ par son estimation empirique $\hat{G}_B^*$ pour obtenir l'intervalle de confiance percentile du bootstrap.

Définition 5. *L'intervalle de confiance percentile du bootstrap est donné par :*

$$\hat{IC}_{perc}^*(1-\alpha) = \left[\hat{\theta}^*_{(\lceil B\frac{\alpha}{2}\rceil)}; \hat{\theta}^*_{(\lceil B(1-\frac{\alpha}{2})\rceil)}\right]. \tag{2.11}$$

En pratique, cet intervalle de confiance est meilleur que l'intervalle de confiance du bootstrap classique, sauf si la distribution de $\hat{\theta}$ est fortement dissymétrique, ou si $\hat{\theta}$ est un estimateur biaisé de θ. Dans ce dernier cas, des corrections prenant en compte le biais existent.

2.4.3 Méthode t-percentile

La méthode du t-percentile repose sur l'existence d'une quantité pivotale (éventuellement asymptotiquement), c'est-à-dire dont la loi ne dépend d'aucune quantité inconnue.

Exemple 4. *Dans un échantillon gaussien $X_1, \ldots, X_n$ de loi $\mathcal{N}(\mu, \sigma^2)$, avec μ et σ inconnus, la quantité suivante est pivotale :*

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{s_n} \sim t_{n-1}$$

où $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Sa loi est une loi de Student à $n-1$ degrés de liberté et ne dépend ni de μ ni de σ.

Dans le cas d'un échantillon i.i.d. non gaussien, la quantité suivante est asymptotiquement normale par le théorème central limite :

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{s_n} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1)$$

On s'intéresse dans cette section aux statistiques de la forme :

$$S_n = \sqrt{n}\,\frac{\hat{\theta} - \theta}{\hat{\sigma}},$$

avec $\hat{\theta} = T(\mathcal{X})$ et $\hat{\sigma} = V(\mathcal{X})$, que l'on suppose asymptotiquement pivotale. On note J_n la fonction de répartition de S_n à n fixé. On peut alors construire l'intervalle de confiance suivant pour θ :

$$IC(1-\alpha) = \left[\hat{\theta} - \frac{\hat{\sigma}}{\sqrt{n}} J_n^{-1}\left(1-\frac{\alpha}{2}\right); \hat{\theta} - \frac{\hat{\sigma}}{\sqrt{n}} J_n^{-1}\left(\frac{\alpha}{2}\right)\right].$$

Sauf dans certains cas particuliers comme celui de l'exemple ci-dessus, J_n est inconnue (on ne connaît que la loi limite, mais pas à n fixé). On va alors construire une approximation bootstrap de J_n en procédant de la façon suivante :

1. construire les échantillons bootstrap $\mathcal{X}_1^*, \ldots, \mathcal{X}_B^*$
2. construire les statistiques bootstrapées $\hat{\theta}_1^* = T(\mathcal{X}_1^*), \ldots, \hat{\theta}_B^* = T(\mathcal{X}_B^*)$ et $\hat{\sigma}_1^* = V(\mathcal{X}_1^*), \ldots, \hat{\sigma}_B^* = V(\mathcal{X}_B^*)$
3. définir les versions bootstrapées de S_n :

$$S_b^* = \sqrt{n}\,\frac{\hat{\theta}_b^* - \hat{\theta}}{\hat{\sigma}_b^*}, \quad b = 1, \ldots, B$$

4. on estime J^*, la fonction de répartition de la statistique bootstrapée S^*, par sa version empirique :

$$\hat{J}_B^*(x) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{S_b^* \leq x}$$

5. puis on approche J_n^{-1} par l'inverse généralisée de $\hat{J}_B^*$, c'est-à-dire par les quantiles empiriques des statistiques bootstrapées

Définition 6. *L'intervalle de confiance du t-bootstrap est donné par :*

$$\hat{IC}_{t-boot}^*(1-\alpha) = \left[\hat{\theta} - \frac{\hat{\sigma}}{\sqrt{n}} S^*_{\lceil B(1-\frac{\alpha}{2})\rceil}; \hat{\theta}^* - \frac{\hat{\sigma}}{\sqrt{n}} S^*_{\lceil B(\frac{\alpha}{2})\rceil}\right]. \tag{2.12}$$

2.5 Tests bootstrap

Un test d'hypothèse statistique, qu'il soit paramétrique ou non paramétrique, repose sur le choix d'une hypothèse nulle H_0, d'une hypothèse alternative H_1 et d'une statistique de test T qui permet de mesurer, d'une certaine façon, la "distance" entre l'hypothèse nulle et les observations. On rejette typiquement H_0 lorsque cette "distance" est trop élevée, ce qui conduit à construire des régions de rejet de la forme $W(\alpha) = \{\omega \in \Omega \mid T(\mathcal{X}(\omega)) > c_\alpha\}$, avec $\mathbb{P}_{H_0}(W(\alpha)) \leq \alpha$.

Pour décider si on rejette l'hypothèse nulle à un niveau α fixé, on a alors deux possibilités : i) déterminer le seuil c_α et rejeter si $T(\mathcal{X}(\omega)) := T(x_1, \ldots, x_n)$ est supérieure au seuil, ou ii) calculer la p-value du test, i.e. la probabilité $\mathbb{P}_{H_0}(T(\mathcal{X}) > t_{obs})$, avec $t_{obs} := T(x_1, \ldots, x_n)$ où les $x_1, \ldots, x_n$ sont les réalisations observées des variables de notre échantillon ; dans ce cas on rejette H_0 si la p-value est inférieure à α. Pour calculer le seuil de rejet ou la p-value, on a besoin de connaître la loi de la

statistique de test sous H_0. Or cette loi n'est pas toujours connue, c'est le cas typiquement lorsqu'on connaît seulement une loi asymptotique de la statistique de test et que la taille de notre échantillon est trop faible.

On va alors utiliser l'approche du bootstrap pour estimer la loi de la statistique de test sous H_0. Toute la difficulté provient de la façon dont il va falloir ré-échantillonner pour se situer sous l'hypothèse nulle.

2.5.1 Tests du bootstrap paramétrique

On présente ici une méthode générale qui s'applique pour des tests paramétriques basés sur un échantillon. Autrement dit, on fait l'hypothèse que la loi de l'échantillon est connue à un (ensemble de) paramètres près, et on veut faire un test sur un (sous-ensemble des) paramètres.

Pour choisir une statistique de test appropriée dans ce contexte, on peut se reposer sur le principe de Neyman-Pearson pour deux hypothèses simples $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$, et utiliser la statistique du test du rapport de vraisemblance :

$$T(X_1, \dots, X_n) = \frac{L(X_1, \dots, X_n; \theta_1)}{L(X_1, \dots, X_n; \theta_0)}.$$

Cette approche fonctionne bien lorsque θ est le seul paramètre inconnu de la loi de l'échantillon. Or dans de nombreuses situations, il y a plusieurs paramètres inconnus, et la spécification des deux hypothèses simples $H_0 : \theta = \theta_0$ et $H_1 : \theta = \theta_1$ ne permet pas d'identifier de façon unique la loi F.

Pour illustrer ce problème et proposer une solution, considérons le cas où les X_i suivent la loi $F(\theta, \eta)$, avec θ et η inconnus, et où on s'intéresse au test $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$. Dans ce cas, le paramètre d'intérêt est θ, et η est appelé *paramètre de nuisance*. Autrement dit, on ne cherche pas à faire de l'inférence sur η, mais on a tout de même besoin d'obtenir de l'information sur ce paramètre pour pouvoir faire de l'inférence sur le paramètre d'intérêt θ. Une solution consiste alors à estimer η et à faire l'approximation : $X_i \sim F(\theta, \hat{\eta})$. De ce fait, on est ramené au cas où la loi de l'échantillon ne dépend que d'un seul paramètre inconnu. On définit alors la statistique de test :

$$\tilde{T}(X_1, \dots, X_n) = \frac{L(X_1, \dots, X_n; \theta_1, \hat{\eta}_1)}{L(X_1, \dots, X_n; \theta_0, \hat{\eta}_0)},$$

où $\hat{\eta}_0$ est l'estimateur du maximum de vraisemblance de η sous H_0 et $\hat{\eta}_1$ est l'estimateur du maximum de vraisemblance de η sous H_1.

On espère notamment que le seuil de rejet et la p-value calculés à partir de $\tilde{T}$ soit suffisamment proche du vrai seuil et de la vraie p-value.

Cette approche se généralise également à des tests de la forme $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. Dans ce cas on définit :

$$\tilde{T}(X_1, \dots, X_n) = \frac{L(X_1, \dots, X_n; \hat{\theta}, \hat{\eta}_1)}{L(X_1, \dots, X_n; \theta_0, \hat{\eta}_0)},$$

où $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ sous H_1.

Supposons donc que l'on dispose d'une statistique de test T, par exemple obtenue à l'aide du test du rapport de vraisemblance. En présence de paramètres de nuisance, l'hypothèse H_0 peut se ré-écrire

sous la forme : $H_0 : F \in \mathcal{F}_0$. Par exemple, dans le cas d'un échantillon gaussien de moyenne μ et de variance σ^2 inconnues, pour lequel on cherche à tester $H_0 : \mu = \mu_0$, $\mathcal{F}_0$ est l'ensemble des lois normales de paramètres μ_0 et de variance σ^2.

La première approximation proposée dans la section précédente consiste à considérer un estimateur $\hat{F}_0$ de la loi F sous H_0. Dans le cas gaussien évoqué ci-dessus, cela revient à considérer pour $\hat{F}_0$ la loi normale de moyenne μ_0 et de variance $\hat{\sigma}^2$, où $\hat{\sigma}^2$ est l'estimateur du maximum de vraisemblance de σ^2 lorsque $\mu = \mu_0$. On considère alors que

$$\mathbb{P}(T > t_{obs} \mid H_0) \approx \mathbb{P}(T > t_{obs} \mid \hat{F}_0).$$

En général, la loi de T sachant $\hat{F}_0$ n'est pas connue, et on va utiliser le bootstrap pour proposer un estimateur de cette loi. On procède alors de la façon suivante :

1. estimer η par maximum de vraisemblance
2. tirer B échantillons bootstrap $\mathcal{X}_1^*, \ldots, \mathcal{X}_B^*$ selon la loi $\hat{F}_0 = F(\theta, \hat{\eta})$. **Attention ici, on ne fait pas un tirage avec remise des observations initiales, on créé de nouveaux échantillons, de même taille que l'échantillon initial, en utilisant la loi $\hat{F}_0$.**
3. sur chaque échantillon bootstrap, calculer $\hat{\theta}_b^*$, $\hat{\eta}_{1,b}^*$ et $\hat{\eta}_{0,b}^*$, respectivement les estimateurs du maximum de vraisemblance de θ et η sous H_1 et de η sous H_0
4. calculer la statistique de test bootstrapée :

$$\tilde{T}_b^* = \frac{L(X_{b,1}^*, \ldots, X_{b,n}^*; \hat{\theta}_b^*, \hat{\eta}_{1,b}^*)}{L(X_{b,1}^*, \ldots, X_{b,n}^*; \theta_0, \hat{\eta}_{0,b}^*)}$$

5. estimer le seuil de rejet de niveau α par le quantile empirique d'ordre $1-\alpha$ des statistiques de test bootstrapées :

$$\hat{c}_\alpha^* = \tilde{T}^*_{\lceil B(1-\alpha) \rceil}$$

et la p-value par :

$$\hat{p}^* = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\tilde{T}_b^* > t_{obs}}$$

2.5.2 Test d'égalité de deux moyennes

Supposons que l'on dispose de deux échantillons i.i.d. $\mathcal{X} = (X_1, \ldots, X_n)$ d'espérance μ_1 et $\mathcal{Y} = (Y_1, \ldots, Y_m)$ d'espérance μ_2. On souhaite tester :

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2.$$

On considère la statistique de test suivante :

$$T(\mathcal{X}, \mathcal{Y}) = \frac{\bar{Y}_m - \bar{X}_n}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}, \tag{2.13}$$

où $s_1^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ et $s_2^2 = \frac{1}{m-1}\sum_{i=1}^m (Y_i - \bar{Y}_m)^2$. Dans le cas gaussien, la loi de T sous H_0 est connue : c'est une loi de Student. Dans le cas général, c'est-à-dire lorsque les lois des deux échantillons sont inconnues, on ne connaît pas la loi de T. On va alors utiliser le bootstrap pour l'approcher.

On procède de la façon suivante :

1. construire les échantillons centrés : $\tilde{\mathcal{X}} = (X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$ et $\tilde{\mathcal{Y}} = (Y_1 - \bar{Y}_m, \ldots, Y_m - \bar{Y}_m)$: de cette façon les deux échantillons ont la même moyenne empirique (0) et on peut donc échantillonner sous H_0
2. construire B échantillons bootstrap en tirant indépendamment un échantillon dans $\tilde{\mathcal{X}}$ et dans $\tilde{\mathcal{Y}}$. On obtient alors $\tilde{\mathcal{X}}_b^*$ et $\tilde{\mathcal{Y}}_b^*$, pour $b = 1 \ldots, B$.
3. construire les statistiques bootstrapées :
$$T_b^* = \frac{\bar{Y}_{b,m}^* - \bar{X}_{b,n}^*}{\sqrt{\frac{s_{b,1}^{*\,2}}{n} + \frac{s_{b,2}^{*\,2}}{m}}}, \tag{2.14}$$
où $s_{b,1}^{*\,2} = \frac{1}{n-1}\sum_{i=1}^n (X_{b,i}^* - \bar{X}_{b,n}^*)^2$, et de même pour $s_{b,2}^{*\,2}$
4. estimer la p-value par la proportion de statistiques bootstrapées qui dépassent la valeur observée de la statistique de test sur les échantillons initiaux :
$$\hat{p}^* = \frac{1}{B}\sum_{b=1}^B \mathbf{1}_{T_b^* \geq t_{obs}}$$

2.5.3 Test d'égalité de deux distributions

Supposons que l'on dispose de deux échantillons i.i.d. $\mathcal{X} = (X_1, \ldots, X_n)$ avec X_i de loi F_1 et $\mathcal{Y} = (Y_1, \ldots, Y_m)$ avec Y_j de loi F_2. On veut tester l'égalité des deux distributions :

$$H_0 : F_1 = F_2 \quad \text{contre} \quad H_1 : F_1 \neq F_2.$$

On suppose de plus que l'égalité des deux distributions peut se traduire par l'égalité de certains paramètres de la loi : par exemple égalité des moyennes, des variances, ... par exemple que l'on a deux échantillons issus de la même loi mais translatée ou en homothétie :

$$F_1(x) = F(x - \theta_1) \text{ et } F_2(x) = F(x - \theta_2) \quad \text{ou} \quad F_1(x) = F\left(\frac{x}{\theta_1}\right) \text{ et } F_2(x) = F\left(\frac{x}{\theta_2}\right).$$

On a alors

$$H_0 : \theta_1 = \theta_2 \quad \text{contre} \quad H_1 : \theta_1 \neq \theta_2.$$

Supposons que l'on dispose d'une statistique de test adaptée (par exemple si le paramètre θ est l'espérance de la loi, on pourra proposer $T(\mathcal{X}, \mathcal{Y}) = \bar{Y}_m - \bar{X}_n$, si θ est la variance de la loi, on pourra proposer $T(\mathcal{X}, \mathcal{Y}) = \frac{s_1^2}{s_2^2}$, où $s_1^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ et $s_2^2 = \frac{1}{m-1}\sum_{i=1}^m (Y_i - \bar{Y}_m)^2$).

Sous H_0, les deux échantillons étant issus de la même loi, on peut alors considérer l'échantillon concaténé $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ de taille $n + m$:

$$Z_1 = X_1, \ldots, Z_n = X_n, Z_{n+1} = Y_1, \ldots, Z_{n+m} = Y_m.$$

On procède ensuite de la façon suivante :

1. on construit B échantillons bootstrap $\mathcal{Z}_1^*, \ldots, \mathcal{Z}_B^*$
2. on pose $\mathcal{X}_b^* = (Z_{b,1}^*, \ldots, Z_{b,n}^*)$ et $\mathcal{Y}_b^* = (Z_{b,n+1}^*, \ldots, Z_{b,n+m}^*)$
3. on construit les statistiques bootstrapées $T_b^*(\mathcal{X}_b^*, \mathcal{Y}_b^*), b = 1 \ldots, B$
4. on estime la p-value :

$$\hat{p}^* = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{T_b^* > t_{obs}}$$

Chapitre 3

Méthodes de validation croisée

La validation croisée, que l'on appelle aussi "cross-validation", regroupe un ensemble de méthodes dont l'objectif est d'évaluer les performances d'un estimateur ou d'une procédure d'estimation sur un échantillon indépendant de celui sur lequel il a été construit. Elles peuvent également être utilisées dans un objectif de sélection de modèles. Comme on ne dispose en général que d'un seul échantillon, plusieurs approches ont été proposées : i) découper l'échantillon initial en deux parties distinctes, l'une servant à construire l'estimateur ou le modèle statistique, et l'autre servant à évaluer ses performances, et ii) découper l'échantillon en k blocs, chaque bloc jouant successivement le rôle d'échantillon indépendant sur lequel les performances de l'estimateur sont évaluées.

Pourquoi ce recours à un échantillon indépendant pour évaluer la qualité d'un estimateur ou d'un modèle ? Par construction, un estimateur est en général défini comme la solution d'un problème de minimisation ou de maximisation sur un échantillon donné (ex. : estimateur des moindres carrés, du maximum de vraisemblance, ...) : il est donc optimal pour cet échantillon précis. Mais comment être sûr que les résultats se généralisent bien, et que de bonnes performances observées ne sont pas dues à une particularité de l'échantillon sur lequel il a été construit ? pour s'en assurer, il faut disposer d'un second

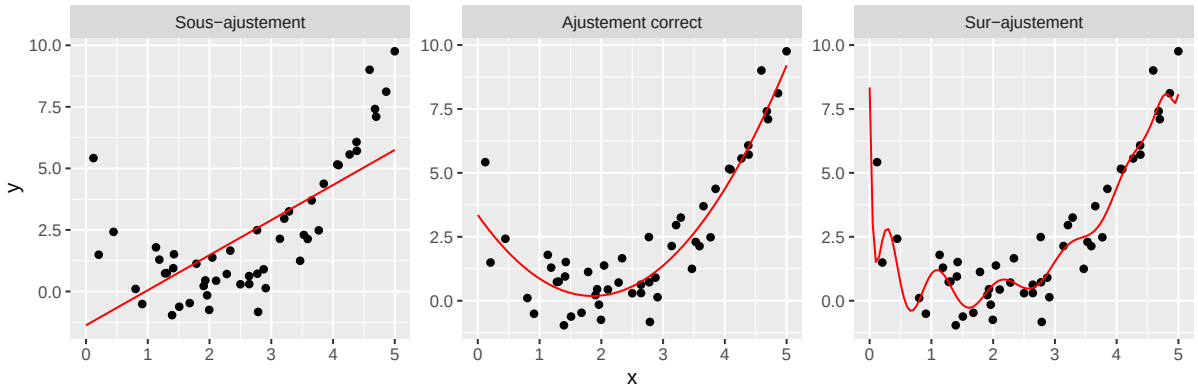


FIGURE 1.5 – Exemple de sous-ajustement (à gauche) et de sur-ajustement (à droite). Au centre, un compromis entre ces deux situations.

échantillon, sur lequel on puisse valider, confirmer, que l'estimateur construit a de bonnes propriétés en général, et pas seulement sur l'échantillon initial. Un problème courant auquel on peut être confronté est celui de *sur-ajustement* : l'estimateur ou le modèle est trop adapté aux données sur lesquelles il a été construit, et fonctionne bien sur cet échantillon, mais sera difficilement généralisable. Un exemple est donné sur la figure 1.5.

Dans la suite, on suppose que l'on dispose d'un échantillon $\mathcal{X} = (X_1, \ldots, X_n)$ i.i.d. de loi F, et que l'on souhaite estimer une quantité $\theta(F)$. On suppose que l'on dispose pour cela d'un estimateur $\hat{\theta}$ et d'un critère permettant d'évaluer les performances de l'estimateur.

3.1 Échantillon apprentissage et échantillon test

Lorsque l'échantillon initial est suffisamment grand, une première idée est de le séparer en deux : un premier sous-échantillon de taille n_a qui va constituer l'échantillon d'apprentissage, et un second sous-échantillon de taille n_t qui va constituer l'échantillon de test ou de validation, avec $n_a + n_t = n$. Ces deux sous-échantillons sont construits aléatoirement à partir de l'échantillon initial.

En pratique, il faut choisir n_a suffisamment grand obtenir un estimateur $\hat{\theta}$ ayant de bonnes propriétés, mais il faut également que n_t ne soit pas trop faible pour éviter que l'estimation des performances de l'estimateur sur l'échantillon test n'ait une variance trop élevée. Parmi les recommandations usuelles, on trouve notamment $n_a = 0.80n$ ou $n_a = 0.75n$.

Cette méthode est facile à mettre en œuvre, mais présente l'inconvénient d'être un peu instable si n_t est faible, car les résultats obtenus sur l'échantillon test peuvent alors varier significativement en fonction du découpage aléatoire base d'apprentissage / base de test. On peut alors procéder à plusieurs découpages aléatoires en base d'apprentissage et base de test, et moyenner les résultats obtenus sur les différents échantillons tests.

3.2 Validation croisée à k blocs

Lorsque l'échantillon initial ne permet pas un découpage base d'apprentissage et base de test d'effectifs suffisant, ou si l'on veut réduire l'effet du découpage aléatoire, on peut avoir recours à une approche basée sur la constitution de plusieurs blocs jouant successivement le rôle d'échantillon test. On moyenne ensuite les résultats obtenus sur les différents échantillons tests. La méthode est présentée dans l'algorithme 1. On appelle cette approche validation croisée à k blocs, ou encore en anglais "k-fold cross validation".

Les choix les plus classiques pour k se situent entre 5 et 15. En prenant k trop élevé, on augmente le temps de calcul, mais également la variance du critère calculé sur les échantillons test car la taille de l'échantillon test est plus faible. À l'inverse, un k trop faible conduit à un biais plus important de l'estimateur car la taille des échantillons d'apprentissage est alors plus faible.

Un cas particulier de la validation croisée à k blocs est celle correspondant à $k = n$, que l'on l'appelle

Algorithme 1 : Validation croisée à k blocs

1 Découper l'échantillon en k blocs de taille similaire (de même taille si k divise n ...)
2 **pour** $i = 1, \ldots, k$ **faire**
3 Construire l'estimateur sur l'échantillon constitué de tous les blocs sauf le i-ème
4 Calculer le critère d'évaluation de l'estimateur sur l'échantillon test constitué du i-ème bloc
5 Calculer la moyenne des k critères d'évaluation

plus communément "leave-one-out cross validation". Elle consiste à écarter à chaque fois une seule observation qui joue alors le rôle d'échantillon test. On peut y trouver une similitude avec la jackknife, qui construisait des sous-échantillons de taille $n - 1$ en mettant de côté successivement chaque observation. Cependant la différence principale réside dans le fait qu'avec le jackknife, la quantité d'intérêt est calculée sur l'échantillon de taille $n - 1$ privé de la i-ème observation, alors que dans la validation croisée "leave-one-out", c'est justement sur cette i-ème observation qui joue le rôle d'échantillon test, que l'on calcule la quantité qui nous intéresse.

Deuxième partie

Méthodes de Monte Carlo

Introduction

En statistique, on est souvent confronté au besoin de calculer des quantités qui s'expriment comme des intégrales : moment d'ordre p, risque quadratique, fonction de vraisemblance dans les modèles à variables latentes (voir partie IV) ou estimateur bayésien (voir partie III), ...

Malheureusement, ces quantités ne sont pas toujours calculables analytiquement, et il est alors nécessaire de se tourner vers des méthodes numériques pour les approcher. Il existe des méthodes déterministes d'approximation d'intégrales, comme par exemple la méthode des trapèzes, ou les méthodes de quadrature. Cependant, ces approches se révèlent souvent peu efficaces en grande dimension, ou lorsque les bornes d'intégration sont infinies. Une alternative à ces méthodes est celle de l'approximation de type Monte Carlo, qui sont particulièrement efficaces en grande dimension.

Une première apparition de l'approche Monte Carlo peut être trouvée à la fin du 18ème siècle, dans ce que l'on appelle le problème de "l'aiguille de Buffon", dont l'objectif est de fournir une approximation du nombre π. La formulation du problème est la suivante : sur un sol composé de lames de parquet parallèles et de même largeur, quelle est la probabilité p qu'une aiguille lancée par terre se retrouve à cheval entre deux lames ? La réponse théorique est $p = \frac{2\,l}{\pi\,t}$, où l est la longueur de l'aiguille et t la largeur des lames. L'expérience pratique consiste alors à lancer un grand nombre d'aiguilles sur le sol, et à calculer la proportion d'aiguilles qui se trouvent entre deux lames. Cette proportion empirique fournit une bonne approximation de la valeur de p.

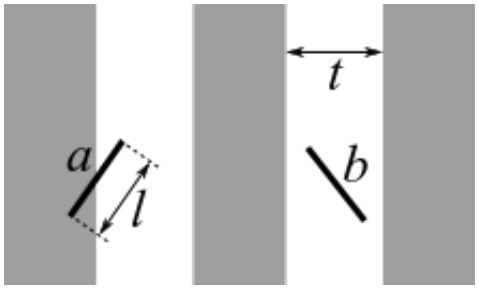


Figure 2.1 – Illustration du problème de l'aiguille de Buffon

Les simulations de type Monte Carlo telles qu'on les connaît et telles qu'on les utilise aujourd'hui ont elles été introduites au milieu des années 1940 par des chercheurs qui travaillaient sur la mise au point de la bombe atomique au laboratoire américain de Los Alamos. Les calculs devaient permettre la simulation de réactions thermonucléaires. Le principe de base repose sur l'utilisation intensive de simulations de quantités aléatoires pour calculer (ou approcher) des quantités déterministes. Le terme "Monte Carlo" a été choisi comme nom de code pour désigner ces travaux classés secret défense, et seraient une référence à la (mauvaise) habitude de l'oncle d'un des chercheurs, qui se rendait souvent

au casino de Monte Carlo.

Le contexte général des méthodes que nous verrons ici est celui de l'estimation d'une intégrale qui se présente sous la forme suivante :

$$\mathbb{E}_f(h(X)) = \int h(x)f(x)dx,$$

où f est la densité de probabilité de la variable aléatoire X.

Chapitre 1

Méthodes de Monte Carlo classiques

Avant de présenter plus formellement les méthodes de Monte Carlo classiques (section 1.2) et l'échantillonnage préférentiel (section 1.3), nous présentons quelques méthodes pour générer des variables aléatoires de loi donnée. En effet, la simulation de variables aléatoires est à la base de toutes les méthodes de Monte Carlo, il est donc important de passer en revue les approches les plus courantes pour la génération de variables aléatoires.

1.1 Génération de variables aléatoires

La plupart des distributions de probabilités standard sont facilement simulables sur les logiciels statistiques ou langages de programmations usuels comme R, Python ou Matlab. Dans ces cas-là, mieux vaut s'en remettre aux fonctions déjà existantes ! À titre d'exemple, voici comment simuler un échantillon de taille $n = 10$ d'une loi exponentielle de paramètre $\lambda = 5$, avec la paramétrisation suivante pour la densité de la loi exponentielle : $f(x) = \lambda e^{-\lambda x}$:

- sous R : `rexp(n=10, rate=5)`
- sous Python : `expon.rvs(scale=5, size=10)` (après avoir fait `from scipy.stats import expon`)
- sous Matlab : `exprnd(mu=1/5,10,1)`

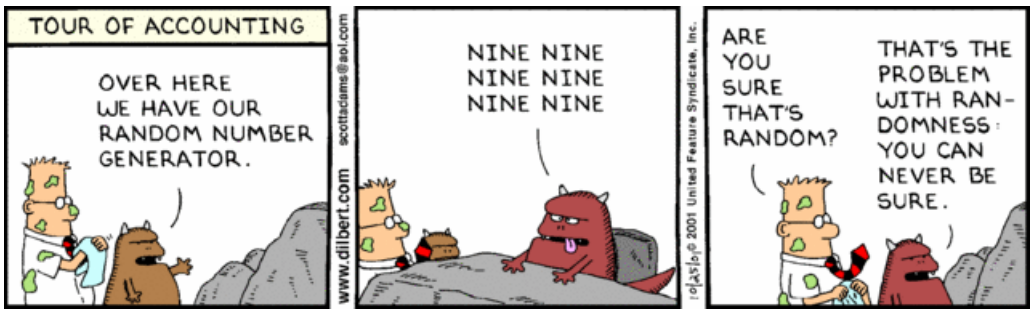


FIGURE 2.2 – Random number generation (Source : Scott Adams)

Pour les distributions non standards, nous présentons deux méthodes de génération de variables aléatoires : la méthode de la fonction inverse et celle de l'acceptation-rejet.

1.1.1 Méthode de la fonction inverse

Cette méthode repose sur le résultat suivant, qui permet d'exprimer n'importe quelle variable aléatoire X à l'aide d'une variable aléatoire de loi uniforme sur l'intervalle $[0, 1]$:

PROPOSITION 1. *Soit F une fonction de répartition, et F^- son inverse généralisée. Soit U une variable aléatoire de loi uniforme sur l'intervalle $[0, 1]$, i.e. $U \sim \mathcal{U}([0, 1])$. Alors $X := F^-(U)$ est une variable aléatoire de fonction de répartition F.*

Démonstration. Notons tout d'abord que si F admet une inverse au sens classique, la preuve est immédiate. En effet, on a alors pour $x \in \mathbb{R}$:

$$\begin{aligned}
\mathbb{P}(X \leq x) &= \mathbb{P}(F^{-1}(U) \leq x) \\
&= \mathbb{P}(F(F^{-1}(U)) \leq F(x)) \quad \text{car } F \text{ est croissante} \\
&= \mathbb{P}(U \leq F(x)) \\
&= F(x),
\end{aligned}$$

par définition de la fonction de répartition de la loi $\mathcal{U}([0, 1])$.

Considérons maintenant le cas où F admet une inverse généralisée, i.e. $F^-(u) = \inf\{x \mid F(x) \geq u\}$. Soient $x \in \mathbb{R}$ et $u \in [0, 1]$. Montrons que $F^-(u) \leq x \Leftrightarrow u \leq F(x)$.

- supposons d'abord que $F^-(u) \leq x$. Par croissance de F, on a $F(F^-(u)) \leq F(x)$. Or on a :

  $$F(F^-(u)) = F(\inf\{x \mid F(x) \geq u\}) \geq u$$

  d'où $u \leq F(x)$.

- supposons maintenant que $u \leq F(x)$. L'inverse généralisé F^- est également une fonction croissance. On a donc $F^-(u) \leq F^-(F(x))$. Or :

  $$F^-(F(x)) = \inf\{y \mid F(y) \geq F(x)\} \leq x,$$

  d'où $F^-(u) \leq x$.

On en déduit :

$$\begin{aligned}
\mathbb{P}(X \leq x) &= \mathbb{P}(F^-(U) \leq x) \\
&= \mathbb{P}(U \leq F(x)) \\
&= F(x)
\end{aligned}$$

□

D'après ce résultat, on peut donc simuler selon n'importe quelle loi de probabilités, en générant tout d'abord un échantillon selon la loi uniforme sur l'intervalle $[0, 1]$, puis en appliquant la transformation $X_i = F^-(U_i)$ à chaque variable U_i générée.

Exemple 5. *Soit X une variable aléatoire suivant la loi de Cauchy. Sa densité est donnée par :*

$$f(x) = \frac{1}{\pi(1+x^2)}$$

et sa fonction de répartition par :

$$F(x) = \frac{1}{\pi}\arctan x + \frac{1}{2}.$$

F étant bijective elle admet une inverse, définie par :

$$F^{-1}(u) = \tan\left(\pi\left(u - \frac{1}{2}\right)\right).$$

Pour générer un échantillon de loi de Cauchy, on peut donc générer un échantillon de loi uniforme sur $[0, 1]$ et appliquer à chaque valeur la transformation F^{-1} ci-dessus. Le code R ci-dessous permet de construire un tel échantillon, dont la distribution empirique est donnée sur la figure 2.3.

```
> u <- runif(1000)
> finv <- function(u){tan(pi*(u-0.5))}
> x <- finv(u)
```

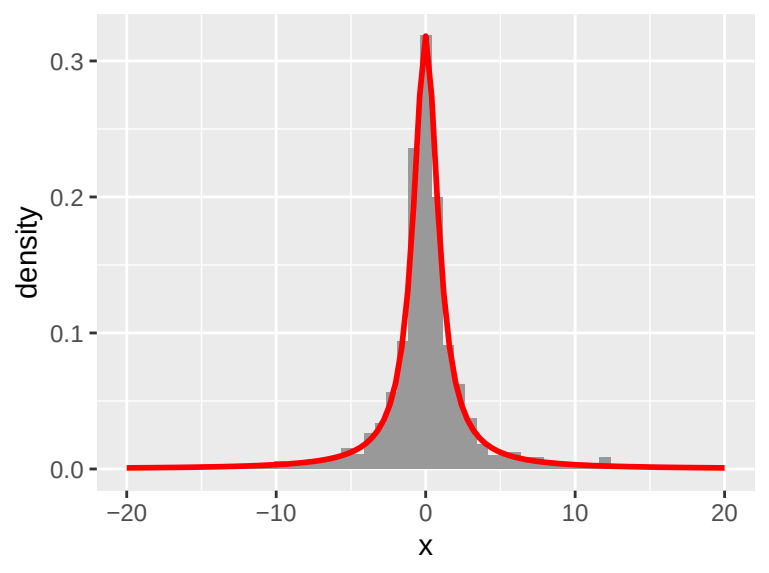


Figure 2.3 – Histogramme d'un échantillon généré par la méthode de la fonction inverse selon la loi de Cauchy, et densité de la loi de Cauchy (en rouge).

Cette méthode fonctionne très bien mais elle requiert le calcul de l'inverse de la fonction de répartition ... ce calcul n'est pas toujours aisé ! Dans certains cas, l'expression analytique de la fonction de

répartition n'est même pas connue (penser à la loi normale ...).

1.1.2 Acceptation-rejet

Les méthodes de type acceptation-rejet permettent de traiter les cas où la loi selon laquelle on veut simuler admet une densité dont on connaît l'expression à une constante multiplicative près. La proposition suivante précise ce résultat dans le cas où on peut majorer la densité de probabilité cible par une constante fois une autre densité de probabilité selon laquelle il est plus facile de simuler.

PROPOSITION 2. *Soit X une variable aléatoire de densité de probabilité f, et soit g une densité de probabilité et M une constante supérieure ou égale à 1 telles que $f(x) \leq Mg(x)$ pour tout x. Alors, pour simuler X selon la loi f, il suffit de :*

1. *simuler Y selon la loi g,*
2. *simuler $U \mid Y = y$ selon la loi uniforme $\mathcal{U}([0, Mg(y)])$,*
3. *poser $X = Y$ si $0 < U < f(Y)$ (on accepte la valeur simulée) et reprendre l'étape 1. sinon (on rejette la valeur simulée)*

Le fait que $f(x) \leq Mg(x)$ pour tout x implique en particulier que le support de f doit être inclus dans le support de g.

Démonstration. Supposons pour simplifier que g est à support sur $\mathbb{R}$. Tout d'abord, on note que la loi jointe du couple (Y, U) s'écrit :

$$f_{(Y,U)}(y,u) = g(y)\frac{1}{Mg(y)}\mathbf{1}_{0\leq u\leq Mg(y)} = \frac{1}{M}\mathbf{1}_{0\leq u\leq Mg(y)}.$$

On a alors :

$$\begin{aligned}
\mathbb{P}(X \leq x) &= \mathbb{P}(Y \leq x \mid U < f(Y)) \\
&= \frac{\mathbb{P}(Y \leq x, U < f(Y))}{\mathbb{P}(U < f(Y))} \\
&= \frac{\int_{-\infty}^{x}\int_{0}^{f(y)} \frac{1}{M} du dy}{\int_{\mathbb{R}}\int_{0}^{f(y)} \frac{1}{M} du dy} \\
&= \frac{\int_{-\infty}^{x} f(y) dy}{\int_{\mathbb{R}} f(y) dy} \\
&= \int_{-\infty}^{x} f(y) dy
\end{aligned}$$

f est donc bien la densité de probabilité de X. □

Le choix de g n'est pas anodin. Si en théorie, la méthode fonctionne quel que soit le choix de g, les temps de calcul peuvent s'allonger si g est "mal choisie". En effet, on peut remarquer que le taux

d'acceptation, c'est-à-dire le nombre de valeurs simulées qui sont finalement acceptées pour former l'échantillon tiré selon f, est donné par :

$$\mathbb{P}(0 < U < f(Y)) = \int_{\mathbb{R}} \int_0^{f(y)} \frac{1}{M} du dy = \frac{1}{M}$$

Autrement dit, on a intérêt à choisir M le plus proche possible de 1, c'est-à-dire g le plus proche possible de f, sinon une grande partie des simulations sont perdues : en effet, en moyenne il faudra simuler M réalisations (Y, U) pour produire une réalisation de X.

Un cas particulier de la méthode d'acceptation-rejet concerne le cas où la densité cible f a un support compact et est bornée par une constante m. Dans ce cas, il suffit de simuler Y selon une loi uniforme sur le support de la loi f, puis de simuler U conditionnellement à l'évènement $\{Y = y\}$ selon une loi uniforme sur l'intervalle $[0, m]$ (autrement dit, indépendamment de la valeur de Y). On conserve ensuite les valeurs de Y simulées telles que $0 < U < f(Y)$.

Exemple 6. *On considère la densité* $f(x) = \frac{1}{8}|x^4 - 5x^2 + 4|\mathbf{1}_{[-2,2]}(x)$. *On a* $f(x) \leq 1/2$, *on peut donc simuler* Y *selon la loi uniforme sur* $[-2, 2]$, *puis simuler* U *selon la loi uniforme sur* $[0, 1/2]$, *et garder les valeurs de* Y *telles que* $0 < U < f(Y)$. *Une illustration de la méthode est présentée sur la figure 2.4, où 1000 couples* (Y, U) *ont été générés. Au final, seuls 524 points ont été acceptés, et notre échantillon de loi* f *ne comporte donc que 524 points. Les autres sont "perdus", ce qui représente ici une perte d'environ la moitié de l'effort de simulation.*
En reprenant les notations de la proposition 2, on a $g(x) = \frac{1}{4}\mathbf{1}_{[-2,2]}(x)$ *et* $f(x) \leq 2g(x)$. *On a donc bien un taux d'acceptation de* $1/M$ *soit 50% seulement des points acceptés.*

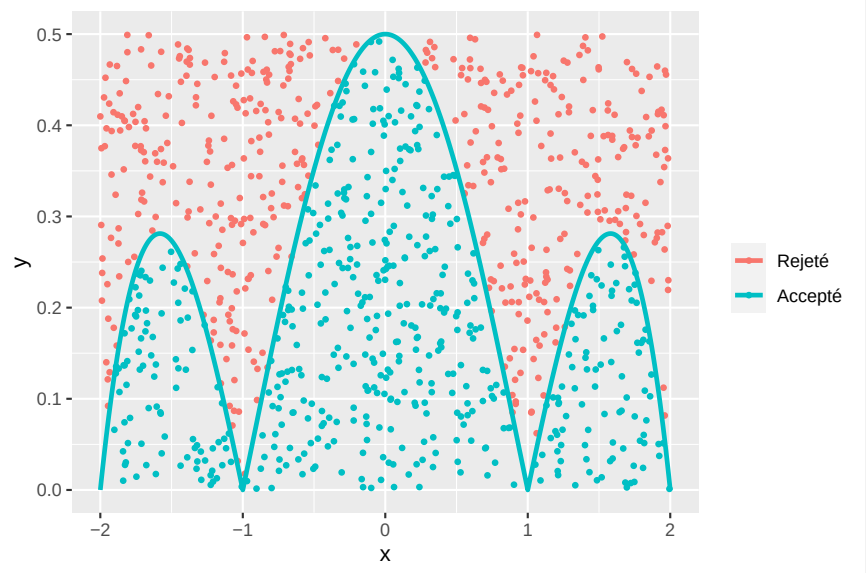


Figure 2.4 – Simulation selon la méthode d'acceptation-rejet : en bleu, les points acceptés par l'algorithme comme étant générés par la loi f, *en rouge les points rejetés. La densité* f *est tracée en bleu.*

Finalement, la figure 2.5 représente l'histogramme de l'échantillon formé par les points acceptés, pour une taille initiale de 10000 et une taille effective (nombre de points acceptés) de 5067. On a tracé la fonction f pour comparaison.

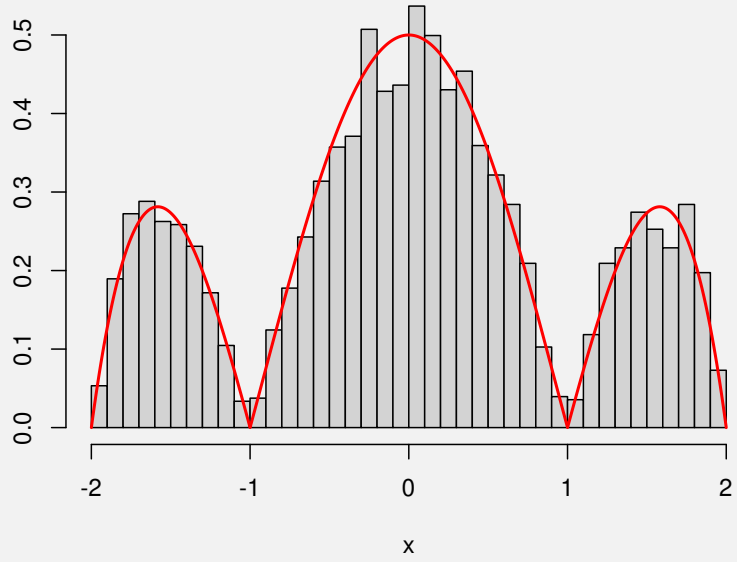


FIGURE 2.5 – Simulation selon la méthode d'acceptation-rejet : en bleu, les points acceptés par l'algorithme comme étant générés par la loi f, en rouge les points rejetés. La densité f est tracée en bleu.

1.2 Monte Carlo classique

Dans cette sextion, on va s'intéresser aux méthodes de Monte Carlo dont l'objectif est d'approcher une intégrale uni ou multi dimensionnelle. Rappelons que l'on cherche à évaluer une intégrale de la forme :

$$\mathbb{E}_f(h(X)) = \int h(x)f(x)dx, \tag{1.1}$$

où f est la densité de probabilité de la variable aléatoire X. Le principe de la méthode de Monte-Carlo classique est simple, et repose sur la loi forte des grands nombres.

DÉFINITION 7 (Monte Carlo classique). *Soit $X_1, \ldots, X_n$ un échantillon simulé de n variables aléatoires i.i.d. de loi f. L'estimateur de Monte Carlo de $\mathbb{E}(h(X))$, est défini par :*

$$\hat{h}_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \tag{1.2}$$

Biais de l'estimateur. Par linéarité de l'espérance, l'estimateur de Monte Carlo est sans biais.

Consistance. Par la loi forte des grands nombres, il est également fortement consistant.

Variance. La précision de l'estimateur peut être choisie aussi fine que l'on souhaite, en augmentant le nombre de simulations, et la vitesse de convergence peut être obtenue à l'aide du théorème central limite, sous l'hypothèse supplémentaire $\mathbb{E}_f(h^2(X)) < \infty$. En effet, dans ce cas on a, en notant $\sigma^2 := \mathrm{Var}(h(X))$:

$$\mathrm{Var}(\hat{h}_n) = \frac{\sigma^2}{n}, \quad \text{avec} \quad \sigma^2 = \int \left(h(x) - \mathbb{E}_f(h(X))\right)^2 f(x)dx,$$

que l'on peut estimer par l'estimateur fortement consistant suivant :

$$v_n = \frac{1}{n}\sum_{i=1}^{n}\left(h(X_i) - \hat{h}_n\right)^2.$$

On peut alors appliquer le théorème central limite et le lemme de Slutsky à la suite de variables aléatoires i.i.d. $h(X_1), \ldots, h(X_n)$:

$$\sqrt{n}\,\frac{\hat{h}_n - \mathbb{E}_f(h(X))}{\sqrt{v_n}} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1)$$

Quelques remarques à ce stade :

- la vitesse de convergence de la méthode de Monte Carlo est de l'ordre de $1/\sqrt{n}$, ce qui est relativement lent par rapport à certaines approches numériques et déterministes en dimension 1 ou 2 (comme la méthode des trapèzes par exemple)
- cependant, la force de ce résultat est qu'il se généralise en dimension supérieure à 1, en appliquant le TCL multidimensionnel : *la vitesse de convergence ne dépend donc pas de la dimension*, contrairement à bon nombre de méthodes déterministes qui deviennent rapidement infaisables en grande dimension
- un autre avantage de la méthode de Monte Carlo est qu'elle ne dépend pas de la régularité de la fonction h, ce qui en fait une méthode de choix lorsque cette fonction est peu régulière (elle doit cependant être L^1)

Grâce au résultat précédent, on peut alors construire un intervalle de confiance pour l'estimation de $\mathbb{E}_f(h(X))$ par $\hat{h}_n$.

Cette méthode est relativement simple à mettre en place si on sait simuler des variables aléatoires selon la loi f. Elle permet d'approcher toute quantité qui s'exprime comme une intégrale, et donc permet en particulier d'approcher le niveau ou la puissance d'un test, de même que sa p-valeur. Ceci peut s'avérer utile lorsque la loi de la statistique de test n'est pas connue ou seulement asymptotiquement.

Exemple 7. *Soit $a > 0$ et X une variable aléatoire de densité de probabilité f. Supposons que l'on souhaite estimer $p = \mathbb{P}(X > a)$. On peut alors générer n variables aléatoires i.i.d. $X_1, \ldots, X_n$ de loi f et estimer p par :*

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{X_i > a}.$$

En effet, par la loi forte des grands nombres, comme les $\mathbf{1}_{X_1>a}, \ldots, \mathbf{1}_{X_n>a}$ sont i.i.d. (de loi de Bernoulli), on a :

$$\hat{p} \xrightarrow[n\to\infty]{} \mathbb{E}(\mathbf{1}_{X_1>a}) = \mathbb{P}(X_1 > a).$$

De plus :

$$v_n = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{1}_{X_i>a} - \hat{p}\right)^2 = \hat{p}(1-\hat{p})$$

Par le théorème central limite, on a donc :

$$\sqrt{n}\,\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1)$$

On en déduit l'intervalle de confiance de niveau $1-\alpha$ suivant pour p :

$$IC(1-\alpha) = \left[\hat{p} - q^{\mathcal{N}(0,1)}_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + q^{\mathcal{N}(0,1)}_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

La différence principale avec un intervalle de confiance basé sur un échantillon de taille n fixé, c'est qu'ici on peut augmenter n autant que l'on veut, ce qui n'est pas le cas d'un échantillon d'observations pour lequel n est en général fixé. On peut donc, avec les méthodes de Monte Carlo, se fixer une précision donnée et en déduire la taille d'échantillonnage nécessaire pour atteindre cette précision.

Exemple 7 (Suite). *Reprenons l'exemple précédent, et supposons que f est la densité d'une loi Gaussienne standard. On cherche à estimer $p = \mathbb{P}(X > 1.96)$. Bien sûr, cette quantité est connue car tabulée (et vaut 0.025 ...). Voyons comment se comporte l'approximation de Monte Carlo dans ce cas.*

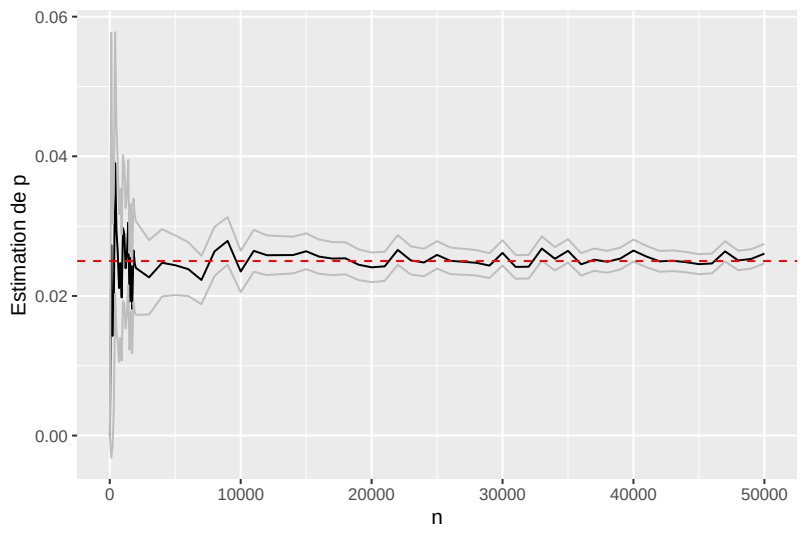


Figure 2.6 – Estimation de $\mathbb{P}(X > 1.96)$ pour X une variable aléatoire Gaussienne standard. En noir, l'estimation de Monte Carlo, en gris les bornes de l'intervalle de confiance, et en rouge la valeur théorique.

La méthode de Monte Carlo classique fonctionne bien mais peut nécessiter une taille d'échantillon importante pour converger. Or, la variance de l'estimateur ne dépend pas seulement de n, mais également de σ, la variance de $h(X)$, et plus précisément de v_n, l'estimateur de σ^2. Dans la section suivante, on va s'intéresser à des approches permettant de diminuer la variance de l'estimateur, permettant ainsi d'obtenir une précision supérieure à celle de l'approche Monte Carlo classique, pour une même taille d'échantillonnage n fixée.

1.3 Échantillonnage préférentiel

1.3.1 Présentation générale

L'idée principale de l'échantillonnage préférentiel ("importance sampling" en anglais) est de simuler non pas directement par rapport à la loi cible f, mais selon une loi *instrumentale* g. En effet, en optimisant le choix de cette distribution on peut améliorer les performances des estimations. C'est le cas notamment lorsqu'on cherche à estimer des probabilités de queue ou d'évènements rares.

Soit g une densité de probabilité dont le support inclus celui de f, c'est-à-dire telle que $f(x) > 0 \Rightarrow g(x) > 0$. Alors on peut ré-écrire (1.1) sous la forme :

$$\mathbb{E}_f(h(X)) = \int h(x)f(x)dx = \int h(x)\,\frac{f(x)}{g(x)}\,g(x)dx = \mathbb{E}_g\left(h(Z)\frac{f(Z)}{g(Z)}\right), \tag{1.3}$$

où Z est une variable aléatoire de loi g. On a adopté ici les notations $\mathbb{E}_f$ et $\mathbb{E}_g$ pour indiquer la loi par rapport à laquelle on calcule l'espérance.

Définition 8 (Échantillonnage préférentiel). *Soit $Z_1, \dots, Z_n$ un échantillon simulé de n variables aléatoires i.i.d. de loi g. L'estimateur par échantillonnage préférentiel de $\mathbb{E}_f(h(X))$, est défini par :*

$$\tilde{h}_n = \frac{1}{n}\sum_{i=1}^{n} h(Z_i)\,\frac{f(Z_i)}{g(Z_i)}. \tag{1.4}$$

Biais. De même que l'estimateur de Monte Carlo classique, l'estimateur par échantillonnage préférentiel est sans biais, sous réserve que le support de f soit inclus dans le support de g.

Consistance. Par la loi forte des grands nombres, on a $\tilde{h}_n \xrightarrow[n\to\infty]{p.s.} \mathbb{E}_g\left(h(Z)\frac{f(Z)}{g(Z)}\right) = \mathbb{E}_f(h(X))$: l'estimateur est fortement consistant.

Définition 9. *On appelle les quantités $w_i = \frac{f(Z_i)}{g(Z_i)}$ les poids d'importance.*

Quel peut être l'intérêt de simuler selon une autre loi que f ? Tout d'abord, il peut être difficile de simuler directement selon f. Dans ce cas, l'échantillonnage préférentiel nous garantit que n'importe quel choix de g peut nous permettre d'estimer l'intégrale initiale, dès lors que le support de g inclut celui

de f. D'autre part, dans le cas d'évènements rares sous la loi f, il peut être plus intéressant de passer par une loi g pour laquelle l'évènement en question a une plus forte probabilité. Dans ce cas, le coût en terme de nombre de simulations est plus faible que si on tirait selon la loi f. Enfin, on peut montrer que dans certains cas, le choix optimal pour la loi d'échantillonnage n'est pas f, et qu'en choisissant une autre loi g on peut réduire la variance de l'estimateur.

EXEMPLE 8. *Un exemple permettant d'illustrer le gain apporté par les méthodes d'échantillonnage préférentiel est celui de l'estimation des probabilités de queues de distribution. Supposons que l'on souhaite estimer* $\mathbb{P}(X \geq 4)$, *pour* $X \sim \mathcal{N}(0,1)$. *Cette probabilité (calculée à l'aide de la fonction* `pnorm()` *sous R) vaut* $3.167124 \cdot 10^{-5}$, *ce qui signifie qu'en pratique on obtient environ 3 réalisations supérieures à 4 sur 100000 simulations de la loi normale centrée réduite ... En utilisant comme loi instrumentale une loi de Student à 1 degré de liberté, cette probabilité est de 0.078, l'évènement est donc beaucoup moins rare. Voyons ce que cela donne en pratique sur un échantillon de taille* $n = 10000$.

```
> ech_MC <- rnorm(10000)
> pchap_MC <- mean(ech_MC > 4)
[1] 0
> ech_IS <- rt(10000,1)
> poids_IS <- dnorm(ech_IS)/dt(ech_IS,1)
> pchap_IS <- mean(poids_IS * (ech_IS > 4))
[1] 3.502817e-05
```

Pour une même taille d'échantillon, on obtient un meilleur estimateur avec l'échantillonnage préférentiel.

1.3.2 Choix de la loi instrumentale

Comment choisir, en pratique, la loi instrumentale utilisée lors de l'échantillonnage préférentiel ? La méthode converge quel que soit le choix de g, pourvu que le support de f soit inclus dans celui de g (sinon on risque de "rater" des parties de probabilité non nulle sous la loi f). Cependant, certains choix s'avèrent plus intéressants que d'autres. Regardons en particulier la variance de l'estimateur par échantillonnage préférentiel. Celle ci est finie si et seulement si :

$$\mathbb{E}_g\left(h^2(X)\frac{f^2(X)}{g^2(X)}\right) < \infty \Leftrightarrow \int h^2(x)\frac{f^2(x)}{g^2(x)}g(x)dx < \infty \Leftrightarrow \int h^2(x)\frac{f^2(x)}{g(x)}dx < \infty \tag{1.5}$$

Une condition suffisante pour que cette espérance soit finie, en plus de la condition $\mathbb{E}(h^2(X)) < \infty$ qui devait déjà être vérifiée pour le Monte Carlo classique, est que le ratio f/g soit borné. Cela correspond aux cas où la loi instrumentale g a des queues de distribution qui sont plus lourdes que celles de la loi cible f. Dans le cas contraire, la variance de l'estimateur a toutes les chances d'être infinie ... Il est alors possible d'utiliser l'estimateur auto-normalisé.

Définition 10 (Échantillonnage préférentiel auto-normalisé). *Soit $Z_1, \dots, Z_n$ un échantillon simulé de n variables aléatoires i.i.d. de loi g. L'estimateur par échantillonnage préférentiel auto-normalisé de $\mathbb{E}_f(h(X))$ est défini par :*

$$\bar{h}_n = \frac{\sum_{i=1}^n h(Z_i)\frac{f(Z_i)}{g(X_i)}}{\sum_{i=1}^n \frac{f(Z_i)}{g(X_i)}} = \frac{\sum_{i=1}^n w_i h(Z_i)}{\sum_{i=1}^n w_i} \tag{1.6}$$

Autrement dit, on renormalise les poids d'importance. Comme le dénominateur converge presque sûrement vers 1 lorsque n tend vers l'infini, l'estimateur (1.6) est aussi un estimateur fortement consistant de $\mathbb{E}_f(h(X))$. Même s'il est biaisé, il a l'avantage de produire un estimateur avec une variance plus faible (voir Robert et Casella (2013)), en particulier parce que la normalisation des poids d'importance permet de les stabiliser. L'autre avantage de cet estimateur est qu'il peut s'utiliser dans les cas où les lois f et g ne sont connues qu'à une constante de normalisation près. En effet, supposons que l'on connaisse $\tilde{f} = af$ et $\tilde{g} = bg$, où f et g sont les densités de probabilités (i.e. normalisées) sous-jacentes. Alors l'estimateur par échantillonnage préférentiel auto-normalisé est défini par :

$$\begin{aligned}
\bar{h}_n &= \frac{\sum_{i=1}^n h(Z_i)\frac{f(Z_i)}{g(X_i)}}{\sum_{i=1}^n \frac{f(Z_i)}{g(X_i)}} \\
&= \frac{\sum_{i=1}^n h(Z_i)\frac{\tilde{f}(Z_i)/a}{\tilde{g}(X_i)/b}}{\sum_{i=1}^n \frac{\tilde{f}(Z_i)/a}{\tilde{g}(X_i)/b}} \\
&= \frac{\sum_{i=1}^n h(Z_i)\frac{\tilde{f}(Z_i)}{\tilde{g}(X_i)}}{\sum_{i=1}^n \frac{\tilde{f}(Z_i)}{\tilde{g}(X_i)}}
\end{aligned}$$

Autrement dit, avec ce choix auto-normalisé, il n'est pas nécessaire de connaître la constante de normalisation de la densité de probabilité. Ceci aura un intérêt pratique très fort dans le cadre des statistiques bayésiennes (voir partie III), où cette situation est courante.

Le théorème suivant, dû à Rubinstein (1981), précise le choix optimal de la loi instrumentale g au sens de la minimisation de la variance de l'estimateur.

Théorème 5. *La loi instrumentale qui minimise la variance de l'estimateur par échantillonnage préférentiel est donnée par :*

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(z)|f(z)dz} \tag{1.7}$$

Démonstration.

$$\begin{aligned}
\mathrm{Var}\left(h(X)\frac{f(X)}{g(X)}\right) &= \mathbb{E}_g\left(h^2(X)\frac{f^2(X)}{g^2(X)}\right) - \left(\mathbb{E}_g\left(h(X)\frac{f(X)}{g(X)}\right)\right)^2 \\
&= \mathbb{E}_g\left(h^2(X)\frac{f^2(X)}{g^2(X)}\right) - \left(\int h(x)\frac{f(x)}{g(x)}g(x)dx\right)^2
\end{aligned}$$

$$= \mathbb{E}_g\left(h^2(X)\frac{f^2(X)}{g^2(X)}\right) - \left(\int h(x)f(x)dx\right)^2$$

Le deuxième terme ne dépend pas de g, pour minimiser la variance on doit donc minimiser le premier terme. Par l'inégalité de Jensen [1], on a :

$$\mathbb{E}_g\left(h^2(X)\frac{f^2(X)}{g^2(X)}\right) \geq \left(\mathbb{E}_g\left(|h(X)|\frac{f(X)}{g(X)}\right)\right)^2 = \left(\int |h(x)|f(x)\right)^2.$$

Cette borne inférieure est indépendante de g, et on montre facilement que le choix de g^* énoncé dans le théorème permet d'atteindre cette borne. □

En pratique, ce résultat est inexploitable, car le choix optimal de g dépend de la valeur de l'intégrale que l'on cherche à calculer ... cependant ce résultat nous suggère de choisir une fonction g telle que le ratio $|h|f/g$ soit presque constant et de variance finie.

1.3.3 Taille effective de l'échantillon

L'estimateur de Monte Carlo s'obtient à partir de l'estimateur par échantillonnage préférentiel en prenant $g = f$. Dans ce cas, les poids d'importance w_i sont tous égaux à 1. Plus généralement, on a $\mathbb{E}_g(f(X)/g(X)) = 1$, et donc une façon de mesurer la qualité du choix de g est de comparer la moyenne empirique des poids d'importance avec 1. En pratique, on utilise le critère suivant, appelé *taille effective de l'échantillon* (effective sampling size en anglais).

Définition 11. *On appelle taille effective d'échantillon la quantité :*

$$ESS = \frac{(\sum_{i=1} w_i)^2}{\sum_{i=1}^n w_i^2} = \frac{1}{\sum_{i=1}^n \bar{w}_i^2},$$

où $\bar{w}_i = \frac{w_i}{\sum_{i=1}^n w_i}$ *est le poids d'importance normalisé.*

Si tous les poids sont égaux à 1, chaque simulation contribue de façon égale, et $ESS = 1$. À l'inverse, plus il y a de poids nuls (dégénérés), plus ESS sera petit. Ce critère nous permet donc de mesurer la taille d'un échantillon i.i.d. équivalent à l'échantillon par échantillonnage préférentiel ayant donné les poids w_i. C'est un critère parmi d'autres pour mesurer la qualité de l'estimateur.

1.3.4 Ré-échantillonnage

L'intérêt de l'échantillonnage préférentiel est qu'il ne permet pas seulement d'approcher des intégrales, mais qu'il offre également la possibilité de simuler selon une loi donnée. En effet, après un premier tirage de $Z_1, \ldots, Z_n$ selon la loi instrumentale g, on obtient les poids d'importance w_i, et leurs

1. Soit f une fonction convexe, et X une variable aléatoire telle que $\mathbb{E}(X)$ et $\mathbb{E}(f(X))$ existent. L'inégalité de Jensen nous dit que $\mathbb{E}(f(X)) \geq f(\mathbb{E}(X))$.

versions normalisées $\bar{w}_i$. On peut alors ré-échantillonner selon une loi multinomiale dont les probabilités sont données par les $\bar{w}_i$. Plus précisément, on peut obtenir un échantillon $X_1, \ldots, X_N$ de loi f, où :

$$\forall i = 1, \ldots, N \quad \mathbb{P}(X_i = Z_j) = \bar{w}_j \tag{1.8}$$

En pratique, cela revient à tirer aléatoirement avec remise parmi les $Z_1, \ldots, Z_n$, selon une loi de probabilité non uniforme puisque guidée par les $\bar{w}_i$.

> Remarque. *Les poids normalisés sont biaisés, à cause du dénominateur (la constante de renormalisation). Cependant, ce biais diminue à mesure que n augmente, et en pratique on peut tout de même utiliser cette approche pour obtenir des réalisations selon la loi f.*

Chapitre 2

Chaînes de Markov

Dans ce chapitre, on propose une introduction à la théorie des chaînes de Markov. Cette introduction ne prétend pas être exhaustive, elle est loin d'être complète, mais elle fournit les bases pour comprendre les algorithmes de Monte Carlo par chaînes de Markov qui seront abordés au chapitre 3. En particulier, on ne s'intéressera qu'aux chaînes de Markov à *temps discret*, essentiellement parce qu'en pratique les algorithmes utilisés reposent sur des simulations numériques qui sont, par essence, de nature discrète.

2.1 Introduction

Une chaîne de Markov à temps discret est une suite de variables aléatoires, qui ne sont pas indépendantes les unes des autres, et telle que la probabilité de transition d'un état à un autre dépende uniquement de l'état courant dans lequel se trouve la chaîne. Pour donner une définition plus formelle, on va tout d'abord définir la notion de noyau de transition.

Définition 12 (Noyau de transition). *Soit $(E, \mathcal{E})$ un espace probabilisé. Un noyau de transition K est une fonction de $E \times \mathcal{E}$ dans $\mathbb{R}_+$ telle que :*

(i) $\forall x \in E, K(x, \cdot)$ *est une mesure de probabilité*

(ii) $\forall A \in \mathcal{E}, K(\cdot, A)$ *est mesurable*

Autrement dit, pour chaque $x \in E$, c'est-à-dire pour chaque état possible de la chaîne, $K(x, \cdot)$ définit une mesure de probabilité sur $(E, \mathcal{E})$ telle que pour tout $A \in \mathcal{E}$, l'application $x \mapsto K(x, A)$ est mesurable. Dans la plupart des exemples que nous verrons, l'espace d'états E est continu et le noyau de transition admet une densité : dans ce cas, on appelle aussi noyau la densité de probabilité conditionnelle associée au noyau, i.e. :

$$\mathbb{P}_x(X \in A) := \mathbb{P}(X \in A \mid x) = \int_A k(x, y)dy$$

On peut maintenant introduire la notion de chaîne de Markov.

Définition 13 (Chaîne de Markov). *Soit K un noyau de transition. On appelle chaîne de Markov une séquence de variables aléatoires $X_0, X_1, \ldots, X_n, \ldots$ qui vérifie :*

$$\forall A \in \mathcal{E}, \quad \mathbb{P}(X_n \in A \mid X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n \in A \mid X_{n-1} = x_{n-1}) \tag{2.1}$$

$$= \int_A k(x_{n-1}, y)dy \tag{2.2}$$

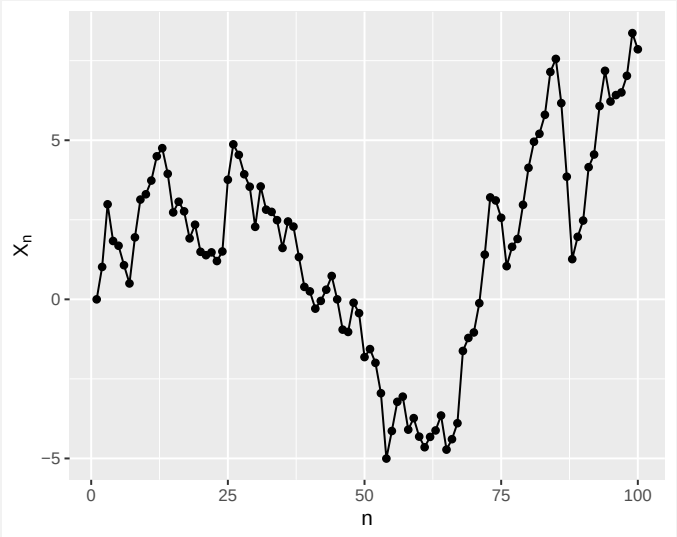
La structure de la chaîne ne dépend donc que de l'état initial x_0 et du noyau de transition. On ne s'intéressera ici qu'aux chaînes de Markov homogènes (en temps), pour lesquelles $\mathbb{P}(X_n \in A \mid X_{n-1} = x) = \mathbb{P}(X_1 \in A \mid X_0 = x)$, c'est-à-dire que la probabilité en (2.1) ne dépend pas de n.

Remarque. *Lorsque l'espace d'états E est discret, le noyau de transition est appelé matrice de transition :*

$$K(x, y) = K_{xy} = \mathbb{P}(X_n = y \mid X_{n-1} = x)$$

Exemple 9. *L'exemple le plus simple de chaîne de Markov est la marche aléatoire simple :*

$$X_n = X_{n-1} + \varepsilon_n, \quad \varepsilon \sim \mathcal{N}(0, 1)$$



Le noyau de transition K nous indique comment passer d'un état de la chaîne à l'état suivant. On peut généraliser cette définition, en introduisant la notion de noyau associé à n transitions, défini par récurrence par :

$$\forall n > 1, \forall x \in E, \forall A \in \mathcal{E}, \quad K^n(x, A) := \int_E K^{n-1}(y, A)k(x, y)dy.$$

On peut interpréter cette définition de la façon (heuristique) suivante : le noyau de transition permettant de passer de l'état x à l'ensemble A en "n étapes" est égal au noyau de transition permettant de passer

de l'état x à l'état y en une étape, multiplié par le noyau de transition permettant de passer de y à A en $n-1$ étapes, ceci en prenant en compte toutes les valeurs possibles pour cette étape intermédiaire en y.

2.2 Mesure invariante

Définition 14 (Mesure invariante). *Une mesure de probabilité π est dite invariante pour une chaîne de Markov de noyau de transition K si :*

$$\forall B \in \mathcal{E}, \pi(B) = \int_E K(x, B)\pi(dx).$$

On parle aussi de loi stationnaire.

Si une telle mesure existe, la chaîne est de la loi stationnaire, car si $X_0 \sim \pi$, alors pour tous les états $n > 0$, on aura également $X_n \sim \pi$. Ceci sera d'une importance cruciale lors de la construction des algorithmes de type MCMC, car les réalisations d'une chaîne de Markov de loi stationnaire seront alors toutes identiquement distribuées.

2.3 Irréductibilité, apériodicité et récurrence

La propriété d'irréductibilité d'une chaîne de Markov est une notion cruciale pour la construction d'algorithmes basés sur les chaînes de Markov.

Définition 15 (Irréductibilité). *Soit μ une mesure de probabilité sur $(E, \mathcal{E})$. Une chaîne de Markov est dite μ-irréductible si :*

$$\forall A \in \mathcal{E} \text{ t.q. } \mu(A) > 0, \exists n \in \mathbb{N}^* \text{ t.q. } \forall x \in E, \quad K^n(x, A) > 0$$

Autrement dit, la chaîne est irréductible s'il est possible d'atteindre n'importe quel ensemble A de mesure non nulle à partir de n'importe quel point x de E en un temps fini. Cette propriété permet notamment de s'affranchir des conditions initiales, car tout sous-ensemble de l'espace d'états est atteignable avec une probabilité non nulle, quel que soit le point de départ de la chaîne.

La notion d'irréductibilité ne suffit pas en pratique, car la chaîne peut malgré tout être contrainte dans ses déplacements par l'existence de "boucles". On a alors besoin de la notion d'*apériodicité* : il n'existe pas de contraintes déterministes sur l'évolution de la chaîne, qui forceraient le passage d'un ensemble à un autre. La définition ci-dessous est due à Roberts et Rosenthal (2004), et sera suffisante dans le contexte des algorithmes MCMC que nous allons voir. Pour une définition plus générale, on pourra voir par exemple Robert et Casella (2013).

Définition 16 (Apériodicité). *Une chaîne de Markov de loi stationnaire π est dite apériodique s'il n'existe pas d'entier $d \geq 2$ et d'ensembles disjoints $A_1, \ldots, A_d$ de $\mathcal{E}$ de mesure non nulle pour π, tels que pour tout $x \in A_i$, $\mathbb{P}(X_n \in A_{i+1} \mid x) = 1$ pour tout $i = 1, \ldots, d-1$ et $\mathbb{P}(X_n \in A_1) = 1$ pour $x \in A_d$.*

La chaîne est dite périodique de période d sinon. Dans ce cas, il existe une chaîne d'ensembles tels que si la chaîne entre dans A_1, elle se retrouve forcée dans ses déplacements : elle ira avec probabilité 1 dans A_2, puis dans A_3, ... jusqu'à A_d où elle retournera alors en A_1.

Les deux notions précédentes sont fondamentales, elles assurent que la chaîne peut aller d'un point x à n'importe quel autre point y à chaque itération. Cependant, cela ne suffit pas pour s'assurer qu'un ensemble A sera visité une infinité de fois. Pour cela, on a besoin de la notion de récurrence. Pour définir proprement cette notion, on va d'abord introduire la notion de nombre de passages.

Définition 17. *Soit $A \in \mathcal{E}$. On définit le nombre de passages en A par :*

$$\eta_A = \sum_{n=1}^{+\infty} \mathbf{1}_{X_n \in A}$$

La notion de nombre de passages nous permet de donner une autre définition équivalente pour l'irréductibilité. Une chaîne de Markov sur $(E, \mathcal{E})$ est $\mu-$irréductible si et seulement si pour tout $x \in E$ et pour tout $A \in \mathcal{E}$ tel que $\mu(A) > 0$, $\mathbb{E}_x(\eta_A) > 0$.

Définition 18 (Récurrence). *Une chaîne μ-irréductible est dite récurrente si pour tout ensemble $A \in \mathcal{E}$ tel que $\mu(A) > 0$, $\mathbb{E}_x(\eta_A) = +\infty$ pour tout $x \in E$.*

Définition 19. *Si une chaîne de Markov irréductible admet une mesure de probabilité invariante, alors la chaîne est dite récurrente positive.*

2.4 Théorème ergodique

Dans la section 1, on a construit des estimateurs pour des quantités s'exprimant comme des intégrales, à l'aide de suites de variables aléatoires i.i.d., et on a utilisé la loi forte des grands nombres pour montrer la forte consistance des estimateurs obtenus. Deux hypothèses sont nécessaires à l'application de la loi forte des grands nombres : d'une part les variables $X_1, \ldots, X_n$ doivent être indépendantes, et d'autre part elles doivent avoir la même loi.

Dans le cadre des chaînes de Markov, il est possible d'obtenir des réalisations issues de la même loi lorsque la chaîne a convergé vers sa distribution stationnaire, mais les variables restent corrélées, ce qui nous empêche d'utiliser la loi forte des grands nombres. Cependant, sous réserve que la chaîne

possède certaines propriétés supplémentaires, le *théorème ergodique* nous permet d'obtenir un résultat similaire à celui de la loi des grands nombres.

Théorème 6 (Théorème egodique). *Une chaîne de Markov $\{X_n\}$ à valeurs dans E de loi stationnaire π est dite* ergodique *lorsqu'elle est apériodique, irréductible et récurrente positive. Elle vérifie alors la propriété suivante, pour toute fonction $h \in L^1(\pi)$:*

$$\frac{1}{M}\sum_{i=1}^{M} h(X_i) \overset{M\to\infty}{\longrightarrow} \int h(x)d\pi(x) \quad p.s.. \tag{2.3}$$

Une fois que l'on a trouvé une chaîne de Markov dont le noyau de transition K admet comme mesure de probabilité invariante la loi d'intérêt, il suffit de vérifier l'apériodicité et l'irréductibilité de la chaîne pour pouvoir appliquer (2.3). Le chapitre suivant propose deux algorithmes permettant de construire de telles chaînes : l'algorithme de Metropolis-Hastings et l'échantillonneur de Gibbs.

Chapitre 3

Algorithmes de Monte Carlo par chaînes de Markov

Dans ce chapitre, on s'intéresse à la construction de chaînes de Markov de loi stationnaire f, pour estimer les intégrales de la forme :

$$\mathbb{E}_f(h(X)) = \int h(x)f(x)dx. \tag{3.1}$$

On peut s'interroger sur l'utilité de telles approches, quand on dispose des méthodes décrites dans les sections 1.2 ou 1.3. On a déjà vu dans cette dernière section, et cela pouvait paraître contre-intuitif, qu'il était parfois préférable de simuler selon une autre loi g dite loi instrumentale plutôt que de simuler directement selon f pour approcher (3.1). Un choix judicieux pour g pouvait en effet permettre de réduire la variance de l'estimateur. Il en est de même avec les algorithmes MCMC : dans certains cas, ils pourront s'avérer plus efficace au sens où la variance de l'estimateur sera plus faible qu'avec un Monte Carlo classique ou un échantillonnage préférentiel. De plus, l'échantillonnage préférentiel requiert un choix adéquat pour la loi instrumentale, et plus la dimension du problème augmente, plus il peut être difficile de bien choisir cette loi g. Les algorithmes MCMC que nous verrons dans ce chapitre reposent également souvent sur une loi instrumentale g, mais comme les simulations sont autorisées à dépendre de la précédente réalisation, cela permettra en pratique un choix plus large pour la loi instrumentale g.

Nous allons voir dans ce chapitre deux algorithmes de type MCMC : l'algorithme de Metropolis-Hastings (section 3.1) et l'échantillonneur de Gibbs (section 3.2).

3.1 Algorithme de Metropolis-Hastings

L'algorithme de Metropolis-Hastings (MH) a d'abord été introduit par Metropolis *et al.* (1953), pour le calcul d'intégrales faisant intervenir des distributions de Boltzmann, et a été généralisé près de 20 ans plus tard par Hastings (1970).

3.1.1 Définition générale

Partant d'une loi cible de densité f, on commence par choisir une densité conditionnelle $q(\cdot|x)$, dont on sait facilement simuler des réalisations aléatoires, ou symétrique (c'est-à-dire telle que $q(y|x) = q(x|y)$). Puis, chaque itération de l'algorithme MH consiste alors à simuler, à partir de l'état courant de la chaîne X_n, un candidat qui sera accepté comme le nouvel état de la chaîne avec une certaine probabilité. Si le candidat est rejeté, la chaîne reste au même endroit, i.e. $X_n = X_{n-1}$. L'algorithme est décrit dans l'encadré 2.

Algorithme 2 : Algorithme de Metropolis-Hastings

initialisation : on initialise la chaîne avec X_0

pour $n = 1, \dots, N$ **faire**

on génère un candidat $Y_n \sim q(\cdot \mid X_{n-1})$

on pose

$$X_n = \begin{cases} Y_n & \text{avec une probabilité } \alpha(X_{n-1}, Y_n) \\ X_{n-1} & \text{avec une probabilité } 1 - \alpha(X_{n-1}, Y_n) \end{cases} \tag{3.2}$$

où

$$\alpha(x, y) = \min\left(1, \frac{f(y)}{f(x)} \frac{q(x \mid y)}{q(y \mid x)}\right). \tag{3.3}$$

On utilise la loi instrumentale $q(\cdot|x)$ pour générer des réalisations de la loi f donc, plus le rapport f/q est faible pour le candidat par rapport à l'état courant de la chaîne, plus la probabilité de le rejeter est forte. Intuitivement, la probabilité d'acceptation $\alpha(x, y)$ permet de faire un compromis entre les deux conditions suivantes : d'une part, on souhaite que l'algorithme se dirige vers des régions de plus forte probabilité sous f, ce qui est contrôlé par le rapport $f(y)/f(x)$ (plus celui-ci est haut, plus on accepte le candidat), et d'autre part on souhaite éviter que l'algorithme ne reste trop longtemps dans une région spécifique de trop forte probabilité sous q, ce qui est contrôlé par le rapport $q(x|y)/q(y|x)$.

REMARQUE. *L'algorithme de Metropolis-Hastings ne dépend de f et de q qu'à travers des ratios : on peut donc se contenter de ne connaître f et /ou g qu'à une constante multiplicative près.*

3.1.2 Propriétés de convergence

Après avoir défini l'algorithme MH, on s'intéresse à ses propriétés. On va notamment montrer que f est une loi stationnaire pour la chaîne de Markov définie par l'algorithme MH, et que la chaîne converge bien en loi vers sa distribution stationnaire. Notons tout d'abord qu'une condition nécessaire pour que la chaîne construite par l'algorithme admette f pour loi stationnaire est que le support de f soit inclus dans le support de q. En effet, si ce n'est pas le cas, certaines parties de probabilité non nulle pour f ne

seront jamais atteinte par la chaîne.

On s'intéresse tout d'abord au noyau de transition de la chaîne de Markov générée par l'algorithme MH. Plaçons-nous au début de l'itération n. La chaîne se trouve à l'état X_{n-1}, et on propose un candidat Y_n. La probabilité qu'un candidat soit accepté, sachant que l'on se trouve en X_{n-1}, est donnée par :

$$\rho(X_{n-1}) = \int \alpha(X_{n-1}, y) q(y \mid X_{n-1}) dy$$

Autrement dit, la probabilité pour que la chaîne reste en X_{n-1} est égale à $1 - \rho(X_{n-1})$. Maintenant, le noyau de transition de la chaîne produite par l'algorithme MH peut s'écrire :

$$K(x, y) = \alpha(x, y) q(y \mid x) + (1 - \rho(x)) \delta_x(y),$$

où δ_x désigne la masse de Dirac au point x. On vérifie qu'il s'agit bien d'un noyau de transition, en particulier, $\int K(x, dy) = 1$, et donc $K(x, \cdot)$ est bien une mesure de probabilité.

Mesure invariante. Le noyau de transition vérifie une propriété fondamentale, appelée condition d'équilibre local, et définie ci-dessous.

Définition 20. *Une chaîne de Markov de noyau de transition K vérifie la condition d'équilibre local s'il existe une fonction f telle que :*

$$K(y, x) f(y) = K(x, y) f(x)$$

Cette condition d'équilibre local fournit en fait une condition suffisante pour que f soit une mesure invariante pour la chaîne de Markov. Dans le cas de l'algorithme MH, le noyau de transition associé à la loi cible f vérifie bien cette condition d'équilibre local. La loi cible f est donc une loi stationnaire pour la chaîne de Markov construite par l'algorithme MH. En effet, on a :

$$\begin{aligned}
\alpha(y, x) q(x \mid y) f(y) &= \min\left(1, \frac{f(x)}{f(y)} \frac{q(y \mid x)}{q(x \mid y)}\right) q(x \mid y) f(y) \\
&= \min\left(\frac{1}{f(x) q(y \mid x)}, \frac{1}{f(y) q(x \mid y)}\right) f(x) q(y \mid x) q(x \mid y) f(y) \\
&= \min\left(\frac{q(x \mid y) f(y)}{f(x) q(y \mid x)}, \frac{q(x \mid y) f(y)}{f(y) q(x \mid y)}\right) f(x) q(y \mid x) \\
&= \min\left(\frac{f(y) q(x \mid y)}{f(x) q(y \mid x)}, 1\right) f(x) q(y \mid x) \\
&= \alpha(x, y) q(y \mid x) f(x)
\end{aligned}$$

On en déduit que $(1 - \rho(x)) f(x) = (1 - \rho(y) f(y)$, et on conclut en remarquant que $\delta_x(y) = \delta_y(x)$.

On a donc montré que la loi cible f est une loi stationnaire pour la chaîne de Markov construite par l'algorithme MH. Mais cela ne suffit pas : il faut montrer que la chaîne ainsi construire *converge* vers cette distribution stationnaire. On pourra alors utiliser le théorème ergodique et approcher l'intégrale (3.1) par une moyenne empirique construite à partir des réalisations de la chaîne.

La chaîne admettant une loi invariante, il nous reste à montrer qu'elle est apériodique, irréductible et récurrente positive. En pratique, on va utiliser une série de résultats qui sont en fait des *conditions suffisantes*, mais pas nécessaires, pour que ces propriétés soient vérifiées. Elles sont suffisantes et plus faciles à utiliser dans le cadre de l'algorithme MH que les définitions initiales.

Irréductibilité. Une chaîne est irréductible si tout ensemble de mesure non nulle peut être atteint à partir de n'importe quel point en un nombre fini d'étapes. Une condition suffisante pour que la chaîne produite par l'algorithme MH soit irréductible est la suivante :

$$q(y \mid x) > 0, \quad \forall (x, y) \in E^2.$$

En effet, dans ce cas tout ensemble de mesure de Lebesgue non nulle peut être atteint en une étape à l'aide de l'algorithme MH.

Apériodicité. Une condition suffisante pour que la chaîne produite par l'algorithme MH soit apériodique est la suivante :

$$\mathbb{P}(X_n = X_{n-1}) > 0.$$

Ceci est vrai par construction. On peut donc rester au même endroit, il n'y a pas de contraintes de déplacements d'un point d'un sous-ensemble vers un autre sous-ensemble, comme dans le cas périodique.

Récurrence. La chaîne est récurrente positive car elle est irréductible et admet une mesure de probabilité invariante.

Les conditions d'applications du théorème ergodique sont réunies : on peut donc utiliser les réalisations d'une chaîne de Markov $\{X_n\}$ construite par l'algorithme MH pour approcher (3.1) :

$$\frac{1}{N} \sum_{n=1}^{N} h(X_n) \xrightarrow[N \to +\infty]{p.s.} \mathbb{E}_f(h(X))$$

3.1.3 Deux cas particuliers

Dans la pratique, le choix de la distribution instrumentale q influence fortement la vitesse de convergence de la chaîne de Markov vers la loi stationnaire f car il dirige l'exploration de l'espace d'états par la chaîne de Markov (Robert et Casella, 2013). Deux types d'approches sont utilisées classiquement pour le choix de la loi q : le cas où q ne dépend pas de l'état courant de la chaîne, et le cas où q est symétrique.

Le cas indépendant. Lorsque la loi instrumentale q est indépendante de la position actuelle de la chaîne X_{n-1}, l'algorithme peut être vu comme une extension des méthodes d'acceptation-rejet. Dans ce cas, l'algorithme s'écrit comme dans l'encadré 3.

Même si les candidats Y_n sont générés indépendamment les uns des autres, les états X_n de la chaîne de Markov ne sont pas indépendants. La probabilité d'accepter le candidat à chaque itération dépend bien de l'état courant de la chaîne.

Algorithme 3 : Algorithme de Metropolis-Hastings indépendent

initialisation : on initialise la chaîne avec X_0

pour $n = 1, \dots, N$ **faire**

on génère un candidat $Y_n \sim q(\cdot)$

on pose

$$X_n = \begin{cases} Y_n & \text{avec une probabilité } \alpha(X_{n-1}, Y_n) \\ X_{n-1} & \text{avec une probabilité } 1 - \alpha(X_{n-1}, Y_n) \end{cases} \tag{3.4}$$

où

$$\alpha(x, y) = \min\left(1, \frac{f(y)}{f(x)} \frac{q(x)}{q(y)}\right). \tag{3.5}$$

La marche aléatoire. Un autre choix usuel pour la loi instrumentale est celui d'une marche aléatoire autour de la valeur actuelle de la chaîne :

$$Y_n = X_{n-1} + \varepsilon_n,$$

où ε_n suit une loi de densité q indépendante de X_{n-1}. Dans ce cas, la loi $q(y|x)$ ne dépend que de la différence $y - x$ et peut s'écrire $q(y - x)$. Lorsque, de plus, la loi considérée est symétrique, c'est-à-dire lorsque $q(u) = q(-u)$, la probabilité d'acceptation se simplifie et on obtient l'algorithme 4.

Algorithme 4 : Algorithme de Metropolis-Hastings par marche aléatoire symétrique

initialisation : on initialise la chaîne avec X_0

pour $n = 1, \dots, N$ **faire**

on génère un candidat $Y_n \sim q(\cdot \mid X_{n-1})$

on pose

$$X_n = \begin{cases} Y_n & \text{avec une probabilité } \alpha(X_{n-1}, Y_n) \\ X_{n-1} & \text{avec une probabilité } 1 - \alpha(X_{n-1}, Y_n) \end{cases} \tag{3.6}$$

où

$$\alpha(x, y) = \min\left(1, \frac{f(y)}{f(x)} \frac{q(x \mid y)}{q(y \mid x)}\right) = \min\left(1, \frac{f(y)}{f(x)} \frac{q(x - y)}{q(y - x)}\right) = \min\left(1, \frac{f(y)}{f(x)}\right).$$

Parmi les choix classiques pour la loi q, on relève la loi uniforme, la loi normale ou la loi de Student, centrées autour de 0. Le choix du support de la loi uniforme, ou de la variance des lois normale et de Student va conditionner la vitesse de convergence de l'algorithme. Si la densité est trop concentrée autour de 0, le candidat généré à chaque étape sera souvent accepté, et l'exploration de l'espace d'états sera lente. Si, au contraire, la densité est trop étendue, les candidats proposés seront souvent rejetés car correspondant à des régions de trop faible probabilité sous f, et la chaîne aura donc tendance à rester

trop longtemps au même point (voir figure 2.7). On peut aussi s'en convaincre en regardant le graphe d'autocorrélation de la chaîne, qui permet de rendre compte de la dépendance entre réalisations successives. Plus la fonction d'autocorrélation décroît rapidement, plus on se retrouve dans une situation "proche" du cas i.i.d.

Pour assurer une bonne exploration de l'espace d'états par la chaîne, on peut s'intéresser au taux d'acceptation des candidats, qui doit donc être ni trop élevé ni trop bas. Ce taux d'acceptation se calcule facilement en regardant la proportion empirique de candidats ayant été acceptés. Des valeurs optimales ont été dérivées pour ce taux d'acceptation, selon l'algorithme utilisé. En terme de recommandations pratiques, on conseille souvent un taux d'acceptation d'environ 0.5 en dimension 1 ou 2 et d'environ 0.25 en grande dimension (Robert *et al.*, 2010).

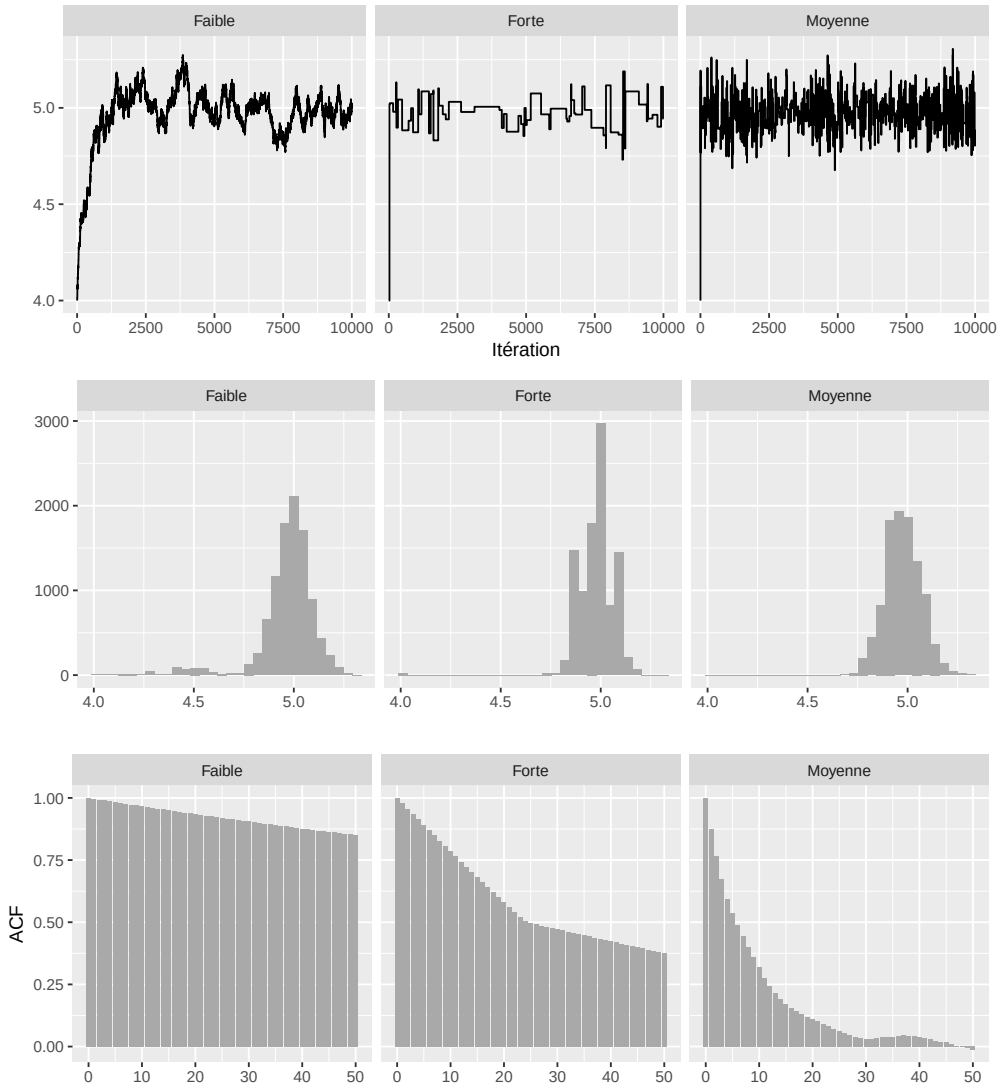


FIGURE 2.7 – Algorithme de Metropolis-Hastings à marche aléatoire gaussienne de moyenne 0 et de variance σ^2, pour une variance σ^2 faible, forte ou moyenne. En haut : évolution de la chaîne au cours du temps, au milieu : histogramme des réalisations de la chaine et en bas : autocorrélation des réalisations.

3.2 Échantillonneur de Gibbs

L'échantillonneur de Gibbs a été introduit par Geman et Geman (1984) puis repris plus tard par Gelfand et Smith (1990), et repose sur l'idée selon laquelle on peut simuler selon une loi multidimensionnelle en simulant chaque composante du vecteur selon la loi conditionnelle associée.

3.2.1 Définition générale

Supposons que l'on a $X = (X_1, \ldots, X_p)$ de loi (multidimensionnelle) f. L'échantillonneur de Gibbs repose sur la décomposition de la loi jointe f en une série de lois univariées, aussi appelée *lois conditionnelles complètes*. Ces densités conditionnes, notées $f_1, \ldots, f_p$, sont définies comme les lois conditionnelles de chaque composante du vecteur sachant les autres composantes. Autrement dit, la densité f_j est donnée par :

$$X_j \mid X_1 = x_1, \ldots, X_{j-1} = x_{j-1}, X_{j+1} = x_{j+1}, \ldots, X_p = x_p \sim f_j(\cdot \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p), \tag{3.7}$$

L'échantillonneur de Gibbs consiste alors en p simulations séquentielles de chaque composante du vecteur X, comme détaillé dans l'algorithme 5.

Algorithme 5 : Échantillonneur de Gibbs

initialisation : on initialise la chaîne avec $X_0 = (X_{0,1}, \ldots, X_{0,p})$

pour $n = 1, \ldots, N$ **faire**

on génère :

$$\begin{aligned}
X_{n,1} &\sim f_1(\cdot \mid x_{n-1,2}, \ldots, x_{n-1,p}),\\
X_{n,2} &\sim f_2(\cdot \mid x_{n,1}, x_{n-1,3}, \ldots, x_{n-1,p}),\\
&\ldots\\
X_{n,p} &\sim f_p(\cdot \mid x_{n,1}, \ldots, x_{n,p-1})
\end{aligned}$$

Cet algorithme, à la différence de l'algorithme MH, a un taux d'acceptation de 1 : à chaque itération, la chaîne de Markov change de position. L'implémentation de cet algorithme est possible dès lors que chaque loi conditionnelle est connue, à une constante de normalisation près (on peut dans ce cas utiliser les méthodes présentées dans le chapitre 1, section 1.1). Si on ne sait pas simuler directement selon ces lois conditionnelles, on peut introduire une étape de type Metropolis-Hastings pour chaque composante.

Un cas particulier est obtenu dans le cas où $p = 2$, i.e. on cherche à échantillonner selon une loi jointe à 2 dimensions. En effet, dans ce cas l'échantillonneur de Gibbs possède de meilleures propriétés que lorsque $p > 2$, notamment en terme de convergence.

Remarque.

- *l'échantillonneur de Gibbs est par essence multidimensionel, et ne peut donc pas s'appliquer aux problèmes en dimension 1*
- *l'algorithme n'est pas non plus adapté aux situations où la taille du vecteur varie. Un exemple typique est celui des modèles de mélange, que nous aborderons dans la partie IV.*

3.2.2 Le théorème de Hammersley-Clifford

L'une des propriétés fondamentales de l'échantillonneur de Gibbs est que la connaissance des lois conditionnelles complètes suffit pour reconstruire la loi jointe. Ce résultat est donné par le théorème de Hammersley-Clifford, que l'on peut démontrer tout d'abord dans le cas plus simple où $p = 2$, puis dans le cas général.

Théorème 7 (Théorème de Hammersley-Clifford).

- *cas $p = 2$: soit (X, Y) un couple de variables aléatoires de densités conditionnelles $f_{X|Y}(x|y)$ et $f_{Y|X}(y|x)$. Alors la loi jointe du couple a pour densité :*

$$f(x, y) = \frac{f_{Y|X}(y|x)}{\int \frac{f_{Y|X}(z|x)}{f_{X|Y}(x|z)} dz}$$

- *cas $p > 2$: soit $X = (X_1, \ldots, X_p)$ un vecteur aléatoire tel que le support de la loi jointe est le produit cartésien des supports des lois marginales. Alors la loi jointe du vecteur X vérifie :*

$$f(x_1, \ldots, x_p) \propto \prod_{j=1}^{p} \frac{f_j(x_j|x_1, \ldots, x_{j-1}, z_{j+1}, \ldots, z_p)}{f_j(z_j|x_1, \ldots, x_{j-1}, z_{j+1}, \ldots, z_p)},$$

pour tout vecteur $(z_1, \ldots, z_p)$ dans le support de f.

Si ce théorème assure que la connaissance des lois conditionnelles complètes suffit à reconstruire la loi jointe, ce n'est pas ce résultat qui nous permettra, en pratique, de montrer que la chaîne de Markov produite par l'échantillonneur de Gibbs admet bien pour loi stationnaire la loi jointe f.

3.2.3 Propriétés de convergence

Le noyau de transition de la chaîne de Markov est donné, pour $(x, y) \in (\mathbb{R}^p)^2$ par :

$$K(x, y) = \prod_{j=1}^{p} f_j(y_j \mid y_1, \ldots, y_{j-1}, x_{j+1}, \ldots, x_p) \tag{3.8}$$

Mesure invariante.

PROPOSITION 3. *La loi jointe f est une mesure invariante pour la chaîne de Markov produite par l'échantillonneur de Gibbs.*

Démonstration. En partant de la définition 14, il faut montrer que pour tout $y \in \mathbb{R}^p$, $\int K(x,y)f(x)dx = f(y)$. Dans la suite, on utiliser la notation g pour désigner la densité jointe d'un sous-vecteur de $\mathbb{R}^p$. On a :

$$\begin{aligned}
\int K(x,y)f(x)dx &= \int \prod_{j=1}^{p} f_j(y_j|y_1,\dots,y_{j-1},x_{j+1},\dots,x_p)f(x_1,\dots,x_p)dx_1\dots dx_p \\
&= \int f_1(y_1|x_2,\dots,x_p)f_2(y_2|y_1,x_3,\dots,x_p)\dots f_p(y_p|y_1,\dots,y_{p-1})f_1(x_1|x_2,\dots,x_p)g(x_2,\dots,x_p)dx_1\dots dx_p \\
&= \int f_1(y_1|x_2,\dots,x_p)f_2(y_2|y_1,x_3,\dots,x_p)\dots f_p(y_p|y_1,\dots,y_{p-1})g(x_2,\dots,x_p)\underbrace{\left(\int f_1(x_1|x_2,\dots,x_p)dx_1\right)}_{=1}dx_2\dots dx_p \\
&= \int \frac{f(y_1,x_2,\dots,x_p)}{g(x_2,\dots,x_p)}f_2(y_2|y_1,x_3,\dots,x_p)\dots f_p(y_p|y_1,\dots,y_{p-1})g(x_2,\dots,x_p)dx_2\dots dx_p \\
&= \int f(y_1,x_2,\dots,x_p)f_2(y_2|y_1,x_3,\dots,x_p)\dots f_p(y_p|y_1,\dots,y_{p-1})dx_2\dots dx_p \\
&= \int g(y_1,x_3,\dots,x_p)f_2(y_2|y_1,x_3,\dots,x_p)\dots f_p(y_p|y_1,\dots,y_{p-1})\int f_2(x_2|y_1,x_3,\dots,x_p)dx_2dx_3\dots dx_p \\
&= \int g(y_1,x_3,\dots,x_p)f_2(y_2|y_1,x_3,\dots,x_p)\dots f_p(y_p|y_1,\dots,y_{p-1})dx_3\dots dx_p
\end{aligned}$$

En poursuivant l'intégration successive par rapport à $x_3,\dots,x_p$, on obtient à la dernière étape :

$$\begin{aligned}
\int K(x,y)f(x)dx &= \int g(y_1,\dots,y_{p-1},x_p)f_{p-1}(y_{p-1}|y_1,\dots,y_{p-1},x_p)\dots f_p(y_p|y_1,\dots,y_{p-1})dx_p \\
&= \int g(y_1,\dots,y_{p-1},x_p)\frac{f(y_1,\dots,y_{p-2},y_{p-1},x_p)}{g(y_1,\dots,y_{p-1},x_p)}\dots f_p(y_p|y_1,\dots,y_{p-1})dx_p \\
&= f_p(y_p|y_1,\dots,y_{p-1})f(y_1,\dots,y_{p-1})\underbrace{\int f_p(x_p|y_1,\dots,y_{p-1})dx_p}_{=1} \\
&= f(y_1,\dots,y_p)
\end{aligned}$$

□

Irréductibilité et récurrence. L'irréductibilité de la chaîne produite par l'échantillonneur de Gibbs s'obtient directement dans le cas (qui peut s'avérer restrictif en pratique) où la loi jointe vérifie la *condition de positivité*, déjà énoncée dans le théorème de Hammersley-Clifford : le support de la loi jointe est égale au produit cartésien des supports des lois marginales. Dans ce cas, la chaîne est irréductible et récurrente.

Cette *condition de positivité* n'est pas anodine : elle permet d'assurer que la chaîne ne se retrouve pas bloquée dans une région de l'espace. Pour le voir, considérons un exemple où cette condition n'est

pas vérifiée : le support de la loi jointe ne peut pas s'écrire comme le produit cartésien des supports des lois marginales. Par exemple, considérons un vecteur aléatoire (X, Y) dont la loi jointe est donnée par :

$$f(x,y) = \frac{1}{2}\left(\mathbf{1}_{[-1,0]\times[-1,0]}(x,y) + \mathbf{1}_{[0,1]\times[0,1]}(x,y)\right),$$

et dont le support est dessiné sur la figure 2.8.

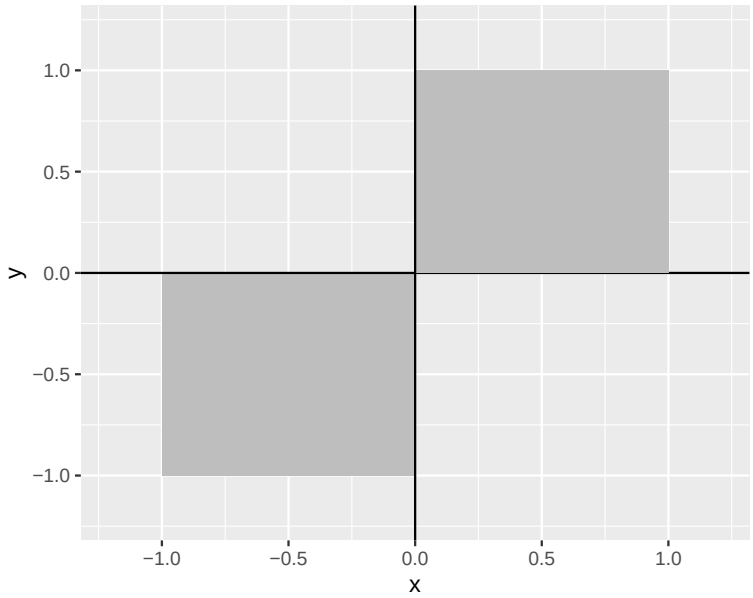


FIGURE 2.8 – Support de la loi jointe $f(x,y) = \frac{1}{2}\left(\mathbf{1}_{[-1,0]\times[-1,0]}(x,y) + \mathbf{1}_{[0,1]\times[0,1]}(x,y)\right)$

Dans un cas comme celui-ci, l'algorithme se retrouve piégé dans l'une des deux zones. En effet, on a alors les lois marginales suivantes pour X et Y respectivement : $f(x) = \frac{1}{2}\mathbf{1}_{[-1,1]}(x)$ et $f(y) = \frac{1}{2}\mathbf{1}_{[-1,1]}(y)$ (autrement dit, les deux coordonnées X et Y sont de loi uniforme sur l'intervalle $[-1, 1]$), ce qui donne les lois conditionnelles complètes :

$$f(y|x) = \frac{\mathbf{1}_{[-1,0]}(x)\mathbf{1}_{[-1,0]}(y) + \mathbf{1}_{[0,1]}(x)\mathbf{1}_{[0,1]}(y)}{\mathbf{1}_{[-1,0]}(x) + \mathbf{1}_{[0,1]}(x)}$$
$$f(x|y) = \frac{\mathbf{1}_{[-1,0]}(y)\mathbf{1}_{[-1,0]}(x) + \mathbf{1}_{[0,1]}(y)\mathbf{1}_{[0,1]}(x)}{\mathbf{1}_{[-1,0]}(y) + \mathbf{1}_{[0,1]}(y)}$$

On voit alors qu'en partant d'un point initial dans le cadrant inférieur par exemple, l'algorithme ne pourra pas s'en échapper. En effet, si $y_0 \in [-1, 0]$, en simulant à l'étape 1 selon la loi conditionnelle complète de X sachant Y, on a $f(x|Y_0 = y_0) = \mathbf{1}_{[-1,0]}(x)$: on se déplace alors dans $[-1, 0]$ pour la coordonnée X. Mais de la même façon, on se déplacera aussi dans $[-1, 0]$ pour la coordonnée Y à cette même étape 1, et ainsi de suite.

Apériodicité. L'apériodicité de la chaîne a été démontrée par exemple par Liu *et al.* (1995).

Les conditions d'application du théorème ergodique sont donc bien réunies. On peut alors utiliser les réalisations d'une chaîne de Markov $\{X_n\}$ construite par l'échantillonneur de Gibbs pour approcher (3.1).

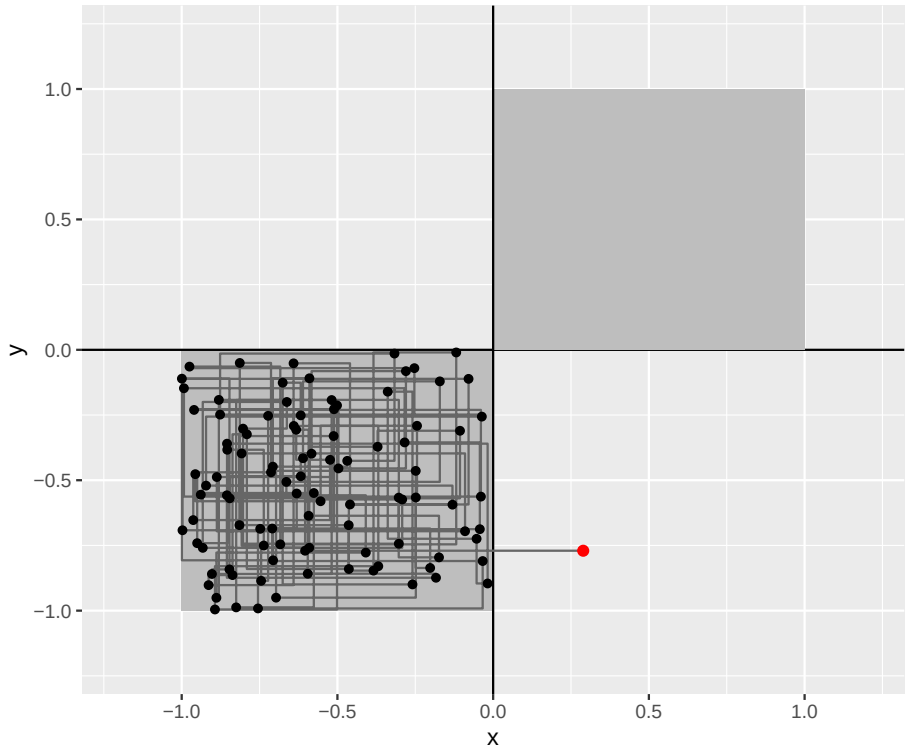


FIGURE 2.9 – Chaîne de Markov générée par un échantillonneur de Gibbs à partir de la loi jointe $f(x,y) = \frac{1}{2}\left(\mathbf{1}_{[-1,0]\times[-1,0]}(x,y) + \mathbf{1}_{[0,1]\times[0,1]}(x,y)\right)$. *Le point de départ de l'algorithme est en rouge, chaque point correspond à une itération de l'algorithme et les étapes intermédiaires liées aux simulations selon les lois conditionnelles complètes sont représentées par les traits pleins.*

3.2.4 Variantes de l'échantillonneur de Gibbs

Échantillonneur de Gibbs avec mise à jour aléatoire. L'échantillonneur de Gibbs tel que présenté dans l'algorithme 5 n'est pas *réversible*. La réversibilité n'est pas une condition nécessaire pour utiliser le théorème ergodique, mais elle apporte de nombreuses propriétés supplémentaires en terme de convergence. Par définition, une chaîne de Markov stationnaire est dite *réversible* si la loi de X_n sachant $\{X_{n-1} = x\}$ est la même que la loi de X_n sachant $\{X_{n+1} = x\}$. Autrement dit, la direction du temps n'a pas d'influence sur la dynamique de la chaîne.

Il existe une variance de l'algorithme 5 qui est réversible, et qui consiste à sélectionner au hasard, à chaque itération de l'algorithme, la composante qui sera mise à jour.

Échantillonneur de Gibbs complété. Dans certains cas, il peut s'avérer plus simple de considérer la loi cible f comme la loi marginale d'une loi sur un espace plus grand. Plus précisément, si on a une densité g telle que :

$$f(x) = \int g(z,x)dz,$$

alors g est appelée *complétion* de f. En pratique, on choisit g de telle sorte qu'il soit plus facile de simuler selon les lois conditionnelles complètes de g, et on implémente l'algorithme 5 sur g au lieu de f. On ré-

Algorithme 6 : Échantillonneur de Gibbs avec mise à jour aléatoire

initialisation : on initialise la chaîne avec $X_0 = (X_{0,1}, \ldots, X_{0,p})$

pour $n = 1, \ldots, N$ **faire**

- on tire au sort un indice j parmi $1, \ldots, p$
- on génère :

$$X_{n,j} \sim f_j(\cdot \mid x_{n-1,1}, \ldots, x_{n-1,j-1}, x_{n-1,j+1}, \ldots, x_{n-1,p})$$

- on pose $X_{n,k} = X_{n-1,k}$ pour tout $k \neq j$

cupère les simulations selon la loi f en ne gardant que les réalisations correspondant aux coordonnées liées à X. Par exemple, si on a f densité sur $\mathbb{R}^p$ et g densité sur $\mathbb{R}^{q+p}$, i.e. $g(z, x) = g(z_1, \ldots, z_q, x_1, \ldots, x_p)$, on garde les p dernières composantes de chaque réalisation de la chaîne de Markov généré par l'échantillonneur de Gibbs complété lancé sur g.

3.3 Diagnostics de convergence

Les algorithmes présentés dans les deux sections précédentes garantissent que la chaîne de Markov obtenue converge vers un régime stationnaire donné par la loi cible f. Cependant, cette convergence théorique ne suffit pas pour savoir en pratique à quel moment l'algorithme peut être stoppé. Dans cette section, nous allons présenter plusieurs outils, essentiellement empirique, qui peuvent aider à identifier les situations de non convergence. Ce ne sont pas des critères absolus, et ils ne doivent pas être interprétés comme des preuves de convergence, mais plutôt comme des "tests" à passer. Vous pouvez, en première approximation, les voir comme des conditions nécessaires mais malheureusement pas suffisantes, de convergence.

Remarque. *Il est courant en pratique d'écarter les premières itérations de la chaîne, considérées comme faisant partie de la période de chauffe, c'est-à-dire du temps nécessaire à la chaîne pour atteindre son régime stationnaire. On parle aussi de période de "burn-in". La taille de cette période de chauffe peut être identifiée visuellement à l'aide d'un graphe d'évolution de la chaîne. En pratique on peut recommander de la prendre au moins égale à 10% de la taille totale de la chaîne.*

3.3.1 Convergence vers la loi stationnaire

La première chose que l'on souhaite vérifier, c'est que la chaîne produite par l'algorithme a bien atteint son régime stationnaire, et que l'on a donc bien des échantillons issus de la loi cible f. Ce point étant difficile à vérifier à la fois d'un point de vue théorique que pratique, on a alors recours à des outils de visualisation graphique, ou à des tests de stationnarité.

Dans l'idéal, il faut lancer plusieurs fois le même algorithme, c'est-à-dire en utilisant le même noyau de transition, mais en prenant des initialisations différentes et suffisamment dispersées les unes par rapport aux autres. On peut alors vérifier graphiquement que les différentes chaînes de Markov obtenues convergent vers la même zone. Si ce n'est pas le cas, il est clair que la convergence n'est pas atteinte, et il faut donc augmenter le nombre d'itérations. En revanche, ce type de diagnostic peut s'avérer limité s'il y a des zones de probabilités non nulles pour f qui n'ont pas été explorées par l'algorithme. La figure 2.10 illustre un tel cas : on cherche à simuler selon la loi de mélange $0.7\mathcal{N}(1,1) + 0.3\mathcal{N}(6,1)$ à l'aide d'un algorithme de Metropolis-Hastings en utilisant comme loi instrumentale la loi $\mathcal{N}(0,1)$ (indépendante donc de l'état actuel de la chaîne). En partant de deux points de départs différents, -2 et 2.5, on obtient les deux chaînes de Markov de la figure en haut à gauche. Aucune des deux chaînes n'a exploré le deuxième mode, elles ne sont donc pas encore dans le régime stationnaire, mais rien ne semble indiquer un problème. Sur le graphe en haut à droite, on a représenté l'histogramme des réalisations auxquels on a superposé la vraie densité (que l'on peut tracer dans ce cas particulier). On voit en effet que le plus petit mode n'a pas été exploré, et qu'il faudrait donc augmenter sensiblement le nombre d'itérations. Cependant, en prenant d'autres initialisations plus dispersées les unes par rapport aux autres, on pourrait détecter ce problème de mode non exploré. En effet, si la chaîne est initialisé autour de 7 par exemple, les candidates proposés seraient toujours centrés autour de 1, et même si ces valeurs sont de plus forte densité sous f que la valeur 7, le ratio $q(X_{n-1})/q(Y_n)$ dans le calcul de la probabilité d'acceptation sera toujours très faible si X_{n-1} est autour de 7. La chaîne restera donc bloquée longtemps à son point de départ, indiquant alors un problème de convergence. En changeant de loi instrumentale pour une marche aléatoire gaussienne de variance 1, on obtient les graphes du bas. Cette fois, le régime stationnaire semble bien atteint.

On peut également utiliser des critères plus quantitatifs pour tester la stationnarité. Une possibilité lorsqu'on ne dispose que d'une seule réalisation de l'algorithme (et donc d'une seule chaîne) consiste à comparer les distributions de deux sous-parties de la chaîne à l'aide par exemple d'un test de Kolmogorov-Smirnov. Ce test étant utilisable pour des échantillons i.i.d. il faut cependant faire attention à la corrélation entre les réalisations successives de la chaîne. Afin d'obtenir des réalisations les plus non corrélées possibles, on peut en pratique élaguer la chaîne en ne retenant qu'une partie de la trajectoire, par exemple, une réalisation toutes les dix itérations. Pour identifier la bonne fréquence de sous-échantillonnage, on peut s'aider du graphe d'autocorrélation, dont un exemple est donné sur la figure 2.11. On voit notamment que la corrélation entre deux états X_t et X_{t+s} devient négligeable lorsque $s = 30$. On pourra alors construire la chaîne $X_0, X_{30}, X_{60}, \ldots$ pour obtenir des échantillons "presque" indépendants. Pour revenir au test de Kolmogorov-Smirnov, on peut tester la stationnarité d'une chaîne de taille N en comparant les distributions de deux sous-parties de la chaîne, par exemple $X_0, X_s, X_{2s}, \ldots, X_{ks}$ et $X_{ks}, \ldots, X_{Ks}$ avec $k = \lfloor N/(2s) \rfloor$ et $K = \lfloor N/s \rfloor$. Dans l'exemple de la figure 2.10, le test appliqué aux deux échantillons ainsi formés ne permet pas de rejeter l'hypothèse nulle d'égalité des distributions, et ce, pour toutes les chaînes de Markov obtenues quelle que soit la loi instrumentale. Là encore, on attire donc l'attention sur le fait que ce test ne permet pas de détecter un problème lié à la non exploration

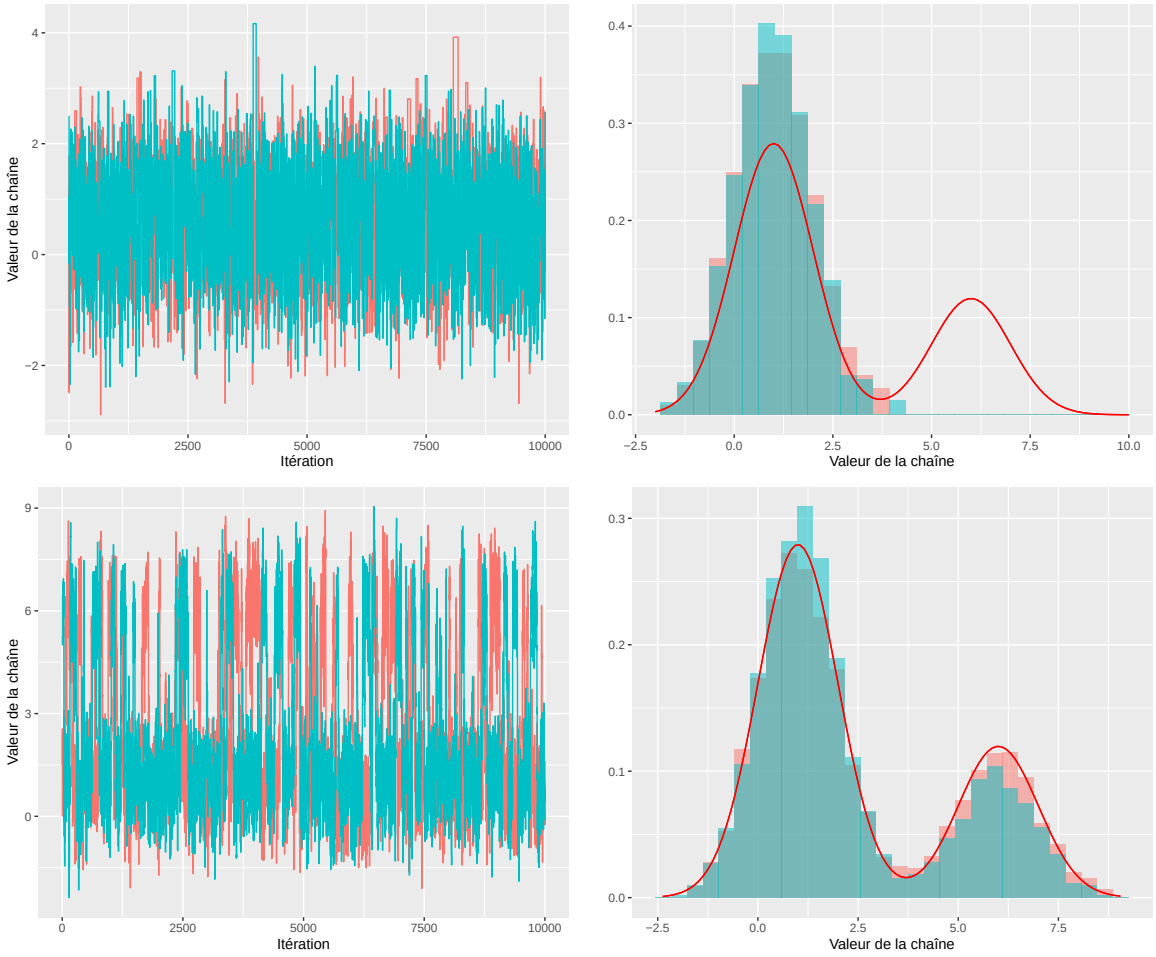


FIGURE 2.10 – Évolution de deux chaînes de Markov par algorithme de Metropolis-Hastings : en haut, avec une loi normale centrée réduite comme loi instrumentale et en bas avec une marche aléatoire gaussienne de variance 1 comme loi instrumentale. A gauche, on a représenté l'évolution de deux chaînes indépendantes parties de deux initialisations différentes, et à droite on a représenté l'histogramme des réalisations de la chaîne, en superposant la densité cible.

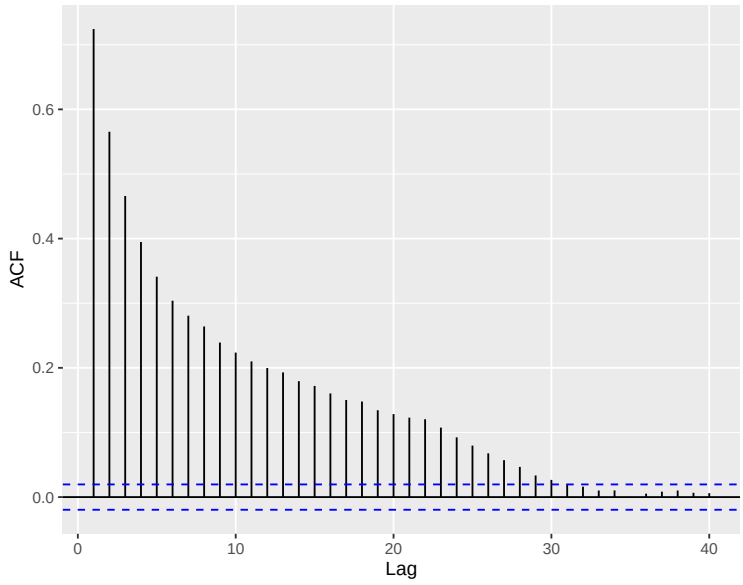


FIGURE 2.11 – Fonction d'autocorrélation

d'une partie du support de f.

3.3.2 Convergence en moyenne empirique

En application du théorème ergodique, si la chaîne a atteint son régime stationnaire, alors la moyenne empirique $\frac{1}{N}\sum_{n=1}^{N} h(X_n)$ devrait converger vers l'espérance théorique h_f. Cette convergence peut se voir graphiquement, en traçant l'évolution de la moyenne empirique en fonction des itérations et en vérifiant que celle-ci se stabilise bien autour d'une valeur unique.

Au-delà de cet outil diagnostic graphique il existe un critère, très utilisé en pratique et dû à Gelman et Rubin (1992). Ce critère repose sur la comparaison de M chaînes de Markov indépendantes $\{X_n^{(m)}\}, m = 1, \ldots, M$, obtenues en parallèle les unes des autres, et est parfois utilisé comme critère d'arrêt. Plus précisément, on s'intéresse à la variance entre chaînes B_N et à la variance inter chaînes W_N définies par :

$$B_N = \frac{1}{M-1}\sum_{m=1}^{M}\left(\bar{h}_m - \bar{h}\right)^2, \qquad W_N = \frac{1}{M-1}\sum_{m=1}^{M}\left(\frac{1}{N-1}\sum_{n=1}^{N}\left(h\left(X_n^{(m)}\right) - \bar{h}_m\right)^2\right),$$

où :

$$\bar{h}_m = \frac{1}{N-1}\sum_{n=1}^{N} h\left(X_n^{(m)}\right), \quad \bar{h} = \frac{1}{M}\sum_{m=1}^{M}\bar{h}_m.$$

On peut alors construire un estimateur de la variance de h_f en utilisant l'information contenue dans les M chaînes :

$$\hat{\sigma}_N^2 = \frac{N-1}{N}W_N + B_N. \tag{3.9}$$

Or, lorsque toutes les chaînes ont atteint le régime stationnaire, c'est-à-dire lorsque N tend vers l'infini, les deux quantités $\hat{\sigma}_N^2$ et W_N sont équivalentes (et B_N est nulle, car toutes les chaînes devraient

retourner la même estimation pour la variance). Gelman et Rubin proposent alors un critère basé sur la comparaison de ces deux quantités, et qui, dans le cas où la loi cible est approximativement gaussienne, suit une loi de Student à ν degrés de libertés. On a :

$$R_N^2 = \frac{\hat{\sigma}_N^2 + B_N/M}{W_N} \frac{\nu_N}{\nu_N - 2}, \tag{3.10}$$

où $\nu_N = 2(\hat{\sigma}_N^2 + \frac{B_N}{M})^2/W_N$ est un estimateur du degré de liberté de la loi de Student sous-jacente. La quantité R_N^2 suit approximativement une loi de Fisher $\mathcal{F}(1, \nu_N)$, et tend vers 1 lorsque N tend vers l'infini.

On peut alors en déduire un critère d'arrêt, ou un test de convergence, en étudiant l'évolution de R_N en fonction de N. Ce critère est implémenté par exemple dans le package `R coda`, qui propose également d'autre critères de convergence et des outils graphiques de diagnostic de convergence.

Troisième partie

Introduction aux statistiques bayésiennes

Introduction

Au point de départ de toute analyse statistique se trouve un échantillon $X = (X_1, \dots, X_n)$, que l'on suppose en général constitué de variables aléatoires i.i.d.. En statistique paramétrique, on suppose que la loi commune de l'échantillon est connue à un ensemble de paramètres près. C'est-à-dire que l'on suppose qu'elle appartient à une famille de lois indexées par un paramètre θ : on considère alors la famille de lois $\mathcal{P} = \{\mathbb{P}_\theta; \theta \in \Theta\}$. L'objectif de l'analyse statistique est d'obtenir de l'information sur θ à partir des observations $X_1, \dots, X_n$.

L'objet fondamental en statistique mathématique, qui nous permet de construire des estimateurs et des tests d'hypothèses ayant de bonnes propriétés, c'est la *fonction de vraisemblance*. Elle est au cœur des statistiques. Cette vraisemblance est définie comme *la loi jointe des observations*, et est vue comme une fonction de θ.

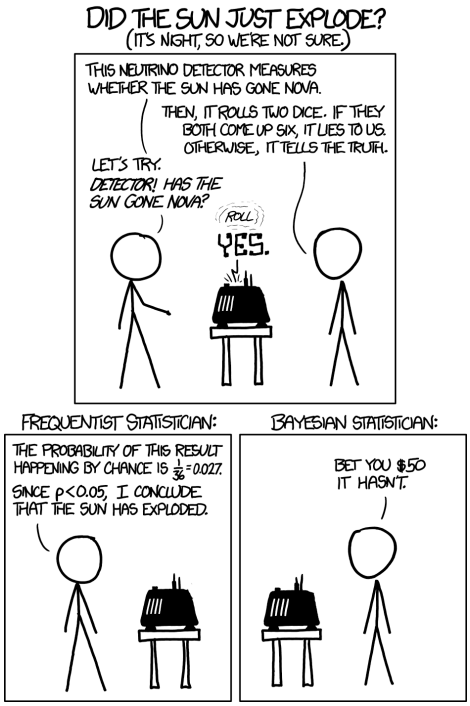


Figure 3.1 – Illustration (caricaturale ...) de la différence entre les approches fréquentiste et bayésienne. Source : xkcd

L'approche classique, ou plutôt historique, consiste à considérer θ comme une quantité déterministe, inconnue, que l'on cherche à approcher à l'aide d'une quantité aléatoire (un estimateur). On s'intéresse alors aux propriétés de cet estimateur aléatoire : quel est son biais, sa variance, sa distance à θ. Cette approche est celle qui est enseignée au premier semestre dans le cours de statistique mathématique. On l'appelle *l'approche fréquentiste*.

Une autre approche consiste à considérer θ comme étant lui-même une variable (ou un vecteur) aléatoire. L'objectif est alors d'obtenir de l'information sur la *distribution* de θ à partir des observations. C'est ce que l'on appelle *l'approche bayésienne*. L'approche bayésienne utilise aussi la fonction de vraisemblance, qui s'interprète alors comme la loi conditionnelle des observations sachant θ. On a cependant besoin d'introduire une autre quantité clé : la loi *a priori*, qui est définie comme la loi marginale de θ. L'objectif final de l'analyse bayésienne est d'obtenir de l'information sur la loi de θ conditionnellement aux observations : c'est ce que l'on appelle la *loi a posteriori*.

Chapitre 1

L'approche bayésienne

On suppose que l'on dispose d'un échantillon $\mathcal{X} = (X_1, \ldots, X_n)$ de variables aléatoires i.i.d. de loi $\mathbb{P}_\theta$, avec $\theta \in \Theta \subset \mathbb{R}^d$ et où $X_i \in E$. Le plus souvent, on aura $E = \mathbb{R}^p$ ou $E = \mathbb{N}^p$. Dans la suite, on supposera que $\mathbb{P}_\theta$ admet une densité par rapport à la mesure de référence sur $(E, \mathcal{E})$, c'est-à-dire par rapport à la mesure de Lebesgue sur $\mathbb{R}^p$ si $E = \mathbb{R}^p$ ou par rapport à la mesure de comptage si $E = \mathbb{N}^p$. On note alors $f(\cdot \mid \theta)$ cette densité.

1.1 Définitions et notations

Dans l'approche bayésienne, on munit l'ensemble Θ d'une mesure de probabilité, censée représenter, ou mesurer, l'incertitude entourant le paramètre θ. On supposera de plus que cette mesure de probabilité admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$.

Définition 21 (Loi a priori). *On appelle loi a priori toute mesure de probabilité π sur l'ensemble Θ.*

La définition d'un modèle bayésien repose donc sur le choix d'un modèle statistique pour l'échantillon, c'est-à-dire d'une loi $\mathbb{P}_\theta$ et d'une loi a priori pour θ. Dans le cadre bayésien, la loi $\mathbb{P}_\theta$ s'interprète comme la loi conditionnelle de X_i sachant θ. La vraisemblance s'interprète alors comme la loi conditionnelle de $\mathcal{X}$ sachant θ. À partir de ces deux lois, vraisemblance et loi a priori, on peut définir la loi jointe des observations et de θ :

$$\begin{aligned}(E^n \times \Theta, \otimes_{i=1}^n \mathcal{E} \times \mathcal{T}) &\to [0,1] \\ (x_1, \ldots, x_n, \theta) &\mapsto g(x_1, \ldots, x_n, \theta) = L(x_1, \ldots, x_n; \theta)\pi(\theta)\end{aligned}$$

Dans le cas où les observations sont i.i.d., la vraisemblance est donnée par

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n f(x_i \mid \theta).$$

Un modèle bayésien peut donc s'écrire sous forme hiérarchique, avec tout d'abord la spécification de

la loi conditionnelle des observations sachant θ, puis la spécification de la loi marginale de θ :

$$X_i \mid \theta \sim f(\cdot \mid \theta) \quad \textit{(vraisemblance)}$$
$$\theta \sim \pi \quad \textit{(loi a priori)}$$

Définition 22 (Hyperparamètres). *Les paramètres de la loi a priori sont appelés hyperparamètres. Ils sont considérés comme fixes et connus.*

Dans un modèle bayésien, la fonction de vraisemblance ne donne donc pas la loi *marginale* de $\mathcal{X}$, mais bien sa loi conditionnelle sachant θ. La loi marginale de $\mathcal{X}$ s'obtient en intégrant la loi jointe par rapport à θ. Elle admet une densité par rapport à la mesure de référence sur $(E^n, \otimes_{i=1}^n \mathcal{E})$ donnée par :

$$f_X(x_1, \ldots, x_n) = \int L(x_1, \ldots, x_n; \theta)\pi(\theta)d\theta$$

Une fois le modèle bayésien défini, on cherche, comme dans le cadre de l'analyse fréquentiste, à faire de l'inférence sur θ (estimation, tests, ...). Pour cela, on s'intéresse à la *loi a posteriori* de θ, c'est-à-dire la loi conditionnelle de θ sachant les observations $\mathcal{X}$.

Définition 23 (Loi a posteriori). *La loi a posteriori est la loi conditionnelle de θ sachant $\mathcal{X}$. On la note $p(\theta \mid \mathcal{X} = x)$, ou plus simplement par abus de notation et selon les contextes, $p(\theta \mid \mathcal{X})$ ou $p(\theta \mid x)$.*

La philosophie de l'analyse bayésienne est la suivante : la loi a priori π renferme notre connaissance a priori du phénomène étudié et de la façon dont peut être distribué le paramètre θ, la vraisemblance permet de modéliser le phénomène observé, et la loi a posteriori permet de "mettre à jour" nos connaissances sur θ à la lumière de l'expérience aléatoire et des observations récoltées. Le théorème de Bayes nous permet de calculer la loi a posteriori à partir de la vraisemblance et de la loi a priori :

Théorème 8. *La loi a posteriori admet une densité de probabilité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ qui est donnée par :*

$$p(\theta \mid x) = \frac{L(x_1, \ldots, x_n; \theta)\pi(\theta)}{f_X(x_1, \ldots, x_n)}, \tag{1.1}$$

où $x = (x_1, \ldots, x_n)$.

En pratique, il ne sera pas toujours nécessaire de calculer explicitement le dénominateur, c'est-à-dire la loi marginale de $\mathcal{X}$. En effet, cette quantité ne dépend pas de θ et correspond en fait à la constante de normalisation du numérateur. C'est la raison pour laquelle la loi a posteriori est souvent définie ou connue *à une constante de normalisation près*. On utilise alors la notation suivante pour désigner le fait que la loi a posteriori est proportionnelle à la loi jointe :

$$p(\theta \mid x) \propto L(x_1, \ldots, x_n; \theta)\pi(\theta).$$

Dans ce cas, toutes les méthodes vues dans les chapitres précédents et ne nécessitant qu'une connaissance à une constante multiplicative près des densités de probabilité prennent alors tout leur sens.

Exemple 10. *Considérons une pièce de monnaie dont on souhaite estimer la probabilité qu'elle tombe sur pile. Pour cela, on réalise n lancers indépendants et pour chaque lancer i, on note X_i la variable aléatoire qui vaut 1 si la pièce tombe sur pile, et 0 sinon. On note θ la probabilité d'obtenir pile. On a :*

$$X_i \mid \theta \sim \mathcal{B}(\theta)$$

Dans un cadre bayésien, on définit donc la loi conditionnelle de X_i sachant θ. On doit ensuite définir une loi a priori pour θ. Cette loi correspond à notre connaissance a priori du phénomène. Comme θ est une quantité variant entre 0 et 1, il nous faut définir une mesure de probabilité sur $[0,1]$. Plusieurs possibilités s'offrent à nous. Supposons que nous n'avons aucune information particulière sur cette pièce de monnaie. On va alors choisir un a priori dit non informatif, c'est-à-dire qui ne privilégie aucune région particulière de Θ. On peut par exemple prendre une loi uniforme sur $[0,1]$. On a alors :

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

Il nous reste à définir la loi a posteriori de θ. Commençons par déterminer la loi marginale de X. On peut tout d'abord remarquer que l'on ne perd aucune information si l'on considère la variable aléatoire $X = \sum_{i=1}^{n} X_i$. Le problème peut alors se reformuler de la façon suivante :

$$\begin{aligned} X \mid \theta &\sim \mathcal{B}(n,\theta) \\ \theta &\sim \mathcal{U}([0,1]) \end{aligned}$$

La vraisemblance est donnée par $L(x;p) = \mathbb{P}(X = x \mid \theta = t) = \binom{n}{x} t^x (1-t)^{n-x}$ et la loi marginale de X est donnée par :

$$\begin{aligned} \mathbb{P}(X = x) &= \int \mathbb{P}(X = x \mid \theta = t)\mathbf{1}_{[0,1]}(t)dt \\ &= \int \binom{n}{x} t^x (1-t)^{n-x}\mathbf{1}_{[0,1]}(t)dt \\ &= \int \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} t^x (1-t)^{n-x}\mathbf{1}_{[0,1]}(t) \\ &= \frac{1}{n+1}\int \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} t^x (1-t)^{n-x}\mathbf{1}_{[0,1]}(t) \\ &= \frac{1}{n+1}. \end{aligned}$$

On reconnaît en effet l'intégrale de la densité d'une loi Beta$(x+1, n-x+1)$. On a alors la loi a posteriori suivante :

$$p(t \mid X = x) = \frac{\mathbb{P}(X = x \mid \theta = t)\mathbf{1}_{[0,1]}(t)}{\mathbb{P}(X = x)}$$

$$
\begin{aligned}
&= \frac{\frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} t^x (1-t)^{n-x} \mathbf{1}_{[0,1]}(t)}{\frac{1}{n+1}} \\
&= \frac{(n+1)\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} t^x (1-t)^{n-x} \mathbf{1}_{[0,1]}(t) \\
&= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} t^x (1-t)^{n-x} \mathbf{1}_{[0,1]}(t)
\end{aligned}
$$

La loi a posteriori de $\theta \mid X$ est donc une loi Beta$(X+1, n-X+1)$. Comment interpréter ce résultat ? Sachant que $X = x$, c'est-à-dire sachant que le nombre de pièces qui sont tombées sur pile est x (et donc $n - x$ est le nombre de pièces qui sont tombées sur face), la probabilité p de tomber sur pile suit une loi Beta$(x+1, n-x+1)$. De plus, la loi uniforme sur $[0,1]$ est aussi la loi Beta$(1,1)$. Autrement dit, la loi a priori de θ est aussi une loi Beta, et les paramètres correspondants sont mis à jour dans la loi a posteriori en fonction du nombre de pièces tombant sur pile ou face. Rappelons que le mode d'une loi Beta(a,b) est donné par $\frac{a-1}{a+b-2}$. Le mode de la loi a posteriori est donc $\frac{X}{n} = \frac{1}{n}\sum_{i=1}^n X_i$, c'est-à-dire la proportion empirique de pièces tombant sur pile.

La figure 3.2 illustre l'influence des observations sur la loi a posteriori. Le mode de la loi a posteriori est localisé en $\frac{1}{n}\sum_{i=1}^n X_i$, la proportion empirique du nombre de "piles" sur les n lancers. À titre d'illustration, on a comparé deux tailles d'échantillon : $n = 10$ et $n = 100$.

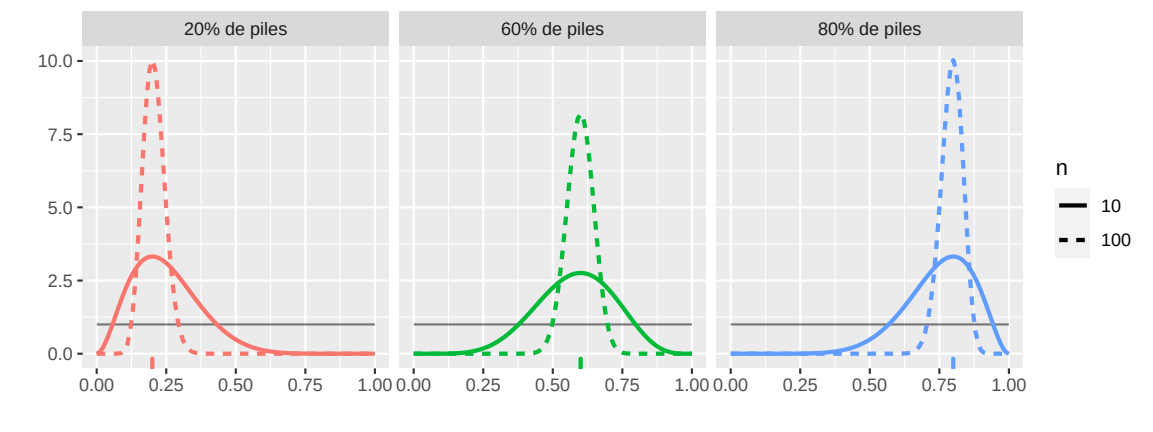


FIGURE 3.2 – Densités des lois a posteriori dans le modèle beta-binomial en fonction du nombre de "pile" observés sur $n = 10$ lancers. Les traits sur l'axe des abscisses correspondent à la proportion empirique de piles sur les 10 lancers (i.e. $\frac{1}{n}\sum_{i=1}^n X_i$), et la densité de la loi a priori est représentée en gris.

Le cadre bayésien fournit également les outils pour prédire la valeur d'une variable aléatoire, une fois l'inférence effectuée. Ceci s'avère particulièrement utile en régression, par exemple. Supposons par exemple que l'on souhaite prédire la valeur d'une variable aléatoire Y qui dépend de X à travers sa loi (dont on suppose pour simplifier l'écriture qu'elle admet la densité $g(y \mid \theta, x)$). On peut alors définir la *distribution prédictive* de Y par :

$$
g(y \mid x) = \int g(y \mid \theta, x) p(\theta \mid x) d\theta. \tag{1.2}
$$

1.2 Inférence à partir de la loi a posteriori

1.2.1 Estimateurs ponctuels

Dans le paradigme bayésien, la quantité d'intérêt que l'on cherche à estimer, c'est la loi a posteriori. Cependant, on peut être amené à considérer des estimateurs ponctuels de θ, définis à l'aide de cette loi a posteriori. On peut citer notamment :

- la *moyenne a posteriori*, définie comme l'espérance de la loi a posteriori : $\int_\Theta \theta\, p(\theta \mid x) d\theta$
- la *médiane a posteriori*, définie comme la médiane de la loi a posteriori. Plus généralement, on peut définir les quantiles de la loi a posteriori (en utilisant si besoin l'inverse généralisée de la fonction de répartition de la loi a posteriori)
- le *maximum a posteriori* (MAP), défini comme le mode de la loi a posteriori : $\tilde{\theta} = \arg\max_\theta p(\theta \mid x)$

Ces estimateurs ponctuels sont parfois utilisés pour résumer l'information contenue dans la loi a posteriori. Nous verrons dans la section 1.3 quelques justifications théoriques pour le choix de ces estimateurs ponctuels. Le MAP correspond au maximum de la loi jointe, et ne nécessite pas de connaître ou de calculer la loi marginale f_X. Comme la loi jointe s'écrit comme le produit de la vraisemblance et de la loi a priori, le MAP peut se voir comme un maximum de vraisemblance pénalisé, où le terme de pénalisation correspond à la loi a priori. Les propriétés asymptotiques de l'estimateur du maximum de vraisemblance, en particulier la consistance et la normalité asymptotique, sont préservées pour l'estimateur du MAP. En fait, on peut montrer que lorsque la taille de l'échantillon n tend vers l'infini, le MAP converge vers l'estimateur du maximum de vraisemblance. En effet, dans ce cas l'information apportée par l'échantillon devient alors prépondérante par rapport à l'information apportée par la loi a priori, et dans ce cas c'est la vraisemblance qui a le plus de poids dans le calcul de la loi jointe. Si le MAP est asymptotiquement équivalent à l'estimateur du maximum de vraisemblance, il a l'avantage d'être défini et calculable pour toute taille d'échantillon n.

1.2.2 Région de crédibilité

La loi a posteriori peut également être utilisée pour fournir des régions de crédibilité dont la définition est donnée ci-dessous.

Définition 24 (Région de crédibilité). *Une région de crédibilité de niveau* $1-\alpha$ *pour la loi a posteriori* $p(\theta \mid x)$ *est un ensemble aléatoire* R *tel que :*

$$\int_R p(\theta \mid x) d\theta \geq 1 - \alpha$$

En fréquentiste, une région de confiance de niveau $1-\alpha$ est un ensemble aléatoire qui a une probabilité $1-\alpha$ de contenir la (vraie) valeur inconnue du paramètre θ que l'on cherche à estimer.

En bayésien, une région de crédibilité de niveau $1-\alpha$ est un ensemble aléatoire tel que la probabilité pour que θ prenne ses valeurs dans cet ensemble est égale à $1-\alpha$.

En effet, en bayésien, θ est une variable aléatoire, on peut donc lui associer la probabilité d'appartenir à un ensemble donné, alors qu'en fréquentiste θ est déterministe : la valeur inconnue est dans la région de confiance ou n'y est pas, mais cela n'a pas de sens de calculer la probabilité pour qu'une quantité déterministe appartienne à un ensemble donné.

Il y a plusieurs façons de construire des régions de crédibilité, les deux plus courantes étant celle basée sur les quantiles a posteriori et celle basée sur la construction de la région de plus forte densité a posteriori.

Méthode basée sur les quantiles. Pour simplifier la présentation, on suppose que $\theta \in \mathbb{R}$, mais les résultats se généralisent facilement au cas $\mathbb{R}^d$. On cherche alors à construire un *intervalle de crédibilité*. Notons $F_{\theta|x}$ la fonction de répartition associée à la loi a posteriori $p(\theta \mid x)$, et q_α^{post} le quantile d'ordre α associé à la loi a posteriori, autrement dit, $q_\alpha^{post} = F^-_{\theta|x}(\alpha)$ (inverse généralisée dans le cas général, inverse au sens classique si la loi a posteriori admet une densité par rapport à la mesure de Lebesgue). On peut proposer par exemple l'intervalle symétrique suivant :

$$I_{\text{quant}}(\alpha) = \left[q_{\alpha/2}^{post}; q_{1-\alpha/2}^{post}\right] \tag{1.3}$$

On peut également construire des intervalles de crédibilité non symétrique avec la même approche, en choisissant des quantiles d'ordre différent.

Highest Posterior Density (HPD) - région de plus forte densité a posteriori. La méthode précédente basée sur les quantiles est très simple à mettre en œuvre, mais elle ne fournit pas nécessairement la région de crédibilité de niveau $1-\alpha$ qui soit la plus petite possible. Ceci est particulièrement vrai si la loi a posteriori est asymétrique. Or, comme en fréquentiste, on peut chercher à minimiser la taille de la région de crédibilité, pour un niveau $1-\alpha$ fixé.

Définition 25 (Région de plus forte densité a posteriori). *Une région de plus forte densité a posteriori (ou HPD), de niveau $1-\alpha$ est une région R définie par :*

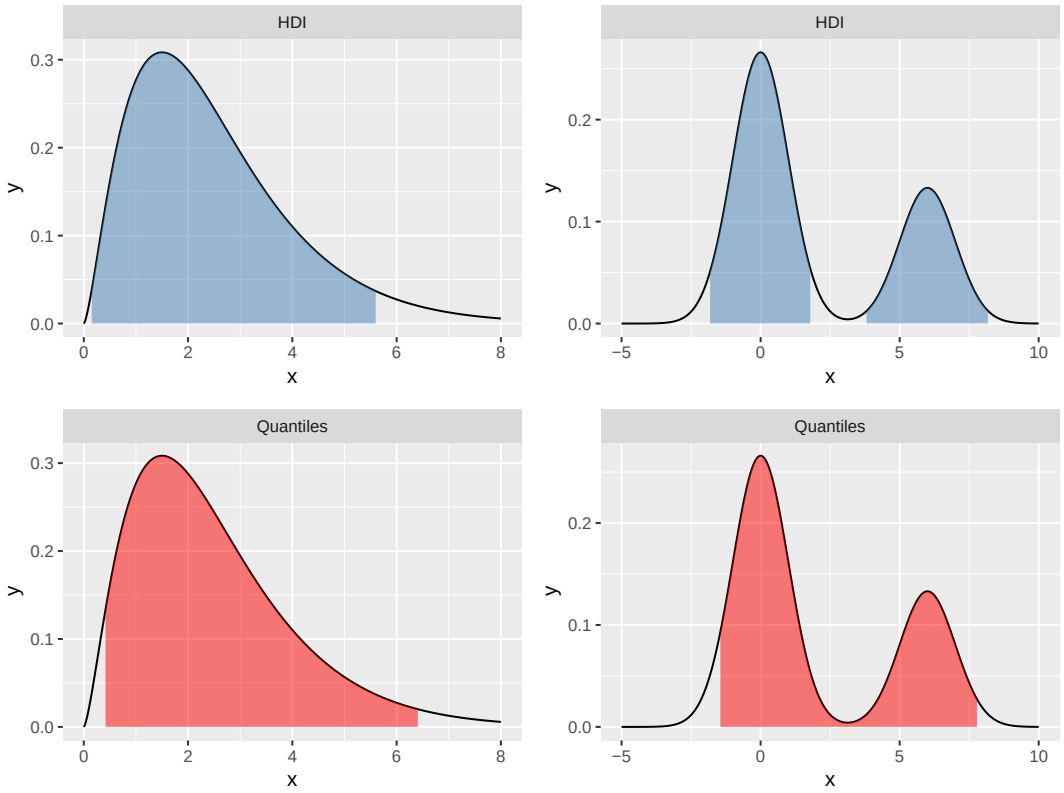
$$R = \{\theta; p(\theta \mid x) \geq k\},$$

où k est le plus grand nombre tel que :

$$\int_{\theta; p(\theta|x)\geq k} p(\theta \mid x) d\theta = 1-\alpha.$$

Si la loi a posteriori est unimodale et si $\theta \in \mathbb{R}$, la région de plus forte densité est un intervalle. Si la loi est multimodale, on peut se retrouver avec des régions de confiance définies comme des unions d'ensemble. Une illustration est donnée sur la figure 3.3.

Cette méthode fournit les régions de crédibilité de plus petite taille pour un niveau $1-\alpha$ donné, mais a l'inconvénient d'être parfois difficile à calculer en pratique.



(a) *loi unimodale* (b) *Loi bimodale*

FIGURE 3.3 – Illustration de la différence entre intervalle de crédibilité basé sur la région de plus forte densité (en haut) et sur la méthode des quantiles (en bas).

1.3 Lien avec la théorie de la décision

L'objectif de cette section est de proposer un cadre théorique pour valider ou justifier le choix des estimateurs bayésiens. L'analyse bayésienne peut en effet s'inscrire dans un cadre plus large qui est celui de la théorie de la décision. Dans certains cas, en fonction des critères de décision que l'on se fixe, on peut montrer que ce sont les estimateurs issus d'une analyse bayésienne qui sont optimaux.

1.3.1 Cadre général

On dispose souvent, que ce soit dans le cadre fréquentiste ou bayésien, de plusieurs estimateurs pour un paramètre donné. On peut alors se demander comment choisir le "meilleur" estimateur, en ayant pris soin de définir ce que l'on entend par "meilleur". Un critère que l'on a déjà rencontré pour comparer des estimateurs est le risque quadratique. De façon générale, le cadre de la théorie de la décision consiste à se fixer ce que l'on appelle une fonction de perte, à partir de laquelle on pourra définir un critère à minimiser pour identifier le meilleur estimateur au sens de la fonction de perte préalablement définie.

Définition 26 (Fonction de perte). *Une fonction de perte est une fonction mesurable* $\ell : \Theta \times \Theta \to \mathbb{R}^+$ *telle que* $\ell(\theta, \theta') = 0$ *si et seulement si* $\theta = \theta'$.

Les exemples les plus classiques de fonction de perte sont :

- la fonction de *perte quadratique* : $\ell(\theta, \theta') = (\theta - \theta')^2$
- la fonction de *perte en valeur absolue* : $\ell(\theta, \theta') = |\theta - \theta'|$
- la fonction de *perte 0-1* : $\ell(\theta, \theta') = \mathbf{1}_{\theta=\theta'}$. Celle-ci est surtout utilisée en classification, pour des observations binaires.

À partir de la fonction de perte, on peut définir la fonction de risque pour un estimateur $\hat{\theta}$.

Définition 27 (Fonction de risque). *La fonction de risque de l'estimateur* $\hat{\theta} = T(\mathcal{X})$ *pour la fonction de perte* ℓ *est donnée par :*

$$\begin{aligned} R : \Theta &\to \mathbb{R}^+ \\ \theta &\mapsto R(\theta, T(\mathcal{X})) = \mathbb{E}_\theta(\ell(\theta, T(\mathcal{X}))) = \int_E \ell(\theta, T(x)) dP_\theta(x), \end{aligned}$$

où P_θ *est la loi jointe de l'échantillon.*

Autrement dit, la fonction de risque correspond à la perte moyenne engendrée par le choix de l'estimateur $T(\mathcal{X})$ pour approcher θ.

1.3.2 Critères de décision

Un critère simple et intuitif pour comparer des estimateurs entre eux sur la base de la fonction de risque est le suivant : si on peut trouver un estimateur $\hat{\theta}_2$ qui est toujours meilleur (au sens où le risque est toujours plus faible) que l'estimateur $\hat{\theta}_2$, quelque soit la valeur de θ, alors on peut d'ores et déjà éliminer $\hat{\theta}_1$.

Définition 28. *Un estimateur* $T(\mathcal{X})$ *est dit inadmissible s'il existe un estimateur* $\tilde{T}(\mathcal{X})$ *tel que :*

$$\begin{aligned} \forall \theta \in \Theta, \quad & R(\theta, \tilde{T}(\mathcal{X})) \leq R(\theta, T(\mathcal{X})) \\ \exists \theta' \in \Theta, \quad & R(\theta', \tilde{T}(\mathcal{X}) < R(\theta', T(\mathcal{X})) \end{aligned}$$

Un estimateur est dit admissible s'il n'est pas inadmissible.

Même si la notion d'admissibilité paraît judicieuse, en pratique beaucoup d'estimateurs sans intérêt particulier sont admissibles. C'est le cas notamment des estimateurs constants. Si par exemple la loi $\mathbb{P}_\theta$ est absolument continue par rapport à la mesure de Lebesgue, alors l'estimateur $T(\mathcal{X}) \equiv \theta_0$, pour

un $\theta_0 \in \Theta$ est admissible. On va alors introduire deux autres notions de risque, que l'on cherchera à minimiser : le risque minimax et le risque de Bayes.

Risque et estimateurs minimax.

Définition 29 (Risque maximal). *Le risque maximal d'un estimateur $T(\mathcal{X})$ est défini par :*

$$R_{\max}(T(\mathcal{X})) = \sup_{\theta \in \Theta} R(\theta, T(\mathcal{X})).$$

On cherche ensuite à minimiser ce risque maximal, ce qui nous conduit à introduire la notion de risque minimax.

Définition 30 (Risque et estimateur minimax). *Le risque minimax est défini comme le minimum des risques maximaux :*

$$R_M = \inf_{T(\mathcal{X})} R_{\max}(T(\mathcal{X})).$$

Un estimateur $T(\mathcal{X})$ est dit minimax si son risque maximal est égal au risque minimax, i.e. si

$$R_{\max}(T(\mathcal{X})) = R_M.$$

Un estimateur minimax est un estimateur qui minimise le pire risque possible.

Risque et estimateurs de Bayes.
Pour définir le risque de Bayes, on besoin de se placer dans un cadre bayésien, c'est-à-dire de définir une loi a priori π sur Θ.

Définition 31 (Risque de Bayes). *Le risque de Bayes d'un estimateur $T(\mathcal{X})$ pour la loi a priori π est défini par :*

$$R_B(\pi, T(\mathcal{X})) = \int_\Theta R(\theta, T(\mathcal{X}))\pi(\theta)d\theta$$

On s'intéresse donc au risque moyen de l'estimateur selon la distribution a priori de θ. On peut ré-écrire ce risque sous la forme :

$$\begin{aligned} R_B(\pi, T(\mathcal{X})) &= \int_\Theta \left(\int_E \ell(\theta, T(x)) dP_\theta(x) \right) \pi(\theta) d\theta \\ &= \int_\Theta \int_E \ell(\theta, T(x))\ dP_\theta(x)\pi(\theta)d\theta \\ &= \mathbb{E}_{(\mathcal{X},\theta)}(\ell(\theta, T(\mathcal{X}))) \end{aligned}$$

Le risque de Bayes s'interprète donc comme l'espérance de la fonction de perte par rapport à la loi jointe du couple $(\mathcal{X}, \theta)$.

Définition 32 (Estimateur de Bayes). *Le risque de Bayes associé à la loi a priori π est donné par :*

$$R_B(\pi) = \inf_{T(\mathcal{X})} R_B(\pi, T(\mathcal{X})).$$

Un estimateur $T(\mathcal{X})$ est dit de Bayes si son risque est égal au risque minimal ci-dessus, i.e. si :

$$R_B(\pi, T(\mathcal{X})) = R_B(\pi).$$

1.3.3 Construction d'estimateurs de Bayes

L'avantage du risque de Bayes est qu'il est souvent possible de donner une expression ou une construction explicite pour les estimateurs de Bayes associés à une loi a priori et à une fonction de perte données. On mentionne ici les cas les plus courants. On définit tout d'abord le risque a posteriori.

Définition 33 (Risque a posteriori). *Le risque a posteriori d'un estimateur $T(\mathcal{X})$ pour la loi a priori π est défini par :*

$$\rho(\pi, T(\mathcal{X}) \mid \mathcal{X} = x) = \int_\Theta \ell(\theta, T(x)) p(\theta \mid x) d\theta.$$

Il s'agit donc de l'espérance du risque par rapport à la loi a posteriori. On notera plus simplement $\rho(\pi, T(\mathcal{X}) \mid \mathcal{X})$.

Le risque a posteriori *dépend des observations* : on raisonne conditionnellement aux données. Le théorème suivant nous permet de faire le lien entre risque a posteriori et estimateur de Bayes.

Théorème 9. *L'estimateur qui minimise, pour chaque valeur $\mathcal{X}$, le risque a posteriori, est un estimateur de Bayes. Autrement dit, l'estimateur suivant :*

$$T^*_\pi(\mathcal{X}) = \arg\min_T \rho(\pi, T(\mathcal{X}) \mid \mathcal{X})$$

est un estimateur de Bayes.

Démonstration. La preuve repose sur le théorème de Fubini. Supposons pour simplifier l'écriture que dP_θ admet une densité par rapport à la mesure de référence sur E. On a alors :

$$\begin{aligned}
R_B(\pi, T(\mathcal{X})) &= \int_\Theta \int_E \ell(\theta, T(x)) f(x \mid \theta) dx \, \pi(\theta) d\theta \\
&= \int_E \int_\Theta \ell(\theta, T(x)) f(x \mid \theta) \pi(\theta) d\theta dx \\
&= \int_E \int_\Theta \ell(\theta, T(x)) \frac{f(x \mid \theta)\pi(\theta)}{f_X(x)} d\theta \, f_X(x) dx \\
&= \int_E \int_\Theta \ell(\theta, T(x)) p(\theta \mid x) d\theta \, f_X(x) dx \\
&= \int_E \rho(\pi, T(\mathcal{X}) \mid \mathcal{X} = x) \, f_X(x) dx
\end{aligned}$$

Par définition de T_π^*, on a $\rho(\pi, T(\mathcal{X}) \mid \mathcal{X} = x) \geq \rho(\pi, T_\pi^*(\mathcal{X}) \mid \mathcal{X} = x)$ pour tout estimateur T. Donc on a aussi :

$$R_B(\pi, T(\mathcal{X})) \geq R_B(\pi, T_\pi^*(\mathcal{X})).$$

□

Ce résultat est fondamental, car il nous permet en pratique de minimiser le risque a posteriori pour trouver des estimateurs de Bayes. Ce calcul est souvent plus facile, car on se retrouve avec une intégrale sur Θ plutôt que sur $\Theta \times E$.

Estimateur de Bayes pour la perte quadratique.
L'estimateur de Bayes pour la fonction de perte quadratique, associé à une loi a priori π telle que $\int_\Theta \theta^2 \pi(\theta) d\theta < \infty$, est l'estimateur de la moyenne a posteriori :

$$T_\pi^*(\mathcal{X}) = \mathbb{E}(\theta \mid X) = \int_\Theta \theta \, p(\theta \mid X) \, d\theta$$

Estimateur de Bayes pour la perte en valeur absolue.
L'estimateur de Bayes pour la fonction de perte en valeur absolue, associé à une loi a priori π telle que $\int_\Theta |\theta| \pi(\theta) d\theta < \infty$, est l'estimateur de la médiane a posteriori.

Remarque. *L'estimateur du mode a posteriori ne peut pas s'exprimer comme la solution d'un problème de minimisation d'un risque.*

1.3.4 Principaux résultats

Dans cette section, on présente les principaux résultats reliant les différentes notions vues dans les sections précédentes.

Théorème 10. *Pour toute loi a priori π, on a $R_B(\pi) \leq R_M$*

Démonstration. Par définition du risque de Bayes associée à une loi a priori π, on a

$$R_B(\pi) = \inf_{T(X)} \int_\Theta R(\theta, T(\mathcal{X})) \pi(\theta) d\theta.$$

Or :

$$\begin{aligned}
\int_\Theta R(\theta, T(\mathcal{X})) \pi(\theta) d\theta &\leq \sup_\theta R(\theta, T(\mathcal{X})) \int_\Theta \pi(\theta) d\theta \\
&\leq \sup_\theta R(\theta, T(\mathcal{X})) \\
&\leq R_{\max}(T(\mathcal{X}))
\end{aligned}$$

En prenant l'infimum de chaque côté de l'inégalité, on obtient le résultat cherché. □

Le risque de Bayes est donc toujours plus petit que le risque minimax. En effet, le risque minimax peut être vu comme un risque "pessimiste", au sens où il cherche à minimiser le risque obtenu dans le pire des cas. Dans les résultats suivants, on explore l'admissibilité des estimateurs de Bayes.

Théorème 11 (Admissibilité des estimateurs de Bayes). *Soit $T(\mathcal{X})$ un estimateur de Bayes pour une loi a priori π, unique à équivalence près (deux estimateurs sont dits équivalents si leurs fonctions de risque sont égales en tous points). Alors $T(\mathcal{X})$ est admissible.*

Démonstration. Supposons par l'absurde que $T(\mathcal{X})$ est inadmissible. Alors il existe un autre estimateur $T'(\mathcal{X})$ tel que $R(\theta, T'(\mathcal{X})) \leq R(\theta, T(\mathcal{X})$ pour tout $\theta \in \Theta$, et il existe un $\theta' \in \Theta$ tel que $R(\theta, T'(\mathcal{X}) < R(\theta, T(\mathcal{X}))$. On a alors :

$$R_B(\pi, T'(\mathcal{X})) = \int R(\theta, T'(\mathcal{X}))\pi(\theta)d\theta \leq \int R(\theta, T(\mathcal{X}))\pi(\theta)d\theta = R_B(\theta, T(\mathcal{X}))$$

Or $T(\mathcal{X})$ est un estimateur de Bayes, donc $R_B(\pi, T(\mathcal{X})) = R_B(\pi)$, i.e. il atteint l'infimum du risque de Bayes. Donc $T'(\mathcal{X})$ est aussi un estimateur de Bayes. Or on a supposé $T(\mathcal{X})$ unique à équivalence près. On a donc nécessairement $R(\theta, T(\mathcal{X})) = R(\theta, T'(\mathcal{X}))$ pour tout $\theta \in \Theta$, ce qui contredit l'existence d'un θ' tel que le risque soit strictement plus faible pour T' par rapport à T. □

Une condition suffisante pour que deux estimateurs T et T' soient équivalents pour la fonction de perte quadratique est que la loi marginale de $\mathcal{X}$ domine toutes les lois conditionnelles de $\mathcal{X}$ sachant θ. Faisons maintenant le lien entre estimateur admissible et estimateur minimax.

Théorème 12. *Un estimateur $T(\mathcal{X})$ admissible et de risque constant est un estimateur minimax.*

Démonstration. Comme T est de risque constant, on a $R(\theta, T(\mathcal{X})) = R_{\max}(T(\mathcal{X}))$ pour tout $\theta \in \Theta$. Supposons alors que T n'est pas minimax. Cela signifie qu'il existe un autre estimateur T' tel que $R_{\max}(T'(\mathcal{X})) < R_{\max}(T(\mathcal{X}))$. En particulier, cela implique :

$$R(\theta, T'(\mathcal{X})) \leq R(\theta, T(\mathcal{X})), \quad \forall \theta \in \Theta$$

Mais dans ce cas, on contredit l'hypothèse selon laquelle $T(\mathcal{X})$ est admissible. □

Enfin, on fait le lien entre estimateur de Bayes et estimateur minimax.

Théorème 13. *Soit $T(\mathcal{X})$ un estimateur de Bayes pour une loi a priori π, dont la fonction de risque est majorée par son risque de Bayes, i.e. tel que :*

$$\forall \theta \in \Theta, \; R(\theta, T(\mathcal{X})) \leq R_B(\pi, T(\mathcal{X})).$$

Alors $T(\mathcal{X})$ est un estimateur minimax.

Démonstration. Là encore, on raisonne par l'absurde. Supposons que $T(\mathcal{X})$ ne soit pas minimax. On peut alors trouver un estimateur $T'(\mathcal{X})$ tel que :

$$R_{\max}(T'(\mathcal{X})) < R_{\max}(T(\mathcal{X})).$$

Or par hypothèse, on a $R_{\max}(T(\mathcal{X})) \leq R_B(\pi, T(\mathcal{X}))$. De plus, on a montré (voir preuve du théorème 10) que $R_B(\pi, T(\mathcal{X})) \leq R_{\max}(T(\mathcal{X}))$. Donc on a :

$$R_B(\pi, T'(\mathcal{X})) \leq R_{\max}(T'(\mathcal{X})) < R_{\max}(T(\mathcal{X})) \leq R_B(\pi, T(\mathcal{X})) \tag{1.4}$$

Ceci contredit l'hypothèse selon laquelle $T(\mathcal{X})$ est un estimateur de Bayes. □

Exemple 11. *On se place dans le modèle de l'exemple 10. On suppose $X \mid \theta \sim \mathcal{B}(n, \theta)$ et on choisit comme loi a priori pour θ la loi uniforme sur $[0, 1]$. On a montré que la loi a posteriori, donc la loi de $\theta \mid X$ est une loi Beta$(X + 1, n - X + 1)$. Considérons la fonction de perte quadratique, i.e. $\ell(\theta, \theta') = (\theta - \theta')^2$. On a vu que l'estimateur de la moyenne a posteriori était un estimateur de Bayes associé à ℓ. On peut donc proposer :*

$$\tilde{\theta} = \frac{X + 1}{n + 2}.$$

Calculons le risque de cet estimateur :

$$\begin{aligned}
R(\theta, \tilde{\theta}) &= \mathbb{E}(\ell(\theta, \tilde{\theta})) \quad \textit{(espérance par rapport à la loi conditionnelle de } X \textit{ sachant } \theta) \\
&= \mathbb{E}(\theta^2 - 2\theta\tilde{\theta} + \tilde{\theta}^2) \\
&= \theta^2 - 2\theta\,\mathbb{E}\left(\frac{X + 1}{n + 2}\right) + \mathbb{E}\left[\left(\frac{X + 1}{n + 2}\right)^2\right] \\
&= \theta^2 - \frac{2\theta(n\theta + 1)}{n + 2} + \frac{n\theta(1 - \theta) + n^2\theta^2 + 2n\theta + 1}{(n + 2)^2} \\
&= \frac{(n - 4)\theta(1 - \theta) + 1}{(n + 2)^2}
\end{aligned}$$

Le risque maximal de $\tilde{\theta}$ est défini par :

$$\begin{aligned}
R_{\max}(\tilde{\theta}) &= \sup_{\theta \in [0,1]} R(\theta, \tilde{\theta}) \\
&= \frac{n}{4(n + 2)^2}
\end{aligned}$$

Son risque de Bayes, pour la loi a priori uniforme sur $[0, 1]$, est donné par :

$$R_B(\pi, \tilde{\theta}) = \int R(\theta, \tilde{\theta})\pi(\theta)d\theta = \int_0^1 \frac{(n - 4)\theta(1 - \theta) + 1}{(n + 2)^2}\, d\theta = \frac{1}{6(n + 2)}$$

Chapitre 2

Choix de la loi a priori

Le choix d'une loi a priori est au cœur de la définition du modèle bayésien. Ce choix n'est donc pas anodin et ne doit pas être pris à la légère. Certains choix permettent de faciliter le calcul de la loi a posteriori, c'est le cas notamment des loi a priori dites *conjuguées*, d'autres ont l'avantage d'être invariantes par changement de paramétrisation. C'est souvent sur ce choix de la loi a priori que se concentrent l'essentiel des critiques contre l'approche bayésienne, car de ce choix découle toute l'inférence. L'influence de la loi a priori peut, selon les cas, se révéler faible ou au contraire majeure. En pratique il est toujours possible de choisir la loi a priori de façon à influencer et orienter les résultats dans la direction voulue. Pour cette raison, il est important que le choix de la loi a priori soit correctement justifié. Ce choix ne doit pas être uniquement motivé par des raisons pratiques de faisabilité des calculs ou de simplification des estimateurs, même si cela peut rentrer en compte dans certains cas.

2.1 Prise en compte de l'information a priori

Une première source d'information que l'on peut utiliser pour construire la loi a priori d'un modèle bayésien est celle de notre connaissance du phénomène étudié. On peut savoir par exemple que le paramètre à estimer est positif, ou compris entre 0 et 1, … ces contraintes peuvent être prises en compte dans la loi a priori. Lorsque les paramètres du modèle statistiques ont une signification physique, biologique, médicale, … il est également possible de tenir compte de la connaissance provenant d'études antérieures ayant traité du même sujet. Ceci peut permettre de construire des loi a priori "subjectives" qui prennent en compte l'information dont on dispose en amont de l'expérience aléatoire. En pratique, tenir compte de l'avis d'experts pour le choix d'une loi a priori n'est pas sans danger … on peut notamment mentionner l'existence de biais cognitifs, ou la difficulté à prendre en compte des avis d'experts divergents.

Au-delà de ces considérations subjectives, il existe d'autres critères plus "objectifs" permettant de choisir une loi a priori, qui sont abordés dans les sections suivantes.

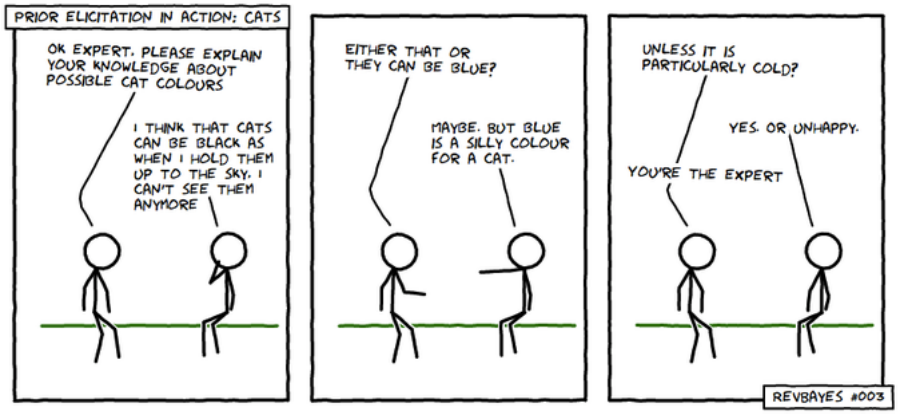


FIGURE 3.4 – De la difficulté de construire une loi a priori à partir de dires d'experts ... (source : RevBayes)

2.2 Lois a priori impropres

Il peut arriver que l'information dont on dispose a priori, ou même que des critères théoriques et plus objectifs conduisent à considérer une mesure sur Θ qui n'est pas une mesure de probabilité, et qui n'est même pas de masse totale finie, i.e. telle que :

$$\int_\Theta \pi(\theta)d\theta = +\infty.$$

Ce choix peut être justifié par des raisons subjectives. Prenons par exemple le cas d'un échantillon i.i.d. dont la loi commune est donnée par $f(x \mid \theta) = f(x-\theta)$. Autrement dit, θ est un paramètre de localisation. Si on ne dispose d'aucune information a priori , il peut sembler raisonnable de ne privilégier aucune région de l'espace et de choisir un priori qui soit proportionnel à la mesure de Lebesgue sur $\mathbb{R}$. Parfois, ce choix de loi impropre s'avère le seul raisonnable ou justifiable pour obtenir un a priori non informatif (voir aussi Section 2.4). De plus, même si cela peut s'avérer contre-intuitif, il est tout à fait possible que la loi a posteriori associée à un a priori impropre soit une mesure de probabilité. Dans ce cas, l'utilisation de ce type d'a priori n'empêche pas de conduire une analyse bayésienne. Dans ce cours, on considérera que l'utilisation d'un a priori impropre ne peut se faire qu'à condition que la loi a posteriori associée soit une mesure de probabilité.

EXEMPLE 12. *Soit X une variable aléatoire gaussienne de moyenne inconnue θ et de variance connue et égale à 1. On considère un a priori impropre pour θ, sous la forme $\pi(\theta) = 1$ pour $\theta \in \mathbb{R}$. La vraisemblance est donnée par :*

$$L(x;\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right),$$

la pseudo-loi marginale (ce n'est pas une mesure de probabilité) est donnée par :

$$f_X(x) = \int_{\mathbb{R}} L(x,\theta)\pi(\theta)d\theta = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right) d\theta = 1,$$

et la loi a posteriori de θ est donnée par :

$$p(\theta \mid x) = L(x;\theta)\pi(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\theta - x)^2\right).$$

La loi a posteriori est donc une loi normale $\mathcal{N}(X, 1)$. On aurait même pu définir la loi a priori comme étant égale à une constante c quelconque, sans que cela n'affecte les résultats.

2.3 Lois conjuguées

Les familles de lois conjuguées permettent d'obtenir des expressions explicites pour la loi a posteriori, qui appartiennent alors à la même famille que la loi a priori. Leur popularité tient donc au fait qu'elles permettent de simplifier les calculs, rendant les lois a posteriori explicites. Avant l'essor des outils de calculs numériques, elles constituaient souvent l'unique possibilité de réaliser une analyse bayésienne.

Un autre argument avancé en leur faveur est celui d'une plus grande objectivité, puisque le choix de la vraisemblance détermine entièrement le choix de la loi a priori, qui n'est pas dicté par des éléments subjectifs extérieurs. Cependant, on pourrait aussi leur reprocher d'influencer le choix du statisticien, qui pourrait être tenté de choisir une famille de lois conjuguées pour simplifier les calculs. Ceci ne doit pas être l'unique raison du choix d'une loi a priori donnée.

La notion de famille conjuguée implique à la fois les lois a priori et a posteriori, mais également la vraisemblance : c'est le choix particulier d'une loi a priori pour une vraisemblance donnée, qui aboutit à une loi a posteriori conjuguée.

Définition 34 (Lois conjuguées). *Une famille de lois $\mathcal{F}$ sur Θ est dite conjuguée par rapport à la vraisemblance $L(x \mid \theta)$ si, pour toute loi a priori $\pi \in \mathcal{F}$, la loi a posteriori $p(\theta \mid x)$ appartient aussi à $\mathcal{F}$.*

Le tableau 2.1 résume les principales familles de lois conjuguées pour un paramètre θ unidimensionnel. Nous allons montrer certains de ces résultats dans les sections suivantes.

2.3.1 Vraisemblance gaussienne

Soient $X_1, \ldots, X_n$ des variables i.i.d. gaussiennes centrées et de variance inconnue $\mathcal{N}(0, \sigma^2)$. La vraisemblance s'écrit :

$$L(X_1, \ldots, X_n; \sigma^2) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} X_i^2\right).$$

La vraisemblance, vue comme une fonction de σ^2, s'écrit donc sous la forme $C(\sigma^2)^{-\alpha}e^{-\beta/\sigma^2}$. Pour trouver une loi conjuguée pour σ^2, il suffirait donc de trouver une loi a priori dont la densité a la même forme, i.e. une loi a priori de la forme $\pi(\sigma^2) \propto (\sigma^2)^{-a}e^{-b/\sigma^2}$. La loi a posteriori étant proportionnelle

Table 2.1 – Principales familles de lois conjuguées, pour un échantillon i.i.d. $X_1, \dots, X_n$.

Famille	Vraisemblance $L(x \mid \theta)$	Loi a priori $\pi(\theta)$	Loi a posteriori $p(\theta \mid x)$
Normale	$\mathcal{N}(\theta, s^2)$ (s^2 connue)	$\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}\left(\frac{s^2\mu+\tau^2\sum_{i=1}^n X_i}{s^2+n\tau^2}, \frac{s^2\tau^2}{s^2+n\tau^2}\right)$
Normale	$\mathcal{N}(\mu, 1/\theta^2)$ (μ connu)	$\mathcal{G}(a, b)$	$\mathcal{G}(a+n/2, b+\sum_{i=1}^n(\mu-X_i)^2/2)$
Gamma	$\mathcal{G}(k, \theta)$ (k connu)	$\mathcal{G}(a, b)$	$\mathcal{G}(a+nk, b+\sum_{i=1}^n X_i)$
Bernoulli	$\mathcal{B}(\theta)$	$\text{Beta}(a, b)$	$\text{Beta}(a+\sum_{i=1}^n X_i, b+n-\sum_{i=1}^n X_i)$
Poisson	$\mathcal{P}(\theta)$	$\mathcal{G}(a, b)$	$\mathcal{G}(a+\sum_{i=1}^n X_i, b+n)$
Binomiale négative	$\mathcal{NB}(m, \theta)$ (m connu)	$\text{Beta}(a, b)$	$\text{Beta}(a+\sum_{i=1} X_i, b+nm)$

au produit de la vraisemblance et de la loi a priori, on aurait alors $p(\sigma^2 \mid x) \propto (\sigma^2)^{-(a+\alpha)} e^{-(b+\beta)/\sigma^2}$, i.e. la loi a posteriori serait bien dans la même famille de lois que la loi a priori.

Ceci nous conduit à choisir comme loi a priori pour σ^2 une loi *inverse-Gamma* (voir A). On a alors le modèle bayésien suivant :

$$\begin{aligned} X_i \mid \sigma^2 &\sim \mathcal{N}(0, \sigma^2) \\ \sigma^2 &\sim \mathcal{IG}(a, b) \end{aligned}$$

On obtient alors la loi a posteriori suivante pour σ^2 :

$$\begin{aligned} p(\sigma^2 \mid x) &\propto \underbrace{(\sigma^2)^{-n/2} e^{-\left(\frac{\sum_{i=1}^n X_i^2}{2}\right)/\sigma^2}}_{\text{vraisemblance}} \underbrace{(\sigma^2)^{-a-1} e^{-b/\sigma^2} \mathbf{1}_{\mathbb{R}^+}(\sigma^2)}_{\text{a priori}} \\ &\propto (\sigma^2)^{-a-n/2-1} e^{-\left(b+\frac{\sum_{i=1}^n X_i^2}{2}\right)/\sigma^2} \end{aligned}$$

Donc la loi a posteriori de σ^2 sachant $X_1, \dots, X_n$ est une loi inverse-Gamma $\mathcal{IG}\left(a+n/2, b+\frac{\sum_{i=1}^n X_i^2}{2}\right)$.

Plaçons-nous maintenant dans le cas où la moyenne de la loi normale est aussi inconnue, i.e. :

$$X_i \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2).$$

Dans ce cas, la vraisemblance s'écrit :

$$L(X_1, \dots, X_n; \mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2\right).$$

On remarque qu'il n'est pas possible de "séparer" la vraisemblance en un terme qui ne dépendrait que de μ et un terme qui ne dépendrait que de σ^2. Si c'était le cas, on pourrait reprendre le raisonnement fait dans le cas où seule la variance est inconnue et identifier des lois a priori conjuguées séparément pour μ et pour σ^2. On aurait envie de proposer une loi inverse-Gamma pour σ^2, et une loi normale pour μ, à cause de l'expression $\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2\right) = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (\mu-X_i)^2\right)$. La présence de σ^2 nous suggère alors de choisir une loi normale pour la loi a priori de μ *conditionnellement* à σ^2. Cela nous conduit au modèle bayésien suivant :

$$
\begin{aligned}
X_i \mid \mu, \sigma^2 &\sim \mathcal{N}(\mu, \sigma^2) \\
\mu \mid \sigma^2 &\sim \mathcal{N}\left(a, \frac{\sigma^2}{b}\right) \\
\sigma^2 &\sim \mathcal{IG}(c, d)
\end{aligned}
$$

On obtient alors la loi a posteriori suivante :

$$
\begin{aligned}
\mu \mid \sigma^2, X_1, \ldots, X_n &\sim \mathcal{N}\left(\frac{\sum_{i=1}^n X_i + ab}{n+b}, \frac{\sigma^2}{n+b}\right) \\
\sigma^2 \mid X_1, \ldots, X_n &\sim \mathcal{IG}\left(c + \frac{n}{2}, d + \frac{\sum_{i=1}(X_i - \bar{X}_n)^2}{2} + \frac{nb}{2(n+b)}(\bar{X}_n - a)^2\right)
\end{aligned}
$$

2.3.2 Vraisemblance exponentielle

Soient $X_1, \ldots, X_n$ des variables i.i.d. de loi exponentielle de paramètre θ inconnu. La vraisemblance s'écrit :

$$
L(X_1, \ldots, X_n; \theta) = \theta^n e^{-\theta \sum_{i=1}^n X_i} \mathbf{1}_{\mathbb{R}^+}(\min X_i).
$$

La vraisemblance s'écrit donc sous la forme d'une expression faisant intervenir θ à une certaine puissance, et une exponentielle qui dépend de θ à travers un terme multiplicatif négatif. Cela nous fait penser à la loi Gamma, qui a la même expression. On choisit donc un a priori Gamma pour θ, pour le modèle bayésien suivant :

$$
\begin{aligned}
X_i \mid \theta &\sim \mathcal{E}(\theta) \\
\theta &\sim \mathcal{G}(a, b)
\end{aligned}
$$

La loi a posteriori est donnée par :

$$
\begin{aligned}
p(\theta \mid x) &\propto \theta^n e^{-\theta \sum_{i=1}^n X_i} \theta^{a-1} e^{-b\theta} \mathbf{1}_{\mathbb{R}^+}(\theta) \\
&\propto \theta^{a+n-1} e^{-(b+\sum_{i=1}^n X_i)\theta} \mathbf{1}_{\mathbb{R}^+}(\theta)
\end{aligned}
$$

On reconnaît une loi Gamma $\mathcal{G}(a+n, b+\sum_{i=1}^n X_i)$.

2.3.3 Vraisemblance binomiale

On a déjà vu un exemple de loi conjuguée faisant intervenir une vraisemblance binomiale, dans l'exemple 10. Supposons le modèle suivant :

$$
\begin{aligned}
X_i \mid \theta &\sim \mathcal{B}(\theta) \\
\theta &\sim \text{Beta}(a, b)
\end{aligned}
$$

avec $X_1, \ldots, X_n$ n variables aléatoires i.i.d. de loi de Bernoulli de paramètre θ. La vraisemblance est donnée par :

$$
L(X_1, \ldots, X_n; \theta) = \binom{n}{\sum_{i=1}^n X_i} \theta^{\sum_{i=1}^n X_i} (1-\theta)^{n - \sum_{i=1}^n X_i}
$$

On a donc une vraisemblance qui fait intervenir θ à une certaine puissance, et $1-\theta$ à une certaine puissance. Cela nous fait penser effectivement à la loi Beta. La loi a posteriori est alors :

$$\begin{aligned} p(\theta \mid x) &\propto \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}\theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{[0,1]}(\theta) \\ &\propto \theta^{a+\sum_{i=1}^n X_i-1}(1-\theta)^{n+b-\sum_{i=1}^n X_i-1}\mathbf{1}_{[0,1]}(\theta) \end{aligned}$$

On reconnaît la loi Beta$(a+\sum_{i=1}^n X_i, n+b-\sum_{i=1}^n X_i)$.

2.4 A priori non informatifs

Lorsqu'aucune information a priori n'est disponible pour dériver une loi a priori subjective, ou déterminer les valeurs des hyperparamètres d'une loi a priori conjuguée, on peut se tourner vers ce que l'on appelle un *a priori non informatif*. Comme son nom l'indique, ce type de loi a priori a pour objectif d'influencer le moins possible l'analyse, sans apporter d'information supplémentaire par rapport à celle apportée par la vraisemblance. L'avantage apporté par cette approche est qu'elle permet d'utiliser le paradigme bayésien même lorsqu'aucune information a priori n'est disponible. Il existe plusieurs moyens de construire une loi a priori non informative. Nous en décrivons plusieurs dans les sections suivantes.

2.4.1 Loi a priori de Laplace

L'idée de Laplace est de choisir des loi a priori uniformes. Si l'espace des paramètres Θ n'est pas compact, cela conduit à des loi a priori impropres, ce qui n'est pas forcément un problème tant que la loi a posteriori reste bien définie comme une mesure de probabilité. Un argument qui vient souvent contredire ce choix de loi a priori uniforme est celui de non-invariance de la loi a priori à reparamétrisation près. Autrement dit, si on considère une nouvelle paramétrisation du problème sous la forme $\eta = h(\theta)$, choisir un a priori uniforme pour θ ne conduit pas, en général, à un a priori uniforme pour η.

Exemple 13. *Reprenons l'exemple 10, et changeons de paramètre pour $\eta = \sqrt{\theta}$, avec θ de loi uniforme sur $[0,1]$. On a alors, pour g mesurable bornée :*

$$\begin{aligned} \mathbb{E}(g(\eta)) &= \int g(\sqrt{(}\theta))\mathbf{1}_{[0,1]}(\theta)d\theta \\ &= \int g(x)2x\mathbf{1}_{[0,1]}(x)dx \end{aligned}$$

La loi de η n'est donc pas uniforme sur $[0,1]$, il s'agit d'une loi Beta$(2,1)$. Or, si nous n'avons pas d'information que θ, nous n'en avons pas plus sur $\sqrt{\theta}$ ... on aimerait donc une loi a priori qui reste uniforme quel que soit le choix de paramétrisation du problème.

2.4.2 Loi a priori de Jeffreys

Jeffrey a proposé une approche permettant de construire des lois a priori invariantes par changement de paramétrisation. Pour cela, il se base sur l'information de Fisher, définie par :

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial \ln L(X;\theta)}{\partial \theta} \right)^2 \right], \tag{2.1}$$

qui peut se ré-écrire, sous certaines conditions de régularité (voir le cours de statistique mathématique du semestre 1) :

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 \ln L(X;\theta)}{\partial \theta^2} \right], \tag{2.2}$$

Définition 35 (A priori de Jeffreys).

- *si $\theta \in \mathbb{R}$, on appelle a priori de Jeffreys la mesure suivante :*

$$\pi(\theta) \propto I^{1/2}(\theta). \tag{2.3}$$

- *si $\theta \in \mathbb{R}^d, d > 1$, l'a priori de Jeffreys est donné par :*

$$\pi(\theta) \propto \det(I(\theta))^{1/2}. \tag{2.4}$$

***R.Q.** : ce choix peut conduire à un a priori impropre.*

L'information de Fisher représente la quantité d'information apportée par le modèle (c'est-à-dire par la vraisemblance) pour le paramètre θ. Avec un a priori construit de cette façon-là, on privilégie donc les régions de Θ où le modèle apporte le plus d'information.

Exemple 14. *En poursuivant notre exemple fil rouge Beta-binomial, avec $X_1, \ldots, X_n$ i.i.d. de loi de Bernoulli $\mathcal{B}(\theta)$, cherchons la loi a priori de Jeffreys pour θ. On peut ré-écrie le problème à l'aide de $X = \sum_{i=1}^n X_i$, et on a la vraisemblance suivante :*

$$L(X;\theta) = \binom{n}{X} \theta^X (1-\theta)^{n-X}.$$

En dérivant deux fois la log-vraisemblance par rapport à θ on obtient :

$$\frac{\partial^2 \ln L(X;\theta)}{\partial \theta^2} = -\frac{X}{\theta^2} - \frac{n-X}{(1-\theta)^2}$$

D'où :

$$I(\theta) = -\mathbb{E}_\theta \left(-\frac{X}{\theta^2} - \frac{n-X}{(1-\theta)^2} \right) = \frac{n}{\theta(1-\theta)}$$

On en déduit que la loi a priori de Jeffreys est $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$. *Il s'agit donc d'une loi* $Beta(1/2, 1/2)$.

La figure 3.5 représente les deux lois a priori étudiées pour cet exemple : la loi uniforme sur $[0,1]$ *et la loi* $Beta(1/2,1/2)$. *La loi a priori de Jeffreys peut sembler, à première vue, plus informative que la loi uniforme car on met plus de poids sur les valeurs extrêmes, i.e. autour de 0 et de 1. La notion de loi non-informative est en fait à comprendre à travers celle d'invariance par changement de paramétrisation. Considérons maintenant l'estimateur de Bayes pour la fonction de perte quadratique, pour un a priori* $Beta(a,b)$ *i.e. :*

$$\hat{\theta} = \frac{a + \sum_{i=1}^{n} X_i}{a+b+n}.$$

Si on compare cet estimateur à celui du maximum de vraisemblance, $\hat{\theta}_{MV} = \frac{\sum_{i=1}^{n} X_i}{n}$, *on peut interpréter* a *et* b *comme respectivement le nombre de "piles" et de "faces" que l'on rajoute au nombre de "piles" et de "face" observés. En prenant une loi uniforme sur* $[0,1]$, *c'est-à-dire une loi* $Beta(1,1)$, *on "rajoute" donc artificiellement un lancer avec un "pile" et un lancer avec un "face" à nos observations, avant d'en prendre la moyenne empirique. Cela a pour effet de minimiser l'impact des observations extrêmes* $\sum_{i=1}^{n} X_i = 0$ *et* $\sum_{i=1}^{n} X_i = n$, *pour lesquels l'estimateur du maximum de vraisemblance donne 0 et 1 respectivement. Ces deux valeurs extrêmes sont exclues avec un a priori uniforme sur* $[0,1]$. *Avec un a priori de type* $Beta(1/2,1/2)$, *on ne rajoute "qu'un demi lancer" avec pile et "un demi lancer" avec face à nos observations. On augmente donc l'influence des observations, même si les deux estimations extrêmes restent exclues également avec cette loi a priori.*

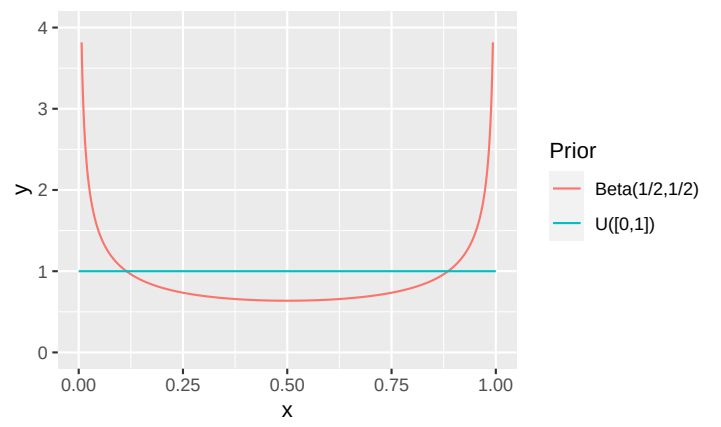


Figure 3.5 – Densités des lois a priori $Beta(1/2,1/2)$ *et* $\mathcal{U}([0,1])$

La proposition suivante nous assure que le choix d'un a priori de Jeffreys permet d'obtenir une mesure invariante par changement de paramètre.

Proposition 4. *Soit* π *la mesure a priori de Jeffreys pour* θ, *et soit* $\eta = g(\theta)$, *avec* g *un* C^1*-difféomorphisme. Alors la mesure image de* π *par* g *est la mesure a priori de Jeffreys pour* η.

Démonstration. L'information de Fisher pour η s'obtient à l'aide du changement de variable $\eta = g(\theta)$ et du jacobien de ce changement de variable. En dimension 1, on a (les résultats se généralisent facilement au cas multidimensionnel) :

$$I(\theta) = (g'(\theta))^2 I(g(\theta)) \Leftrightarrow I(g^{-1}(\eta)) = (g'(g^{-1}(\eta))^2 I(\eta)$$

Notons $\pi^* = \pi \circ g^{-1}$, la mesure image de π par g. On se place dans le cas où π^* est une loi a priori propre, ce qui simplifie les notations. Par définition de la mesure image, on a $\int h(g(\theta))\pi(\theta)d\theta = \int h(\eta)\pi^*(\eta)d\eta$, pour toute fonction h borélienne bornée. De plus :

$$\begin{aligned}
\int h(g(\theta))\pi(\theta)d\theta &= \int h(\eta)\pi(g^{-1}(\eta))\frac{d\eta}{g'(g^{-1}(\eta))} \quad \text{(changement de variable)} \\
&= \int h(\eta)C\, I^{1/2}(g^{-1}(\eta))\frac{d\eta}{g'(g^{-1}(\eta))} \quad \text{par définition de } \pi \\
&= \int h(\eta)C\, I^{1/2}(\eta)d\eta
\end{aligned}$$

On en déduit que $\pi^*(\eta) = C\, I^{-1/2}(\eta)$: il s'agit bien de la loi a priori de Jeffreys pour η. □

Chapitre 3

Loi a posteriori - simulation et propriétés asymptotiques

Nous avons vu dans le chapitre précédent quelques critères pour choisir une loi a priori dans un modèle bayésien. Une fois cette loi a priori choisie, et en association avec la vraisemblance des observations, on obtient une loi a posteriori définie de façon unique. Dans certains cas, celle-ci est connue de façon explicite (c'est le cas par exemple des familles de loi conjuguées). Cependant, il existe de nombreuses situations où l'expression de cette loi n'est pas accessible analytiquement. Dans ce cas, on pourra avoir recours à des algorithmes permettant de générer des réalisations de cette loi (voir section 3.1). Dans la section 3.2, on s'intéresse aux propriétés asymptotiques de la loi a posteriori, ce qui permettra notamment de décrire les propriétés asymptotiques des estimateurs ponctuels obtenus à partir de la loi a posteriori et les propriétés des intervalles de crédibilité déjà évoqués en section 1.2.

3.1 Simulation selon la loi a posteriori

On rappelle que la loi a posteriori est donnée par la formule (1.1) :

$$p(\theta \mid x) = \frac{L(x_1, \ldots, x_n; \theta)\pi(\theta)}{f_X(x_1, \ldots, x_n)}, \tag{3.1}$$

où L est la vraisemblance, π la loi a priori et f_X la loi marginale des observations. Or cette dernière s'obtient en calculant l'intégrale de la loi jointe (c'est-à-dire du numérateur dans l'équation ci-dessus) par rapport à θ. Dans un certain nombre de situations, cette intégrale est difficile à calculer analytiquement, en particulier lorsque la dimension de θ augmente.

On peut alors avoir recours à toutes les approches vues dans la partie II qui ne nécessitent de connaître la loi cible qu'à une constante de normalisation près. Les méthodes les plus utilisées dans le cadre bayésien sont l'échantillonnage préférentiel auto-normalisé et, surtout, les algorithmes MCMC. Que deviennent ces algorithmes dans le cadre spécifique des statistiques bayésiennes ?

3.1.1 Metropolis-Hastings

La loi cible est la loi a posteriori, qui vérifie :

$$p(\theta \mid x) \propto L(X_1, \ldots, X_n; \theta)\pi(\theta)$$

C'est une densité de probabilité pour la variable aléatoire θ, et les observations $X_1, \ldots, X_n$ sont fixées. Par conséquent, un algorithme MCMC fournit des réalisations de la variable aléatoire θ. Cela conduit à l'algorithme de Metropolis-Hastings suivant :

Algorithme 7 : Algorithme de Metropolis-Hastings pour simuler selon une loi a posteriori

initialisation : on initialise la chaîne avec θ_0

pour $k = 1, \ldots, N$ **faire**

on génère un candidat $Y_k \sim q(\cdot \mid \theta_{k-1})$

on pose

$$\theta_k = \begin{cases} Y_k & \text{avec une probabilité } \alpha(\theta_{k-1}, Y_k) \\ \theta_{k-1} & \text{avec une probabilité } 1 - \alpha(\theta_{k-1}, Y_k) \end{cases} \tag{3.2}$$

où

$$\alpha(\theta_{k-1}, Y_k) = \min\left(1, \frac{p(Y_k \mid x)}{p(\theta_{k-1} \mid x)} \frac{q(\theta_{k-1} \mid Y_k)}{q(Y_k \mid \theta_{k-1})}\right) \tag{3.3}$$

$$= \min\left(1, \frac{L(X_1, \ldots, X_n; Y_n)\pi(Y_k)}{L(X_1, \ldots, X_n; \theta_{n-1})\pi(\theta_{k-1})} \frac{q(\theta_{k-1} \mid Y_k)}{q(Y_k \mid \theta_{k-1})}\right) \tag{3.4}$$

L'utilisation de lois a priori impropres, ou définies à une constante de normalisation près, ne pose pas de problème pour l'utilisation de l'algorithme de Metropolis-Hastings. Une fois que l'on a obtenu un échantillon de N réalisations selon la loi a posteriori, il est possible d'approcher les estimateurs ponctuels et les intervalles de crédibilité à l'aide par exemple de la moyenne empirique ou des quantiles empiriques.

3.1.2 Échantillonneur de Gibbs

Si l'algorithme de Metropolis-Hastings est le plus utilisé dans le contexte bayésien, l'échantillonneur de Gibbs peut s'avérer particulièrement utile et efficace, notamment en cas de modèle hiérarchique. Un modèle hiérarchique est un modèle dans lequel la loi a priori est définie comme une succession de lois conditionnelles.

Définition 36 (Modèle bayésien hiérarchique). *Un modèle bayésien hiérarchique est un modèle bayésien dans lequel la loi a priori se décompose en m distributions conditionnelles et une distribution mar-*

ginale $\pi_1(\theta \mid \theta_1), \pi_2(\theta_1 \mid \theta_2), \ldots, \pi_n(\theta_{m-1} \mid \theta_m), \pi_{m+1}(\theta_m)$ *vérifiant :*

$$\pi(\theta) = \int \pi_1(\theta \mid \theta_1)\pi_2(\theta_1 \mid \theta_2) \ldots \pi_m(\theta_{m-1} \mid \theta_m)\pi_{m+1}(\theta_m) d\theta_1 d\theta_2 \ldots d\theta_{m-1} d\theta_m \tag{3.5}$$

Les paramètres θ_i *sont appelés* hyperparamètres de niveau i.

Ce type de modèle permet notamment de prendre en compte une incertitude sur la loi a priori de θ, en incluant l'incertitude autour des hyperparamètres de la loi a priori. Ils permettent également de simplifier l'écriture d'un modèle faisant intervenir un grand nombre de paramètres Un exemple de modèle hiérarchique est donné ci-dessous.

Exemple 15. *On s'intéresse à la pollution d'un plan d'eau. Pour cela, on effectue des mesures de concentrations en polluant sur n différentes localisations du plan d'eau. Les conditions environnementales pouvant varier localement, on suppose que la concentration en polluant du i-ème point de mesure suit une loi normale dont la moyenne dépend du site de mesure :*

$$X_i \mid \theta_i \sim \mathcal{N}(\theta_i, 1).$$

On suppose ensuite que les variations locales sur la concentration moyenne de polluant peuvent se modéliser par la loi suivante :

$$\begin{aligned} \theta_i \mid \mu &\sim \mathcal{N}(\mu, \sigma^2) \\ \mu &\sim \mathcal{N}(\beta, \tau^2), \end{aligned}$$

avec σ^2, β *et* τ *connus. Les paramètres du modèle, dont on va chercher la loi a posteriori, sont les* $\theta_1, \ldots, \theta_n, \mu$. *L'hyperparamètre de niveau 1 est* σ^2 *et les hyperparamètres de niveau 2 sont* β *et* τ^2.

L'avantage de ce type de modèle est qu'il permet également de décomposer la loi a posteriori. Par exemple, si on a

$$X \mid \theta \sim f(x \mid \theta), \quad \theta \mid \theta_1 \sim \pi_1(\theta \mid \theta_1), \quad \theta_1 \sim \pi_2(\theta_1),$$

la loi a posteriori peut s'écrire :

$$p(\theta \mid x) = \int p(\theta \mid \theta_1, x) p(\theta_1 \mid x) d\theta_1,$$

avec :

$$\begin{aligned} p(\theta \mid \theta_1, x) &= \frac{f(x \mid \theta)\pi_1(\theta \mid \theta_1)}{f_1(x \mid \theta_1)} \\ f_1(x \mid \theta_1) &= \int f(x \mid \theta)\pi_1(\theta \mid \theta_1) d\theta \\ p(\theta_1 \mid x) &= \frac{f_1(x \mid \theta_1)\pi_2(\theta_1)}{f_X(x)} \\ f_X(x) &= \int f_1(x \mid \theta_1)\pi_2(\theta_1) d\theta_1 \end{aligned}$$

En pratique, cela signifie que l'on peut simuler selon la loi a posteriori en simulant d'abord selon la loi conditionnelle de θ_1 sachant x, puis selon la loi conditionnelle de θ sachant (θ_1, x). Dans certains cas cela peut s'avérer plus simple que de simuler directement selon la loi a posteriori. Un autre résultat intéressant des modèles bayésiens hiérarchiques concerne les lois conditionnelles complètes. On a la proposition suivante :

PROPOSITION 5. *Dans un modèle bayésien hiérarchique, la distribution conditionnelle de θ_j sachant x et θ_i, pour tout $j \neq i$ et $i = 0, \dots, m$, est donnée par :*

$$p(\theta_j \mid \theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_m, \theta, x) = p(\theta_j \mid \theta_{j-1}, \theta_{j+1}),$$

avec la convention $\theta_0 = \theta$ et $\theta_{m+1} = 0$. Autrement dit, la loi conditionnelle complète pour les hyperparamètres de niveau j ne dépend que des hyperparamètres de niveau $j-1$ et $j+1$.

Ce résultat permet de simplifier drastiquement les calculs et les expressions des lois conditionnelles complètes. L'échantillonneur de Gibbs est alors l'outil idéal pour traiter ce type de modèle. On traite un exemple complet ci-dessous.

EXEMPLE 16 (Modèle à effet aléatoire). *On s'intéresse au modèle à effet aléatoire suivant :*

$$\begin{aligned} X_{ij} &= \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \alpha_i &\sim \mathcal{N}(\mu, \gamma^2) \end{aligned}$$

Les observations sont les X_{ij}, $i = 1, \dots, n$, $j = 1, \dots, n_i$. On suppose alors le modèle bayésien suivant :

$$\begin{aligned} X_{ij} \mid \alpha_i, \sigma^2 &\sim \mathcal{N}(\alpha_i, \sigma^2), \quad i = 1, \dots, n, j = 1, \dots, K \\ \alpha_i \mid \mu, \gamma^2 &\sim \mathcal{N}(\mu, \gamma^2), \quad i = 1, \dots, n \\ \mu \mid \sigma_u^2 &\sim \mathcal{N}(\mu_0, \sigma_u^2) \\ \sigma^2 &\sim \mathcal{IG}(a_1, b_1) \\ \gamma^2 &\sim \mathcal{IG}(a_2, b_2) \\ \sigma_u^2 &\sim \mathcal{IG}(a_3, b_3) \end{aligned}$$

Les observations sont les X_{ij}, $i = 1, \dots, n$, $j = 1, \dots, K$, les paramètres d'intérêt sont $\alpha_1, \dots, \alpha_n, \mu, \sigma^2, \gamma^2, \sigma_u^2$ et les hyperparamètres sont $\mu_0, a_1, b_1, a_2, b_2, a_3, b_3$.

On cherche la loi a posteriori des paramètres, c'est-à-dire la loi conditionnelle du vecteur de paramètres $(\alpha_1, \dots, \alpha_n, \mu, \sigma^2, \gamma^2, \sigma_u^2)$ sachant $X_{11}, \dots, X_{nK}$. Pour cela, on commence par écrire la loi jointe

des paramètres et des observations. Celle-ci s'écrit :

$$f(\alpha_1,\dots,\alpha_n,\mu,\sigma_u^2,\gamma^2,\sigma^2,X_{11},\dots,X_{nK}) = \frac{1}{(2\pi)^{nK/2}(\sigma^2)^{nK/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{K}(X_{ij}-\alpha_i)^2\right)\times$$
$$\frac{1}{(2\pi)^{n/2}(\gamma^2)^{n/2}}\exp\left(-\frac{1}{2\gamma^2}\sum_{i=1}^{n}(\alpha_i-\mu)^2\right)\times\frac{1}{(2\pi)^{1/2}(\sigma_u^2)^{1/2}}\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_u^2}\right)\times$$
$$\frac{b_1^{a_1}}{\Gamma(a_1)}\sigma^{-2(a_1-1)}e^{-\frac{b_1}{\sigma^2}}\mathbf{1}_{\mathbb{R}^+}(\sigma^2)\frac{b_2^{a_2}}{\Gamma(a_2)}\gamma^{-2(a_2-1)}e^{-\frac{b_2}{\gamma^2}}\mathbf{1}_{\mathbb{R}^+}(\gamma_2)\frac{b_3^{a_3}}{\Gamma(a_3)}\sigma_u^{-2(a_3-1)}e^{-\frac{b_3}{\sigma_u^2}}\mathbf{1}_{\mathbb{R}^+}(\sigma_u^2)$$

Pour obtenir les lois conditionnelles complètes pour chaque variable, il faut intégrer cette loi jointe par rapport à toutes les autres variables. En pratique, cela revient à ré-écrire l'expression de la densité jointe en fonction de la variable dont on cherche la loi conditionnelle, toutes les autres variables étant considérées comme des constantes. Pour α_i, on obtient par exemple :

$$\begin{aligned}
f(\alpha_i \mid (X_{ij})_{i=1,\dots,n,j=1,\dots,K},(\alpha_j)_{j\neq i},\mu,\sigma_u^2,\gamma^2,\sigma^2) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{K}(\alpha_i-X_{ij})^2-\frac{1}{2\gamma^2}(\alpha_i-\mu)^2\right)\\
&\propto \exp\left(-\frac{1}{2\sigma^2}(K\alpha_i^2-2\alpha_iK\bar{X}_i)-\frac{1}{2\gamma^2}(\alpha_i^2-2\alpha_i\mu)\right)\\
&\propto \exp\left(-\frac{1}{2}\left[\left(\frac{1}{\sigma^2/K}+\frac{1}{\gamma^2}\right)\alpha_i^2-2\alpha_i\left(\frac{\bar{X}_i}{\sigma^2/K}+\frac{\mu}{\gamma^2}\right)\right]\right)\\
&\propto \exp\left(-\frac{1}{2\frac{\sigma^2\gamma^2}{K\gamma^2+\sigma^2}}\left(\alpha_i^2-2\alpha_i\left(\frac{\gamma^2}{K\gamma^2+\sigma^2}\bar{X}_i+\frac{\sigma^2}{K\gamma^2+\sigma^2}\mu\right)\right)\right)\\
&\propto \exp\left(-\frac{1}{2\frac{\sigma^2\gamma^2}{K\gamma^2+\sigma^2}}\left(\alpha_i-\left(\frac{\gamma^2}{K\gamma^2+\sigma^2}\bar{X}_i+\frac{\sigma^2}{K\gamma^2+\sigma^2}\mu\right)\right)^2\right)
\end{aligned}$$

Autrement dit, on reconnaît une loi normale $\mathcal{N}\left(\frac{\gamma^2}{K\gamma^2+\sigma^2}\bar{X}_i+\frac{\sigma^2}{K\gamma^2+\sigma^2}\mu;\frac{\sigma^2\gamma^2}{K\gamma^2+\sigma^2}\right)$.

On peut procéder de même pour les autres variables, et on obtient les résultats suivants :

$$\mu \mid (X_{ij})_{i=1,\dots,n,j=1,\dots,K},(\alpha_j)_{i=1,\dots,n},\sigma_u^2,\gamma^2,\sigma^2 \sim \mathcal{N}\left(\frac{\gamma^2}{\gamma^2+n\sigma_u^2}\mu_0+\frac{n\sigma_u^2}{\gamma^2+n\sigma_u^2}\bar{\alpha};\frac{\sigma_u^2\gamma^2}{\gamma^2+n\sigma_u^2}\right)$$
$$\sigma^2 \mid (X_{ij})_{i=1,\dots,n,j=1,\dots,K},(\alpha_j)_{i=1,\dots,n},\sigma_u^2,\mu,\gamma^2, \sim \mathcal{IG}\left(\frac{nK}{2}+a_1;\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{K}(X_{ij}-\alpha_i)^2+b_1\right)$$
$$\gamma^2 \mid (X_{ij})_{i=1,\dots,n,j=1,\dots,K},(\alpha_j)_{i=1,\dots,n},\sigma_u^2,\mu,\sigma^2 \sim \mathcal{IG}\left(\frac{n}{2}+a_2;\frac{1}{2}\sum_{i=1}^{n}(\alpha_i-\mu)^2+b_2\right)$$
$$\sigma_u^2 \mid (X_{ij})_{i=1,\dots,n,j=1,\dots,K},(\alpha_j)_{i=1,\dots,n},\mu,\gamma^2,\sigma^2 \sim \mathcal{IG}\left(\frac{1}{2}+a_3;\frac{1}{2}(\mu-\mu_0)^2+b_3\right)$$

Il est alors possible d'utiliser un échantillonneur de Gibbs pour simuler selon la loi a posteriori jointe. Le code R et les sorties graphiques sont présentées ci-dessous.

```
X <- data("Energy") # du package "mcsm"

n <- 2
Xbar1 <- mean(X[,1])
Xbar2 <- mean(X[,2])
K <- nrow(X)

# hyperparamètres
mu0 <- 500; a1 <- 10; b1 <- 10; a2 <- 10; b2 <- 10; a3 <- 10; b3 <- 10

N <- 10000
# initialisation
alpha_1 <- rep(0,N); alpha_2 <- rep(0,N); mu <- rep(0,N)
sigma2 <- rep(0,N); gamma2 <- rep(0,N); sigmamu2 <- rep(0,N)

sigma2[1] <- 1/rgamma(1,shape=a1,rate=b1)
gamma2[1] <- 1/rgamma(1,shape=a2,rate=b2)
sigmamu2[1] <- 1/rgamma(1,shape=a3,rate=b3)
mu[1] <- rnorm(1,mu0,sqrt(sigmamu2[1]))
alpha_1[1] <- rnorm(1,mu,sqrt(gamma2[1]))
alpha_2[1] <- rnorm(1,mu,sqrt(gamma2[1]))

sh1 <- (n*K)/2 + a1
sh2 <- n/2 + a2
sh3 <- 1/2 + a3

for (i in 2:N){
  varalpha <- (gamma2[i-1]*sigma2[i-1]/(K*gamma2[i-1] + sigma2[i-1]))
  D <- K*gamma2[i-1] + sigma2[i-1]
  alpha_1[i] <- rnorm(1, (K*gamma2[i-1]/D)*Xbar1 + (sigma2[i-1]/D)*mu[i-1],
                          sqrt(varalpha))
  alpha_2[i] <- rnorm(1, (K*gamma2[i-1]/D)*Xbar2 + (sigma2[i-1]/D)*mu[i-1],
                          sqrt(varalpha))
  meanalpha <- 0.5*(alpha_1[i] + alpha_2[i])
  mu[i] <- rnorm(1, (gamma2[i-1]/(gamma2[i-1] + n*sigmamu2[i-1]))*mu0 +
           (n*sigmamu2[i-1]/(gamma2[i-1] + n*sigmamu2[i-1]))*meanalpha,
           sqrt((gamma2[i-1]*sigmamu2[i-1]/(gamma2[i-1] + n*sigmamu2[i-1]))))
```

```
  rate1 <- 0.5*sum((X[,1]-alpha_1[i])^2) + 0.5*sum((X[,2]-alpha_2[i])^2) + b1
  rate2 <- 0.5*(alpha_1[i]-mu[i])^2 + 0.5*(alpha_2[i]-mu[i])^2 + b2
  sigma2[i] <- 1/rgamma(1, shape=sh1, rate=rate1)
  gamma2[i] <- 1/rgamma(1, shape=sh2, rate=rate2)
  sigmamu2[i] <- 1/rgamma(1, shape=sh3, rate=0.5*(mu[i]-mu0)^2 + b3)
}
```

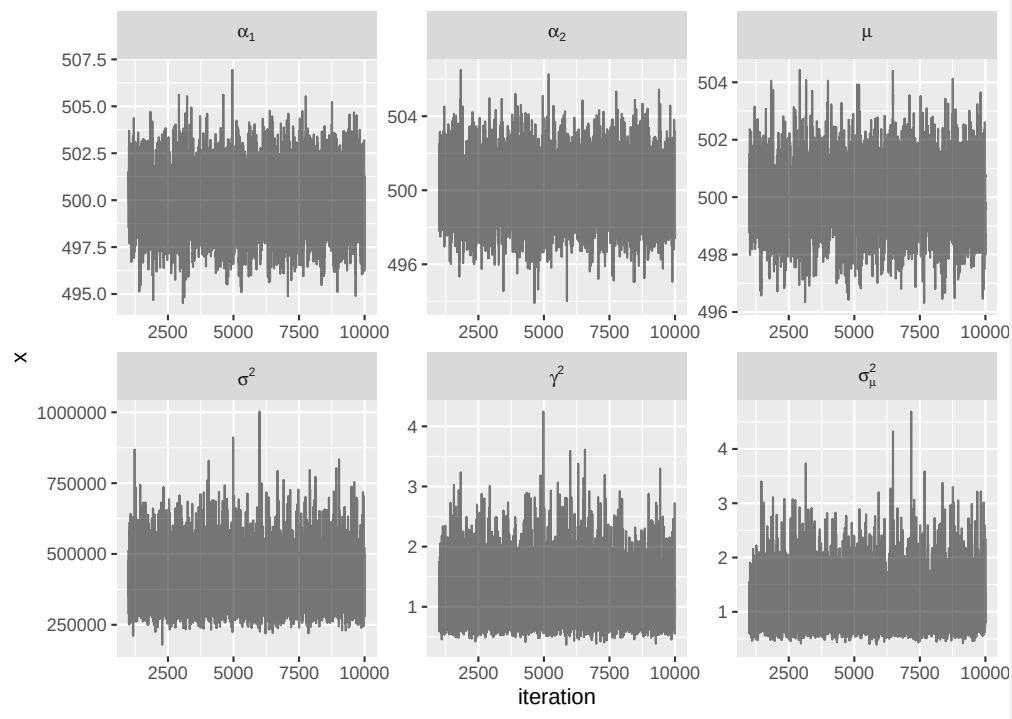


FIGURE 3.6 – Evolution des trajectoires de chaque composante par échantillonneur de Gibbs, après un temps de chauffe de 1000 itérations.

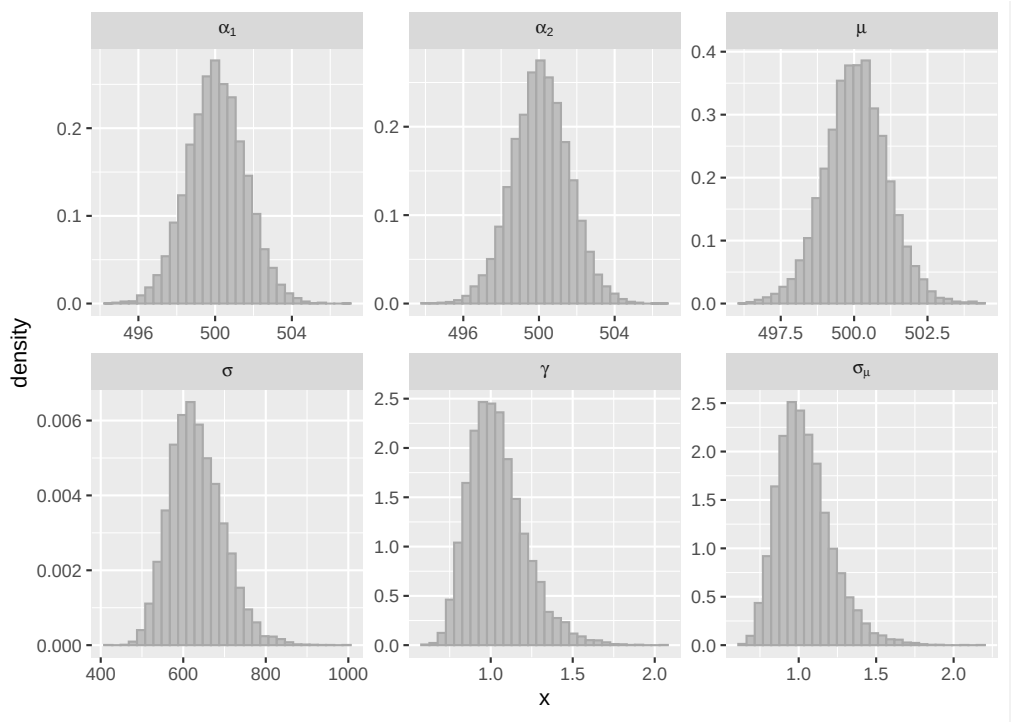


Figure 3.7 – Loi a posteriori marginale de chaque paramètre.

3.2 Propriétés asymptotiques

Dans cette section, on va s'intéresser aux propriétés asymptotiques de la loi a posteriori.

3.2.1 Consistance de la loi a posteriori

L'approche bayésienne peut être utilisée dans un cadre fréquentiste, c'est-à-dire lorsque l'on suppose qu'il existe une vraie valeur unique θ^* telle que les observations $X_1, \ldots, X_n$ soient i.i.d. de loi f_{θ^*}. Dans ce cas, on peut se demander si la loi a posteriori que l'on obtient peut nous permettre d'obtenir ou de définir des estimateurs pour θ^* ayant de bonnes propriétés asymptotiques.

Prenons l'exemple de la famille de lois conjuguées associées à une vraisemblance gaussienne (voir Tableau 2.1). La moyenne a posteriori est donnée par :

$$\mathbb{E}(\theta \mid x) = \frac{s^2\mu + \tau^2 \sum_{i=1}^n X_i}{s^2 + n\tau^2}$$

et la variance a posteriori par :

$$\mathrm{Var}(\theta \mid x) = \frac{s^2\tau^2}{s^2 + n\tau^2}.$$

Sur cet exemple, on remarque que, sous l'hypothèse selon laquelle les X_i sont distribués selon la loi $\mathcal{N}(\theta^*, s^2)$, la moyenne a posteriori est un estimateur consistant de θ^*. De plus, la variance a posteriori tend vers 0 lorsque n tend vers l'infini : on a donc une concentration de la loi a posteriori autour de θ^* lorsque la taille de l'échantillon tend vers l'infini.

Ce comportement est-il spécifique de cet exemple ? Ou est-ce une caractéristique que l'on retrouve de façon générale ? Commençons par donner une définition de la consistance de la loi a posteriori, que l'on utilisera dans la suite. La loi a posteriori dépend de la taille de l'échantillon n, on s'intéressera donc à la *séquence* de lois a posteriori indicées par n. Pour rendre plus claire la dépendance de la loi a posteriori en n, on notera $p_n(\cdot \mid x)$ la densité associée à la loi a posteriori.

Définition 37. *Soit $X_1, \dots, X_n$ un échantillon i.i.d. de loi $\mathbb{P}_{\theta^*}$. Une séquence de lois a posteriori $p_n(\cdot \mid x)$ est dite consistante en θ^* lorsque*

$$p_n(\mathcal{V}_{\theta^*} \mid x) \xrightarrow[n\to+\infty]{\mathbb{P}_{\theta^*}} 1$$

pour tout voisinage $\mathcal{V}_{\theta^*}$ de θ^*. La convergence a lieu en probabilité, sous la loi $\mathbb{P}_{\theta^*}$.

Autrement dit, la loi a posteriori se concentre autour de θ^* lorsque n tend vers l'infini. On notera également que $p_n(\mathcal{V}_{\theta^*} \mid x)$ converge en probabilité vers 1 si et seulement si $\mathbb{E}_{\theta^*}(p_n(\mathcal{V}_{\theta^*} \mid x)) \to 1$ (où $p_n(\mathcal{V}_{\theta^*} \mid x)$.

Peut-on trouver des exemples de loi a posteriori qui ne sont pas consistantes ? Oui, et c'est relativement "facile" à trouver. Pour cela, il suffit de choisir une loi a priori dont le support ne contient pas la vraie valeur θ^*. En effet, dans ce cas le support de la loi a posteriori est lui aussi contraint et ne contiendra pas θ^*. La convergence en probabilité de la définition n'est donc pas possible, car il existera toujours un voisinage de θ^* de probabilité nulle sous la loi a posteriori.

En fait, on a un résultat général qui nous donne les conditions suffisantes pour obtenir la consistance de la loi a posteriori. En particulier, il suffit que la loi a priori soit non nulle dans un voisinage de θ^*.

Théorème 14 (Bernstein-von Mises). *Soit $X_1, \dots, X_n$ un échantillon i.i.d. dont la loi commune $\mathbb{P}_\theta$, $\theta \in \Theta \subset \mathbb{R}^d$ admet une densité f_θ par rapport à une mesure dominante, et supposons le modèle régulier*[a]*. Soit π une loi a priori pour θ admettant une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$. On suppose de plus que π est non nulle et continue en θ^*. Soit $\hat{\theta}_{MV}$ l'estimateur du maximum de vraisemblance de θ. Alors :*

$$\int \left| p_n(\theta \mid x) - f_N\left(\theta; \hat{\theta}_{MV}; \frac{1}{n} I(\theta^*)^{-1}\right) \right| d\theta \xrightarrow[n\to+\infty]{} 1, \tag{3.6}$$

où $f_N(\cdot; \mu; \Gamma)$ est la densité d'une loi normale de moyenne μ et de matrice de covariance Γ.

a. voir le cours de Statistique mathématique du premier semestre pour la définition d'un modèle régulier : il s'agit en particulier d'un modèle pour lequel on peut définir l'information de Fisher, ce qui requiert des hypothèses de régularité sur la fonction de vraisemblance. La plupart des modèles étudiés sont des modèles réguliers. C'est le cas notamment pour les modèles de la famille exponentielle. Un exemple typique de modèle non régulier est celui où le support de la loi dépend du paramètre inconnu.

La preuve de ce théorème est admise. En pratique, cela signifie que lorsque n est suffisamment grand, *et sous un certain nombre d'hypothèses, cachées ici sous le terme "modèle régulier"*, on peut appro-

cher la loi a posteriori par une loi normale centrée sur le maximum de vraisemblance et de variance proportionnelle à l'inverse de la matrice d'information de Fisher en θ^*. Autrement dit, en se plaçant dans le cadre fréquentiste où il existe une vraie valeur inconnue θ^*, l'approche bayésienne est asymptotiquement équivalent à l'approche par maximum de vraisemblance, et ceci quelque soit le choix de la loi a priori (sous réserve que le support de la loi a priori contienne un voisinage de θ^*).

On peut déjà "apercevoir" ce résultat sur la figure 3.2, où l'on a tracé la loi a posteriori pour deux valeurs de n, $n = 10$ et $n = 100$. On voit que la loi a posteriori se concentre plus, lorsque n augmente, autour du maximum de vraisemblance représenté par un trait vertical sur l'axe des abscisses. La figure 3.8 illustre le phénomène pour un modèle avec vraisemblance exponentielle $\mathcal{E}(\theta)$ et loi a priori Gamma. On a simulé des échantillons de taille n, pour n variant de 10 à 1000, et pour une vraie valeur de $\theta^* = 10$.

3.2.2 Niveau de confiance des régions de crédibilité

Si on se place dans un cadre fréquentiste, on peut également se demander si les régions de crédibilité définies à partir de la loi a posteriori sont également des régions de confiance pour le vrai paramètre θ^*. C'est en fait une conséquence du théorème de Bernstein-von Mises.

Pour simplifier la présentation, on suppose $\theta \in \mathbb{R}$. Le premier résultat que l'on obtient nous donne la convergence de l'intervalle de crédibilité basé sur les quantiles vers l'intervalle de confiance asymptotique construit à partir de l'estimateur du maximum de vraisemblance.

Proposition 6. *On suppose les hypothèses nécessaires au théorème de Bernstein-von Mises vérifiées. Soit $\alpha \in [0,1]$. Soit $I_{quant}(\alpha) = \left[q^{post}_{\alpha/2}(n); q^{post}_{1-\alpha/2}(n)\right]$ l'intervalle de crédibilité de niveau $1-\alpha$ basé sur les quantiles de la loi a posteriori. Alors on a :*

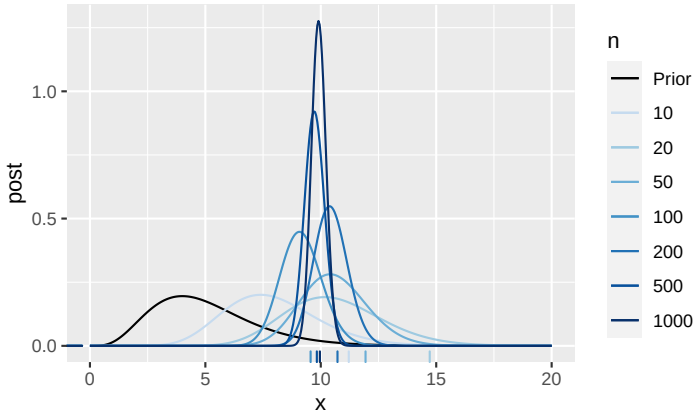
$$q^{post}_{\alpha/2}(n) \xrightarrow[n\to+\infty]{\mathbb{P}} \hat{\theta}_{MV} - q^{\mathcal{N}(0,1)}_{1-\alpha/2}\frac{1}{\sqrt{nI(\theta^*)}}$$

$$q^{post}_{1-\alpha/2}(n) \xrightarrow[n\to+\infty]{\mathbb{P}} \hat{\theta}_{MV} + q^{\mathcal{N}(0,1)}_{1-\alpha/2}\frac{1}{\sqrt{nI(\theta^*)}}$$

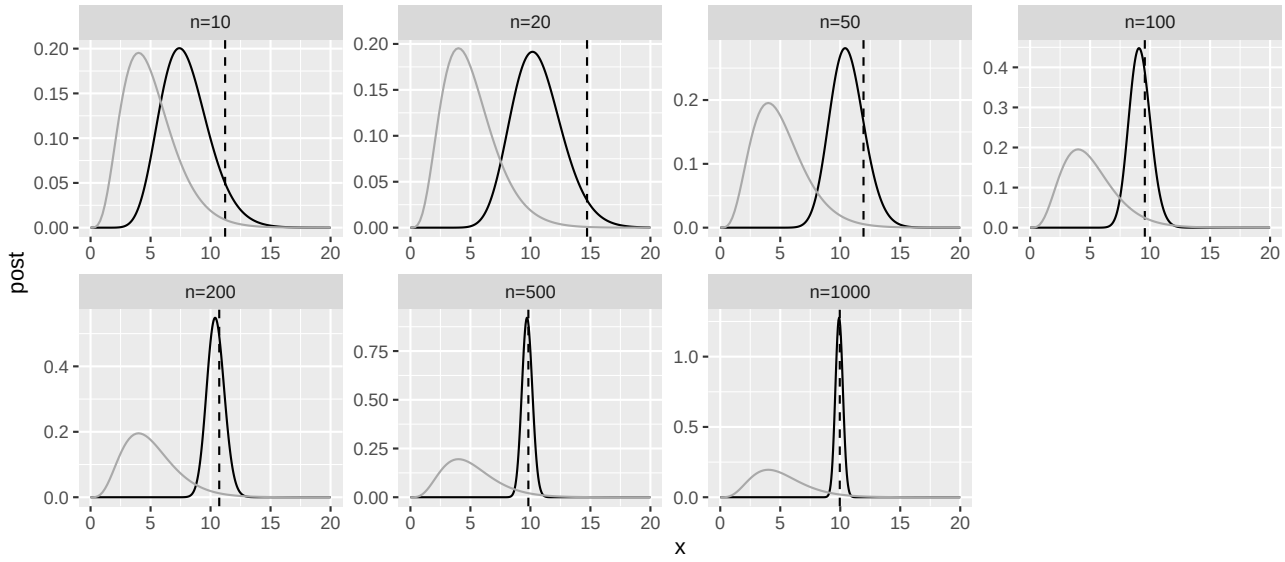
Là encore, la démonstration est admise. Elle découle de la convergence de la loi a posteriori vers la loi normale centrée réduite. Les bornes de l'intervalle de crédibilité bayésien basé sur les quantiles de la loi a posteriori convergent donc en probabilité vers les bornes de l'intervalle de confiance basé sur la loi asymptotique de l'estimateur du maximum de vraisemblance (EMV). Cela présente un intérêt pratique qui peut être non négligeable, même pour une approche purement fréquentiste. En effet, l'intervalle de confiance asymptotique de l'EMV dépend de $I(\theta^*)$. Or cette quantité est inconnue, car θ^* n'est pas connue. Une approche classique consiste alors à remplacer $I(\theta^*)$ par $I(\hat{\theta}_{MV})$, qui est un estimateur consistant de $I(\theta^*)$ lorsque les conditions de régularité du théorème sont vérifiées. Ceci dit, cette quantité peut être malgré tout difficile à calculer. Avec l'approche bayésienne, aucune estimation de l'information de Fisher n'est nécessaire : les quantiles de la loi a posteriori suffisent et donnent directement les "bonnes bornes" de l'intervalle de confiance asymptotique de l'EMV.

Des deux résultats précédents, on déduit la proposition suivante, qui fait le lien entre région de crédibilité et région de confiance.

Proposition 7. *On suppose les hypothèses nécessaires au théorème de Bernstein-von Mises vérifiées. Soit* $\alpha \in [0,1]$. *Soit* $I_{quant}(\alpha) = \left[q^{post}_{\alpha/2}(n); q^{post}_{1-\alpha/2}(n)\right]$ *l'intervalle de crédibilité de niveau* $1-\alpha$ *basé sur les quantiles de la loi a posteriori. Alors* $I_{quant}(\alpha)$ *est un intervalle de confiance asymptotique de niveau* $1-\alpha$.



(a)



(b)

FIGURE 3.8 – Illustration de la convergence de la loi a posteriori vers une loi centrée sur l'estimateur du maximum de vraisemblance. Figure (a) : les traits en abscisse représentent la valeur de l'estimateur du maximum de vraisemblance calculé sur l'échantillon de taille n. Figure (b) : les lignes en pointillés représentent la valeur de l'estimateur du maximum de vraisemblance, et la courbe en gris correspond à la loi a priori.

Quatrième partie

Algorithme Espérance-Maximisation

Introduction

L'algorithme Espérance-Maximisation (EM) est un algorithme itératif dont l'objectif est d'obtenir le maximum (local) de vraisemblance d'un modèle pour lequel la vraisemblance n'a pas une expression explicite. Il est notamment utilisé pour les modèles dits à variables latentes, c'est-à-dire contenant des variables non observées. Dans un modèle à variables latentes (on parle aussi de modèle à variables cachées ou manquantes), on dispose d'un échantillon i.i.d. $X_1, \dots, X_n$ de variables observées, et d'un ensemble $Z_1, \dots, Z_n$ de variables non observées. On s'intéresse plus particulièrement au cas où la loi des X_i ne s'exprime pas facilement, mais où la loi jointe des (X_i, Z_i) a une expression beaucoup plus simple. Ceci signifie que, si on observait à la fois les X_i et les Z_i, on obtiendrait une fonction de vraisemblance qui serait "facile" à maximiser.

Exemple 17 (Données censurées). *Un exemple typique de modèle à variables cachées est celui des données censurées. Cette situation arrive fréquemment dans les études cliniques, lorsqu'on s'intéresse par exemple à la survie des patients après administration d'un traitement donné. Une étude clinique est en effet limitée dans le temps : on ne suit pas tous les patients jusqu'à observer leurs dates de décès, mais on les suit sur une période de temps fixée, par exemple 5 ans. A la fin de l'étude, on observe donc les durées de vie des patients selon deux catégories : si le patient i est décédé au cours de l'étude, on observe sa durée de vie post traitement, si le patient est encore en vie à la fin de l'étude, on observe à la place la durée de l'étude. Autrement dit, en notant Y_i la durée de vie du patient i, X_i la durée de vie observée et C la date de censure (d'arrêt de l'étude) on a :*

$$X_i = \begin{cases} Y_i & si\ Y_i \leq C \\ C & sinon \end{cases} \tag{1}$$

Les variables observées sont donc les $X_1, \dots, X_n$ et les variables cachées sont les (Y_i) pour lesquels $Y_i > C$.

Un autre exemple de données censurées survient dans le cas où on mesure une quantité à l'aide d'un appareil de mesure ayant un certain seuil de détection. En-dessous ou au-dessus de ce seuil, selon le type, on n'observera pas la vraie valeur, celle-ci étant tronquée au niveau du seuil de détection.

Chapitre 1

Présentation générale

L'algorithme EM a été introduit à la fin des années 1970 par l'article fondateur de (Dempster *et al.*, 1977), dans lequel sont énoncés les principes généraux, et qui a donné son nom à l'algorithme. D'autres auteurs ont développé avant eux des algorithmes similaires, mais dans des cas particuliers. Ainsi, la plus ancienne référence à un algorithme de type EM revient à Newcomb (1886) pour l'estimation des paramètres d'un modèle de mélange gaussien. Il est particulièrement adapté aux cas où la vraisemblance des données complètes s'écrit plus simplement que la vraisemblance des données observées, et repose sur l'idée suivante : lorsque l'on se trouve en présence de données manquantes, une première intuition est d'estimer ou de remplacer ces données manquantes, puis d'estimer les paramètres du modèle à l'aide des données « augmentées ».

Les modèles à variables latentes sont des modèles pour lesquels on peut définir deux ensembles de variables : $X_1, \ldots, X_n$, les variables observées et $Z_1, \ldots, Z_n$ les variables latentes, cachées ou manquantes. On appelle aussi les X_i *données observées*, les Z_i *données manquantes* et les couples (X_i, Z_i) *données complètes*.

1.1 Définitions et notations

On note $\mathbf{X} = (X_1, \ldots, X_n)$ le vecteur des données observées et $\mathbf{Z} = (Z_1, \ldots, Z_n)$ le vecteur des données latentes. Dans la suite, on note $f(\cdot; \theta)$ la densité jointe des (X_i) et $g(\cdot; \theta)$ la densité jointe des couples (X_i, Z_i), avec $\theta \in \Theta \subset \mathbb{R}^p$:

$$\mathbf{X} \sim f(\mathbf{x}; \theta)$$
$$(\mathbf{X}, \mathbf{Z}) \sim g(\mathbf{x}, \mathbf{z}; \theta)$$

La plupart du temps, les modèles à variables latentes sont décrits sous la forme hiérarchique suivante :

$$X_i \mid Z_i \sim f_{X|Z}(x; z, \theta)$$
$$Z_i \sim f_Z(z; \theta)$$

On a alors $g(x, z; \theta) = f_{X|Z}(x; z, \theta) f_Z(z; \theta)$ et

$$f(x; \theta) = \int g(x, z; \theta) dz = \int f_{X|Z}(x; z, \theta) f_Z(z; \theta) dz. \tag{1.1}$$

On définit la *vraisemblance observée* et la *vraisemblance complète* respectivement comme la loi jointe des $X_i, \ldots, X_n$ et la loi jointe des $(X_1, Z_1), \ldots, (X_n, Z_n)$:

$$L_{obs}(\mathbf{X}; \theta) = f(\mathbf{X}; \theta) \tag{1.2}$$

$$L_{comp}(\mathbf{X}, \mathbf{Z}; \theta) = g(\mathbf{X}, \mathbf{Z}; \theta) \tag{1.3}$$

En cas de variables i.i.d., la vraisemblance peut s'exprimer comme un produit de densités. En utilisant l'écriture de f dans (1.1), on pourrait alors utiliser une approche de type Monte Carlo pour évaluer la vraisemblance observée à l'aide de $n \times M$ réalisations $\tilde{Z}_{i1}, \ldots, \tilde{Z}_{iM}$ i.i.d. selon la loi f_Z, pour un θ fixé :

$$\hat{L}_{obs}(\mathbf{X}; \theta) = \prod_{i=1}^{n} \left(\frac{1}{M} \sum_{k=1}^{M} f_{X|Z}(X_i; \tilde{Z}_{ik}) \right). \tag{1.4}$$

Cette approche permet d'approcher la valeur de la vraisemblance en un point θ fixé, mais n'est pas utilisable pour maximiser la fonction : il faudrait faire cette approximation pour plusieurs valeurs de θ, bien choisies pour recouvrir l'ensemble des valeurs possibles, et à chaque fois cela implique l'approximation de n intégrales ...

1.2 Idée générale

L'algorithme EM s'utilise lorsque la vraisemblance complète est plus facile à maximiser que la vraisemblance observée. L'idée principale est de trouver un moyen de relier ces deux fonctions, et de maximiser la vraisemblance complète dans l'objectif de maximiser la vraisemblance observée. On note $\ell_{obs}(\theta)$ la log-vraisemblance observée et $\ell_{comp}(\theta)$ la log-vraisemblance complète :

$$\ell_{obs}(\theta) = \log L_{obs}(\mathbf{X}; \theta)$$
$$\ell_{comp}(\theta) = \log L_{comp}(\mathbf{X}, \mathbf{Z}; \theta)$$

Si on note $h(\cdot \mid \mathbf{X}, \theta)$ la densité de la loi conditionnelle des variables latentes sachant les variables observées, i.e. la loi conditionnelle de $\mathbf{Z}$ sachant $\mathbf{X}$, on a :

$$h(\mathbf{Z} \mid \mathbf{X}; \theta) = \frac{g(\mathbf{X}, \mathbf{Z}; \theta)}{f(\mathbf{X}; \theta)}$$

On peut maintenant faire le lien entre la log-vraisemblance complète et la log-vraisemblance observée. En effet on a :

$$\log f(\mathbf{X}; \theta) = \log g(\mathbf{X}, \mathbf{Z}; \theta) - \log h(\mathbf{Z} \mid \mathbf{X}; \theta).$$

D'où :

$$\ell_{obs}(\theta) = \ell_{comp}(\theta) - \log h(\mathbf{Z} \mid \mathbf{X}; \theta). \tag{1.5}$$

On aimerait obtenir une expression qui ne dépende pas des Z_i, ces derniers n'étant pas observés. Comme $\mathbf{Z}$ est inconnu, l'idée est d'intégrer par rapport à la loi conditionnelle de $\mathbf{Z}$ sachant $\mathbf{X}$, afin d'obtenir une expression qui ne dépende plus de $\mathbf{Z}$. En remarquant que la log-vraisemblance observée est $\sigma(\mathbf{X})$-mesurable, on obtient pour tout $\theta' \in \Theta$:

$$\int \ell_{obs}(\theta) h(z \mid \mathbf{X}; \theta') dz = \int \ell_{comp}(\theta) h(z \mid \mathbf{X}; \theta') dz - \int \log h(z \mid \mathbf{X}; \theta) h(z \mid \mathbf{X}; \theta') dz$$

$$\ell_{obs}(\theta) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta'}[\ell_{comp}(\theta)] - \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta'}[\log h(\mathbf{Z} \mid \mathbf{X}; \theta)],$$

où $\mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta'}$ représente l'espérance par rapport à la loi conditionnelle de $\mathbf{Z}$ sachant $\mathbf{X}$ en θ'. On peut donc exprimer la log-vraisemblance observée comme la différence entre deux quantités. Nous allons étudier plus en détails ces deux termes. Nous posons :

$$Q(\theta, \theta') = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta'}[\ell_{comp}(\theta)]$$
$$H(\theta, \theta') = -\mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta'}[\log h(\mathbf{Z} \mid \mathbf{X}; \theta)]$$

On a donc, pour tout $\theta' \in \Theta$:

$$\ell_{obs}(\theta) = Q(\theta, \theta') + H(\theta, \theta'). \tag{1.6}$$

Notons ici la distinction entre θ', qui est fixe, et qui permet de calculer l'espérance conditionnelle (on a besoin de fixer une valeur précise pour faire l'intégration), et θ, qui est une variable libre : on cherche à maximiser l'expression en fonction de θ.

Examinons maintenant ces deux fonctions Q et H, et commençons par la fonction H. Par définition, on a :

$$\begin{aligned} H(\theta; \theta') - H(\theta'; \theta') &= -\mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta'}[\log h(\mathbf{Z} \mid \mathbf{X}; \theta) - \log h(\mathbf{Z} \mid \mathbf{X}; \theta')] \\ &= -\mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta'}\left[\log \frac{h(\mathbf{Z} \mid \mathbf{X}; \theta)}{h(\mathbf{Z} \mid \mathbf{X}; \theta')}\right] \end{aligned}$$

En utilisant la concavité de la fonction logarithme et l'inégalité de Jensen [1], on en déduit :

$$\begin{aligned} H(\theta; \theta') - H(\theta'; \theta') &\geq -\log \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta'}\left[\frac{h(\mathbf{Z} \mid \mathbf{X}; \theta)}{h(\mathbf{Z} \mid \mathbf{X}; \theta')}\right] \\ &\geq -\log \int \frac{h(\mathbf{z} \mid \mathbf{X}; \theta)}{h(\mathbf{z} \mid \mathbf{X}; \theta')} h(\mathbf{z} \mid \mathbf{X}; \theta') d\mathbf{z} \\ &\geq -\log \int h(\mathbf{z} \mid \mathbf{X}; \theta) d\mathbf{z} \\ &\geq 0 \end{aligned}$$

En particulier, on a $H(\theta; \theta') = H(\theta; \theta)$ si et seulement si $\theta = \theta'$.

L'algorithme EM procède alors de façon itérative : à partir d'une initialisation θ^0, l'itération m de l'algorithme consiste à décomposer la log-vraisemblance observée selon l'équation (1.6), à partir de la valeur courante des paramètres :

$$\ell_{obs}(\theta) = Q(\theta, \theta^{m-1}) + H(\theta, \theta^{m-1}). \tag{1.7}$$

1. Soit f une fonction convexe, et X une variable aléatoire telle que $\mathbb{E}(X)$ et $\mathbb{E}(f(X))$ existent. L'inégalité de Jensen nous dit que $\mathbb{E}(f(X)) \geq f(\mathbb{E}(X))$.

D'après l'inégalité précédente, à chaque itération de l'algorithme EM, on est donc assuré d'augmenter la fonction H. Il suffit donc de s'intéresser à la fonction Q et à la quantité $Q(\theta;\theta^{m-1})$. Chaque itération de l'algorithme EM se divise alors en deux étapes, l'une dite d'*espérance* (étape E) qui consiste à calculer la quantité $Q(\theta;\theta^{m-1})$ en fonction de la valeur courante de l'estimation θ^{m-1}, et une seconde de *maximisation* (étape M), qui consiste à maximiser la fonction Q. L'objectif est de maximiser la log-vraisemblance observée en maximisant Q.

1.3 Description de l'algorithme

En partant d'une valeur initiale θ^0, l'itération m de l'algorithme consiste à réaliser successivement les deux étapes E et M décrites ci-dessous. Une illustration de l'algorithme est présentée sur la figure 4.1.

Étape E (Espérance)
On calcule la quantité $Q(\theta;\theta^{m-1})$, en fonction de la valeur courante de θ^{m-1} :

$$Q(\theta;\theta^{m-1}) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta^{m-1}}[\ell_{comp}(\theta)] \tag{1.8}$$

Étape M (Maximisation)
On maximise la fonction Q par rapport à θ, et on met à jour le vecteur de paramètres de la façon suivante :

$$\theta^m = \arg\max_{\theta\in\Theta} Q(\theta;\theta^{m-1}). \tag{1.9}$$

Par construction de l'algorithme EM, celui-ci produit une séquence monotone, comme énoncé dans le théorème ci-dessous.

Théorème 15 (Monotonicité de l'algorithme EM). *La séquence $(\theta^m)_m$ obtenue avec l'algorithme EM vérifie :*

$$\ell_{obs}(\theta^{m+1}) \geq \ell_{obs}(\theta^m).$$

De plus, on a $\ell_{obs}(\theta^{m+1}) = \ell_{obs}(\theta^m)$ si et seulement si $Q(\theta;\theta^{m+1}) = Q(\theta;\theta^m)$.

Exemple 18. *Nous allons illustrer l'algorithme EM sur un exemple très basique. Supposons que l'on dispose de deux pièces de monnaies A et B, pour lesquelles la probabilité de tomber sur pile est respectivement θ_A et θ_B. On cherche à estimer θ_A et θ_B. Pour cela, on met en place l'expérience suivante : pour $i = 1,\dots,n$, on tire au sort équitablement l'une des deux pièces A ou B, on effectue 10 lancers avec la pièce sélectionnée et on note X_i le nombre de piles obtenus. Si on garde la trace de la pièce choisie à chaque tour i, nos données sont les couples (X_i, Z_i) où $Z_i = 1$ si le dé A a été choisi au tour i et $Z_i = 0$*

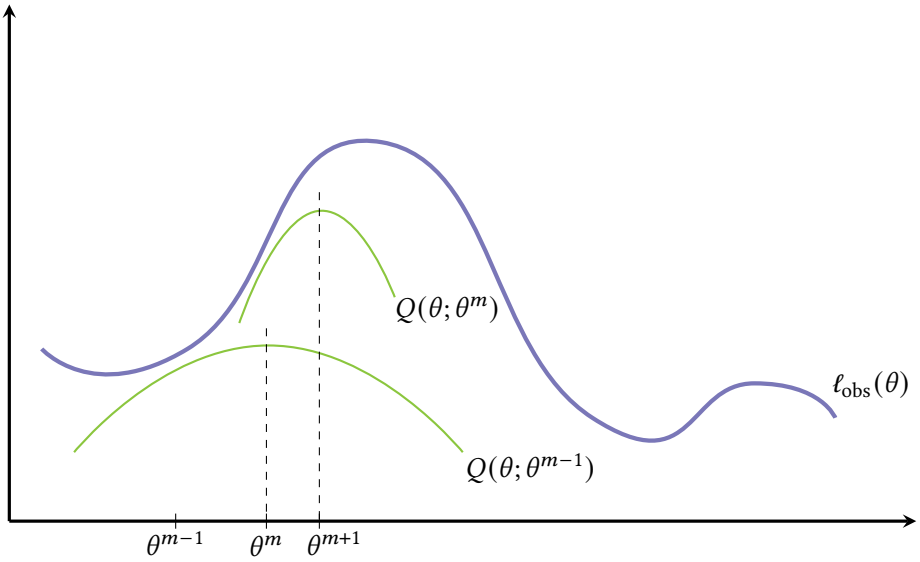


FIGURE 4.1 – Illustration de l'étape m de l'algorithme EM dans le cas où $\theta \in \mathbb{R}$. À partir de θ^{m-1}, on obtient la fonction $Q(\theta;\theta^{m-1})$ que l'on maximise afin d'obtenir θ^m. Puis, on recalcule la fonction Q à partir de la nouvelle valeur de θ^m, et ainsi de suite jusqu'à convergence de l'algorithme.

sinon. Dans ce cas, la vraisemblance s'écrit :

$$\begin{aligned}
L(x_1,\dots,x_n,z_1,\dots,z_n;\theta_A,\theta_B) &= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i, Z_i = z_i) \\
&= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i, Z_i = 1)^{z_i} \mathbb{P}(X_i = x_i, Z_i = 0)^{1-z_i} \\
&= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid Z_i = 1)^{z_i} \mathbb{P}(Z_i = 1)^{z_i} \mathbb{P}(X_i = x_i \mid Z_i = 0)^{1-z_i} \mathbb{P}(Z_i = 0)^{1-z_i} \\
&= \prod_{i=1}^{n} \left(\binom{10}{x_i} \theta_A^{x_i} (1-\theta_A)^{10-x_i} \right)^{z_i} \left(\binom{10}{x_i} \theta_B^{x_i} (1-\theta_B)^{10-x_i} \right)^{1-z_i}
\end{aligned}$$

La vraisemblance peut se décomposer en deux termes, l'un ne dépendant que de θ_A et l'autre ne dépendant que de θ_B. En passant au logarithme, on a donc :

$$\ell(\theta_A) \propto \sum_{i=1}^{n} (X_i \log(\theta_A) + (10 - X_i)\log(1-\theta_A)) Z_i$$

$$\ell(\theta_B) \propto \sum_{i=1}^{n} (X_i \log(\theta_B) + (10 - X_i)\log(1-\theta_B))(1 - Z_i)$$

On en déduit les estimateurs du maximum de vraisemblance suivants :

$$\hat{\theta}_A = \frac{1}{10 \sum_{i=1}^{n} Z_i} \sum_{i=1}^{n} X_i Z_i$$

$$\hat{\theta}_B = \frac{1}{10\sum_{i=1}^{n}(1-Z_i)}\sum_{i=1}^{n} X_i(1-Z_i)$$

Autrement dit, on estime θ_A (resp. θ_B) par le nombre de fois où la pièce A (resp. B) est tombée sur pile dans l'ensemble des lancers réalisés avec la pièce A (resp. B).

Que se passe t-il maintenant si on a oublié de noter la pièce ayant servi à effectuer les 10 lancers au tour i ? Dans ce cas, les $Z_1, \ldots, Z_n$ ne sont pas observées : ce sont les variables manquantes du modèle. Pour mettre en place l'algorithme EM dans ce contexte, on a besoin de construire la fonction Q qui intervient dans chacune des deux étapes E et M de chaque itération de l'algorithme. On se place à l'itération m de l'algorithme. On note $\theta = (\theta_A, \theta_B)$.

Etape E

$$Q(\theta, \theta^{m-1}) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta^{m-1}}[\ell_{comp}(\theta)]$$

La log-vraisemblance complète $\ell_{comp}(\theta) = \log L(x_1, \ldots, x_n, z_1, \ldots, z_n; \theta_A, \theta_B)$. Comme tout à l'heure, elle se décompose donc en deux termes, l'un lié à θ_A et l'autre lié à θ_B. On a :

$$\ell_{comp}(\theta) = C_A + \sum_{i=1}^{n}(X_i \log(\theta_A) + (10 - X_i)\log(1-\theta_A))Z_i + C_B + \sum_{i=1}^{n}(X_i \log(\theta_B) + (10 - X_i)\log(1-\theta_B))(1-Z_i)$$

Il nous faut maintenant calculer l'espérance de cette quantité par rapport à la loi conditionnelle de $\mathbf{Z}$ sachant $\mathbf{X}$, en θ^{m-1}. Or les Z_i sont des variables aléatoires valant 0 ou 1, donc des variables de Bernoulli. On a alors $Z_i \mid X_i = k \sim \mathcal{B}(p(k,\theta))$, où le paramètre de la loi de Bernoulli p est donné par :

$$\begin{aligned}
p(k,\theta) &= \mathbb{P}(Z_i = 1 \mid X_i = k) \\
&= \frac{\mathbb{P}(Z_i = 1, X_i = k)}{\mathbb{P}(X_i = k)} \\
&= \frac{\mathbb{P}(X_i = k \mid Z_i = 1)\mathbb{P}(Z_i = 1)}{\mathbb{P}(X_i = k, Z_i = 1) + \mathbb{P}(X_i = k, Z_i = 0)} \\
&= \frac{\mathbb{P}(X_i = k \mid Z_i = 1)\mathbb{P}(Z_i = 1)}{\mathbb{P}(X_i = k \mid Z_i = 1)\mathbb{P}(Z_i = 1) + \mathbb{P}(X_i = k \mid Z_i = 0)\mathbb{P}(Z_i = 0)} \\
&= \frac{\binom{10}{k}\theta_A^k(1-\theta_A)^{10-k}1/2}{\binom{10}{k}\theta_A^k(1-\theta_A)^{10-k}1/2 + \binom{10}{k}\theta_B^k(1-\theta_B)^{10-k}1/2} \\
&= \frac{\theta_A^k(1-\theta_A)^{10-k}}{\theta_A^k(1-\theta_A)^{10-k} + \theta_B^k(1-\theta_B)^{10-k}}
\end{aligned}$$

Finalement, on a :

$$Q(\theta, \theta^{m-1}) = C + \sum_{i=1}^{n}(X_i \log(\theta_A) + (10 - X_i)\log(1-\theta_A))\mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta^{m-1}}[Z_i] +$$

$$\sum_{i=1}^{n}(X_i \log(\theta_B) + (10 - X_i)\log(1-\theta_B))(1 - \mathbb{E}_{Z|X,\theta^{m-1}}[Z_i])$$
$$= C + \sum_{i=1}^{n}(X_i \log(\theta_A) + (10 - X_i)\log(1-\theta_A))p(X_i, \theta^{m-1}) +$$
$$\sum_{i=1}^{n}(X_i \log(\theta_B) + (10 - X_i)\log(1-\theta_B))(1 - p(X_i, \theta^{m-1}))$$

Là encore on peut décomposer cette fonction en deux termes, l'un ne dépendant que de θ_A et l'autre ne dépendant que de θ_B. On obtient alors des expressions très proches de ce que l'on avait pour $\ell(\theta_A)$ et $\ell(\theta_B)$:

$$Q(\theta_A, \theta^{m-1}) \propto \sum_{i=1}^{n}(X_i \log(\theta_A) + (10 - X_i)\log(1-\theta_A))p(X_i, \theta^{m-1})$$
$$Q(\theta_B, \theta^{m-1}) \propto \sum_{i=1}^{n}(X_i \log(\theta_B) + (10 - X_i)\log(1-\theta_B))(1 - p(X_i, \theta^{m-1}))$$

Etape M

On veut maintenant maximiser $Q(\theta_A, \theta^{m-1})$ et $Q(\theta_B, \theta^{m-1})$ en θ_A et θ_B. En dérivant chaque expression, on obtient :

$$\theta_A^m = \frac{1}{10\sum_{i=1}^{n} p(X_i, \theta^{m-1})} \sum_{i=1}^{n} X_i p(X_i, \theta^{m-1})$$
$$\theta_B^m = \frac{1}{10\sum_{i=1}^{n}(1 - p(X_i, \theta^{m-1}))} \sum_{i=1}^{n} X_i (1 - p(X_i, \theta^{m-1}))$$

On obtient des expressions très proches de ce que l'on avait obtenu pour les estimateurs du maximum de vraisemblance dans le cas où toutes les variables étaient observées. Ici, la variable indicatrice Z_i est remplacée par la probabilité pour que cette variable soit égale à 1. On pondère donc chaque observation X_i par la probabilité qu'elle corresponde au lancer de la pièce A ou B.

1.4 Convergence

La convergence de l'algorithme sous des conditions générales de régularité a été étudiée par Dempster *et al.* (1977) ; Wu (1983). La convergence de la séquence (θ^m) produite par l'algorithme EM vers le maximum de vraisemblance n'est pas garantie, et en général, dans la plupart des applications, la convergence a lieu vers un point stationnaire de la vraisemblance, qui peut être un maximum local ou global, ou un point-selle. Sous certaines conditions supplémentaires de régularité, Wu (1983) a montré que l'on peut s'assurer de la convergence vers un maximum local. Cependant, ces conditions peuvent être difficiles à vérifier dans la pratique, et l'algorithme peut alors se retrouver bloqué à un point stationnaire de la vraisemblance qui ne soit ni un maximum global, ni même un maximum local. Dans ces cas-là, une perturbation aléatoire du vecteur de paramètres peut permettre à l'algorithme de s'en éloigner. Il

s'agit d'un des avantages des versions stochastiques de l'algorithme EM, que nous verrons au chapitre 3.

Le théorème suivant assure la convergence de l'algorithme vers un point stationnaire de la log-vraisemblance.

THÉORÈME 16. *Si la fonction $Q(\theta;\theta')$ est continue en θ et en θ', alors la séquence $(\theta^m)_m$ produite par l'algorithme EM converge vers un point stationnaire θ^* de la log-vraisemblance, et la séquence $\ell_{obs}(\theta^m)$ converge de façon monotone vers $\ell_{obs}(\theta^*)$.*

En pratique l'algorithme EM converge donc, au mieux, vers un maximum local de la vraisemblance :

- il est recommandé de lancer l'algorithme plusieurs fois avec des initialisations différentes : si les différentes réalisations convergent vers la même valeur de θ, on pourra conclure qu'il s'agit probablement d'un maximum global de la vraisemblance, si la convergence a lieu vers différentes valeurs, il faut alors identifier celle correspondant à la valeur la plus élevée pour ℓ_{obs}.
- on peut définir un critère d'arrêt pour l'algorithme, soit basé sur la fonction Q, soit sur la séquence d'estimateurs. Par exemple, on peut se fixer un seuil ε et une norme $\|\cdot\|$ sur $\mathbb{R}^p$, et arrêter l'algorithme lorsque :
$$\|\theta^m - \theta^{m-1}\| < \varepsilon.$$

Chapitre 2

Modèles de mélange

2.1 Introduction et exemples

Les modèles de mélange sont des modèles très courants en statistique. Ils permettent de modéliser des échantillons regroupant plusieurs sous-populations. Ils peuvent également être utilisés dans le contexte de la classification non supervisée. Les modèles de mélange appartiennent à la famille des modèles à variables latentes, pour lesquels l'algorithme EM peut être utilisé pour obtenir le maximum de vraisemblance. Voici quelques exemples où les modèles de mélange interviennent :

- on mesure la taille en cm chez n adultes, mais on a oublié de noter le sexe de l'individu. Si on suppose que la taille d'un individu est distribuée selon une loi normale dont la moyenne dépend du sexe, on se retrouve alors avec deux sous-populations, c'est-à-dire un mélange de deux lois normales dont les moyennes seront différentes selon le sexe de l'individu. On parle alors de modèle de mélange gaussien. Graphiquement, cela peut se visualiser en traçant l'histogramme des données, qui doit laisser apparaître deux modes.

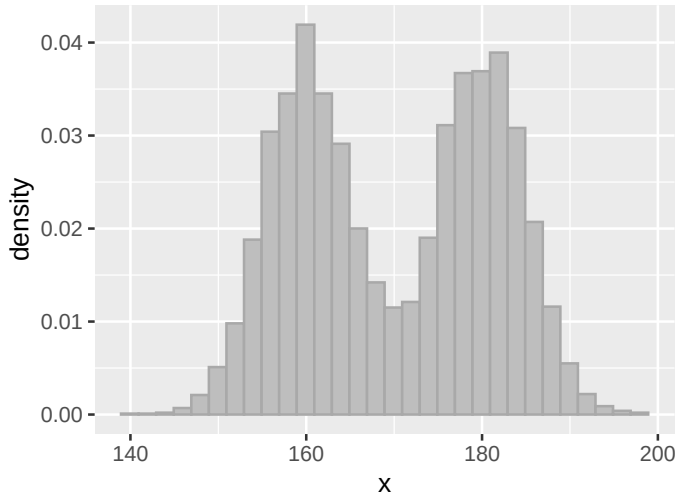


Figure 4.2 – Distribution de la taille dans une population adulte (hommes et femmes mélangés)

- on s'intéresse à l'abondance d'une espèce dans un milieu naturel, et pour cela on effectue des relevés sur différents sites, en comptant le nombre d'individus observés. Sur certains de ces sites, aucun individu de l'espèce ne sera observé : cela peut être dû au fait que l'espèce n'est pas présente sur ce site (ce que l'on appelle un "zéro structurel"), ou au fait qu'aucun individu n'a été repéré lors de l'étude (ce que l'on appelle un "zéro aléatoire", dû au fait que la probabilité de n'observer aucun individu est non nulle). Dans ce cas, on obtient un mélange entre une loi de comptage (par exemple une loi de Poisson) et un Dirac en 0. On parle de modèle à inflation de zéros. Graphiquement, cela peut se visualiser en traçant le barplot des données : on aperçoit un nombre plus élevé qu'attendu de zéros.

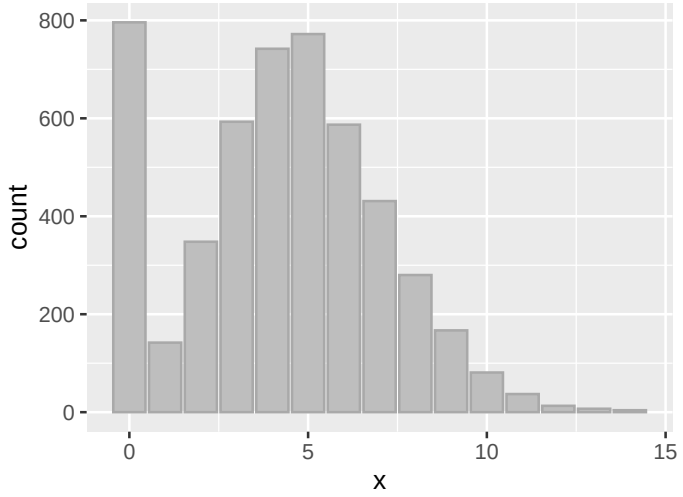


FIGURE 4.3 – Données de comptage à inflation de zéros (exemple avec une loi de Poisson comme loi de comptage).

2.2 Notations et définitions

Un modèle de mélange, on l'a vu, permet de modéliser plusieurs sous-populations. Notons K le nombre de sous-populations différentes dans l'échantillon. On a alors besoin de deux variables pour définir la distribution de la i-ème observation : Z_i, qui permet d'identifier la sous-population à laquelle appartient l'observation, puis X_i, qui décrit la loi de cette observation. La variable aléatoire Z_i est une variable aléatoire discrète pouvant prendre K valeurs possibles. On pose :

$$\mathbb{P}(Z_i = k) = \pi_k, \ k = 1, \dots, K,$$

avec $\sum_{k=1}^{K} \pi_k = 1$.

Supposons pour simplifier que $X_i \in \mathbb{R}$ ou $X_i \in \mathbb{N}$ (les résultats se généralisent pour des observations multidimensionnelles). On note $f_k(\cdot; \phi_k)$ la densité conditionnelle de X_i sachant que l'on est dans la population k, c'est-à-dire :

$$X_i \mid Z_i = k \sim f_k(\cdot; \phi_k),$$

où $f_k(\cdot;\phi_k)$ est une densité par rapport à une mesure de référence μ (la mesure de Lebesgue si $X_i \in \mathbb{R}$, la mesure de comptage si $X_i \in \mathbb{N}$). On note F_k la fonction de répartition associée à la densité $f_k(\cdot;\phi_k)$. La loi jointe de (X_i, Z_i) est donnée par :

$$\begin{aligned}\mathbb{P}(X_i \leq x, Z_i = k) &= \mathbb{P}(X_i \leq x \mid Z_i = k)\mathbb{P}(Z_i = k) \\ &= \pi_k F_k(x).\end{aligned}$$

La densité jointe par rapport à la mesure produit sur $\mathbb{R} \times \mathbb{N}$ est $\pi_k f_k$. On peut ré-écrire la densité jointe sous la forme :

$$\pi_k f_k(x) = \prod_{l=1}^{K} (\pi_k f_k(x))^{\mathbf{1}_{l=k}}$$

La loi marginale de X_i est donnée par :

$$\begin{aligned}\mathbb{P}(X_i \leq x) &= \sum_{k=1}^{K} \mathbb{P}(X_i \leq x, Z_i = k) \\ &= \sum_{k=1}^{K} \pi_k F_k(x).\end{aligned}$$

Autrement dit, la loi marginale de X_i admet aussi une densité par rapport à la mesure μ, donnée par :

$$f(x;\theta) = \sum_{k=1}^{K} \pi_k f_k(x;\phi_k), \tag{2.1}$$

avec $\theta = (\phi_1, \ldots, \phi_K, \pi_1, \ldots, \pi_K)$. Finalement, on définit la loi *conditionnelle* de Z_i sachant que $X = x$, par :

$$p_k(x;\theta) = \mathbb{P}_\theta(Z_i = k \mid X_i = x) = \frac{\pi_k f_k(x;\phi_k)}{f(x;\theta)} = \frac{\pi_k f_k(x;\phi_k)}{\sum_{k=1}^{K} \pi_k f_k(x;\phi_k)}.$$

Une densité de probabilité qui s'écrit comme dans (2.1) est appelée une *densité de mélange*. Les paramètres d'une loi de mélange sont les paramètres de chaque composante, c'est-à-dire les $\phi_1, \ldots, \phi_K$, ainsi que les probabilités $\pi_1, \ldots, \pi_K$.

2.3 Estimation

On s'intéresse maintenant à l'estimation de $\theta = (\pi_1, \ldots, \pi_K, \phi_1, \ldots, \phi_K)$. On est dans un contexte de modèle à variables latentes, pour lequel le problème de maximisation de la vraisemblance n'admet pas de solution explicite. On va donc mettre en place un algorithme EM. Parlons tout d'abord d'identifiabilité.

2.3.1 Identifiabilité

L'identifiabilité d'un modèle permet de s'assurer qu'il est possible de retrouver la vraie valeur du paramètre θ dont dépend la loi des observations. On dit d'un modèle statistique de loi $\mathbb{P}_\theta, \theta \in \Theta$ qu'il est identifiable si et seulement si :

$$\forall \theta, \theta' \in \Theta, \mathbb{P}_\theta = \mathbb{P}_{\theta'} \Rightarrow \theta = \theta'$$

La plupart des modèles statistiques que l'on a pu rencontrer jusque là étaient identifiables. Malheureusement, ce n'est pas le cas du modèle de mélange. En effet, il est toujours possible d'ajouter des composantes dans le mélange, associées à des poids π_k nuls, tout en conservant la même loi de mélange. Ou encore, de diviser une population en deux sous-populations ayant la même loi f_k. Par exemple, si on s'intéresse au modèle de mélange gaussien à deux composantes :

$$\pi_1 = 0.2, \pi_2 = 0.8, \quad f_1(x;\phi_1) = f_{\mathcal{N}}(x;0,1), f_2(x;\phi_2) = f_{\mathcal{N}}(x;2,1)$$

Celui-ci peut se ré-écrire :

$$\pi_1 = 0.2, \pi_2 = 0.8, \pi_3 = 0, \quad f_1(x;\phi_1) = f_{\mathcal{N}}(x;0,1), f_2(x;\phi_2) = f_{\mathcal{N}}(x;2,1), f_3(x;\phi_3) = f_{\mathcal{N}}(x;20,1) \quad (2.2)$$

ou encore :

$$\pi_1 = 0.2, \pi_2 = 0.6, \pi_3 = 0.2, \quad f_1(x;\phi_1) = f_{\mathcal{N}}(x;0,1), f_2(x;\phi_2) = f_{\mathcal{N}}(x;2,1), f_3(x;\phi_3) = f_{\mathcal{N}}(x;2,1). \quad (2.3)$$

On peut donc commencer par introduire deux contraintes sur les paramètres pour éviter ce genre de situation. En particulier, on supposera :

$$\pi_k > 0, \forall k = 1, \ldots, K \quad \text{et} \quad \phi_k \neq \phi_l,\ k \neq l.$$

Cela ne suffit pas encore cependant, car on peut obtenir la même loi de mélange en permutant les composantes. Autrement dit, le modèle ci-dessous est équivalent au modèle :

$$\pi_1 = 0.8, \pi_2 = 0.2, \quad f_1(x;\phi_1) = f_{\mathcal{N}}(x;2,1), f_2(x;\phi_2) = f_{\mathcal{N}}(x;0,1).$$

Ce phénomène est connu sous le nom de *label switching*. En pratique, on se contente alors de l'identifiabilité à permutation des paramètres près.

2.3.2 Algorithme EM

On rappelle que notre objectif est d'estimer $\theta = (\phi_1, \ldots, \phi_K, \pi_1, \ldots, \pi_K)^t$. Pour cela, on va utiliser l'algorithme EM pour obtenir l'estimateur de maximum de vraisemblance de θ.

Donnons les expressions des log-vraisemblances observée et complète dans le contexte des modèles de mélange. On a :

$$\ell_{obs}(\theta) = \sum_{i=1}^{n} \ln f(X_i;\theta) \quad (2.4)$$

$$= \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k f_k(X_i;\phi) \right) \quad (2.5)$$

$$\ell_{comp}(\theta) = \sum_{i=1}^{n} \ln f(X_i, Z_i;\theta) \quad (2.6)$$

$$= \sum_{i=1}^{n} \ln \prod_{k=1}^{K} (\pi_k f_k(X_i;\phi_k))^{\mathbf{1}_{Z_i=k}} \tag{2.7}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbf{1}_{Z_i=k} \ln (\pi_k f_k(X_i;\phi_k)) \,. \tag{2.8}$$

En partant d'une valeur initiale θ^0, l'itération m de l'algorithme EM consiste alors à réaliser successivement les deux étapes E et M décrites ci-dessous.

Étape E (Espérance)

On calcule la quantité $Q(\theta;\theta^{m-1})$, en fonction de la valeur courante de $\theta^{m-1} = (\phi_1^{m-1},\ldots,\phi_K^{m-1},\pi_1^{m-1},\ldots,\pi_K^{m-1})$. Comme $f_k(\mathbf{X}_i;\theta_k)$ est $\sigma(\mathbf{X})$-mesurable, on a :

$$\begin{aligned}
Q(\theta;\theta^{m-1}) &= \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta^{m-1}}\left[\ell_{\text{comp}}(\theta)\right] \\
&= \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta^{m-1}}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbf{1}_{Z_i=k} \ln (\pi_k f_k(X_i;\phi_k))\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta^{m-1}}[\mathbf{1}_{Z_i=k}] \ln (\pi_k f_k(\mathbf{X}_i;\phi_k)) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{P}_{\mathbf{Z}|\mathbf{X},\theta^{m-1}}(Z_i=k) \ln (\pi_k f_k(\mathbf{X}_i;\phi_k)) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{P}_{\theta^{m-1}}(Z_i=k \mid X_i=x) \ln (\pi_k f_k(\mathbf{X}_i;\phi_k)) \,.
\end{aligned}$$

Finalement, on obtient la fonction Q suivante :

$$Q(\theta;\theta^{m-1}) = \sum_{i=1}^{n}\sum_{k=1}^{K} p_k(X_i;\theta^{m-1}) \ln (\pi_k f_k(\mathbf{X}_i;\phi_k)) \tag{2.9}$$

$$Q(\theta;\theta^{m-1}) = \sum_{i=1}^{n}\sum_{k=1}^{K} p_k(X_i;\theta^{m-1}) \ln \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} p_k(X_i;\phi^{m-1}) \ln f_k(X_i;\phi_k), \tag{2.10}$$

avec :

$$p_k(X_i;\theta^{m-1}) = \frac{\pi_k^{m-1} f_k(X_i;\phi_k^{m-1})}{\sum_{k=1}^{K} \pi_k^{m-1} f_k(X_i;\phi_k^{m-1})}.$$

Étape M (Maximisation)

L'étape M consiste à maximiser la fonction Q par rapport à θ. Comme la fonction Q peut se décomposer en deux parties, l'une ne dépendant que de $\pi = (\pi_1,\ldots,\pi_K)$ et l'autre ne dépendant que de $\phi = (\phi_1,\ldots,\phi_K)$, la maximisation peut se faire indépendamment pour π et pour ϕ. On peut même décomposer la partie ne dépendant que de ϕ en K termes, chacun ne dépendant que de ϕ_k. En maximisant (2.9) par rapport à p_k on obtient :

$$\pi_k^m = \frac{1}{n}\sum_{i=1}^{n} p_k(X_i;\theta^{m-1}) \tag{2.11}$$

Pour la maximisation par rapport à ϕ_k, cela dépend de la loi f_k. Dans le cas gaussien, c'est-à-dire lorsque

$$f_k(x;\phi_k) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right),$$

on obtient :

$$\mu_k^m = \frac{\sum_{i=1}^n p_k(X_i;\theta^{m-1})\, X_i}{\sum_{i=1}^n p_k(X_i;\theta^{m-1})} \tag{2.12}$$

$$\sigma_k^{2,m} = \frac{\sum_{i=1}^n p_k(X_i;\theta^{m-1})\, (X_i-\mu_k^m)^2}{\sum_{i=1}^n p_k(X_i;\theta^{m-1})} \tag{2.13}$$

Démonstration. Pour les π_k, pour tenir compte de la contrainte $\sum_{k=1}^K \pi_k = 1$, on utilise la méthode des multiplicateurs de Lagrange. Le Lagrangien associé est $\mathcal{L}(\pi_1,\dots,\pi_K) = Q(\theta;\theta^{m-1}) + \lambda(\sum_{k=1}^K \pi_k - 1)$. On dérive par rapport à π_k :

$$\frac{\partial \mathcal{L}(\pi_1,\dots,\pi_K)}{\partial \pi_k} = \sum_{i=1}^n \frac{p_k(X_i;\theta^{m-1})}{\pi_k} + \lambda$$

$$\text{d'où} \quad \frac{\partial \mathcal{L}(\pi_1,\dots,\pi_K)}{\partial p_k} = 0 \Leftrightarrow \pi_k^m = -\frac{\sum_{i=1}^n p_k(X_i;\theta^{m-1})}{\lambda}$$

On obtient alors, sous la contrainte $\sum_{k=1}^K \pi_k = 1$:

$$\begin{aligned}
\lambda &= -\sum_{k=1}^K \sum_{i=1}^n p_k(X_i;\theta^{m-1}) \\
&= -\sum_{i=1}^n \sum_{k=1}^K p_k(X_i;\theta^{m-1}) \\
&= -\sum_{i=1}^n 1 \\
&= -n
\end{aligned}$$

D'où :

$$\pi_k^m = \frac{1}{n}\sum_{i=1}^n p_k(X_i;\theta^{m-1})$$

Pour les μ_k, on a :

$$\begin{aligned}
\frac{\partial Q(\theta;\theta^{m-1})}{\partial \mu_k} &= -\frac{1}{2\sigma_k^2}\sum_{i=1}^n -2\, p_k(X_i;\theta^{m-1})(X_i-\mu_k) \\
&= \frac{1}{\sigma_k^2}\sum_{i=1}^n p_k(X_i;\theta^{m-1})(X_i-\mu_k)
\end{aligned}$$

$$\begin{aligned}
\text{d'où} \quad \frac{\partial Q(\theta;\theta^{m-1})}{\partial \mu_k} = 0 &\Leftrightarrow \sum_{i=1}^n p_k(X_i;\theta^{m-1})X_i - \mu_k \sum_{i=1}^n p_k(X_i;\theta^{m-1}) = 0 \\
&\Leftrightarrow \mu_k^m = \frac{\sum_{i=1}^n p_k(X_i;\theta^{m-1})\, X_i}{\sum_{i=1}^n p_k(X_i;\theta^{m-1})}
\end{aligned}$$

Pour les variances on a :

$$\frac{\partial Q(\theta;\theta^{m-1})}{\partial \sigma_k^2} = \sum_{i=1}^{n} p_k(X_i;\theta^{m-1}) \left(-\frac{1}{2\sigma_k^2} + \frac{(X_i-\mu_k)^2}{2\sigma_k^4} \right)$$

$$\text{d'où} \quad \frac{\partial Q(\theta;\theta^{m-1})}{\partial \sigma_k^2} = 0 \Leftrightarrow \sigma_k^2 \sum_{i=1}^{n} p_k(X_i;\theta^{m-1}) = \sum_{i=1}^{n} p_k(X_i;\theta^{m-1})(X_i-\mu_k)^2$$

$$\Leftrightarrow \sigma_k^{2,m} = \frac{\sum_{i=1}^{n} p_k(X_i;\theta^{m-1})\ (X_i-\mu_k^m)^2}{\sum_{i=1}^{n} p_k(X_i;\theta^{m-1})}$$

□

2.3.3 Approche bayésienne

On peut aussi mettre en place une approche bayésienne pour estimer $\theta = (\pi, \phi)$. Dans ce cas, il faut définir une loi a priori π_θ sur Θ. On a alors :

$$\begin{aligned} X_i \mid Z_i = k; \theta &\sim f_k(\cdot;\phi_k) \\ \mathbb{P}(Z_i = k \mid \theta) &= \pi_k, \\ \theta &\sim \pi_\theta(\cdot). \end{aligned}$$

On cherche la loi a posteriori, c'est-à-dire la loi conditionnelle de θ sachant les observations $X_1, \ldots, X_n$ (on rappelle que les Z_i ne sont pas observées). On a :

$$\begin{aligned} p(\theta \mid x) &\propto f(\mathbf{X};\theta)\pi_\theta(\theta) \\ &\propto \prod_{i=1}^{n} f(X_i;\theta)\pi_\theta(\theta) \\ &\propto \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k f_k(X_i;\phi_k) \right) \pi_\theta(\theta) \end{aligned}$$

Cette loi a posteriori ne peut pas s'écrire de façon explicite, ce qui implique en particulier que les estimateurs ponctuels usuels tels que la moyenne ou la médiane a posteriori n'ont pas d'expressions analytiques. On a alors besoin de se tourner vers des méthodes permettant de simuler des réalisations selon cette loi a posteriori.

L'écriture du modèle de mélange sous forme de modèle hiérarchique nous incite à utiliser un échantillonneur de Gibbs. Ce qui intéressant ici, c'est que l'on va se servir du fait que la loi jointe des (X_i, Z_i) est plus simple à manipuler que la loi jointe des X_i. En particulier, on va utiliser l'échantillonneur de Gibbs pour simuler des réalisations du couple $(Z, \theta \mid X)$. En ne s'intéressant qu'aux coordonnées correspondant à θ, on obtiendra des réalisations selon la loi a posteriori, i.e. selon la loi conditionnelle de θ sachant X. Pour utiliser l'échantillonneur de Gibbs on a besoin de connaître les lois conditionnelles complètes, c'est-à-dire les lois de $\mathbf{Z}$ sachant $(\mathbf{X}, \theta)$ et de θ sachant $(\mathbf{Z}, \mathbf{X})$.
Pour la première, on a, pour $\mathbf{z} = (z_1, \ldots, z_n) \in \{1, \ldots, K\}^n$:

$$\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathbf{X} = \mathbf{x}, \theta) = \frac{f_{X,Z,\theta}(\mathbf{x}, \mathbf{z}, \theta)}{f_{X,\theta}(\mathbf{x}, \theta)}$$

$$
\begin{aligned}
&= \frac{f_{X|Z,\theta}(\mathbf{x};\mathbf{z},\theta) f_{Z|\theta}(\mathbf{z};\theta)\pi_\theta(\theta)}{f_{X,\theta}(\mathbf{x},\theta)} \\
&\propto f_{X|Z,\theta}(\mathbf{x};\mathbf{z},\theta) f_{Z|\theta}(\mathbf{z};\theta) \\
&\propto \prod_{i=1}^{n} f_{z_i}(x_i,\phi_{z_i})\pi_{z_i}
\end{aligned}
$$

On voit en particulier que la loi conditionnelle complète jointe des $(Z_1,\dots,Z_n)$ s'écrit comme un produit de n termes, chacun ne dépendant que de z_i : cela signifie que, conditionnellement à $X_1,\dots,X_n$ et θ, les Z_i sont indépendants. On en déduit :

$$
\mathbb{P}(Z_i = k \mid \mathbf{X} = \mathbf{x}, \theta) \propto \pi_k f_k(x_i,\phi_k)
$$

et donc :

$$
\mathbb{P}(Z_i = k \mid \mathbf{X} = \mathbf{x}, \theta) = \frac{\pi_k f_k(x_i,\phi_k)}{\sum_{k=1}^{K} \pi_k f_k(x_i,\phi_k)}
$$

Toutes les quantités qui interviennent dans cette expression sont explicites : on peut simuler très facilement selon cette loi conditionnelle discrète.

Pour la deuxième loi conditionnelle complète, on a :

$$
\begin{aligned}
f_{\theta|Z,X}(\theta \mid \mathbf{z},\mathbf{x}) &= \frac{f_{X,Z,\theta}(\mathbf{x},\mathbf{z},\theta)}{f_{X,Z}(\mathbf{x},\mathbf{z})} \\
&= \frac{f_{X|Z,\theta}(\mathbf{x};\mathbf{z},\theta) f_{Z|\theta}(\mathbf{z};\theta)\pi_\theta(\theta)}{f_{X,Z}(\mathbf{x},\mathbf{z})} \\
&\propto f_{X|Z,\theta}(\mathbf{x};\mathbf{z},\theta) f_{Z|\theta}(\mathbf{z};\theta)\pi_\theta(\theta) \\
&\propto \pi_\theta(\theta) \prod_{i=1}^{n} f_{z_i}(x_i,\phi_{z_i})\pi_{z_i}
\end{aligned}
$$

Cette loi dépend de la loi a priori. Un choix classique consiste à choisir deux lois a priori indépendantes, l'une pour les $\pi_1,\dots,\pi_K$, et l'autre pour $\phi_1,\dots,\phi_K$. Dans ce cas-là, la loi a priori s'écrit comme un produit de deux termes, l'un ne dépendant que de $\pi_1,\dots,\pi_K$, et l'autre ne dépendant que de $\phi_1,\dots,\phi_K$. On peut alors également décomposer la loi conditionnelle complète ci-dessus en un produit de deux termes :

$$
f_{\theta|Z,X}(\theta \mid \mathbf{z},\mathbf{x}) = f_{\pi|Z,X}(\pi \mid \mathbf{z},\mathbf{x}) f_{\phi|Z,X}(\phi \mid \mathbf{z},\mathbf{x})
$$

$$
f_{\pi|Z,X}(\pi \mid \mathbf{z},\mathbf{x}) \propto \pi_\pi(\pi) \prod_{i=1}^{n} \pi_{z_i} \quad f_{\phi|Z,X}(\phi \mid \mathbf{z},\mathbf{x}) \propto \pi_\phi(\phi) \prod_{i=1}^{n} f_{z_i}(x_i,\phi_{z_i})
$$

Pour la loi a priori sur π, il est classique de choisir une loi de Dirichlet d'ordre K. Cette loi est une généralisation de la loi Beta au cas multidimensionnel. En effet, cette loi est à valeurs dans le simplexe de dimension $K-1$, c'est-à-dire que chaque composante d'un vecteur suivant une loi de Dirichlet est comprise entre 0 et 1, et la somme des composantes est égale à 1. Marginalement, chaque composante suit une loi Beta. C'est une loi usuelle, que l'on peut retrouver sur la plupart des logiciels classiques de statistique. La loi conditionnelle complète est alors aussi une loi de Dirichlet :

$$
\pi \sim \mathcal{D}(\alpha_1,\dots,\alpha_K)
$$

$$\pi \mid \mathbf{X}, \mathbf{Z} \sim \mathcal{D}(\alpha_1 + n_1, \ldots, \alpha_K + n_K),$$

où $n_k = \sum_{i=1}^{n} \mathbf{1}_{Z_i=k}$ est le nombre d'observations dans la k-ème composante.

Pour la loi a priori sur ϕ, on peut là encore décomposer en autant de termes qu'il y a de composantes dans le mélange. Cela permet en effet d'exploiter le fait qu'on a cette même décomposition dans le terme $\prod_{i=1}^{n} f_{z_i}(x_i, \phi_{z_i})$. Autrement dit, on va choisir des lois a priori indépendantes pour chaque $\phi_k, k = 1, \ldots, K$. De cette façon, on obtient également une décomposition de la loi $f_{\phi|Z,X}(\phi \mid \mathbf{z}, \mathbf{x})$ en un produit de K termes, chacun ne dépendant que de ϕ_k. Dans le cas d'un modèle de mélange gaussien, on peut choisir des lois a priori gaussiennes pour chaque moyenne μ_k et des lois inverse-Gamma pour les paramètres de variance σ_k^2 :

$$\mu_k \mid \sigma_k^2 \sim \mathcal{N}\left(m_k, \frac{\sigma_k^2}{\tau_k}\right)$$
$$\sigma_k^2 \sim \mathcal{IG}\left(\frac{a_k}{2}, \frac{b_k}{2}\right)$$

où m_k, τ_k, a_k, b_k sont les hyperparamètres. On récupère ainsi des lois conditionnelles complètes conjuguées :

$$\mu_k \mid \sigma_k^2, \mathbf{X}, \mathbf{Z} \sim \mathcal{N}\left(\frac{m_k\tau_k + \sum_{i,Z_i=k} X_i}{\tau_k + n_k}, \frac{\sigma_k^2}{\tau_k + n_k}\right)$$
$$\sigma_k^2 \mid \mathbf{X}, \mathbf{Z} \sim \mathcal{IG}\left(\frac{a_k + n_k}{2}, \frac{b_k + \sum_{i,Z_i=k}(X_i - \mu_k)^2}{2}\right)$$

Chapitre 3

Extensions

L'algorithme EM nécessite, à chaque itération, deux opérations : d'une part le calcul de la fonction Q, qui s'écrit comme une espérance, et d'autre part la maximisation de cette fonction Q. On a vu jusqu'à présent des exemples où ces deux étapes pouvaient s'écrire analytiquement. Que faire si on ne parvient pas à écrire l'espérance selon la loi conditionnelle des variables latentes sachant les observations ? et si la maximisation de la fonction Q n'est pas possible analytiquement ? Plusieurs extensions de l'algorithme originel ont été proposées pour pallier à ces problèmes.

3.1 Maximisation difficile ou non explicite

Lorsque l'étape de maximisation n'est pas faisable explicitement, on peut tout d'abord se tourner vers des méthodes de maximisation numérique. Il s'agit alors d'utiliser des algorithmes d'optimisation de type Newton-Raphson, ou quasi-Newton, pour maximiser la fonction Q.

Lorsque le paramètre θ est multidimensionnel, il est aussi possible de remplacer l'étape de maximisation globale par une succession d'étapes de maximisation conditionnelles, où chaque coordonnée de θ est maximisée conditionnellement aux autres coordonnées. C'est l'algorithme ECM (pour "Expectation Conditional Maximization"), que l'on peut traduire en français par "Espérance - Maximisation Conditionnelle".

Une autre approche est possible, c'est celle de l'algorithme EM généralisé. Cette méthode a été proposée en même temps que l'algorithme EM, et est présentée notamment dans l'ouvrage de référence de McLachlan et Krishnan (2007). Il s'agit de remplacer l'étape de maximisation par une étape d'amélioration de la fonction Q. Autrement dit, l'étape M de l'itération m de l'algorithme est remplacée par l'étape suivante : trouver θ^m tel que $Q(\theta^m; \theta^{m-1}) > Q(\theta^{m-1}; \theta^{m-1})$.

3.2 Variants stochastiques

Dans le cas où l'étape E d'espérance ne peut pas se faire explicitement, des extensions ont été proposées, reposant sur une estimation de cette espérance par des méthodes de type Monte-Carlo. On

présente ici deux variants stochastiques de l'algorithme EM : l'algorithme MCEM (pour Monte-Carlo EM (Wei et Tanner, 1990) et l'algorithme SAEM (Kuhn et Lavielle, 2004).

3.2.1 Monte Carlo-EM

Une première approche lorsque le calcul de l'espérance n'est pas explicite consiste à approcher l'intégrale par une approximation de type Monte-Carlo. Cependant, cela nécessite de pouvoir simuler facilement selon la loi conditionnelle de $\mathbf{Z}$ sachant $\mathbf{X}$, ce qui n'est pas toujours le cas dans les problèmes de données incomplètes. Une autre méthode consiste alors à utiliser une approche de type MCMC.

À l'itération m de l'algorithme, on remplace alors l'étape E par une étape de simulation :

Étape S (Simulation)
On génère une chaîne de Markov de taille n_m de loi stationnaire $h(\cdot \mid \mathbf{X}; \theta^{m-1})$. En notant $\left(\mathbf{Z}^1, \ldots, \mathbf{Z}^{n_m}\right)$ les n_m réalisations de la chaîne de Markov, la fonction $Q(\theta; \theta^{m-1})$ peut alors être approchée par

$$\hat{Q}(\theta; \theta^{m-1}) = \frac{1}{n_m} \sum_{k=1}^{n_m} \log L_{comp}(\mathbf{X}, \mathbf{Z}^k; \theta^{m-1}). \tag{3.1}$$

Simulations de Monte-Carlo

On présente ci-dessous quelques exemples d'algorithmes MCMC qui peuvent être utilisés, et notamment le cas de l'algorithme de Metropolis-Hastings. La loi cible est la loi conditionnelle $h(\mathbf{Z} \mid \mathbf{X}; \theta^{m-1})$. On peut choisir comme loi instrumentale la loi marginale $f_Z(\mathbf{Z}; \theta^{m-1})$. En effet, à l'itération $k+1$ de l'algorithme de Metropolis-Hastings, on a alors la simplification suivante dans le calcul de la probabilité d'acceptation de l'algorithme, pour un candidat $\tilde{Z}$ et sachant l'état courant de la chaîne $\mathbf{Z}^k$:

$$\begin{aligned}
\alpha(\tilde{Z}, \mathbf{Z}^k) = \frac{h(\tilde{Z} \mid \mathbf{X}; \theta^{m-1})}{h(\mathbf{Z}^k \mid \mathbf{X}; \theta^{m-1})} \frac{f_Z(\mathbf{Z}^k; \theta^{m-1})}{f_Z(\tilde{Z}; \theta^{m-1})} &= \frac{g(\mathbf{X}, \tilde{Z}; \theta^{m-1}) / f(\mathbf{X}; \theta^{m-1})}{g(\mathbf{X}, \mathbf{Z}^k; \theta^{m-1}) / f(\mathbf{X}; \theta^{m-1})} \frac{f_Z(\mathbf{Z}^k; \theta^{m-1})}{f_Z(\tilde{Z}; \theta^{m-1})} \\
&= \frac{g(\mathbf{X}, \tilde{Z}; \theta^{m-1})}{f_Z(\tilde{Z}; \theta^{m-1})} \frac{f_Z(\mathbf{Z}^k; \theta^{m-1})}{g(\mathbf{X}, \mathbf{Z}^k; \theta^{m-1})} \\
&= \frac{f_{X|Z}(\mathbf{X}; \tilde{Z}, \theta^{m-1})}{f_{X|Z}(\mathbf{X}; \mathbf{Z}^k, \theta^{m-1})}
\end{aligned}$$

La probabilité d'acceptation ne fait donc intervenir que la loi conditionnelle de $\mathbf{X}$ sachant $\mathbf{Z}$, qui est connue.

On peut également proposer un algorithme de Metropolis-Hastings à marche aléatoire. Dans ce cas la probabilité d'acceptation ne fait intervenir que le ratio des densités cibles. En reprenant les calculs fait ci-dessus, on obtient alors :

$$\begin{aligned}
\alpha(\tilde{Z}, \mathbf{Z}^k) = \frac{h(\tilde{Z} \mid \mathbf{X}; \theta^{m-1})}{h(\mathbf{Z}^k \mid \mathbf{X}; \theta^{m-1})} &= \frac{g(\mathbf{X}, \tilde{Z}; \theta^{m-1})}{g(\mathbf{X}, \mathbf{Z}^k; \theta^{m-1})} \\
&= \frac{f_{X|Z}(\mathbf{X}; \tilde{Z}, \theta^{m-1}) f_Z(\tilde{Z}; \theta^{m-1})}{f_{X|Z}(\mathbf{X}; \mathbf{Z}^k, \theta^{m-1}) f_Z(\mathbf{Z}^k; \theta^{m-1})}
\end{aligned}$$

Là encore, la probabilité d'acceptation ne fait intervenir que des densités connues et se calcule donc facilement.

Taille de la chaîne et critère d'arrêt

Lorsque la fonction Q est approchée par des méthodes de type Monte-Carlo ou Monte-Carlo par chaîne de Markov, il faut tenir compte de l'erreur commise en remplaçant Q par $\hat{Q}$. En particulier, la propriété de monotonie de l'algorithme EM, qui garantissait une augmentation de la vraisemblance à chaque itération, n'est plus nécessairement vérifiée.

De plus, l'utilisation d'une taille m_k constante ne permet pas la convergence de l'algorithme, à cause de la persistance de l'erreur de Monte Carlo. Une première approche consiste alors à augmenter la taille de la chaîne afin d'obtenir une meilleure précision au fur et à mesure que l'algorithme EM progresse. L'idée est donc de démarrer avec une chaîne de taille suffisante sans être trop grande, puis d'augmenter la taille de la chaîne à chaque itération, de façon déterministe, pour obtenir des estimations de plus en plus précises à mesure que l'on s'approche du maximum de vraisemblance.

Une autre consiste à retrouver la propriété de monotonie de l'algorithme EM, en utilisant un critère permettant de retrouver cette propriété avec une forte probabilité. Plus précisément, à chaque itération m, on calcule un intervalle de confiance de niveau α pour $\Delta Q_m = Q(\theta^m; \theta^{m-1}) - Q(\theta^{m-1}; \theta^{m-1})$, basé sur l'approximation normale suivante :

$$\sqrt{n_m}\left(\Delta\hat{Q}_m - \Delta Q_m\right) \longrightarrow \mathcal{N}(0, \sigma_Q^2), \tag{3.2}$$

où $\Delta\hat{Q}_m = \hat{Q}(\theta^m; \theta^{m-1}) - \hat{Q}(\theta^{m-1}; \theta^{m-1})$, et où $\hat{Q}$ est calculée selon l'équation (3.1). Si la borne inférieure de cet intervalle de confiance est supérieure à 0, on accepte la nouvelle estimation θ^m, et sinon, on augmente la taille de l'échantillon. Une augmentation géométrique est suggérée par les auteurs, c'est-à-dire de type $n_m \leftarrow n_m + n_m/c$, où $c = 2, 3, \ldots$ D'un point de vue pratique, cela revient à générer un nouvel échantillon de type MCMC que l'on annexe à l'échantillon courant, puis à ré-évaluer l'intervalle de confiance. La procédure est répétée jusqu'à ce que le candidat soit accepté.

L'un des inconvénients de cette approche est qu'elle nécessite l'utilisation d'une méthode adéquate pour estimer la variance σ_Q^2 dans le cas MCMC où les réalisations ne sont pas indépendantes et où le théorème central limite classique ne s'applique plus.

3.2.2 L'algorithme SAEM

Dans l'algorithme MCMC-EM présenté dans la section précédente, les simulations générées ne sont pas conservées d'une itération de l'algorithme à l'autre. De plus, la taille de la chaîne augmente avec le nombre d'itérations, ce qui peut conduire à un allongement du temps de calcul lorsque le modèle est complexe.

Une alternative, basée sur une méthode d'approximation stochastique (Robbins et Monro, 1951) consiste à ré-utiliser les réalisations des itérations précédentes en y associant un facteur de pondération

qui décroît avec la distance à l'itération courante. Plus formellement, l'étape E est remplacée par une étape de simulation et une étape d'approximation :

Étape S (Simulation)
On génère des échantillons i.i.d. ou on génère une chaîne de Markov de taille n_m de loi stationnaire $h(\cdot \mid \mathbf{X}; \theta^{m-1})$.

Étape A (Approximation Stochastique)
À partir d'une séquence décroissante de pas positifs (γ_m), la fonction Q à l'étape $m+1$ est approchée de la façon suivante :

$$\hat{Q}(\theta; \theta^m) = \hat{Q}(\theta; \theta^{m-1}) + \gamma_m \left[\frac{1}{n_m} \sum_{k=1}^{n_m} \log L_{comp}(\mathbf{X}, \mathbf{Z}^k; \theta^{m-1}) - \hat{Q}(\theta; \theta^{m-1}) \right]. \tag{3.3}$$

Convergence de l'algorithme

De même qu'avec l'algorithme MCMC-EM, des hypothèses supplémentaires sont nécessaires pour assurer la convergence de l'algorithme SAEM. La première condition requise pour assurer la convergence de l'algorithme SAEM porte sur le comportement de la séquence $\{\gamma_m\}$. En particulier, celle-ci doit vérifier :

$$\forall m \in \mathbb{N}, \gamma_m \in [0,1], \sum_{m=1}^{\infty} \gamma_m = \infty \text{ et } \exists \lambda \in]1/2, 1] \text{ tel que } \sum_{m=1}^{\infty} \gamma_m^{1+\lambda} < \infty.$$

La méthode converge à une vitesse optimale pour des pas de l'ordre de $\gamma_m \propto m^{-a}$, avec $1/2 < a \leq 1$.

D'autres hypothèses sont nécessaires pour assurer la convergence de cet algorithme, et portent notamment sur la régularité de la log-vraisemblance observée, ou sur les propriétés ergodiques de la chaîne de Markov intervenant dans l'étape de simulation.

Cinquième partie

Annexes

Annexe A

Lois usuelles

A.1 Lois discrètes

Loi de Bernoulli

Soit $p \in [0,1]$. Une variable aléatoire X suit la loi de Bernoulli de paramètre p, notée $\mathcal{B}(p)$, si :

$$\mathbb{P}(X=1) = p, \quad \mathbb{P}(X=0) = 1-p.$$

$$\mathbb{E}(X) = p$$
$$\mathrm{Var}(X) = p(1-p)$$

Loi binomiale

Soit $p \in [0,1]$. Une variable aléatoire X suit la loi binomiale de paramètres n et p, notée $\mathcal{B}(n,p)$ si :

$$\forall k \in \{0,\dots,n\},\ \mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

$$\mathbb{E}(X) = np$$
$$\mathrm{Var}(X) = np(1-p)$$

Loi géométrique

Soit $p \in [0,1]$. Une variable aléatoire X suit la loi géométrique de paramètre p, notée $G(p)$ si :

$$\forall k \in \mathbb{N}^*,\ \mathbb{P}(X=k) = p(1-p)^{k-1}.$$

$$\mathbb{E}(X) = \frac{1}{p}$$
$$\mathrm{Var}(X) = \frac{1-p}{p^2}$$

Loi uniforme sur $\{1, \dots, n\}$

Une variable aléatoire X suit la loi uniforme sur $\{1, \dots, n\}$ si :

$$\forall k \in \{1, \dots, n\},\ \mathbb{P}(X = k) = \frac{1}{n}.$$

$$\mathbb{E}(X) = \frac{n+1}{2}$$
$$\mathrm{Var}(X) = \frac{n^2 - 1}{12}$$

Loi de Poisson

Soit $\lambda > 0$. Une variable aléatoire X suit la loi de Poisson de paramètre λ, notée $\mathcal{P}(\lambda)$ si :

$$\forall k \in \mathbb{N},\ \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\mathbb{E}(X) = \lambda$$
$$\mathrm{Var}(X) = \lambda$$

Loi binomiale négative

Soit $p \in [0, 1]$. Une variable aléatoire X suit la loi binomiale négative de paramètres n et p, notée $\mathcal{NB}(n, p)$ si :

$$\forall k \in \mathbb{N},\ \mathbb{P}(X = k) = \binom{n+k-1}{k} p^n (1-p)^k.$$

$$\mathbb{E}(X) = \frac{n(1-p)}{p}$$
$$\mathrm{Var}(X) = \frac{n(1-p)}{p^2}$$

Loi multinomiale

Soit $n > 0$ et $p_1, \dots, p_K$ tels que $p_k \in [0, 1] \forall k \in \{1, \dots, K\}$ et $\sum_k p_k = 1$. Une variable aléatoire $X = (X_1, \dots, X_K)$ à valeurs dans $\mathbb{N}^K$ suit la loi multinomiale de paramètres n et $(p_1, \dots, p_K)$ si :

$$\mathbb{P}(X_1 = n_1, \dots, X_K = n_K) = \frac{n!}{n_1! \dots n_K!} p_1^{n_1} \dots p_K^{n_K},$$

avec $\sum_k n_k = n$. Chaque composante X_k suit alors une loi binomiale $\mathcal{B}(n, p_k)$. Les composantes ne sont pas indépendantes.

$$\mathbb{E}(X_k) = np_k$$
$$\mathrm{Var}(X_k) = np_k(1-p_k)$$
$$\mathrm{Cov}(X_k, X_l) = -np_kp_l$$

A.2 Lois continues

Loi uniforme

Soient $a < b$. Une variable aléatoire X suit la loi uniforme sur l'intervalle $[a, b]$, notée $\mathcal{U}([a, b])$ si sa densité vérifie :

$$f(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x)$$

Sa fonction de répartition est :

$$F(x) = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } <\leq x < b \\ 1 & \text{si } x > b \end{cases}$$

$$\mathbb{E}(X) = \frac{a+b}{2}$$
$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

Loi normale

Soient $\mu \in \mathbb{R}$ et $\sigma > 0$. Une variable aléatoire X suit la loi normale de paramètres μ et σ^2, notée $\mathcal{N}(\mu, \sigma^2)$ si sa densité vérifie :

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Sa fonction de répartition n'a pas d'expression explicite. On note souvent Φ la fonction de répartition de la loi $\mathcal{N}(0, 1)$.

$$\mathbb{E}(X) = \mu$$
$$\mathrm{Var}(X) = \sigma^2$$

Loi log-normale

Soient $\mu \in \mathbb{R}$ et $\sigma > 0$. Une variable aléatoire X suit la loi log-normale de paramètres μ et σ^2, notée $\mathcal{LN}(\mu, \sigma^2)$ si $\ln X \sim \mathcal{N}(\mu, \sigma^2)$. Sa densité vérifie :

$$f(x) = \frac{1}{\sqrt{2\pi}x\sigma} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

$$\mathbb{E}(X) = e^{\mu+\sigma^2/2}$$
$$\mathrm{Var}(X) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$$

Loi de Cauchy

Soient $a > 0$. Une variable aléatoire X suit la loi de Cauchy de paramètres θ et ν, notée $C(\theta, \nu)$ si sa densité vérifie :

$$f(x) = \frac{1}{\pi} \frac{\nu}{(x-\theta)^2 + \nu^2}$$

Sa fonction de répartition est :

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan \frac{x-\theta}{\nu}$$

La loi de Cauchy n'admet pas d'espérance, ni de variance (ni aucun moment d'ordre supérieur).

Loi exponentielle

Soit $\lambda > 0$. Une variable aléatoire X suit la loi exponentielle de paramètre λ, notée $\mathcal{E}(\lambda)$ si sa densité vérifie :

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}_{\mathbb{R}^+}(x).$$

Sa fonction de répartition est :

$$F(x) = (1 - e^{-\lambda x}) \mathbf{1}_{\mathbb{R}^+}(x).$$

$$\mathbb{E}(X) = \frac{1}{\lambda}$$
$$\mathrm{Var}(X) = \frac{1}{\lambda^2}$$

Loi du chi-deux

Soient $\nu > 0$. Une variable aléatoire X suit la loi du chi-deux à ν degrés de liberté, notée χ^2_ν si sa densité vérifie :

$$f(x) = \frac{1}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}} \mathbf{1}_{\mathbb{R}^+}(x)$$

$$\mathbb{E}(X) = \nu$$
$$\mathrm{Var}(X) = 2\nu$$

Si $X \sim \mathcal{N}(0,1)$, alors $X^2 \sim \chi^2_1$. Si $X_1, \ldots, X_n$ sont i.i.d. de loi $\mathcal{N}(0,1)$, alors $\sum_{i=1}^n X_i \sim \chi^2_n$.

Loi Gamma

Soient $k > 0$ et $\theta > 0$. Une variable aléatoire X suit la loi Gamma de paramètres (k, θ), notée $\mathcal{G}(k, \theta)$ si sa densité vérifie :

$$f(x) = \frac{\theta^k}{\Gamma(k)} x^{k-1} e^{-\theta x} \mathbf{1}_{\mathbb{R}^+}(x).$$

$$\mathbb{E}(X) = \frac{k}{\theta}$$

$$\mathrm{Var}(X) = \frac{k}{\theta^2}$$

Si $X_i \sim \mathcal{G}(k_1, \theta)$ etles (X_i) sont mutuellement indépendantes, alors $\sum_{i=1}^n X_i \sim \mathcal{G}(\sum_{i=1}^n k_k, \theta)$. Si $X \sim \mathcal{G}(1, \theta)$, alors $X \sim \mathcal{E}(\theta)$. Si $X \sim \mathcal{G}(\frac{\nu}{2}, \frac{1}{2})$, alors $X \sim \chi^2_\nu$.

Loi inverse-Gamma

Soient $a > 0$ et $b > 0$. Une variable aléatoire X suit la loi inverse-Gamma de paramètres a et b, notée $\mathcal{IG}(a, b)$ si sa densité vérifie :

$$f(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-\frac{b}{x}} \mathbf{1}_{\mathbb{R}^+}(x)$$

$$\mathbb{E}(X) = \frac{b}{a-1} \quad \text{pour } a > 1$$
$$\mathrm{Var}(X) = \frac{b^2}{(a-1)^2(a-2)} \quad \text{pour } a > 2$$

Loi Beta

Soient $a > 0$ et $b > 0$. Une variable aléatoire X suit la loi Beta de paramètres a et b, notée $\mathrm{Beta}(a, b)$ si sa densité vérifie :

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \mathbf{1}_{[0,1]}(x)$$

$$\mathbb{E}(X) = \frac{a}{a+b}$$
$$\mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Bibliographie

I. Castillo : Introduction aux statistiques bayésiennes, 2017. Polycopié de cours - Université Pierre et Marie Curie.

B. Delyon, M. Lavielle et E. Moulines : Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1) :94–128, 1999.

A. Dempster, N. M. Laird et D. Rubin : Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38, 1977.

B. Efron : Bootstrap methods : Another look at the jackknife. *Ann. Statist.*, 7(1) :1–26, 01 1979. URL https://doi.org/10.1214/aos/1176344552.

A. E. Gelfand et A. F. M. Smith : Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410) :398–409, 1990.

A. Gelman, G. O. Roberts, W. R. Gilks *et al.* : Efficient metropolis jumping rules. *Bayesian statistics*, 5(599-608) :42, 1996.

A. Gelman et D. B. Rubin : Inference from iterative simulation using multiple sequences. *Statistical science*, p. 457–472, 1992.

S. Geman et D. Geman : Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6) :721–741, 1984.

P. Hall : *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

W. K. Hastings : Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109, 04 1970.

E. Kuhn et M. Lavielle : Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM : Probability and Statistics*, 8 :115–131, 2004.

J. S. Liu, W. H. Wong et A. Kong : Covariance structure and convergence rate of the gibbs sampler with various scans. *Journal of the Royal Statistical Society : Series B (Methodological)*, 57(1) :157–169, 1995.

G. McLachlan et T. Krishnan : *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, 2nd édn, 2007.

X.-L. Meng et D. B. Rubin : Maximum likelihood estimation via the ecm algorithm : A general framework. *Biometrika*, 80(2) :267–278, 1993.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller et E. Teller : Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6) :1087–1092, 1953.

S. Newcomb : A generalized theory of the combination of observations so as to obtain the best result. *American journal of Mathematics*, p. 343–366, 1886.

H. Robbins et S. Monro : A stochastic approximation method. *Ann. Math. Statist.*, 22 :400–407, 1951.

C. Robert : *The Bayesian choice : from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

C. Robert et G. Casella : *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

C. P. Robert, G. Casella et G. Casella : *Introducing monte carlo methods with r*, vol. 18. Springer, 2010.

G. O. Roberts, A. Gelman, W. R. Gilks *et al.* : Weak convergence and optimal scaling of random walk metropolis algorithms. *Annals of Applied probability*, 7(1) :110–120, 1997.

G. O. Roberts et J. S. Rosenthal : General state space markov chains and mcmc algorithms. *Probab. Surveys*, 1 :20–71, 2004.

G. O. Roberts, J. S. Rosenthal *et al.* : Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4) :351–367, 2001.

R. Y. Rubinstein : *Simulation and the Monte Carlo Method*. John Wiley & Sons, Inc., USA, 1st édn, 1981.

J. Shao et D. Tu : The jackknife and bootstrap. *Springer Series in Statistics, New York*, 85(486-492) :8, 1995.

G. C. Wei et M. A. Tanner : A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411) :699–704, 1990.

C. F. J. Wu : On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1) :95–103, 1983.