ANALYSE BAYÉSIENNE: UN DÉMARRAGE EN DOUCEUR

Ehssan Ghashim[4], Patrick Boily[1,2,3,4]

Résumé

L'analyse bayésienne est parfois dénigrée par les analystes de données, en partie à cause de l'élément d'arbitraire perçu associé à la sélection d'une loi a priori significative pour un problème spécifique et des (précédentes) difficultés liées à la production de la loi a posteriori dans toutes les situations sauf les plus simples. D'autre part, nous avons entendu dire que "si les analystes de données classiques ont besoin d'un grand nombre d'astuces intelligentes pour exploiter leurs données, les Bayésiens n'ont jamais vraiment besoin que d'une seule règle." Avec l'avènement des méthodes d'échantillonnage numérique efficace, les analystes de données modernes ne peuvent pas hésiter à ajouter la flèche bayésienne à leur carquois. Dans ce bref rapport, nous présentons les concepts de base qui sous-tendent l'analyse bayésienne, ainsi qu'un petit nombre d'exemples bien connus qui illustrent les points forts de l'approche.

Mots-clés

Inférence bayésienne, analyse de donnés bayésienne, distributions a priori d'entropie maximale, méthodes MCMC.

Reconnaissance de financement

Certaines sections de ce rapport ont été financées par l'entremise d'un octroi de l'Université d'Ottawa visant le développement de matériel pédagogique en français (2019-2020).

[1]Département de mathématiques et de statistique, Université d'Ottawa, Ottawa
[2]Sprott School of Business, Carleton University, Ottawa
[3]Data Action Lab, Ottawa
[4]Idlewyld Analytics and Consulting Services, Wakefield, Canada
Courriel: pboily@uottawa.ca

Table des matières

1. Introduction

La statistique bayésienne est un système qui permet de décrire l'incertitude épistémologique en utilisant le langage mathématique de la probabilité. L'inférence bayésienne, quant à elle, est le processus qui consiste à ajuster un modèle de probabilité à un ensemble de données et à résumer le résultat par une loi de probabilités sur les paramètres du modèle et sur les quantités non observées (i.e. prédictions).

1.1 Historique

En 1763, Thomas Bayes publie un article sur le problème de l'induction, c'est-à-dire le fait d'argumenter du spécifique au général. En langage et notation moderne, Bayes voulait utiliser des données binomiales (r succès sur n tentatives) afin d'apprendre la chance sous-jacente θ que chaque tentative réussisse. Sa contribution principale a été d'utiliser une loi probabiliste qui représente l'incertitude sur θ. Cette loi mesure l'incertitude "épistémologique," due au manque de connaissance, plutôt qu'une probabilité "aléatoire" découlant de l'imprévisibilité essentielle des événements futurs comme on la retrouve dans les jeux de hasard.

Dans ce cadre, une probabilité représente un "degré de croyance" (une conviction) à propos d'une proposition; il est possible que la probabilité d'un événement soit enreg-

istrée différemment par deux observateurs différents basée sur des informations de base auxquelles ils ont accès.

Les statistiques bayésiennes modernes reposent également sur la formulation des lois de probabilités pour exprimer l'incertitude sur des quantités inconnues. Il peut s'agir des paramètres sous-jacents d'un système (induction) ou des observations à venir (prédiction).

1.2 Théorème de Bayes

Le théorème de Bayes fournit une expression pour la probabilité conditionnelle de A sachant B, c'est-à-dire:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

Le théorème de Bayes peut être considéré comme un moyen d'actualiser de façon cohérente notre incertitude à la lumière de nouvelles preuves. L'utilisation d'une loi de probabilités comme "langage" pour exprimer notre incertitude n'est pas un choix arbitraire: elle peut en fait être déterminée à partir de principes plus profonds de raisonnement logique ou de comportement rationnel.

Exemple 1. Considérons une clinique médicale.

- Supposons que A représente l'événement "Le patient a une maladie hépatique." Les données obtenues par le passé suggèrent que 10% des patients qui se présentent à la clinique souffrent d'une maladie hépatique $P(A) = 0.10$.
- L'événement B représente le test décisionnel "Le patient est alcoolique." Disons que 5% des patients de la clinique sont alcooliques: $P(B) = 0.05$.
- L'événement $B \mid A$ représente alors le scénario où un patient est alcoolique, étant donné qu'il est atteint d'une maladie hépatique: disons que $P(B \mid A) = 0.07$.

Selon le théorème de Bayes, la probabilité qu'un patient soit atteint d'une maladie hépatique en supposant qu'il soit alcoolique est donc la suivante:

$$P(A \mid B) = \frac{0.07 \times 0.10}{0.05} = 0.14$$

Bien qu'il s'agisse d'une augmentation notable par rapport 10% précédemment suggéré par les expériences passées, il demeure peu probable qu'un patient en particulier souffre d'une maladie hépatique.

Le théorème de Bayes avec des événements multiples

Soient D certaines données observées et A, B, et C des événements mutuellement exclusifs (et exhaustifs) conditionnels à D. On note que

$$\begin{aligned} P(D) &= P(A \cap D) + P(B \cap D) + P(C \cap D) \\ &= P(D \mid A)P(A) + P(D \mid B)P(B) + P(D \mid C)P(C). \end{aligned}$$

Selon le théorème de Bayes, nous obtenons alors

$$\begin{aligned} P(A \mid D) &= \frac{P(D \mid A)P(A)}{P(D)} \\ &= \frac{P(D \mid A)P(A)}{P(D \mid A)P(A) + P(D \mid B)P(B) + P(D \mid C)P(C)}. \end{aligned}$$

En général, s'il y a n événements exclusifs et exhaustifs $A_1, ..., A_n$, on a, pour tout $i \in \{1, ..., n\}$:

$$P(A_i \mid D) = \frac{P(A_i)P(D \mid A_i)}{\sum_{k=1}^{n} P(A_k)P(D \mid A_k)}$$

Le dénominateur est tout simplement $P(D)$, la **distribution marginale** des données. Notez que si les événements A_i représentent des portions de la ligne réelle continue, la somme est remplacée par une intégrale.

Exemple 2. Selon l'Enquête sociale générale de 1996 au Canada, chez les hommes de 30 ans ou plus:

- 11% des répondants se situant dans le quartile de revenu le plus bas étaient des diplômés collégiaux.
- 19% des répondants se situant dans le deuxième quartile de revenu le plus bas étaient des diplômés collégiaux.
- 31% des répondants se situant dans le tRDIsième quartile de revenu le plus bas étaient des diplômés collégiaux.
- 53% des répondants se situant dans le quartile de revenu le plus élevé étaient des diplômés collégiaux.

Quelle est la probabilité qu'un diplômé collégial se situe dans le quartile de revenu le moins élevé?

Soient $Q_i, i = 1, 2, 3, 4$ des événements correspondant aux quartiles de revenu (c-à-dire $P(Q_i) = 0.25$) et D l'événement qu'un homme de plus de 30 ans est un diplômé collégial. Alors

$$\begin{aligned} P(Q_1 \mid D) &= \frac{P(D \mid Q_1)P(Q_1)}{\sum_{k=1}^{4} P(Q_k)P(D \mid Q_k)} \\ &= \frac{(0.11)(0.25)}{(0.11 + 0.19 + 0.31 + 0.53)(0.25)} = 0.09. \end{aligned}$$

1.3 Les fondements de l'inférence bayésienne

Les méthodes statistiques bayésiennes débutent avec des convictions a priori déjà existantes, et elles les mettent à jour en utilisant les données observées afin de fournir des convictions a posteriori sur lesquelles on peut établir des décisions inférentielles:

$$\underbrace{P(\theta \mid D)}_{\text{a posteriori}} = \underbrace{P(\theta)}_{\text{a priori}} \times \underbrace{P(D \mid \theta)}_{\text{vraisemblance}} / \underbrace{P(D)}_{\text{évidence}},$$

où l'évidence est donnée par

$$P(D) = \int P(D \mid \theta)P(\theta)d\theta.$$

Dans la langue vernaculaire de l'analyse bayésienne des données (ABD),

- la **distribution a priori**, $P(\theta)$, représente la force de conviction envers θ (sans tenir compte des données observées D);
- la **distribution a posteriori**, $P(\theta \mid D)$, représente la force de conviction envers θ lorsque les données observées D sont prises en compte;
- la **vraisemblance**, $P(D \mid \theta)$, représente la probabilité que les données observées D soient générées par le modèle avec valeurs de paramètres θ, et
- l'**évidence**, $P(D)$, est la probabilité d'observer les données D selon le modèle, déterminée en additionnant (ou en intégrant) toutes les valeurs possibles des paramètres et pondérée par la conviction envers ces valeurs de paramètre.

Exemple 3. *Application aux neurosciences.* Les neuroscientifiques cognitifs étudient les zones du cerveau qui sont actives pendant des tâches mentales particulières. Dans de nombreuses situations, les chercheurs observent qu'une certaine région du cerveau est active et en déduisent qu'une fonction cognitive particulière est donc en cours d'exécution; [41] rappelle que de telles conclusions ne sont pas nécessairement fermes et doivent être faîtes en tenant compte de la règle de Bayes.

Le même article présente le tableau de fréquence des études précédentes qui ont porté sur toute tâche liée à la langue (en particulier le traitement phonologique et sémantique) et si une **région d'intérêt** particulière (RDI) dans le cerveau a été activée ou non:

	Langue (L)	**Autre** ($\overline{L}$)
Activée (A)	166	199
Non Activée ($\overline{A}$)	703	2154

Supposons qu'une nouvelle étude soit menée et qu'elle constate que le RDI est activé (A). Si la probabilité a priori que la tâche implique un traitement du langage est $P(L) = 0.5$, quelle est la probabilité a posteriori, $P(L \mid A)$, étant donné que le RDI est activé?

On se sert tout simplement de la règle de Bayes:

$$\begin{aligned} P(L \mid A) &= \frac{P(A \mid L)P(L)}{P(A \mid L)P(L) + P(A \mid \overline{L})P(\overline{L})} \\ &= \frac{(166/(166+703))0.5}{(166/(166+703))0.5 + (199/(199+2154))0.5} \\ &= 0.693 \end{aligned}$$

On remarque que la probabilité a posteriori d'impliquer des processus linguistiques est légèrement plus élevée que la probabilité a priori.

Exercices

Exercice 1. (En 1975, un référendum national a eu lieu pour déterminer si le Royaume-Uni (R.U.) evrait rester membre de la Communauté économique Européenne (C.E.E)). Supposons que 52% des électeurs soutiennent le Parti travailliste et que 48% soutiennent le Parti conservateur. Supposons que 55% des électeurs du Parti travailliste et 85% des électeurs du Parti conservateur préférrent que le R.U. reste dans la C.E.E. Quelle est la probabilité qu'une personne votant "Oui" (en faveur de rester dans la C.E.E.) soit un électeur du Parti travailliste? [3]

Exercice 2. Étant donné les statistiques suivantes, quelle est la probabilité qu'une femme ait un cancer du sein si le résultat de sa mammographie est positif? [20]

- 1% des femmes de plus de 50 ans ont un cancer du sein.
- 90% des femmes atteintes d'un cancer du sein ont un résultat positif à la mammographie.
- 8% des femmes auront des résultats faussement positifs.

2. Méthodes bayésiennes et analyse des données

La caractéristique essentielle des méthodes bayésiennes est l'utilisation explicite de la probabilité pour quantifier l'incertitude dans les inférences statistiques.

2.1 Les trois étapes de l'analyse bayésienne des données

Le processus de l'analyse bayésienne des données (ABD) peut être idéalisé en le divisant en trois étapes:

1. Mise en place d'un modèle de probabilité complet (la **distribution a priori**) – une loi conjointe de probabilité pour toutes les quantités observables et non-observables dans un problème; le modèle doit être cohésif avec ce qui est connu du problème scientifique sous-jacent et avec la démarche de saisie des données.
2. Le conditionnement sur les données observées (**nouvelles observations**) – on calcule et interprète la distribution a posteriori appropriée (c'est-à-dire la loi conditionnelle de probabilité des quantités non-observées d'intérêt, étant donné les données observées).
3. L'évaluation de l'ajustement du modèle et des implications de la probabilité conditionnelle obtenue (la **distribution a posteriori**) – à quel point le modèle s'adapte-t-il aux données? ses conclusions sont-elles raisonnables? dans quelle mesure les résultats sont-ils sensibles aux hypothèses de modélisation faites à l'étape 1? Selon les réponses, on peut modifier ou redévelopper le modèle et répéter les 3 étapes.

L'essence des méthodes bayésiennes consiste à identifier les convictions sur les résultats probables a priori, et de les mettre à jour en fonction des **données observées**.

Par exemple, si le taux de réussite actuel d'une stratégie de jeu de hasard est de 5%, il est raisonnable de s'attendre à ce qu'une petite modification de la stratégie améliore ce taux de 5% points de pourcentage, mais il est très probable que la modification n'aura qu'un petit effet; de même, il est improbable que le nouveau taux de réussite atteigne 30% (après tout, ce n'est qu'une petite modification).

Lorsque les données entrent, nous commençons à mettre à jour nos croyances. Si les données entrantes indiquent une amélioration du taux de réussite, nous déplaçons vers le haut notre estimation a priori de l'effet; plus nous recueillons de données, plus nous sommes confiants par rapport à l'estimation de l'effet de la stratégie et plus nous pouvons laisser de côté la distribution a priori dans notre compréhension des effets de la stratégie.

La distribution des effets résultante fournit de l'information **a posteriori** – c'est une loi de probabilité décrivant l'effet probable de la stratégie.

3. Les lois a priori

Lorsque l'on spécifie un modèle, on doit nécessairement fournir une distribution a priori pour les paramètres inconnus. Ces distributions jouent un rôle crucial dans l'inférence bayésienne grâce à l'équation de mise à jour

$$P(\theta \mid D) \propto P(\theta) \times P(D \mid \theta).$$

Dans l'approche bayésienne, toutes les quantités inconnues sont décrites de façon probabiliste, même avant que les données n'aient été observées.

Le choix d'une distribution a priori est **subjectif** (la décision d'utiliser une distribution spécifique est laissée à l'entière discrétion des analystes). Mais le choix de la distribution a priori **n'est ni plus ni moins subjectif que le choix de la vraisemblance, la sélection d'un échantillon, le cadre d'estimation, ou la statistique utilisée pour la réduction des données**.

Le choix d'une distribution a priori peut cependant affecter considérablement les conclusions a posteriori, surtout lorsque la taille de l'échantillon est petite.

Nous examinons maintenant quelques méthodes générales permettant de détermines les lois a priori.

3.1 Lois a priori conjuguées

La distribution a posteriori du vecteur θ pourrait très bien ne pas avoir de forme analytique – c'est le principal défi des méthodes bayésiennes.

Plus précisément, il est souvent difficile, voir même mathématiquement impossible, de produire les distributions a posteriori marginales (c'est-à-dire, exprimées à l'aide d'un unique paramètre, et non des distributions conjointes) à partir d'une distribution a posteriori de grande dimension. Il y a cependant quelques exceptions permettant d'obtenir facilement des observation de la distribution a posteriori en utilisant une **distribution a priori conjuguée**.

La conjugaison est une propriété commune d'une distre distribution a priori et d'un modèle de vraisemblance; la distribution a posteriori correspondante prend la même forme que la distribution a priori, mais avec un ou des paramètre(s) différent(s).

Le tableau ci-dessous présente quelques vraisemblances communes et leur distribution a priori conjuguée (une liste détaillée se trouve dans [27]).

Vraisemblance	A priori	Hyperparamètres
Bernouilli	Beta	$\alpha > 0, \beta > 0$
Binomiale	Beta	$\alpha > 0, \beta > 0$
Poisson	Gamma	$\alpha > 0$, $\beta > 0$
Normale – μ	Normale	$\mu \in \mathbb{R}$, $\sigma^2 > 0$
Normale – σ^2	Gamma Inverse	$\alpha > 0$, $\beta > 0$
Exponentielle	Gamma	$\alpha > 0$, $\beta > 0$

Par exemple, si la probabilité de s succès après n tentatives (la vraisemblance) est donnée par

$$P(s, n \mid q) = \frac{n!}{s!(n-s)!} q^s (1-q)^{n-s}, \quad q \in [0, 1],$$

et la distribution a priori de q suit une distribution Beta(α, β), $\alpha, \beta > 0$, c'est-à-dire que

$$P(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha, \beta)},$$

pour $q \in [0, 1]$, alors la distribution a posteriori de q étant donné s succès après n tentatives suit une loi Beta($\alpha+s, \beta+n-s$):

$$P(q \mid s, n) = \frac{P(s, n \mid q) \times P(q)}{P(s, n)} = \frac{q^{\alpha+s-1}(1-q)^{\beta+n-s-1}}{B(\alpha+s, \beta+n-s)}$$

pour $q \in [0, 1]$.

Les distributions a priori conjuguées sont mathématiquement pratiques et assez flexibles, selon les hyperparamètres spécifiques qui sont utilisés, mais elles **reflètent des connaissances a priori très spécifiques**. On déconseille leur utilisation, à moins de posséder réellement ce savoir.

3.2 Lois a priori non-informatives

Une distribution a priori non-informative est une distribution sur laquelle très peu d'éléments explicatifs sont fournis au sujet des paramètres inconnus. Ces distributions sont très utiles dans la perspective du bayésianisme traditionnel qui cherche à atténuer la critique fréquentiste de la **subjectivité intentionnelle**. Ces a priori fournissent intentionnellement très peu d'informations spécifiques sur le(s) paramètre(s).

La distribution a priori non-informative classique est celle donnée par la distribution uniforme sur une région finie. Par exemple, si les données suive une distribution de Bernoulli avec paramètre θ, la distribution uniforme sur θ est

$$P(\theta) = 1, \quad 0 \leq \theta \leq 1.$$

Cette approche est sensée lorsque θ a un support fini. Mais pour des données suivant une loi normale $N(\mu, 1)$, par exemple, la distribution a priori uniforme sur $-\infty < \mu < \infty$ est impropre car l'intégrale de $P(\mu) = a$ sur $\mathbb{R}$ diverge pour tout $a \neq 0$; cependant, un tel choix pourrait toujours être acceptable tant que la distribution a posteriori résultante est standardisée (c'est-à-dire que l'intégrale de la loi a posteriori converge sur son support).

Comme il y a plusieurs cas où une distribution a priori impropre mène a une distribution a posteriori aussi impropre, un soin est justifié.

On justifie souvent la décision d'utiliser des lois a priori non-informatives par le principe qu'il faut "laisser les données parler d'elles-mêmes," de sorte que les inférences ne soient pas affectées par des préjugés qui ne sont pas soutenu par les données actuelles.

3.3 Lois a priori informatives

Les a priori informatifs sont ceux qui insèrent **délibérément** les informations que les chercheurs ont sous la main. Cela semble être une approche raisonnable puisque les connaissances scientifiques antérieures devraient jouer un rôle dans l'inférence statistique . Cependant, il y a deux exigences importantes pour les chercheurs:

1. déclaration claire de spécifications a priori, et
2. une analyse de sensibilité détaillée pour montrer l'effet de ces a priori par rapport aux types non informés.

La transparence est nécessaire pour éviter l'écueil courant de le **(data fishing en anglais ou triturage de données)**; l'analyse de sensibilité peut donner une idée de la valeur informative exacte de la distribution a priori. Mais d'où viennent les a priori informatifs, en premier lieu? En général, ces a priori sont dérivés de:

- l'intuition du chercheur, des études effectuées par le passé, des articles publiées;
- interviewer des experts du domaine;
- la commodité avec la conjugaison, et
- non paramétriques et d'autres sources dérivées de données.

L'information a priori qui provenant des études précédents n'ont pas besoins d'être en accord. Une stratégie utile consiste à construire des spécifications a priori à partir **d'écoles de pensée concurrentes** afin de contraster les a priori qui en résultent et produire des déclarations éclairées sur la force relative de chacun d'eux.

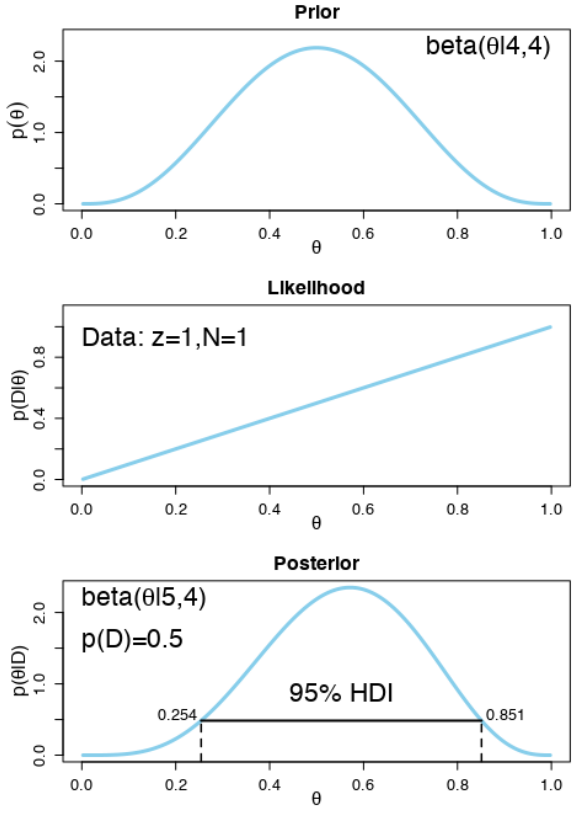
Exemple 4. Influence de la distribution a priori. Nous avons noté précédemment qu'une vraisemblance de Bernoulli et un a priori bêta forment un ensemble de la distribution a priori conjugués. Pour cet exercice, nous utilisons la fonction R `BernBeta()` définie dans [10] (notez que la fonction renvoie les valeurs bêta a posteriori à chaque appel, donc les valeurs renvoyées peuvent être réintroduites dans la distribution a priori lors du prochain appel de fonction).

(a) Commencez par une loi a priori qui exprime une certaine incertitude de l'équité d'une pièce de monnaies:$Beta(\theta \mid 4,4)$. Lancer la pièce de monnaie une fois et supposer qu'une face est obtenue. Quelle est la loi a posteriori de l'incertitude dans l'équité de la pièce θ?

Solution: à l'invite de commande R, tapez:

$> post = BernBeta(c(4,4), c(1))$

Cette fonction utilise la relation de conjugaison de la Section 3.1 pour déterminer la loi a posteriori Beta pour l'incertitude de l'équité de la pièce de monnaies étant donné les paramètres de la Beta a priori et les données observées en supposant une vraisemblance de Bernouilli. (1 représente un F(ace) sur le lancement, 0 un P(ile) Cependant, nous savons sur des bases théoriques que la distribution a posteriori suit une loi $\text{Beta}(\theta \mid 4+1, 4+1-1) = \text{Beta}(\theta \mid 5,4)$.



La marque sur l'axe y des ordonnées de la loi a pos-

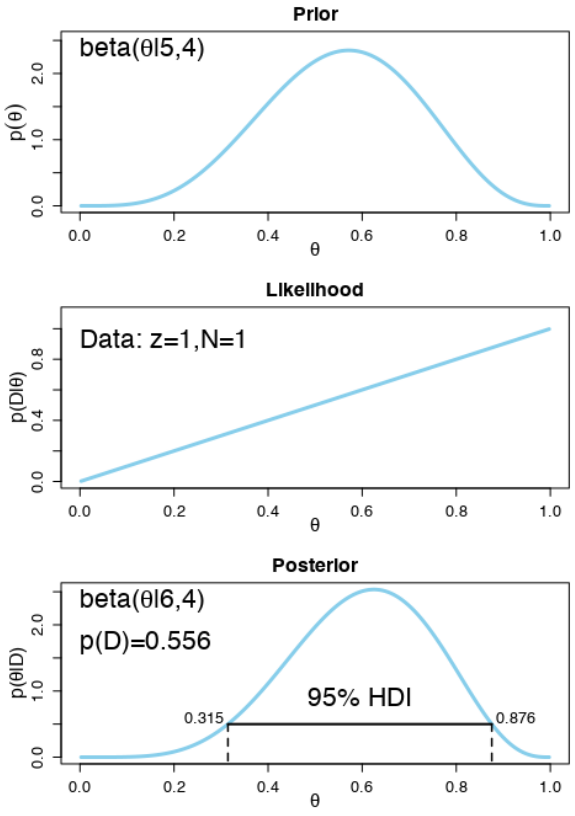
teriori fournit les paramètres a posteriori (ils sont également donnés en tapant `show(post)` à l'invite de commande).

(b) Utilisons les paramètres a posteriori du lancement précédent comme une a priori pour le lancement suivant. Supposons que nous retournions à nouveau et obtenions une F . Quel est la nouvelle a posteriori sur l'incertitude de l'équité de la pièce de monnaies?

Solution: à l'invite de commande R, tapez

$> post = BernBeta(post, c(1))$

La loi a posteriori est Beta($\theta \mid 6, 4$), qui est indiquée ci- dessous.



(c) En utilisant le plus récent a posteriori comme a priori pour le lancement suivant, retournez une troisième fois et obtenez à nouveau une F. Quel est le nouveau a posteriori?

Solution: dans ce cas, nous savons que la distribution a posteriori pour l'équité de la pièce suit une loi Beta($\theta \mid 7, 4$) (nous ne fournirons ni le code ni la sortie, cette fois- ci!). Est-ce que 3 F d'affilée vous font réfléchir? Est-ce qu'il a suffisamment de preuves pour suggérer que $\theta \neq 0.5$ (c'est-à-dire que la pièce n'est pas juste)? Et si vous retourniez 18 F d'affilée à partir de ce moment??

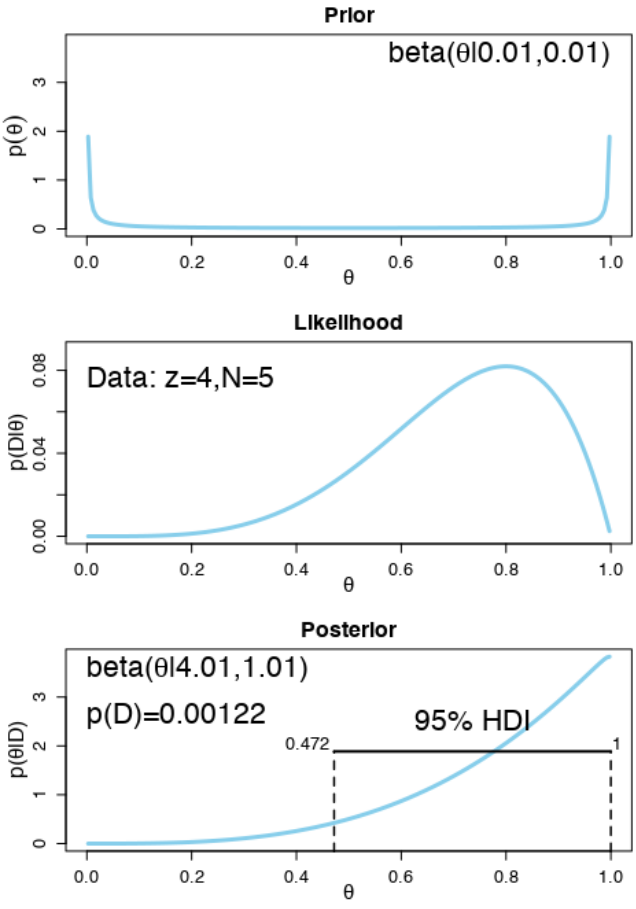
Lorsqu'on travaille sur un problème, il peut être facile de s'écarter et de se confondre avec la notation. Dans ces cas, il est utile de revenir à la définition de chacun des termes du théorème de Bayes. (c.à.d $P(\theta \mid D)$, $P(D)$, $P(D \mid \theta)$, etc.).

Exemple 5. *Un a priori inhabituel.* Supposons qu'une amie possède une pièce de monnaie dont nous savons qu'elle provient d'un magasin de magie; en conséquence, nous croyons que la pièce est fortement biaisée dans l'une ou l'autre des deux directions (il pourrait s'agir d'une pièce de monnaie truquée dont les deux faces seraient F, par exemple), mais nous ne savons pas laquelle elle favorise. Nous exprimerons la croyance de cet a priori comme une loi Beta. Disons que notre amie retourne la pièce de monnaie cinq fois; ce qui donne 4 F et 1 P. Quelle est la loi a posteriori de l'équité de la pièce de monnaie θ?

Solution: à l'invite de commande R, tapez

```
> post = BernBeta(c(1,1)/100,c(1,1,1,1,0))
```

donnant la distribution a posteriori ci-dessous.



Le code ci-dessus utilise un a priori donné par Beta($\theta \mid 0.01, 0.01$). Cet a priori traduit notre conviction que la pièce est fortement biaisée (bien que nous ne sachions

pas dans quelle direction se situe le biais avant de voir les données). Le choix de 0.01 est arbitraire, dans un sens; 0.1 aurait également fonctionné, par exemple.

La loi a posteriori est donc Beta($\theta \mid 4.01, 1.01$) qui, comme indiqué ci-dessus, a son mode essentiellement à 1.0, et non près de la moyenne ≈ 0.8. La pièce est-elle vraiment biaisée? Dans quelle direction? Comment votre réponse changerait-elle si vous n'aviez aucune raison de soupçonner que la pièce était biaisée en premier lieu?

3.4 Lois a priori d'entropie maximale

Que les a priori soient non informatifs ou informatives, nous recherchons la loi qui encode le mieux l'état a priori des connaissances à partir d'un ensemble des lois d'essai.

Considérons un espace discret X de cardinalité M avec une densité de probabilité $P(X) = (p_1, ..., p_M)$. **L'entropie** d'un p désigné par $H(p)$, est donné par

$$H(p) = -\sum_{i=1}^{M} p_i \log p_i, \text{[1]} \quad \text{avec } 0 \cdot \log(0) = 0.$$

Le principe de l'entropie maximale (MaxEnt) précise que, étant donné une classe de loi d'essai avec contraintes, la distribution a priori optimale est la loi d'essai ayant la plus grande entropie. Par exemple, la contrainte la plus fondamentale est que p se trouve dans le simplexe de probabilité, c'est-à-dire $\sum_i p_i = 1$ et $p_i \geq 0$ pour tous i dans le cas discret, ou $\int_\Omega p(Z)dZ = 1$ et $P(Z) \geq 0$ sur Ω dans le cas continu.

Exemple 6. Sans contraintes, le principe de MaxEnt donne une a priori qui résout le problème d'optimisation:

$$\begin{aligned} \max \quad & -p_1 \log p_1 - \cdots - p_M \log p_M \\ \text{s.t.} \quad & p_1 + \cdots + p_M = 1 \text{ et } p_1, \ldots, p_M \geq 0 \end{aligned}$$

En utilisant la méthode de multiplicateur de Lagrange, cette optimisation se réduit à

$$p^* = \text{argmax}_p \{H(p) - \lambda(p_1 + \cdots + p_M - 1)\},$$

dont la solution est $p^* \propto$ constant. Donc, sans contrainte supplémentaire, la loi uniforme est l'entropie maximale a priori.

Exemple 7. *Utilisation de MaxEnt pour créer un a priori pour l'inférence bayésienne*. "La blague sur New York est que vous ne pouvez jamais prendre un taxi, mais quand vous n'en avez pas besoin, alors vous trouvez des taxis partout" (citation et exemple tirés du tutoriel de S.DeDeo sur les méthodes d'entropie maximale [29]). Comment pourrions-nous utiliser l'analyse bayésienne pour prédire le temps d'attente des taxis? À divers moments, dirigez-vous vers la rue, dites «J'ai besoin d'un taxi!» Et notez le temps que vous avez mis avant qu'un taxi soit disponible. Peut-être que les observations (en quelques minutes) ressemblent à ceci

$$6, 3, 4, 6, 2, 3, 2, 6, 4, 4.$$

Que pouvez-vous conclure sur le temps d'attente pour un taxi à New York? Dans le meilleur des cas, un taxi nous attend à l'approche du trottoir ($j = 0$), tandis que dans le pire des cas (une apocalypse zombie, par exemple?), Aucun taxi ne vient jamais ($j \to \infty$). Mais peut-on dire autre chose?

Pour utiliser MaxEnt dans cette situation, nous devons trouver- parmi toutes les loi d'essai qui auraient pu générer les temps d'attente - observés-celui qui a l'entropie la plus élevée. Malheureusement, il existe une infinité de ces lois. Nous pouvons limiter la recherche en incluant une contrainte stipulant que la valeur attendue des lois de l'essai devrait être la même que la moyenne de l'échantillon, à savoir 4.

Les deux contraintes se traduisent par

$$g_1(p) = \sum_{j=0}^{\infty} j \cdot p_j - 4 = 0 \quad \text{and} \quad g_2(p) = \sum_{j=0}^{\infty} p_j - 1 = 0,$$

où $p_j$est la probabilité de devoir attendre j minutes pour un taxi.

La méthode de multiplicateur de Lagrange réduit le problème à la résolution

$$\text{argmax}_p \{\{H(p) - \lambda_1 g_1(p) - \lambda_2 g_2(p)\}.$$

Pour cela, il faut résoudre l'équation du gradient

$$\nabla_p H(p) = \lambda_1 \nabla_p g_1(p) + \lambda_2 \nabla_p g_2(p),$$

ce qui donne lieu à des équations de la forme

$$-(\ln p_j + 1) = \lambda_1 j + \lambda_2, \quad j = 0, 1, \ldots,$$

ou simplement $p_j = \exp(-\lambda_1 j)\exp(-1 - \lambda_2)$ pour $j = 0, 1, \ldots$ Sachant que

$$1 = \sum_{j=0}^{\infty} p_j = \exp(-1 - \lambda_2) \sum_{j=0}^{\infty} \exp(-\lambda_1 j),$$

alors on a

$$\exp(1 + \lambda_2) = \sum_{j=0}^{\infty} \exp(-\lambda_1 j) = \frac{1}{1 - \exp(-\lambda_1)}, \quad (1)$$

en supposant que $|\exp(-\lambda_1)| < 1$. De même,

$$4 = \sum_{j=0}^{\infty} j p_j = \exp(-1 - \lambda_2) \sum_{j=0}^{\infty} j \exp(-\lambda_1 j),$$

[1] Dans le cas d'une densité continue $P(X_1, \ldots, X_n)$ sur un domaine $\Omega \subseteq \mathbb{R}^n$,l'entropie est donnée par $H(p) = -\int_\Omega p(Z)\log(p(Z))dZ$.

de sorte que

$$4\exp(1+\lambda_2) = \sum_{j=0}^{\infty} j\exp(-\lambda_1 j) = \frac{\exp(-\lambda_1)}{(1-\exp(-\lambda_1))^2}. \quad (2)$$

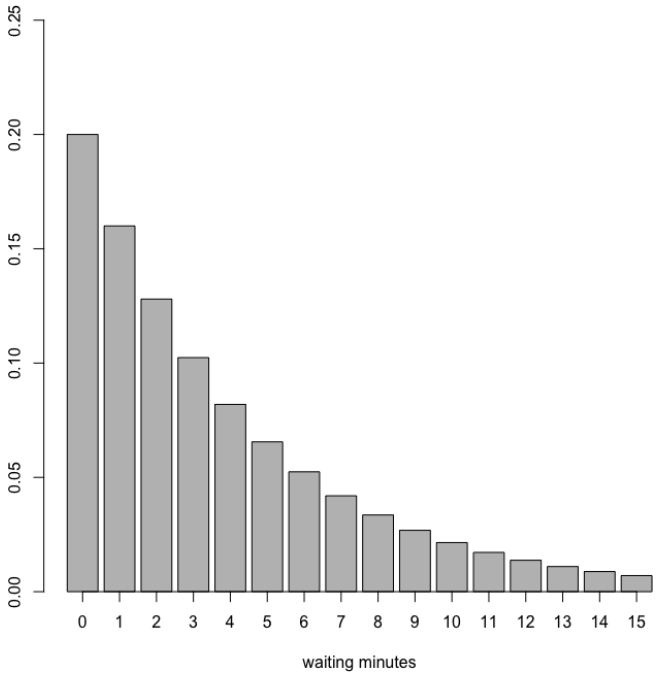
En substituant (1) à (2) et en résolvant pour λ_1, on vois que $\lambda_1 = \ln(5/4)$. En substituant ce résultat à (1), on obtient $\exp(-1-\lambda_2) = \frac{1}{5}$, alors, on a

$$p_j = \exp(-1-\lambda_2)\exp(-\lambda_1 j) = \frac{1}{5}\left(\frac{4}{5}\right)^j, j = 0,\ldots$$

Il est facile de voir que cela définit une loi; une "vérification" est fournie par le code suivant.

```
pmf_maxent <- function(x,lambda=4/5)
    (1-lambda)*(\lambda)^x
sum(pmf_maxent(0:100)) #tester s'il est
    une distribution
mp <- barplot(pmf_maxent(0:15),
    ylim=c(0,.25), xlab="Temps d'attente
    (en minutes)")
axis(1,at=mp,labels=paste(0:15))
```

Cette loi (voir ci-dessous) pourrait être utilisée comme a priori dans une analyse bayésienne de la situation. Notez que certaines informations sur les données (dans ce cas, uniquement la moyenne de l'échantillon) sont utilisées pour définir la distribution a priori MaxEnt.



Exercices

Exercice 3. Dans cet exercice, vous étudierez l'effet possible du choix de la distribution a priori sur les conclusions.

(a) Supposons que vous possédez une pièce de monnaie que vous savez qu'elle a été frappée par le gouvernement fédéral et pour laquelle vous n'avez aucune raison de soupçonner de l'altération de toute sorte. Votre croyance a priori sur l'équité de la pièce est donc fort. Vous lancez la pièce 10 fois et enregistrez 9 F (ace). Quelle est la probabilité que vous prévoyez d'obtenir 1 F au 11ème lancement? Expliquez soigneusement votre réponse; justifiez votre choix de préalable. Comment votre réponse changerait-elle (le cas échéant) si vous utilisez un point de vue fréquentiste?.

(b) Un mystérieux inconnu vous tend une pièce différente, celle-ci faite d'un matériau étrange au toucher, sur laquelle on trouve les mots "l'Association de Global Tricksters" Vous lancez la pièce 10 fois et enregistrez à nouveau $9F$. Expliquez soigneusement votre réponse; justifier votre choix de la distribution a priori. Remarque: utilisez la distribution a priori de l'exemple 5.

4. Les lois a posteriori

La loi a posteriori est utilisée pour estimer une variété des paramètres de modèle d'intérêt , tels que la moyenne, la médiane, le mode, et ainsi de suite.

Il est possible de construire des **intervalles / régions crédibles** directement à partir de la distribution a posteriori (contrairement à "la confiance" intervalles d'inférence fréquentiste).

étant donné une loi a posteriori sur un paramètre θ, un $1-\alpha$ intervalle crédible $[L,U]$ est l'intervalle tel que

$$P(L \le \theta \le U \mid D) \ge 1-\alpha.$$

Parce que la distribution a posteriori est une loi complète sur les paramètres, il est possible de faire toutes sortes de déclarations probabilistes sur leurs valeurs, telles que:

- "Je suis 95% sûr que la vraie valeur du paramètre est plus grande que 0.5."
- Il y a 50% de chances que θ_1 soit plus grand que θ_2.
- etc.

La meilleure approche consiste à construire l'intervalle crédible des valeurs θ en utilisant **l'intervalle HPD** (HPD pour Highest Posterior Density, soit densité a posteriori la plus forte), c'est-à-dire à définir une région C_k dans l'espace des paramètres avec

$$C_k = \{\theta : P(\theta \mid D) \ge k\},$$

où k est le plus grand nombre tel que

$$\int_{C_k} P(\theta \mid D)\, d\theta = 1-\alpha.$$

L'effet est ordinairement de trouver la plus petite région C_k (en mesure) à satisfaire au critère.

La valeur k peut être considérée comme la hauteur d'une ligne horizontale (ou hyperplan, dans le cas des a posteriori multivariées) superposée sur la distribution a posteriori et dont l'intersection(s) avec cette dernière définit une région sur laquelle l'intégrale de la distribution a posteriori est $1-\alpha$. Dans la plupart des cas, la valeur k peut être trouvée numériquement.

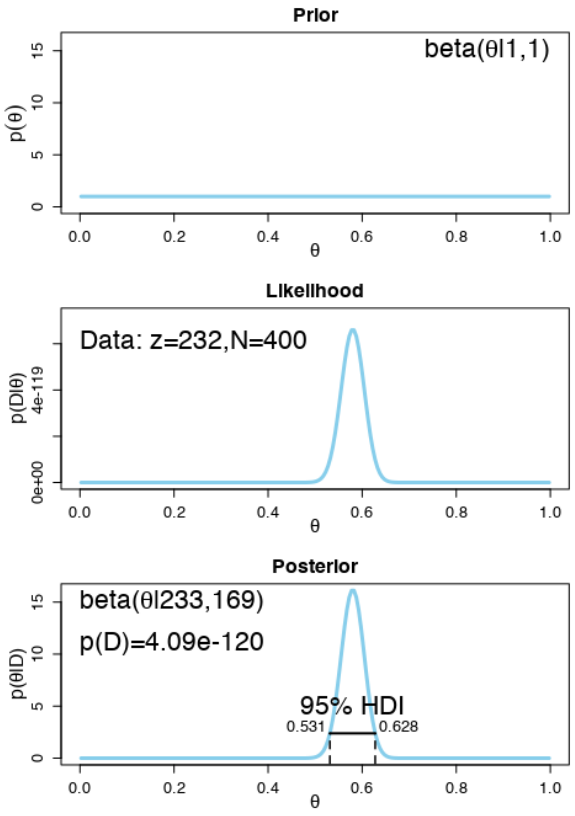
Exemple 8. *HPDs, élections et collecte itérative de données.* C'est une année d'élection et vous voulez savoir si la population en général préfère le candidat A ou le candidat B. Un sondage récemment publié indique que parmi 400 personnes échantillonnées au hasard, 232 ont préféré le candidat A, tandis que les autres ont préféré le candidat B.

(a) Supposons qu'avant la publication du sondage, vous pensiez que la préférence générale suit une loi uniforme. Quel est l'HPD de 95% sur votre croyance après avoir appris le résultat du sondage?

Solution: notons la préférence pour le candidat A par 1, et la préférence pour le candidat B par 0. On peut utiliser la fonction R `BernBeta()` comme dans l'Exemple 4. à l'invite de commande R, tapez

```
> post=BernBeta(c(1,1),c(rep(1,232),rep(0,168)))
```

donnant une a postériori avec un 95% HPD de 0.531 à 0.628 pour la probabilité du candidat A.



(b) D'après le sondage, est-il crédible de croire que la population est également divisée dans ses préférences entre des candidats?

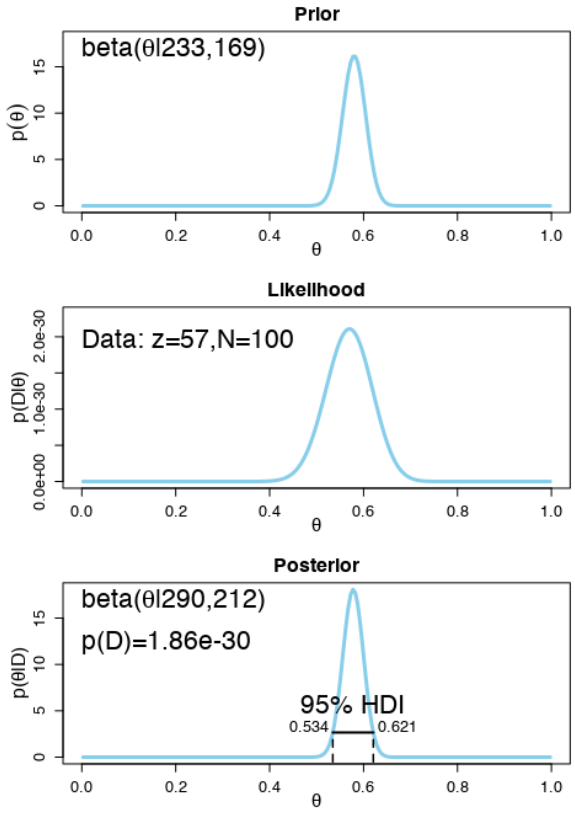
Solution: l'intervalle HPD de la Partie (a) montre que $\theta = 0.5$ ne fait pas partie des valeurs crédibles, il n'est donc pas crédible de croire que la population est également divisée dans ses préférences (au niveau de 95%).

(c) Vous souhaitez effectuer un sondage complémentaire pour affiner votre estimation de la préférence de la population. Dans le sondage complémentaire, vous échantillonnez au hasard 100 personnes et vous constatez que 57 d'entre elles préfèrent le candidat A. En supposant que l'opinion des gens n'a pas changé entre les sondages, quel est l'intervalle HPD de 95% sur la distribution a posteriori?

Solution: à l'invite de commande R, tapez

```
> post=BernBeta(post,c(rep(1,57),rep(0,43)))
```

donne la figure ci-dessous. Le 95% intervalle HPD est a un peu plus étroit pour la préférence du candidat A, de 0.534 à 0.621.



(d) Sur la base du sondage complémentaire, est-il crédible de croire que la population est également divisée dans ses préférences entre les candidats?

Solution: L'intervalle HPD de la partie (c) exclut $\theta = 0.5$; le sondage complémentaire et le sondage initial suggèrent que la population n'est pas divisée également (et préfère en fait le candidat A).

4.1 Les méthodes de Monte-Carlo par Chaîne de Markov (MCMC)

La véritable puissance de l'inférence bayésienne se fait surtout sentir lorsque les spécifications du modèle conduisent à un a posteriori qui ne peut être manipulé analytiquement; dans ce cas, il est habituellement possible de recréer un ensemble synthétique (ou simulé) de valeurs qui partagent les propriétés d'un a posteriori donné. Ces processus sont connus sous le nom des **simulations de Monte Carlo**. Une **Chaîne de Markov** est un ensemble ordonné et indexé de variables aléatoires (un processus stochastique) dans lequel les valeurs des quantités à un état donné ne dépendent, de façon probabiliste, que des valeurs des quantités à l'état précédent. Les méthodes de **Monte-Carlo par Chaîne de Markov** (MCMC) sont une classe d'algorithmes pour l'échantillonnage à partir d'une loi de probabilité basée sur la construction d'une chaîne de Markov avec la loi désirée comme loi d'équilibre. L'état de la chaîne après un certain nombre d'étapes est alors utilisé comme échantillon de la loi souhaitée. La qualité de l'échantillon s'améliore en fonction du nombre d'étapes.

Les techniques MCMC sont souvent appliquées pour résoudre des problèmes d'intégration et d'optimisation dans des espaces de grandes dimensions. Ces deux types de problèmes jouent un rôle fondamental dans l'apprentissage machine (en anglais machine learning)la physique, les statistiques, l'économétrie et l'analyse des décisions. Par exemple, compte tenu des variables $\theta \in \Theta$ et des données data D, les problèmes d'intégration suivants (généralement insolubles) sont principales à l'inférence bayésienne:

- **La normalisation** - pour obtenir la distribution a posteriori $P(\theta \mid D)$ étant donnée la distribution a priori $P(\theta)$ et la vraisemblance $P(D \mid \theta)$, le facteur de normalisation (dénominateur) du théorème de Bayes doit être calculé

$$P(\theta \mid D) = \frac{p(\theta)P(D \mid \theta)}{\int P(D \mid \theta)P(\theta)d\theta}.$$

- **La marginalisation** - étant donnée la distribution a posteriori conjointe de (θ, x), on est souvent intéressés par la distribution a posteriori marginale

$$P(\theta \mid D) = \int P(\theta, x \mid D)dx.$$

- **l'espérance**- l'objectif final de l'analyse est souvent pour obtenir des statistiques sommaires de la forme

$$E(f(\theta)) = \int_{\Theta} f(\theta)P(\theta \mid D)d\theta$$

pour une fonction d'intérêt (c-à-d $f(\theta) = \theta$ (la moyenne), ou $f(\theta) = (\theta - E(\theta))^2$ (la variance)).

L'algorithme de Metropolis-Hastings (MH)

L'algorithme de Metropolis-Hastings (MH) est un type spécifique de processus de Monte Carlo; il fait probablement partie des dix algorithmes qui ont récemment eu la plus grande influence sur le developpement et la pratique de la science et de l'ingénierie.

MH génère une marche aléatoire (c'est-à-dire qu'elle génère une succession d'échantillons a posteriori) de façon à ce que chaque étape de la marche soit **complètement indépendante** des étapes précédentes; la décision de rejeter ou d'accepter l'étape proposée est également indépendante de l'histoire de la marche.

Tout processus pour lequel l'étape actuelle est indépendante (oubliée) des états précédents, à savoir

$$P(X_{n+1} = x \mid X_1 = x_1, \ldots, X_n = x_n) = P(X_{n+1} = x \mid X_n = x_n)$$

pour tout n, X_j et x_j, $j = 1, \ldots, n$, est appelé un **processus de Markov (du premier ordre)**, et une succession de ces étapes est une chaîne de Markov(du premier ordre).

MH utilise une loi candidate ou de proposition pour la distribution a posteriori, disons q$(\cdot, \theta)$, où θ est un vecteur de paramètres qui est fixé par les paramètres de réglage de l'appel utilisateur; MH construit ensuite une chaîne de Markov en proposant une valeur pour θ à partir de cette loi candidate, puis en acceptant ou en rejetant cette valeur (avec une certaine probabilité).

Théoriquement, la loi de proposition peut être presque n'importe quelle loi, mais en pratique il est recommandé de choisir des lois (vraiment) simples: une normale si le paramètre d'intérêt peut être n'importe quel nombre réel (par exemple μ), ou une log-normale si elle a un support positif (disons σ^2).

L'algorithme de **Metropolis-Hastings**(MH) simule des échantillons à partir d'une loi de probabilité en utilisant la fonction de densité conjointe complète et des lois de proposition (indépendantes) pour chacune des variables d'intérêt. L'algorithme MH passe au travers des étapes suivantes à chaque itération en recommençant ces étapes pour i allant de 0 à N.
La première étape consiste à initialiser la valeur de l'échantillon pour chaque variable aléatoire (souvent obtenue par échantillonnage à partir de la loi a priori de la variable). La boucle principale de l'Algorithme 1 comprend trois composantes:

Algorithm 1: L'algorithme de Metropolis-Hastings

1 Générer: $x^{(0)} \sim q(x)$

2 *Propose:* $x^* \sim q(x^{(i)} \mid x^{(i-1)})$

3 *Calculer la probabilité d'acceptation:*

$$\alpha(x^* \mid x^{(i-1)}) = min\left\{1, \frac{q(x^{(i-1)} \mid x^*)\pi(x^*)}{q(x^* \mid x^{(i-1)})\pi(x^{(i-1)})}\right\}$$

4 *Prendre:*

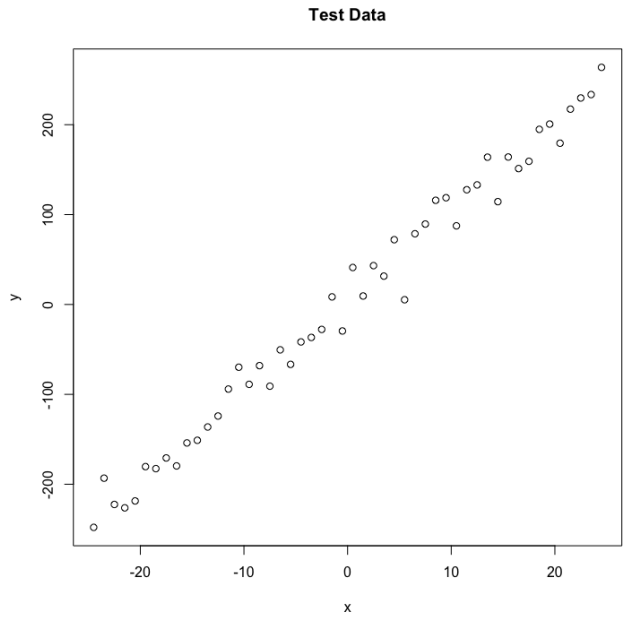
$$x^{(i)} = \begin{cases} x^* & \text{avec probabilité } \alpha \\ x^{(i-1)} & \text{avec probabilité } 1-\alpha \end{cases}$$

- **générer un échantillon de candidats** x^* à partir de la loi de proposition $q(x^{(i)} \mid x^{(i-1)})$;
- **calculer la probabilité d'acceptation** via la fonction d'acceptation $\alpha(x^* \mid x^{(i-1)})$ sur la base de la loi de proposition et de la densité conjointe complète $\pi(\cdot)$;
- **accepter l'échantillon candidat** avec la probabilité α,la probabilité d'acceptation, sinon le **rejeter**.

Exemple 9. L'algorithme MH et la régression linéaire simple. Les **données d'essai** pour cet exemple sont générées à l'aide du code suivant.

```
t.A <- 10 # la vrai valeur de la pente
t.B <- 0 # la vrai valeur de
    l'ordonn\'ee \'a l'origine
t.sd <- 20 # La vrai valeur de la
    variance de l'erreur al\'eatoire
s.Size <- 50 # la taille de
    l'\'echantillon
# cr\'eer des valeur ind\'ependantes de x
x <- (-(s.Size-1)/2):((s.Size-1)/2)
# cr\'eer des valeur d\'ependantes selon
    l'\'equation ax + b + N(0,sd)
y <- t.A * x + t.B +
    rnorm(n=s.Size,mean=0,sd=t.sd)
plot(x,y, main="Donn\'ees d'essai")
```

Remarquez que les valeurs x sont équilibrées autour de zéro pour "décorréler" la pente et l'ordonnée à l'origine. Le résultat devrait ressembler à la graphique ci-dessous.

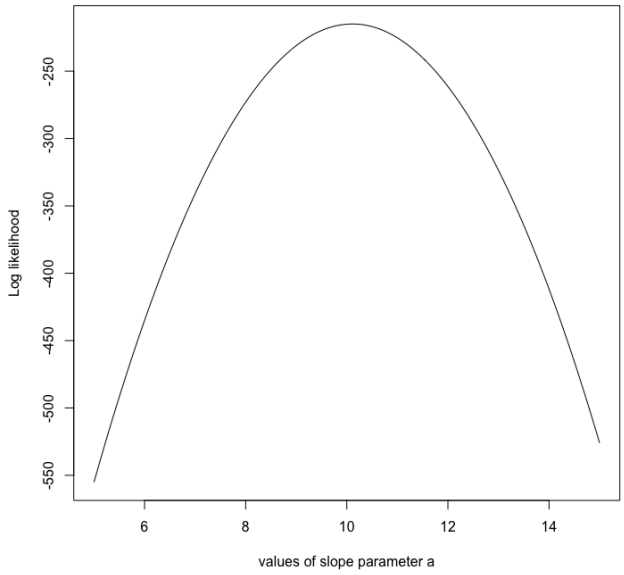


Définition du modèle statistique. L'étape suivante consiste à spécifier le modèle statistique. Nous savons déjà que les données étaient créées avec une relation linéaire $y = ax + b$ ainsi qu'un modèle d'erreur normal $N(0, sd)$ avec l'écart-type sd, donc nous pourrions aussi bien utiliser le même modèle pour l'ajustement et voir si nous pouvons récupérer nos valeurs de paramètres originales. Notez cependant qu'en général, le modèle de génération est inconnu.

Dérivation de la fonction de vraisemblance à partir du modèle. Un modèle linéaire de la forme $y = ax + b + N(0, sd)$ prend la paramètres (a, b, sd) comme entrées. Le résultat devrait être la probabilité d'obtenir les données de test dans le cadre de ce modèle: dans ce cas, il suffit de calculer la différence entre les prédictions $y = ax + b$ t le y observé, puis de rechercher la probabilité (à l'aide de la fonction *dnorm*) que de tels écarts se produisent.

```
likehd <- function(param){
    a = param[1]
    b = param[2]
    sd = param[3]
    pred = a*x + b
    singlelikelihoods = dnorm(y, mean =
        pred, sd = sd, log = T)
    sumll = sum(singlelikelihoods)
    return(sumll) }
# Exemple: Faire la graphe du profil de
    la vraisemblance de la pente a
s.values <-
    function(x){return(likehd(c(x, t.B,
    t.sd)))}
s.likehds <- lapply(seq(1/2*t.A,
    3/2*t.A, by=.05), s.values)
plot(seq(1/2*t.A, 3/2*t.A, by=.05),
    s.likehds , type="l", xlab = "Les
    valeurs du parametre de pente a",
    ylab = "Log de la vraisemblance")
```

Pour illustrer, les dernières lignes du code tracent la Vraisemblance pour une gamme de valeurs paramétriques du paramètre de pente a. Le résultat devrait ressembler au graphe ci-dessous.



Définir les a priori. Dans l'analyse bayésienne, l'étape suivante est toujours nécessaire: nous devons spécifier une loi a priori pour chacun des paramètres du modèle. Pour simplifier les choses, nous utiliserons une loi uniforme pour les trois paramètres et nous utiliserons des lois normales pour chacun d'entre eux.[2]

```
# Prior distribution
prior <- function(param){
    a = param[1]
    b = param[2]
    sd = param[3]
    aprior = dunif(a, min=0, max=2*t.A,
        log = T)
    bprior = dnorm(b, mean=t.B, sd = 5,
        log = T)
    sdprior = dunif(sd, min=0,
        max=2*t.sd, log = T)
    return(aprior+bprior+sdprior)
}
```

L'a posteriori. Le produit de la distribution a priori par la vraisemblance est la quantité réelle avec laquelle MCMC travaille (il n'est pas, strictement parlant, la distribution a posteriori car il n'est pas normalisé).

```
posterior <- function(param){
   return (likehd(param) + prior(param))
}
```

Appliquer l'algorithme MH. L'une des applications les plus fréquentes de le MH (comme dans cet exemple) est l'échantillonnage par la densité a posteriori en statistiques bayésienne.[3] L'objectif de l'algorithme est de sauter dans l'espace des paramètres, mais de manière à avoir la probabilité d'atterrir en un point soit proportionnel à la fonction dont nous échantillonnons (c'est ce que l'on appelle généralement la fonction cible). Dans ce cas, la fonction cible est la fonction a posteriori définie précédemment.

Pour ce faire, il faut

1. commencer par un vecteur de paramètre aléatoire;
2. choisir un nouveau vecteur de paramètre proche de l'ancienne valeur sur la base d'une certaine densité de probabilité (la fonction de proposition), et
3. sauter à ce nouveau point avec une probabilité $\alpha = \min\{1, g(\text{new})/g(\text{old})\}$, où g est la fonction cible.

La loi des vecteurs de paramètres (MH visites) converge vers la loi cible g.

```
######## MH ################
proposalfunction <- function(param){
   return(rnorm(3,mean = param, sd=
       c(0.1,0.5,0.3)))
}

run_metropolis_MCMC <-
    function(startvalue, iterations){
   chain = array(dim = c(iterations+1,3))
   chain[1,] = startvalue
   for (i in 1:iterations){
      proposal =
          proposalfunction(chain[i,])

      probab = exp(posterior(proposal) -
          posterior(chain[i,]))
      if (runif(1) < probab){
         chain[i+1,] = proposal
      }else{
         chain[i+1,] = chain[i,]
      }
   }
   return(chain)
}

startvalue = c(4,0,10)
chain = run_metropolis_MCMC(startvalue,
    10000)
burnIn = 5000
acceptance =
    1-mean(duplicated(chain[-(1:burnIn),]))
```

[2] Nous travaillons avec les logarithmes de toutes quantités, de sorte que la vraisemblance est une somme et non un produit comme ce serait habituellement le cas.

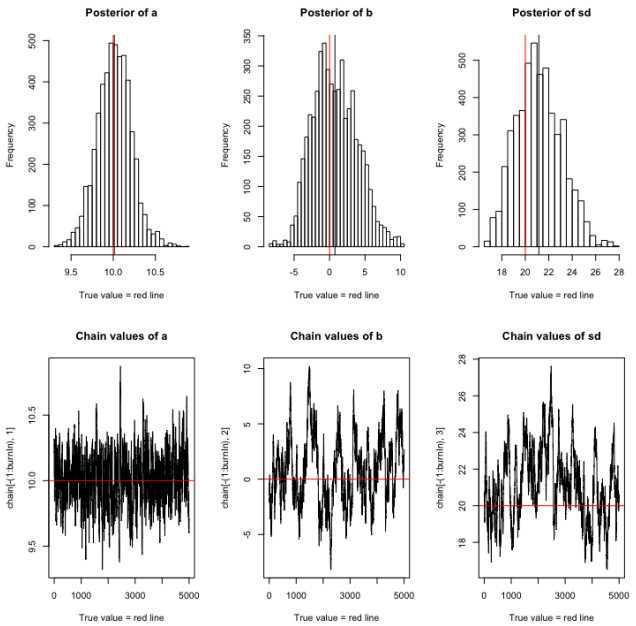
[3] L'algorithme peut être utilisé pour échantillonner à partir de n'importe quelle fonction intégrable.

Les premières étapes de l'algorithme peuvent être biaisées par le processus d'initialisation; ils sont généralement rejetés pour l'analyse (c'est ce qu'on appelle **le temps de rodage**).

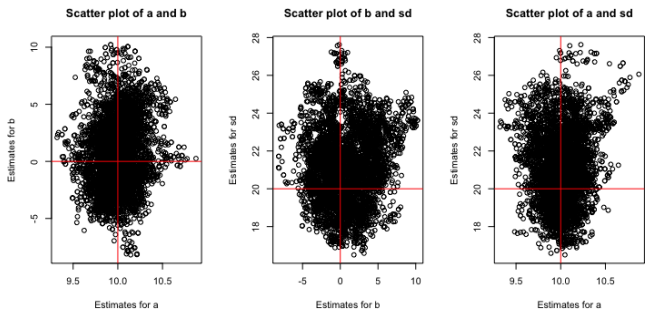
Un résultat intéressant à étudier est le taux d'acceptation: combien de fois une proposition a-t- elle été rejetée par le critère d'acceptation du MH? Le taux d'acceptation peut être influencé par la fonction de proposition: en général, plus la fonction proposée est proche de la dernière plus le taux d'acceptation est élevé. Les taux d'acceptation très élevés, cependant, ne sont habituellement pas bénéfiques, car cela implique que l'algorithme «reste» dans le même voisinage ou point, ce qui se traduit par un sondage sous-optimal de l'espace des paramètres (il y a un **mélange** très faible). Des taux d'acceptation compris entre 20% et 30% sont considérés comme optimaux pour des applications typiques [25].

Enfin, nous pouvons tracer les résultats.

```
### Summary: ######################
 par(mfrow = c(2,3))
hist(chain[-(1:burnIn),1],nclass=30, ,
    main="la distribution a posteriori
    de a", xlab="La vrais valeur = La
    ligne rouge")
abline(v = mean(chain[-(1:burnIn),1]))
abline(v = t.A, col="red")
hist(chain[-(1:burnIn),2],nclass=30,
    main="la distribution a posteriori
    de b", xlab="La vrais valeur = La
    ligne rouge")
abline(v = mean(chain[-(1:burnIn),2]))
abline(v = t.B, col="red")
hist(chain[-(1:burnIn),3],nclass=30,
    main="la distribution a posteriori
    de sd", xlab="La vrais valeur = La
    ligne rouge")
abline(v = mean(chain[-(1:burnIn),3]))
abline(v = t.sd, col="red")
plot(chain[-(1:burnIn),1], type = "l",
    xlab="La vrais valeur = La ligne
    rouge" , main = "Les valeurs de
    cha\^{\i}ne de a",)
abline(h = t.A, col="red")
plot(chain[-(1:burnIn),2], type = "l",
    xlab="La vrais valeur = La ligne
    rouge" , main = "Les valeurs de
    cha\^{\i}ne de b",)
abline(h = t.B, col="red")
plot(chain[-(1:burnIn),3], type = "l",
    xlab="La vrais valeur = La ligne
    rouge" , main = "Les valeurs de
    cha\^{\i}ne de sd",)
abline(h = t.sd, col="red")
# for comparison:
summary(lm(y~x))
```



Les graphes résultants devraient ressembler à celles qui ont été vues dans la colonne de droite: la ligne supérieure indique les estimations a posteriori pour la pente a, l'ordonnée à l'origine b et l'écart type de l'erreur sd;la ligne inférieure montre la chaîne de Markov des valeurs des paramètres. Nous récupérons (plus ou moins) les paramètres d'origine qui ont été utilisés pour créer les données, et il y a une certaine zone autour des valeurs a posteriori les plus élevées qui montre également un certain appui par les données, qui est l'équivalent bayésien des intervalles de confiance.
Les lois a posteriori ci-dessus sont des lois **marginales**, les lois conjoints sont indiquées ci- dessous.



à titre de comparaison, la fonction `lm()` dans R donne les estimations suivantes: a : 9.9880 (se: 0.2092), b : 0.5840 (se: 3.0185), and $sd = 21.34$ (48 d.f.).

Exercices

Exercice 4. Un groupe d'adultes fait une expérience simple d'apprentissage:lorsqu'ils voient les deux mots *radio* et *océan* apparaissent simultanément sur un écran d'ordinateur, on leur demande d'appuyer sur la touche F du clavier; chaque fois que les mots *radio* et *montagne* apparaissent à l'écran, on leur demande d'appuyer sur la touche J. Après plusieurs répétitions d'exercices, deux nouvelles tâches sont

introduites: dans la première, le mot *radio* apparaît tout seul et les participants sont invités à donner la meilleure réponse (F ou J) en fonction de ce qu'ils ont appris auparavant; dans la seconde, les mots *océan* et *montagne* apparaissent simultanément et les participants sont à nouveau invités à donner la meilleure réponse. Ceci est répété avec 50 personnes. Les données montrent que, pour le premier test, 40 participants ont répondu par F et 10 par J; tandis que pour le deuxième test, 15 ont répondu par F et 35 par J. Les gens sont-ils biaisés vers F ou vers J pour l'un des deux tests? Pour répondre à cette question, supposez une a priori uniforme et utilisez un intervalle HPD de 95% pour décider quels biais peuvent être déclarés crédibles.

5. L'incertitude

Selon [12],

> la caractéristique principale de l'inférence bayésienne est la quantification directe de l'incertitude.

L'approche bayésienne Pour modéliser l'incertitude particulièrement utile lorsque:

- les données disponibles sont limitées;
- il y a une certaine inquiétude concernant le sur- ajustement("overfitting," en anglais);
- certains faits ont plus de chances d'être vrais que d'autres, mais cette information n'est pas contenue dans les données, ou
- la vraisemblance précise de certains faits est plus importante que la seule détermination du fait le plus probable (ou le moins probable).

L'exemple suivant représente une approche bayésienne pour faire face à l'incertitude de **Paradoxe des deux enveloppes**.

Exemple 10. On vous a donné deux enveloppes qui sont impossible à distinguer, chacune contenant un chèque, l'une étant deux fois plus élevée que l'autre. Vous pouvez choisir une enveloppe et garder l'argent qu'elle contient. Ayant choisi une enveloppe à volonté, mais avant de l'inspecter, vous avez la possibilité d'échanger les enveloppes. Devriez-vous échanger? Quel est le résultat attendu en le faisant? Expliquez comment ce jeu mène à un cycle infini.

Solution: Soit V la valeur (inconnue) trouvée dans l'enveloppe après la première sélection . L'autre enveloppe contient alors soit $\frac{1}{2}V$ ou soit $2V$, tous deux avec une probabilité de 0.5, et la valeur prévue d'échange est

$$E[\text{échange}] = 0.5 \times \frac{1}{2}V + 0.5 \times 2V = \frac{5}{4}V > V;$$

et il semble donc que cet échange est avantageux. Que la valeur (encore inconnue) du chèque dans la nouvelle enveloppe soit W. Le même argument montre que la valeur attendue de l'échange de **cette** enveloppe est de $\frac{5}{4}W > W$, il serait donc logique d'échanger l'enveloppe une fois de plus, et encore une fois, et ainsi de suite, menant à un cycle infini. Soit V la valeur (incertaine) dans la sélection originale, et W la valeur (également incertaine) dans la deuxième enveloppe. Une bonne résolution nécessite une loi conjointe (a priori) pour V et W. Maintenant, en absence de toute autre information, le plus que nous puissions dire sur cette loi en utilisant le principe de l'entropie maximale est $P(V < W) = P(V > W) = 0.5$.

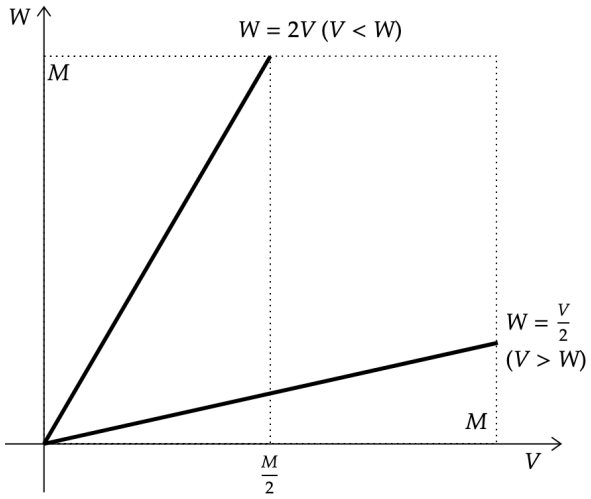
Par définition, si $V < W$, alors $W = 2V$; si, par contre, $V > W$ donc $W = \frac{V}{2}$. Nous montrons maintenant que la valeur attendue dans les deux enveloppes est la même, et donc que l'échange de l'enveloppe n'est pas une meilleure stratégie que de conserver la sélection originale.
En utilisant le théorème de Bayes, nous calculons que

$$\begin{aligned} E[W] &= E[W \mid V < W]P(V < W) + E[W \mid V > W]P(V > W) \\ &= E[2V \mid V < W] \cdot 0.5 + E[0.5V \mid V > W] \cdot 0.5 \\ &= E[V \mid V < W] + 0.25 \cdot E[V \mid V > W], \end{aligned}$$

tandis que

$$\begin{aligned} E[V] &= E[V \mid V < W]P(V < W) + E[V \mid V > W]P(V > W) \\ &= 0.5 \cdot E[V \mid V < W] + 0.5 \cdot E[V \mid V > W]. \end{aligned}$$

Avant de poursuivre, nous devons avoir quelques informations sur la loi conjointe $P(V,W)$ (notez cependant que $E[W]$ ne sera généralement pas égal à $\frac{5}{4}V$, comme cela avait été supposé au début de la solution).
Le domaine Ω e la probabilité conjointe consiste en ces paires (V,W) satisfaisant $V = 2W$ $(V > W)$ ou $W = 2V$ $(V < W)$ pour $0 < V,W < M$, où $M < \infty$ est une limite supérieure de la valeur de chaque chèque. [4]



[4] Dans le pire des cas, M devrait être inférieur à la quantité totale de richesse dont l'humanité a disposé tout au long de son histoire, bien qu'en pratique M devrait être sensiblement plus petit. Il est évident qu'un argument différent devra être présenté dans le cas où $M = \infty$.

Nous avons supposé que le poids de probabilité sur chaque branche de Ω est $1/2$; i nous supposons en outre, disons, que la valeur du chèque est aussi possible d'être l'une des valeurs admissibles sur ces branches, alors la loi conjointe est

$$P(V,W) = \begin{cases} \frac{1}{M} & \text{if } V < W \\ \frac{1}{2M} & \text{if } V > W \\ 0 & \text{otherwise} \end{cases}$$

et les attentes énumérées ci-dessus sont

$$E[V \mid V < W] = \int_{V<W} V \cdot P(V,W)\, d\Omega = \int_0^{M/2} V \cdot \frac{1}{M}\, dV = \frac{M}{8}$$

et

$$E[V \mid V > W] = \int_{V>W} V \cdot P(V,W)\, d\Omega = \int_0^{M} V \cdot \frac{1}{2M}\, dV = \frac{M}{4}.$$

Par conséquent,

$$E[W] = \frac{M}{8} + 0.25 \cdot \frac{M}{4} = \frac{3M}{16}$$

et

$$E[V] = 0.5 \cdot \frac{M}{8} + 0.5 \cdot \frac{M}{4} = \frac{3M}{16},$$

et le fait de changer d'enveloppe ne change pas la valeur prévue du résultat. Il n'y a pas de paradoxe, pas de cycle infini.

Exemple 11. Bayes dans la salle d'audience. Après la mort soudaine de ses deux fils, Sally Clark a été condamnée par un tribunal britannique à la prison à vie en 1996. Entre autres erreurs, le témoin expert Sir Roy Meadow avait mal interprété la faible probabilité de la mort subite des deux bébés comme une faible probabilité de l'innocence de Clark. Après une longue campagne, qui comprenait la réfutation des statistiques de Meadow en utilisant les statistiques bayésiennes, Clark a été libérée en 2003. Bien que l'innocence de Clark n'ait pas pu être prouvée hors de tout doute à l'aide de telles méthodes, sa culpabilité n'a non plus pu être établie hors de tout doute raisonnable et elle a été innocentée. Un article intéressant de la situation peut être trouvé en ligne [39].

6. Pourquoi utiliser les méthodes bayésiennes

Comme nous l'avons vu précédemment, les méthodes bayésiennes présentent un certain nombre de caractéristiques puissantes: elles permettent aux analystes de

- intégrer des connaissances préalables spécifiques sur les paramètres d'intérêt;
- mettre à jour de façon logique les connaissances sur le paramètre après avoir observé les données relatives à l'échantillon;
- faire des déclarations de probabilité formelles sur les paramètres d'intérêt;
- préciser les hypothèses du modèle et vérifier sa qualité et sa sensibilité à propos de ces hypothèses de manière simple;
- fournir des lois de probabilité plutôt que des estimations ponctuelles, et
- traiter les valeurs des données dans l'échantillon comme étant interchangeables.

6.1 Problèmes et solutions

En particulier, les méthodes bayésiennes sont indiquées afin de résoudre un certain nombre de défis problématiques dans l'analyse des données.

1. L'ensemble de données est petit, mais des informations externes connexes sont disponibles: utiliser les informations dans une a priori.
2. Le modèle est extrêmement flexible (modèle à forte variance) et il est donc sujet à un **overfitting** (surajustement): utilisez des a priori avec des niveaux proches de 0 (ce qui est à peu près équivalent au concept de régularisation dans le Machine Learning (l'apprentissage machine).
3. Il est intéressant de déterminer la vraisemblance des valeurs des paramètres, plutôt que de se contenter de produire une " meilleure estimation ": construire la distribution a posteriori complet pour le paramètre(s) et/ou la variable d'intérêt.

6.2 Test A/B Bayésien

Le test A/B est un excellent outil pour décider de mettre en place ou non des caractéristiques incrémentales. Pour effectuer un test A/B, nous divisons les utilisateurs au hasard en un groupe de test et un autre de contrôle, puis fournir la nouvelle caractéristique au groupe de test tout en laissant le groupe de contrôle continuer à découvrir la version actuelle du produit.

Si la procédure de randomisation est appropriée, nous pourrons peut-être attribuer toute différence de résultats entre les deux groupes aux changements que nous mettons en place sans avoir à tenir compte d'autres sources de variation affectant le comportement des utilisateurs. Avant d'agir sur ces résultats, cependant, il est important de comprendre la vraisemblance que toute différence observée soit simplement due au hasard plutôt qu'à une modification du produit.

Par exemple, il est parfaitement possible d'obtenir des rapports F/P différents entre deux pièces de monnaie équitables si nous effectuons seulement une nombre limité des tirages au sort; De la même manière, il est possible d'observer un changement entre les groupes A et B même si le comportement de l'utilisateur sous-jacent est identique.

Exemple 12. (Déduit de [28]) Wakefield Tiles est une entreprise qui vend des carreaux de sol par correspondance.

Elle essaie de devenir un acteur dynamique sur le marché lucratif de Chelsea en offrant un nouveau type de carreaux aux entrepreneurs de la région. Le service de marketing a mené une étude pilote et a essayé deux méthodes de marketing différentes:

- *A* – l'envoi par courrier d'une brochure colorée pour inviter les entrepreneurs à visiter la salle d'exposition de l'entreprise;
- *B* – l'envoi d'une brochure colorée par courrier pour inviter les entrepreneurs à visiter la salle d'exposition de l'entreprise, tout en incluant des échantillons de carreaux gratuits.

Le service marketing a envoyé 16 colis postaux de type *A* et 16 colis postaux de type *B*. Quatre Chelseaites qui ont reçu un colis de type *A* ont visité la salle d'exposition, tandis que 8 de ceux qui ont reçu un colis de type *B* ont fait de même. L'entreprise est consciente que:

- l'expédition de type *A* coûte 30 $ (comprend les frais d'impression et les frais de port);
- l'expédition de type *B* coûte 300$ (comprend en plus le coût des échantillons de tuiles gratuits);
- une visite à la salle d'exposition rapporte, en moyenne, 1000$ de revenus au cours de l'année suivante.

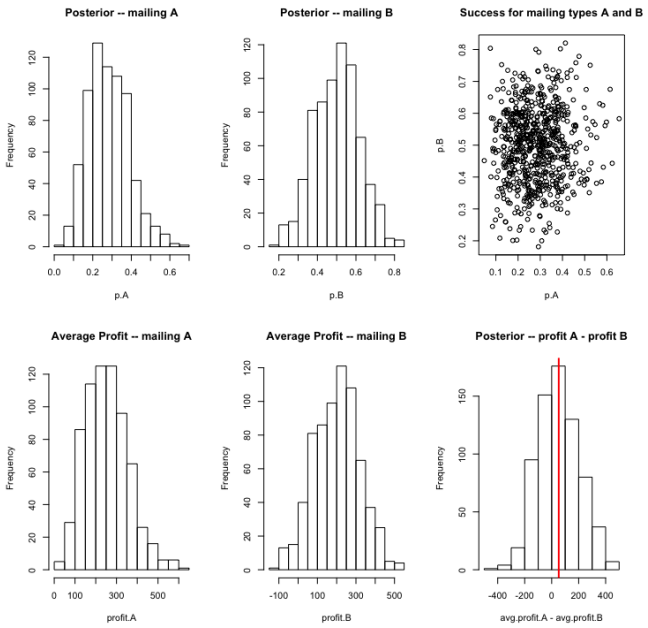
Laquelle des méthodes (A ou B) est la plus avantageuse pour Wakefield Tile?

Solution: la solution bayésienne exige la construction d'une loi a priori et **d'un modèle génératif**; dans le cadre du modèle génératif, nous devrons produire *n* répliques d'échantillons à partir de la loi binomiale (qui peut être fait dans R en utilisant `rbinom(n,size,prob)`).

La loi binomiale simule `n` fois le nombre de "succès" lors des essais de `taille` (envois postaux), où la probabilité de "succès" est `prob`. Une a priori couramment utilisé pour `prob` est la loi uniforme $U(0,1)$, à partir de laquelle nous pouvons échantillonner dans R *via* `runif(1, min = 0, max = 1)`.

```
# Le nombre de r\'ep\'etition de la
    distribution a priori
n.draws <- 200000

# la distribution a priori
# Ce g\'en\`ere la probabilit\'e des
# succ\`es des exp\'editions A et B,
# pour toutes les r\'ep\'etitions
prior <- data.frame(p.A = runif(n.draws,
    0, 1), p.B = runif(n.draws, 0, 1))

# Le mod\`ele g\'en\'eratif
# Cela nous indique le nombre de #
    visiteurs \`a attendre
# pour les exp\'editions des types # A
    et B
generative.model <- function(p.A, p.B) {
  visitors.A <- rbinom(1, 16, p.A)
  visitors.B <- rbinom(1, 16, p.B)
  c(visitors.A = visitors.A, visitors.B
      = visitors.B)
}

# simuler les donn\'ees des
# param\`etres \`a partir de l'a
# priori et du mod\`ele g\'en\'eratif
# Cela va simuler le vrai nombre
# des visiteurs pour chaque
# r\'ep\'etition
sim.data <-
    as.data.frame(t(sapply(1:n.draws,
    function(i) {
  generative.model(prior$p.A[i],
      prior$p.B[i])})))

# On garde seulement les
# probabilit\'es a priori pour
# lequelles le mod\`ele g\'en\'eratif
# correspond aux donn\'ees observ\'ees

posterior <- prior[sim.data$visitors.A
    == 4 & sim.data$visitors.B == 8, ]

# Visualiser les a posteriori
par(mfrow = c(1,3))
hist(posterior$p.A, main = "la
    distribution a posteriori -- la
    probabilit\'e de succ\`es pour
    l'exp\'edition de type A",
    xlab="p.A")
hist(posterior$p.B, main = "L'a
    posteriori -- la probabilit\'e de
    succ\`es pour l'exp\'edition de type
    B", xlab="p.B")
plot(posterior,main = "Nuage de points
    de la probabilit\'e de succ\`es pour
    les exp\'editions de types A et B",
    xlab="p.A", ylab="p.B")
```

Les lois a posteriori de la probabilité de succès pour chaque type d'expédition est illustré dans la figure ci- dessous.



Afin d'estimer le bénéfice moyen pour chaque type d'envoi, nous utilisons les lois a posteriori pour la probabilité de succès.

```
# Calculer le moyen estim\'e du profit
    pour chaque type d'envoi
avg.profit.A <- -30 + posterior$p.A *
    1000
avg.profit.B <- -300 + posterior$p.B *
    1000
hist(avg.profit.A, main = "Average
    Profit -- mailing A",
    xlab="profit.A")
hist(avg.profit.B, main = "Average
    Profit -- mailing B",
    xlab="profit.B")
```

Le bénéfice prévu est donc donné par le code suivant:

```
#Le profit estim\'e total
hist(avg.profit.A - avg.profit.B)
expected.avg.profit.diff <-
    mean(avg.profit.A - avg.profit.B)
abline(v = expected.avg.profit.diff ,
    col = "red", lwd =2)
```

Le profit prévu pour un envoi de type A est d'environ 52$ plus élevé que pour un envoi de type B (vos chiffres peuvent varier). Dans ce contexte, il semble préférable de garder les choses simples.

7. Bilan

Quoi?

- L'analyse bayésienne des données est une méthode flexible qui s'adapte à tout type de modèle statistique.
- Le maximum de vraisemblance est en fait un cas particulier de l'ajustement du modèle bayésien.

Pourquoi?

- Permet de définir des modèles fortement configurables.
- Permet d'inclure des informations provenant de nombreuses sources, comme les données et les connaissances d'experts.
- Quantifie et conserve l'incertitude des estimations et des prévisions des paramètres.

Comment?

- R! En utilisant ABC, MCMCpack, JAGS, STAN, R-inla, Python, etc.

Exercices – Solutions

- **Exercice** 1:

$$P(L \mid Y) = \frac{P(Y \mid L)P(L)}{P(Y)}$$

Notez que

$$P(Y) = P(Y,L)+P(Y,\overline{L}) = P(Y \mid L)P(L)+P(Y \mid \overline{L})P(\overline{L}),$$

Donc

$$P(L \mid Y) = \frac{(.55)(.52)}{(.55)(.52)+(.85)(.48)} = 0.41.$$

- **Exercice** 2:

étape 1: affecter des événements à A ou X. Vous voulez savoir quelle est la probabilité qu'une femme ait un cancer, compte tenu d'une mammographie positive. Pour ce problème, le fait d'avoir un cancer est A et un résultat positif est X.

étape 2: faites une liste des éléments de l'équation (cela facilite le travail sur l'équation effective):

$$P(A) = 0.01, P(\overline{A}) = 0.99, P(X \mid A) = 0.9, P(X \mid \overline{A}).$$

étape 3: insérer les éléments dans l'équation et résoudre. Notez que comme il s'agit d'un test médical, nous avons

$$\frac{0.9 \cdot 0.01}{(0.9)(0.01)+(0.08)(0.99)} = 0.10.$$

La probabilité qu'une femme ait un cancer, en cas de résultat positif au test, est donc de 10%.

- **Exercice 3**

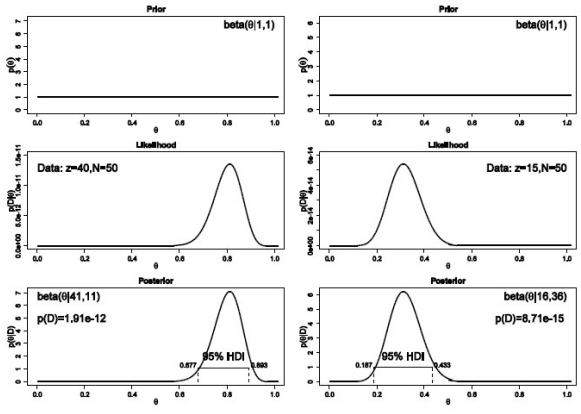
  1. justifier un a priori,on pourrait dire que notre force d'équité est équivalant à avoir déjà vu la pièce de monnaie être retournée 100 fois et avoir obtenu un F dans 50% de ces lancements. Par conséquent, la distribution a priori serait Beta($\theta \mid 50, 50$) (ce n'est pas la seule bonne réponse, bien sûr; vous pourriez plutôt être plus confiants et utiliser, disons, Beta($\theta \mid 500, 500$) si vous supposez que vous avez déjà vu $1,000$ lancement avec 50% de faces).
     L'a posteriori est alors Beta($\theta \mid 50+9, 50+1$), qui a une moyenne de $\frac{59}{59+51} = 0.536$. C'est la probabilité prédite de faces pour le prochain (11th) lancement.
  2. Dans ce cas, nous utilisons une loi Beta($\theta \mid 0.5, 0.5$) a priori, comme celui utilisé dans Exemple 5, car il exprime une croyance que la pièce est soit à biais de face, soit à biais de piles car il exprime une croyance que la pièce est soit à biais de face, soit à biais de piles.

     L'a posteriori est donc Beta($\theta \mid 0.5+9, 0.5+1$), qui a une moyenne de $\frac{9.5}{9.5+1.5} = 0.863$. C'est la probabilité prédite de faces pour le prochain (11ème) lancement. Notez qu'elle est tout a fait différente de la conclusion de la Partie 1.

- **Solution de l'exercice 4**
  Les commandes

```
> post = BernBeta(c(1,1),
c(rep(1,40),rep(0,10)))
> post = BernBeta(c(1,1),
c(rep(1,15),rep(0,35)))
```

donne le graphe ci-dessous.



Dans les deux cas, l'intervale HPD de 95% exclut $\theta = 0.5$, et nous concluons donc que les gens sont en effet biaisés dans leurs réponses, vers F dans le premier cas et vers J dans le second cas.

Références

[1] Bayes, T. [1763], An Essay towards solving a Problem in the Doctrine of Chances, *Phil. Trans. Royal Society London*.

[2] Gill, J. [2002], Bayesian Methods for the Social and Behavioral Sciences, Boca Raton, Florida: CRC Press.

[3] Hitchcock, D. [2014], Introduction to Bayesian Data Analysis, Department of Statistics, University of South Carolina, US, course notes.

[4] Berger, J.O. [1985], Statistical Decision Theory and Bayesian Analysis, Springer-Verlag, New York, 2nd edition.

[5] Robert, C.F. [2006], Le choix bayésien - Principes et pratique, Springer-Verlag France, Paris.

[6] Kruschke, J.K. [2009], Highlighting: A canonical experiment. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 153–185). Elsevier Academic Press.

[7] Sivia, D.S., Skilling, J. [2006], Data Analysis: A Bayesian Tutorial (2nd ed.), Oxford Science.

[8] Silver, N. [2012], The Signal and the Noise, Penguin.

[9] Jaynes, E.T. [2003], Probability Theory: the Logic of Science, Cambridge Press.

[10] Kruschke, J.K. [2011], Doing Bayesian Data Analysis: a Tutorial with R, JAGS, and Stan (2nd ed.), Academic Press

[11] Barber, D. [2012], Bayesian Reasoning and Machine Learning, Cambridge Press.

[12] Gelman, A., Carloin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. [2013], Bayesian Data Analysis (3rd ed.), CRC Press.

[13] Baath, R. [2015], Introduction to Bayesian Data Analysis with R, UseR!

[14] Oliphant, T.E. [2006], A Bayesian perspective on estimating mean variance, and standard-deviation from data, All Faculty Publications 278, BYU.

[15] **Reference Priors and Maximum Entropy** (lecture notes)

[16] **Bayesian Inference** (lecture notes)

[17] **MCMC algorithms for fitting Bayesian models** (lecture notes)

[18] **Bayesian Inference: Metropolis-Hastings Sampling**

[19] **Bayesian Statistics** (scholarpedia article)

[20] **Bayes' Theorem Problems, Definition and Examples** (statisticshowto.com)

[21] **Bayesian AB Testing** (Lyst)

[22] **Introduction to Bayesian Inference** (datascience.com)

[23] **Maximum Entropy Priors** (moreisdifferent.com)

[24] **Maximum Entropy**

[25] **MCMC chain analysis and convergence diagnostics with coda in R** (theoreticalecology.wordpress.com)

[26] **A simple Metropolis-Hastings MCMC in R** (theoreticalecology.wordpress.com)

[27] **Conjugate Priors** (Wikipedia)

[28] **Bayesian A/B Testing for Swedish Fish Incorporated** (tutorial)

[29] **Maximum Entropy Methods Tutorial: A Simple Example: The Taxicab** (video)

[30] **Maximum Entropy Methods Tutorial: MaxEnt applied to Taxicab Example Part 1** (video)

[31] **PYMC3** (documentation)

[32] Chipman, H.A., George, E.I., McCulloch, R.E. [2010], BART: Bayesian additive regression trees, Annals of Applied Statistics 6 (1), 266–298.

[33] Chipman, H.A., George, E.I., McCulloch, R.E. [1998], Bayesian CART model search (with discussion and a rejoinder by the authors), J. Amer. Statist. Assoc. 93 935–960.

[34] Hernandez, B., Raftery, A.E., Pennington, S.R., Parnell, A.C. [2015], Bayesian Additive Regression Trees Using Bayesian Model Averaging, Technical Report no. 636, Department of Statistics, University of Washington.

[35] Kabacoff, R.I. [2011], R in Action, Second Edition: Data analysis and graphics with R.

[36] **BayesTree** (CRAN package)

[37] **BART** (slides)

[38] **Doing Bayesian Data Analysis** (R Code).

[39] **Sally Clark is Wrongly Convicted of Murdering Her Children** (Bayesians Without Borders).

[40] Wilkinson, D.J. [2007], **Bayesian Methods in Bioinformatics and Computational Systems Biology**, Briefings in Bioinformatics, v.8, n.2, 109–116.

[41] Poldrac, R.A. [2006], 'Can cognitive processes be inferred from neuroimaging data?, *Trends Cogn. Sci.* 10(2):59-63.