Statistiques bayésiennes

M. Clertant [1]

Statistiques biomédicales

1. clertant@math.univ-paris13.fr, Lab. Analyse, Géométrie et Applications

Plan

Vraisemblance

Cas discret ($\mathbb{P}(. \mid \theta)$)

La vraisemblance du n-échantillon $(X_1, \ldots, X_n)$, notée $\mathcal{L}_n(.)$, est une fonction du paramètre θ définie par :

$$\mathcal{L}_n(\theta) = \prod_{k=1}^{n} \mathbb{P}(X = Xk \mid \theta) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$

Cas continu (f_θ)

La vraisemblance est :

$$\mathcal{L}_n(\theta) = \prod_{k=1}^{n} f_\theta(X_k).$$

Sous condition d'existence l'estimateur du maximum de vraisemblance (EMV) est :

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta).$$

Dans les calculs on utilise fréquemment la log-vraisemblance, $l_n(\theta) = \log(\mathcal{L}_n(\theta))$, afin de déterminer l'EMV. On minimise $-l_n(\theta)$ explicitement ou à l'aide d'algorithme (gradient, Newton-Raphson).

Exercice : Déterminer le maximum de vraisemblance d'une loi normale $\mathcal{N}(\mu, \sigma^2)$. Les estimateurs obtenus sont-ils biaisés ?

Divergence de Kullback-Leibler et entropie

Soient P et Q deux mesures de probabilité dominées par la mesure μ, avec :

$$\frac{\mathrm{d}P}{\mathrm{d}\mu} = p \quad \text{et} \quad \frac{\mathrm{d}Q}{\mathrm{d}\mu} = q$$

La divergence de Kullback-Leibler entre P et Q est :

$$D_{KL}(P||Q) = \int \log\left(\frac{p}{q}\right) p \,\mathrm{d}\mu.$$

Dans la pratique, Q est souvent une distribution théorique, un modèle, alors que P en est une approximation (les données). La divergence de kullback-Leibler est une pseudo-distance (pas de symétrie, ni d'égalité triangulaire). On a :

$$D_{KL}(P||Q) = \int \log\left(\frac{1}{q}\right) p \,\mathrm{d}\mu - \int \log\left(\frac{1}{p}\right) p \,\mathrm{d}\mu = H(P,Q) - H(P),$$

c'est à dire l'entropie croisée moins l'entropie.

Question : Exprimer la log-vraisemblance d'un échantillon X_1^n de loi de Bernoulli $\mathrm{Be}(q)$ en fonction de l'entropie croisée et montrer que maximiser la log-vraisemblance revient à minimiser une divergence de Kullback-Leibler .

Test UPP

Un test, T, est uniformément plus puissant (UPP) au niveau α si pour tout test T' on a :

$$1-\beta_{\mathcal{L}} = \mathbb{P}_{\mathcal{L}}(T(X_1^n)=1) \leq 1-\beta'_{\mathcal{L}} = \mathbb{P}_{\mathcal{L}}(T'(X_1^n)=1),$$

pour tout $\mathcal{L} \in \mathcal{L}_1$

Théorème (Lemme de Neyman-Pearson)

Soit $X_1^n = (X_1, \ldots, X_n)$ un n-échantillon de vraisemblance $\mathcal{L}_n(\theta)$. On considère les hypothèses $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$. Pour tout seuil α, il existe un test :

$$T(X_1^n) = \begin{cases} 1 & si \ \dfrac{\mathcal{L}_n(\theta_1)}{\mathcal{L}_n(\theta_0)} > c \\ \gamma & si \ \dfrac{\mathcal{L}_n(\theta_1)}{\mathcal{L}_n(\theta_0)} = c \\ 0 & si \ \dfrac{\mathcal{L}_n(\theta_1)}{\mathcal{L}_n(\theta_0)} < c \end{cases}.$$

avec c et γ telle que le test T ait pour seuil de signification α. Il s'agit de l'unique test UPP (pas d'unicité de γ).

Information de Fisher et borne de Cramér-Rao

Soit $X \sim \mathbb{P}_{\theta_0}$ et $f_{\theta_0} = \dfrac{\mathrm{d}\mathbb{P}_{\theta_0}(x)}{\mathrm{d}\mu(x)}$.

On pose : $l(X,\theta) = \log(f_\theta(X))$. Et on considère les dérivées en θ de cette fonction : $l'(X,\theta)$ et $l''(X,\theta)$.

L'information de Fisher est :

$$I(\theta) = \mathbb{E}_\theta\left[(l'(X,\theta))^2\right] = -\mathbb{E}_\theta\left[(l''(X,\theta))\right].$$

Remarque : L'information de Fisher ne dépend pas de X (intégration selon $f_\theta(x)\mathrm{d}\mu(x)$.) Lorsque l'on dispose d'un échantillon X_1^n, l'information de Fisher est $I_n(\theta) = nI(\theta)$.

Démontrer la seconde égalité ci-dessus (valable si la dérivée et l'intégrale peuvent être échangée).

Borne inférieure de Cramér-Rao

Soit X_1^n un échantillon de densité f_θ, et $T(X_1^n)$ un échantillon potentiellement biaisé. La variance de l'estimateur est bornée :

$$Var(T(X_1^n)) \geq \frac{\left(\dfrac{\partial}{\partial\theta}\mathbb{E}_\theta[T(X_1^n)]\right)^2}{nI(\theta)}$$

Si l'estimateur est sans biais, on a : $Var(T(X_1^n)) \geq (nI(\theta))^{-1}$.

Convergence de l'EMV

On dit qu'un estimateur $T(X_1^n)$ de θ est consistant si : $T(X_1^n) \xrightarrow{\mathbb{P}} \theta$

Soient $X_1, \ldots, X_n$ des observations iid suivant la loi f_θ : $X_k \sim f_\theta$.
L'estimateur du maximum de vraisemblance conditionnellement aux X_i est $\hat{\theta}_n$.

Consistance de l'EMV

Sous des conditions de régularité, l'estimateur du maximum de vraisemblance est consistant.

Normalité asymptotique de l'EMV

Sous des conditions de régularité, on a :

$$\sqrt{n}(\hat{\theta}_n - \theta) \underset{n \to +\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, I(\theta)^{-1}),$$

où $I(\theta)$ est l'information de Fisher : $I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log(f_\theta(X))\right]$.

Un théorème très utile en sélection de modèle

On pose : $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$ et $\theta_0 = (\theta_1, \ldots, \theta_{k'}) \in \Theta_0$ avec $k' < k$ de sorte que f_{θ_0} est un sous-modèle de f_θ : $\Theta_0 \subset \Theta$. On note L la fonction de vraisemblance et $\hat{\theta}$ et $\hat{\theta}_0$ les emv respectifs.

Théorème de Wilks

Sous des conditions de régularité et θ étant le vrai paramètre, si l'hypothèse $H_0 : \theta \in \Theta_0$ est vraie, on a :

$$-2 \log \left(\frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \right) \underset{n \to +\infty}{\overset{\mathcal{L}}{\to}} \chi^2(k - k').$$

Transition : Frequentists versus Bayesians

"It will be noticed that in the above description, the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results will tend to $1 - \alpha$.
Consider now the case when a sample is already drawn, and the calculations have given particular limits.
Can we say that in this particular case the probability of the true value falling between these limits is equal to $1 - \alpha$?
The answer is obviously in the negative. The parameter is an unknown constant, and no probability statement concerning its value may be made..."

Neyman, J. (1937). "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability". Philosophical Transactions of the Royal Society A.

Really ?

Plan

Modèle général

Soit X une v.a. provenant d'une expérience statistique et $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$, un modèle de cette expérience.

Modélisation fréquentiste :

$$X \sim \mathbb{P}_\theta.$$

On observe un échantillon $X_1^n = (X_1, \ldots, X_n)$ et on cherche un estimateur $\hat{\theta}_n$ en fonction de cette échantillon.

Modélisation bayésienne :

$$\theta \sim \Pi$$
$$X \mid \theta \sim \mathbb{P}_\theta$$

On cherche l'a posteriori Π_n conditionnellement à l'échantillon X_1^n.

Dans le cadre dominé (celui de ce cours), les lois Π et $\mathbb{P}_\theta$ ont une densité par rapport à une mesure positive σ-finie (mesure de Lebesgue ou de comptage) :

$$\mathrm{d}\mathbb{P}_\theta(x) = p_\theta \mathrm{d}\mu(x) \quad \text{et} \quad \mathrm{d}\Pi(\theta) = \pi \mathrm{d}\nu(\theta).$$

Propriété : Supposons l'application $(x, \theta) \to p_\theta(x)$ mesurable, alors l'application $(x, \theta) \to p_\theta(x)\pi(\theta)$ est une densité par rapport à $\mu \otimes \nu$.

Loi a posteriori

La loi jointe de X et θ a pour densité : $f(x, \theta) = p_\theta(x)\pi(\theta)$.

La formule de la densité conditionnelle entraîne que :

$$\pi(\theta \mid x) = \frac{p_\theta(x)\pi(\theta)}{\int p_\theta(x)\pi(\theta)\mathrm{d}\nu(\theta)}$$

Dans le cas d'un échantillon X_1^n iid sous la loi conditionnelle $\mathbb{P}_\theta$:

$$X_1, \ldots, X_n \mid \theta \sim \bigotimes_{i=1}^{n} P_\theta$$

et la densité jointe de l'échantillon et de θ est :

$$(x_1, \ldots, x_n, \theta) \to \left(\prod_{i=1}^{n} p_\theta(x_i)\right) \pi(\theta).$$

La loi a posteriori conditionnellement à l'échantillon a donc pour densité :

$$\pi_n(\theta) = \pi(\theta \mid x_1^n) = \frac{\prod_{i=1}^{n} p_\theta(x_i)\, \pi(\theta)}{\int \prod_{i=1}^{n} p_\theta(x_i)\, \pi(\theta)\mathrm{d}\nu(\theta)}$$

Deux exemples

Thomas Bayes, *Essay Towards Solving a Problem in the Doctrine of Chances*, 1763 : Un boule de billard lancée sur une ligne de longueur 1 s'arrête uniformément à une distance d. Dans les mêmes conditions, n boules sont lancées. Connaissant le nombre de boules s'étant arrêtées à plus de d et à moins de d, quelle inférence peut-on mener sur d?

Loi normale de variance connue :

$$\theta \sim \mathcal{N}(0,1)$$

$$X_1^n \mid \theta \sim \bigotimes_{i=1}^{n} \mathcal{N}(\theta,1)$$

Déterminer la loi a posteriori $\Pi_n = \Pi(\cdot \mid X_1)$.

Remarque : La densité de la loi a posteriori est proportionnelle à la multiplication de la vraisemblance, $\mathcal{L}_n(\theta)$, par la densité de l'a priori.

$$\pi_n(\theta) \propto \left(\prod_{i=1}^{n} p_\theta(x_i)\right) \pi(\theta) = \mathcal{L}_n(\theta) \times \pi(\theta)$$

Estimateurs classiques basés sur l'a posteriori Π_n

La moyenne :

$$\overline{\theta}_n = \int \theta \, \mathrm{d}\Pi_n(\theta).$$

La matrice de variance-covariance :

$$\overline{\theta}_n = \int \left(\theta - \overline{\theta}_n\right)\left(\theta - \overline{\theta}_n\right)^\top \mathrm{d}\Pi_n(\theta).$$

Le mode :

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \Pi_n(\theta).$$

Soit F_n la fonction de répartition de l'a posteriori Π_n et Q_n la fonction quantile :

$$\forall u \in [0,1],\ Q_n(u) = \inf\{\theta \in \Theta,\ F_n(\theta) \geq u\}.$$

La médiane a posteriori est $Q_n(1/2)$.

Lorsque $\theta \sim \mathcal{N}(0,1)$ et $X_1^n \mid \theta \sim \bigotimes_{i=1}^{n} \mathcal{N}(\theta, 1)$, l'a posteriori est $\mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1}\right)$; la moyenne, le mode et la médiane à posteriori sont alors égales et la variance est $\frac{1}{n+1}$.

Régions de crédibilité

1. Intervalle basé sur des quantiles a priori :
Soit Q_n la fonction quantile a posteriori. Si la fonction de répartition est strictement croissante, on a :

$$\Pi_n([Q_n(\alpha/2), Q_n(1-\alpha/2)]) = \alpha.$$

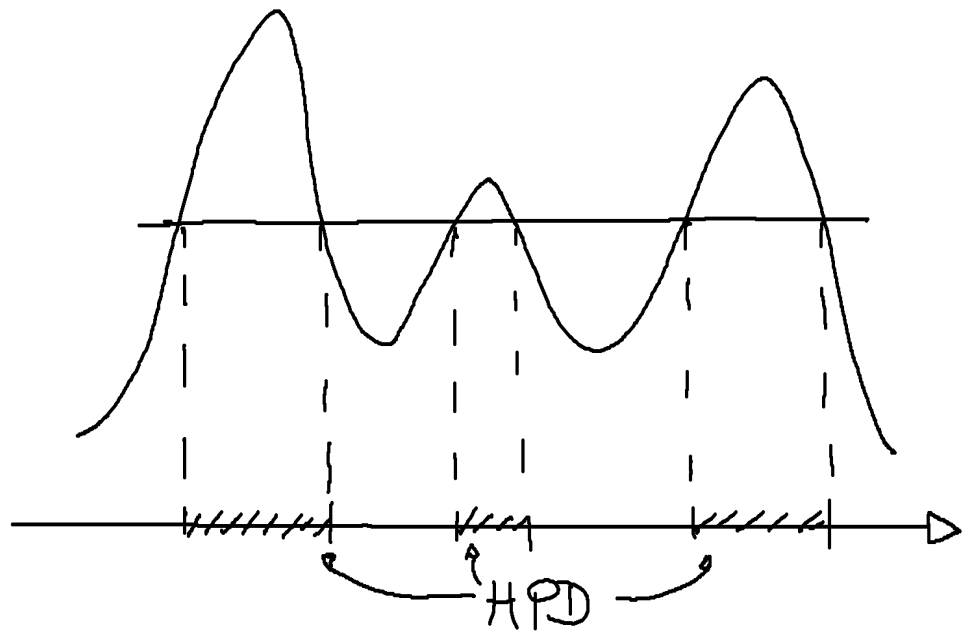
2. Région de plus haute densité (highest posterior density, HPD) : Soit π_n la densité a posteriori. Pour $x \in \mathbb{R}^+$, on pose $R_n(x) = \{\theta \in \Theta, \pi_n(\theta) > x\}$.

La HPD au niveau $1-\alpha$ est :

$$\mathcal{R}_{n,1-\alpha} = R_n(x_\alpha),$$

avec

$$x_\alpha = \sup\{x \in \mathbb{R}^+, \Pi_n\left(R_n(x)\right) \geq 1-\alpha\}.$$



Propriété : Une région HPD est de volume minimal parmi les régions de même niveau de crédibilité.

Plan

Lois a priori conjuguées

On dit qu'une loi a priori est conjuguée avec la vraisemblance d'un modèle si l'a posteriori appartient à la même famille de loi que l'a priori.

Avantages : Calcul explicite des paramètres de la loi a priori ; famille de lois bien connues sous lesquelles on sait échantillonner.

Model $f(x \mid \theta)$	Prior $\Pi(\theta)$	Posterior $\Pi(\theta \mid x)$
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}\left(\frac{\mu\sigma^2 + x\tau^2}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)$
$\mathcal{P}(\theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + x, \beta + 1)$
$\mathcal{B}(n, \theta)$	$\mathrm{Be}(a, b)$	?

TABLE – Quelques exemples et plus ici.

La famille des lois exponentielles (Bernoulli, binomiale, poisson, normale, gamma, χ^2 ...) s'écrit sous la forme :

$$f(x \mid \theta) = h(x) \exp\left(\langle \theta, T(X) \rangle - A(\theta)\right)$$

sa loi a priori conjuguée est :

$$f(x \mid \theta) = k(\mu, \lambda) \exp\left(\langle \theta, \mu \rangle - \lambda A(\theta)\right)$$

Et sinon, pourquoi pas des mélanges de lois ou des lois tronquées ?

Lois conjuguées pour un modèle multivarié

Exercice : Soit $X_i \in \mathbb{R}^d$ tel que

$$\mu \sim \mathcal{N}(\mu', \Sigma')$$
$$X_1^n \mid \theta \sim \bigotimes_{i=1}^{n} \mathcal{N}(\mu, \Sigma).$$

Démontrer le résultat suivant dans le cas d'une observation et pour $\mu' = 0$.

La famille $\mathcal{N}(\mu', \Sigma')$ est conjuguée avec la vraisemblance de X_1^n est on a :

$$\Pi_n \sim \mathcal{N}(\mu_n, \Sigma_n),$$

avec $\Sigma_n = (\Sigma'^{-1} + n\Sigma^{-1})^{-1}$ et $\mu_n = \Sigma_n(\Sigma^{-1}\mu' + n\Sigma^{-1}\overline{X}_n)$

A priori impropres

Dans le cadre d'un modèle bayésien :

$$\theta \sim \Pi(\theta)$$
$$X \mid \theta \sim f(x \mid \theta),$$

On parle de loi a priori impropre (abus de langage), lorsque $\Pi(\Theta) = +\infty$ et $\int f(x \mid \theta)\mathrm{d}\Pi(\theta)$ est finie ps.
La loi a posteriori est alors la loi de probabilité de densité :

$$\pi_n(\theta) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)\mathrm{d}\theta}$$

Vraisemblance normale de moyenne et variance inconnue : Soit $Y \sim \gamma(a, b)$, la loi de Y^{-1} est une inverse-gamma $\gamma^{-1}(a, b)$. La loi normale inverse gamma, notée NIG(a, b, c, d) d'un couple (mu, σ^2) correspond au schéma :

$$\sigma^2 \sim \Gamma^{-1}(c, d)$$
$$\mu \mid \sigma^2 \sim \mathcal{N}(a, \sigma^2/b).$$

La loi est NIG est conjuguée pour $x \mid (\mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2)$. En pratique, on utilise souvent :

$$\mathrm{d}\Pi(\mu, \sigma^2) = \frac{\mathrm{d}\mu\mathrm{d}\sigma^2}{\sigma^2},$$

et

$$\Pi_n \sim \text{NIG}(\overline{X}_n, n, \frac{n}{2}, \frac{n(\overline{X^2}_n) - \overline{X}_n^2}{2}).$$

Invariance- A priori de Jeffreys

Différents principes d'invariance :

- par translation ; a priori impropre constant,
- par changement d'échelle,
- Par re-paramétrisation (comme ci-dessous).

La log-vraisemblance est notée $l(\theta) = \log\left[\mathcal{L}_n(\theta)\right]$. On rappelle que l'information de Fisher est :

$$I(\theta) = \mathbb{E}_\theta(l'(\theta)^2).$$

L'a priori de Jeffreys est :

$$\pi(\theta) \propto \sqrt{I(\theta)}.$$

et si $\theta \in \mathbb{R}^d$, l' a priori de Jeffreys est proportionnel à $\sqrt{\det(I(\theta))}$.

Propriété- Invariance par re-paramétrisation : Soit θ un paramètre et Π, son a priori de Jeffreys. Le nouveau paramètre est $\eta = g(\theta)$ (avec g un difféomorphisme). La mesure image $\Pi \circ g^{-1}$ est l'a priori de Jeffreys pour le modèle paramétré par η.

Exercice : Soit $X \mid \theta \sim \mathcal{B}(\theta)$.
1. Montrer que la loi uniforme $\mathcal{U}_{[0,1]}$ n'est pas un a priori de Jeffreys. On posera : $\theta = \eta^2$ et on calculera l'a priori de η.
2. Calculer l'a priori de Jeffreys de ce modèle.

Les approches bayésiennes empiriques et hiérarchiques

Empirical Bayes : Le modèle dépend d'un paramètre α,

$$\theta \mid \alpha \sim \Pi_\alpha(\theta)$$
$$X \mid \theta \sim P_\theta,$$

et celui s'évalue comme le meilleur candidat de la vraisemblance marginilisé en θ :

$$\alpha^\star = \arg\max_\alpha \int p_\theta(x_1^n) \mathrm{d}\Pi_\alpha(\theta).$$

Hierarchical Bayes : Le paramètre α est doté d'une loi :

$$\alpha \sim L$$
$$\theta \mid \alpha \sim \Pi_\alpha(\theta)$$
$$X \mid \theta \sim P_\theta,$$

La loi marginale de θ est un mélange : $g(\theta) = \int \pi_\alpha(\theta) \mathrm{d}L(\alpha)$.

On peut aussi s'intéresser à α en intégrant en θ (voir ci-dessous).

Exercice- sélection de modèle : Le résultat d'une expérience suit une loi Bernoulli dont le paramètre est soit proche de $\theta = 1/6$, soit proche de $\theta = 1/2$, soit proche $\theta = 5/6$. Proposer un modèle hiérarchique permettant de sélectionner l'une des trois hypothèses.

Plan

Théorème de Doob

Hypothèses :

- Pour tout élément A de la σ-algèbre de $\mathcal{X}$, $\theta \to \mathbb{P}_\theta(A)$ est mesurable.
- Le modèle est identifiable : si $\theta \neq \theta'$, $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$.

Théorème

1. Sous les hypothèses, décrites ci-dessus, il existe $\Theta_0 \in \Theta$ *avec* $\Pi(\Theta_0) = 1$ *tel que l'a posteriori soit consistant en tout* $\theta_0 \in \Theta_0$, *c'est-à-dire, pour tout voisinage* B *de* θ_0 *:*

$$\lim_{n \to +\infty} \Pi(\theta \in B \mid X_1^n) = 1, \quad \mathbb{P}_{\theta_0}\text{-}ps.$$

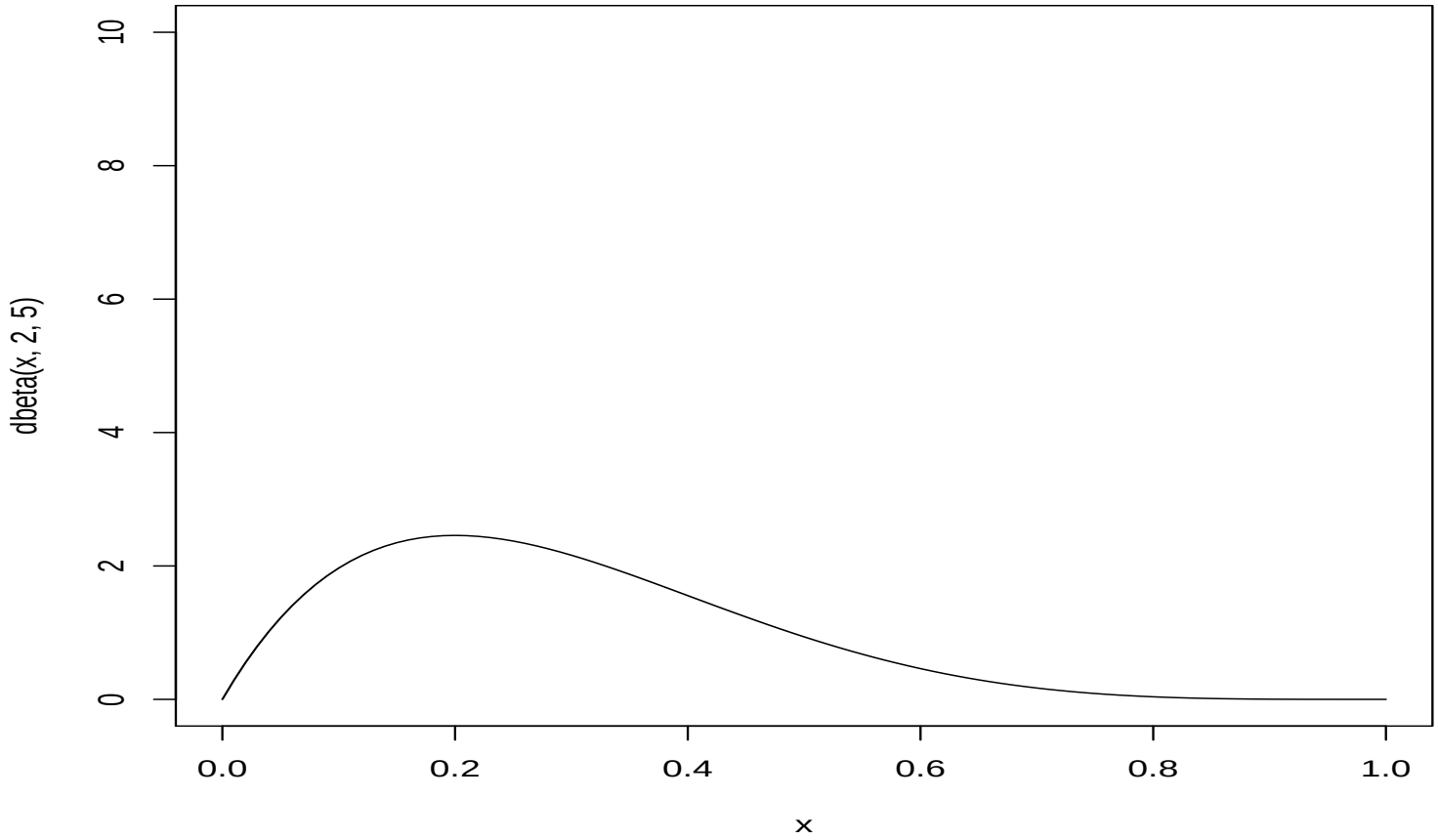
2. Pour $g : \Theta \to \mathbb{R}$ *appartenant à* L^1, *il existe* $\Theta_0 \in \Theta$ *avec* $\Pi(\Theta_0) = 1$ *tel que, pour tout* $\theta_0 \in \Theta_0$:

$$\lim_{n \to +\infty} \mathbb{E}(g(\theta) \mid X_1^n) = g(\theta), \quad \mathbb{P}_{\theta_0}\text{-}ps.$$

Remarque : On ne sait rien des $\theta \in \Theta \setminus \Theta_0$. Cet ensemble est potentiellement grand, en particulier dans le cas bayésien non-paramètrique et alors même que l'a priori est "bien choisi" (Diaconis and Freedman, 1986, *Ann. of Stat.*).
Point de vue fréquentiste de la consistance : Π_n est consistant en un point θ_0 si pour tout voisinage B de θ_0 :

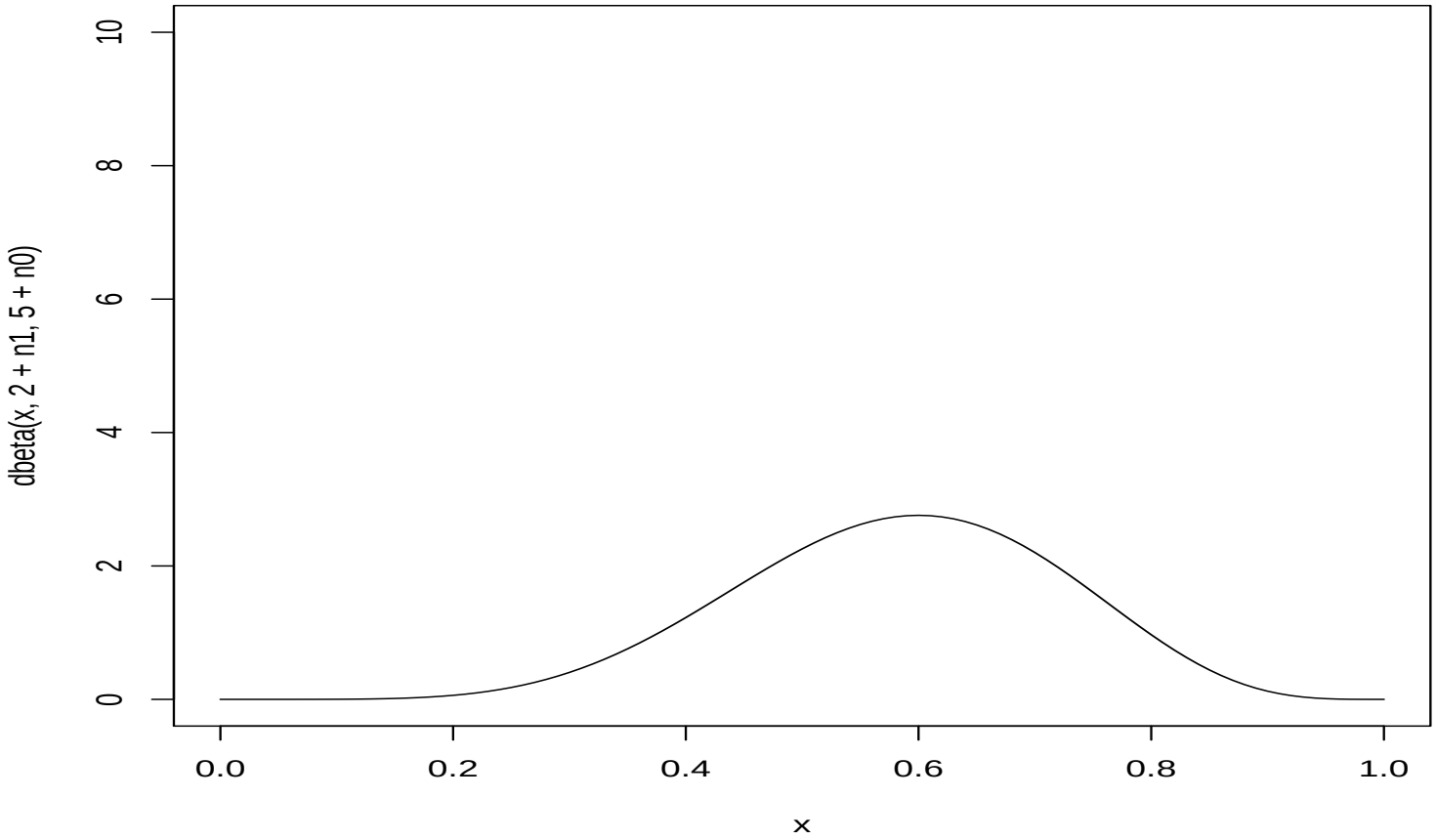
$$\Pi(\theta \in B \mid X_1^n) \to 1 \text{ en probabilité sous } \mathbb{P}_{\theta_0}.$$

Convergence de la loi a posteriori



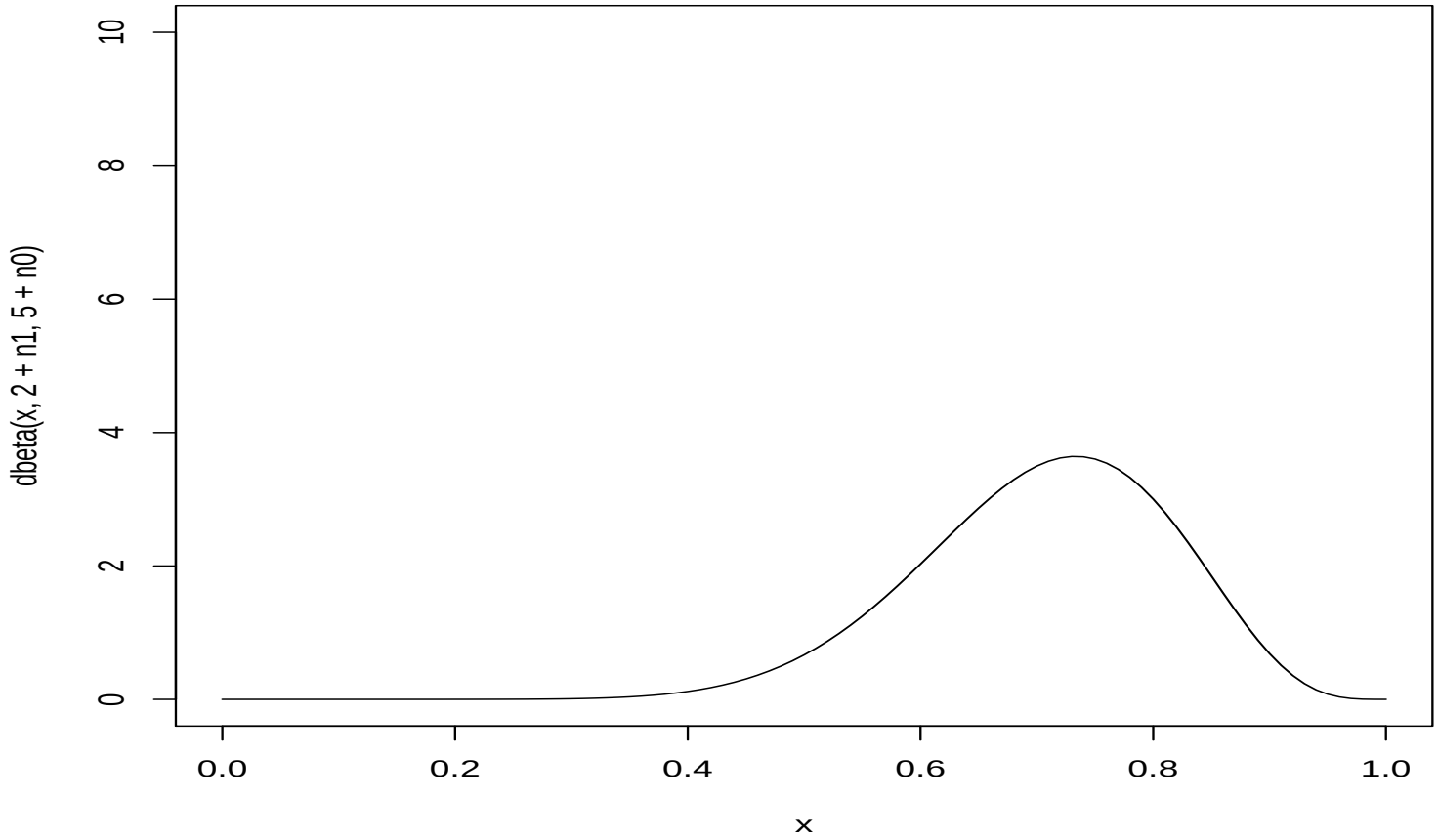
La courbe correspond à l'a priori $\mathrm{Be}(2,5)$; partant d'une loi uniforme, on ajoute 5 pseudo-observations : 4 "0" et 1 "1". Vrai modèle : Bernoulli $\mathcal{B}(0.8)$.

Convergence de la loi a posteriori



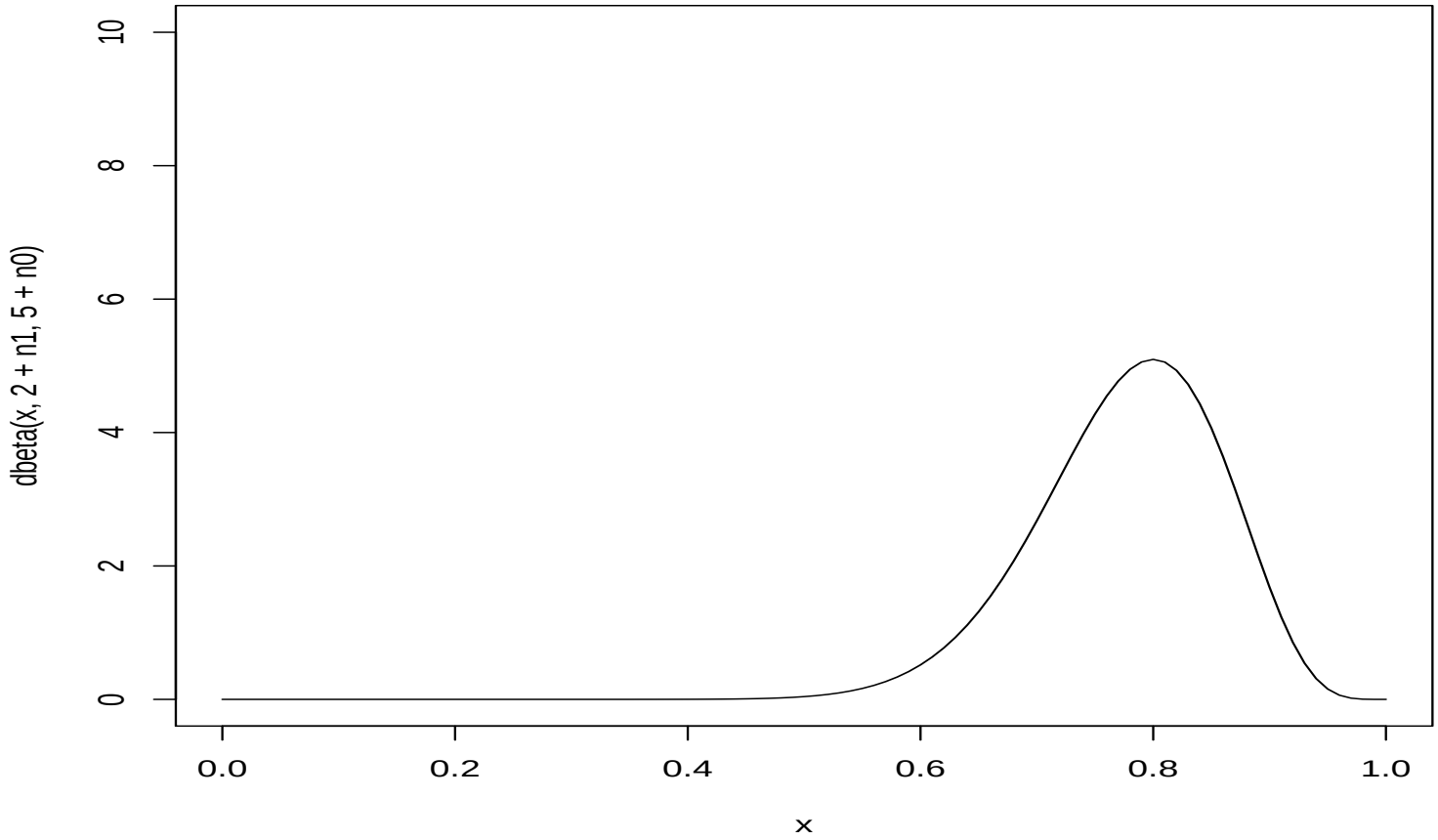
La courbe correspond à l'a posteriori Π_n avec $n = 5$. A priori : Be$(2, 5)$;
Vrai modèle : Bernoulli $\mathcal{B}(0.8)$.

Convergence de la loi a posteriori



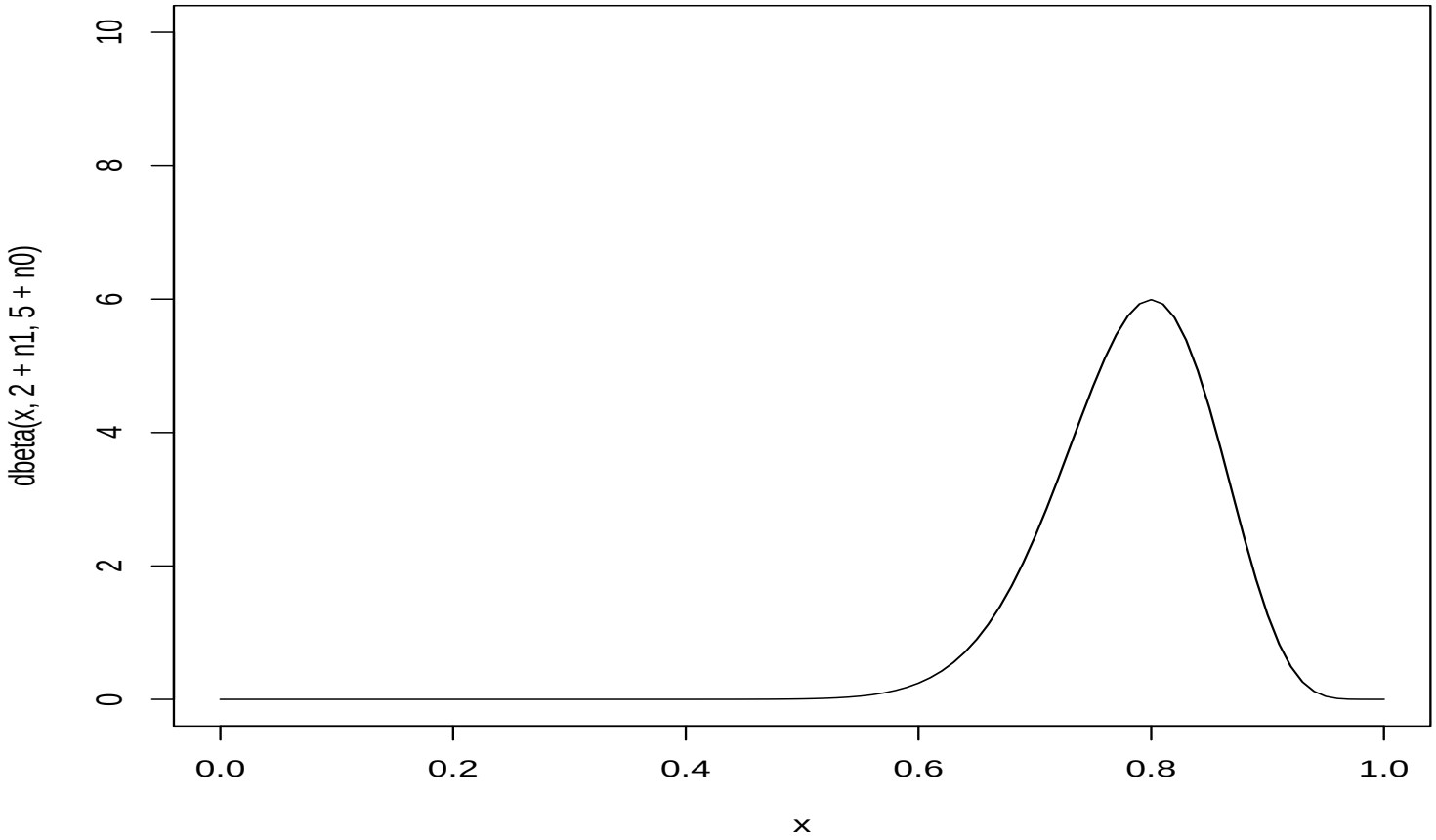
La courbe correspond à l'a posteriori Π_n avec $n = 10$. A priori : $\mathrm{Be}(2, 5)$;
Vrai modèle : Bernoulli $\mathcal{B}(0.8)$.

Convergence de la loi a posteriori



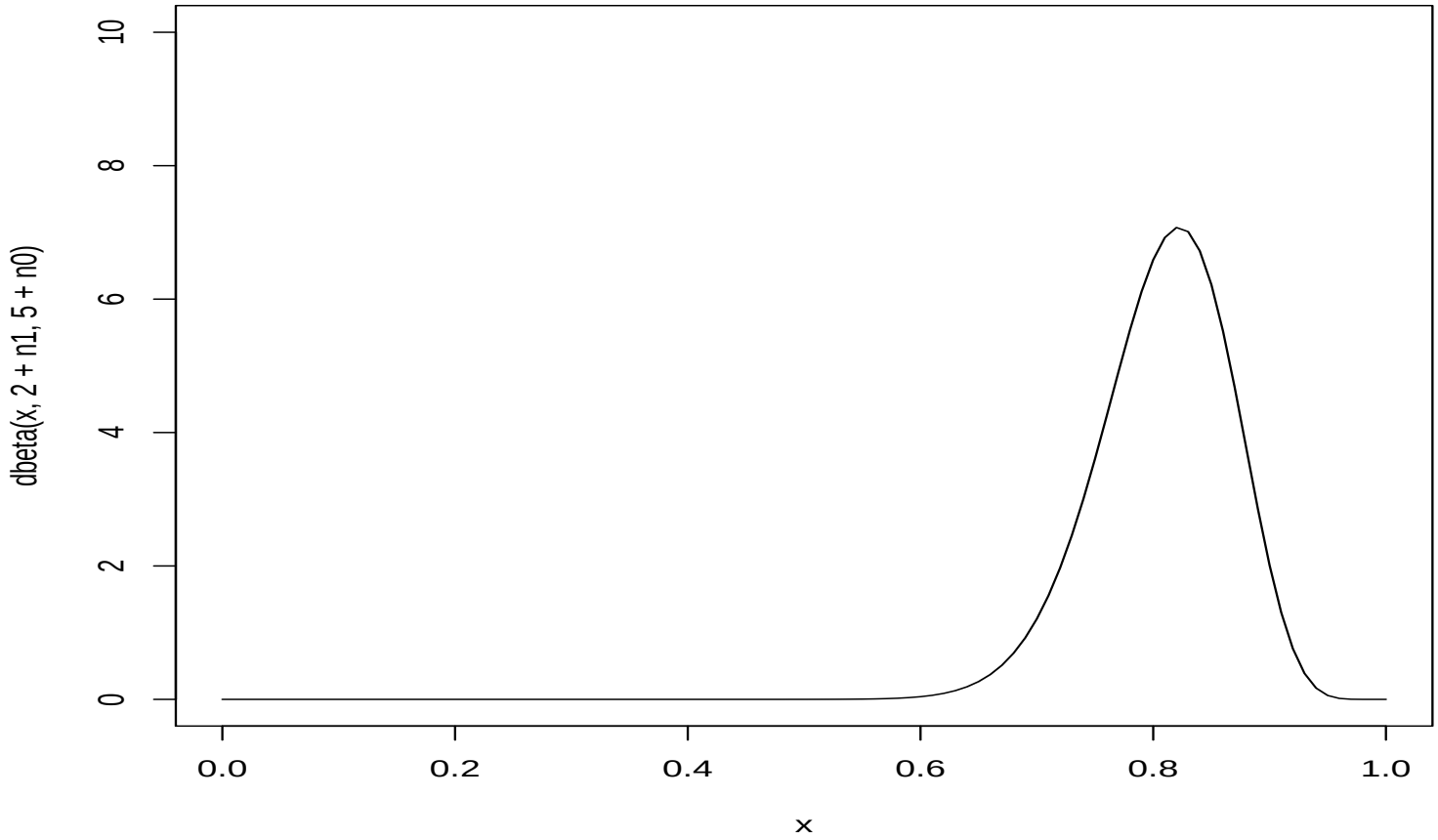
La courbe correspond à l'a posteriori Π_n avec $n = 20$. A priori : $\mathrm{Be}(2, 5)$;
Vrai modèle : Bernoulli $\mathcal{B}(0.8)$.

Convergence de la loi a posteriori



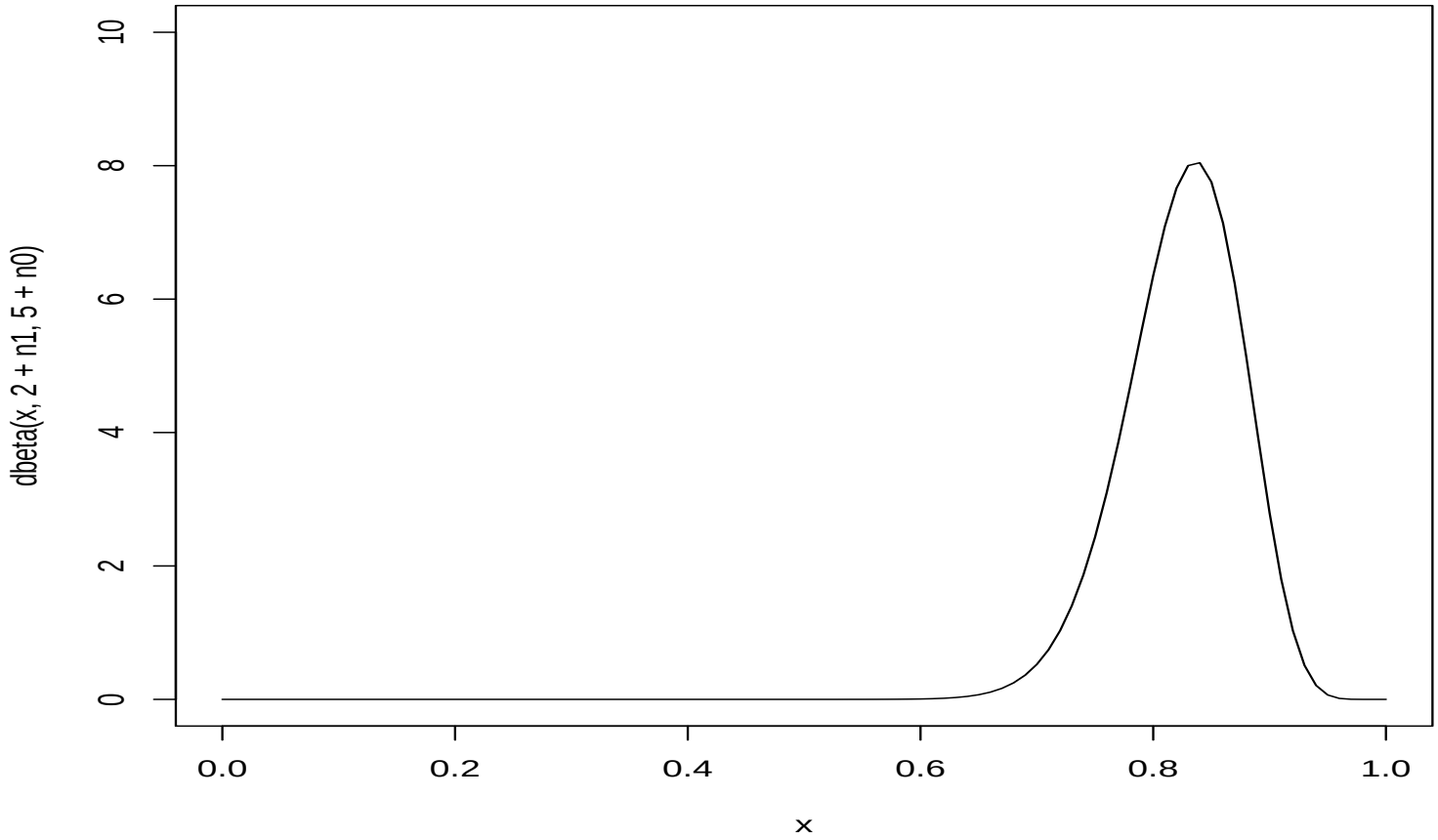
La courbe correspond à l'a posteriori Π_n avec $n = 30$. A priori : $\mathrm{Be}(2, 5)$;
Vrai modèle : Bernoulli $\mathcal{B}(0.8)$.

Convergence de la loi a posteriori



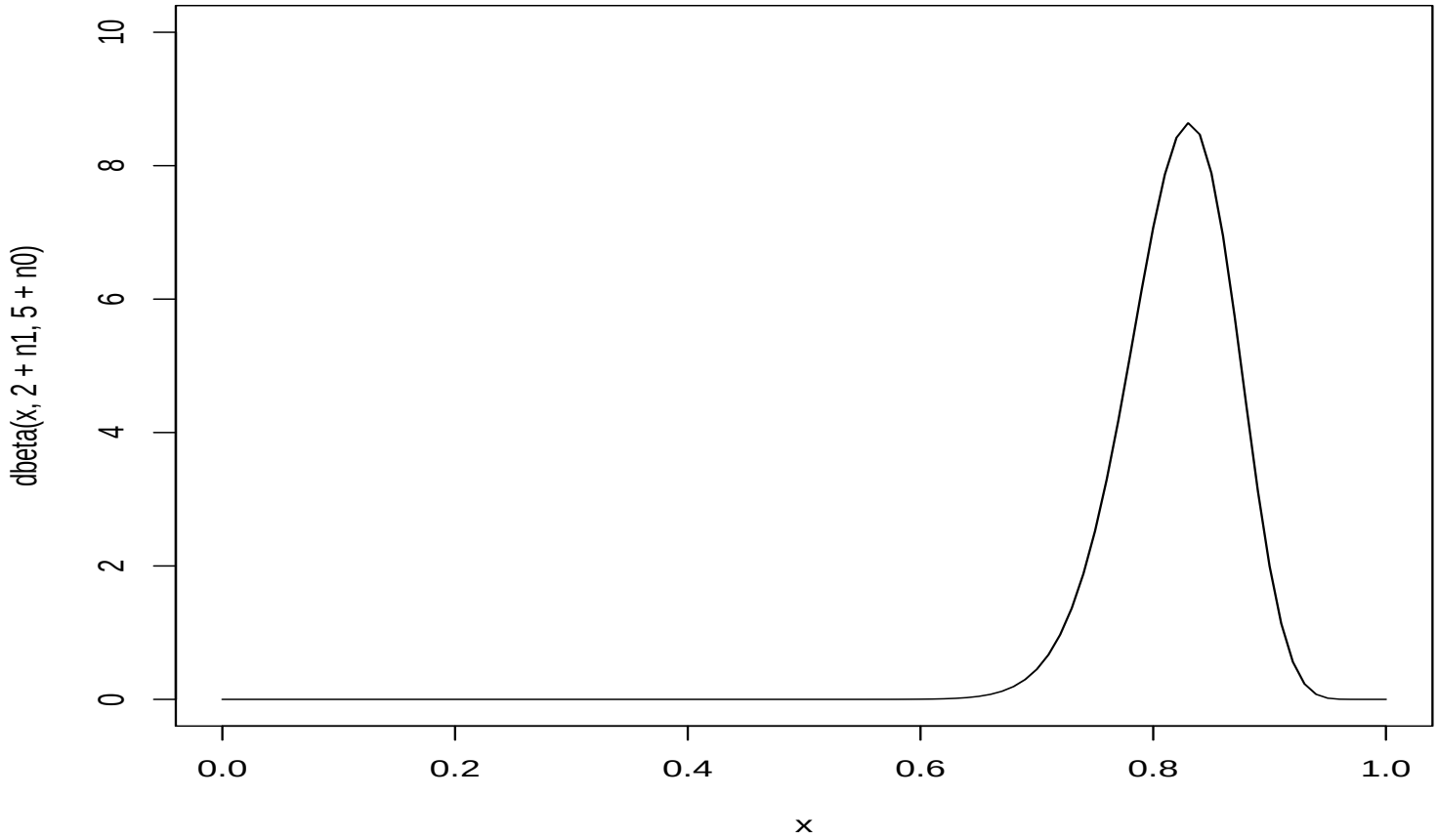
La courbe correspond à l'a posteriori Π_n avec $n = 40$. A priori : $\mathrm{Be}(2, 5)$;
Vrai modèle : Bernoulli $\mathcal{B}(0.8)$.

Convergence de la loi a posteriori



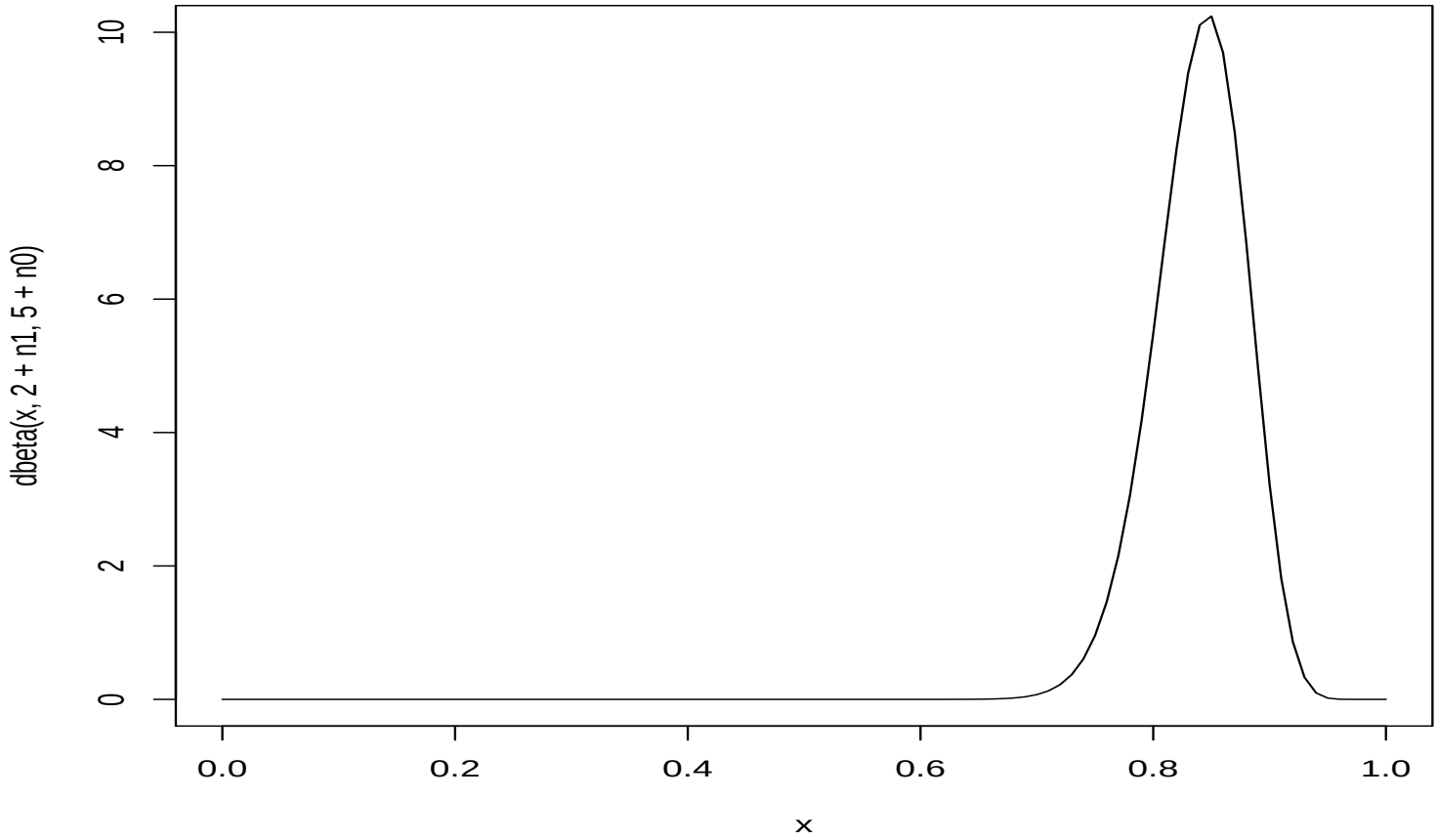
La courbe correspond à l'a posteriori Π_n avec $n = 50$. A priori : Be$(2, 5)$;
Vrai modèle : Bernoulli $\mathcal{B}(0.8)$.

Convergence de la loi a posteriori



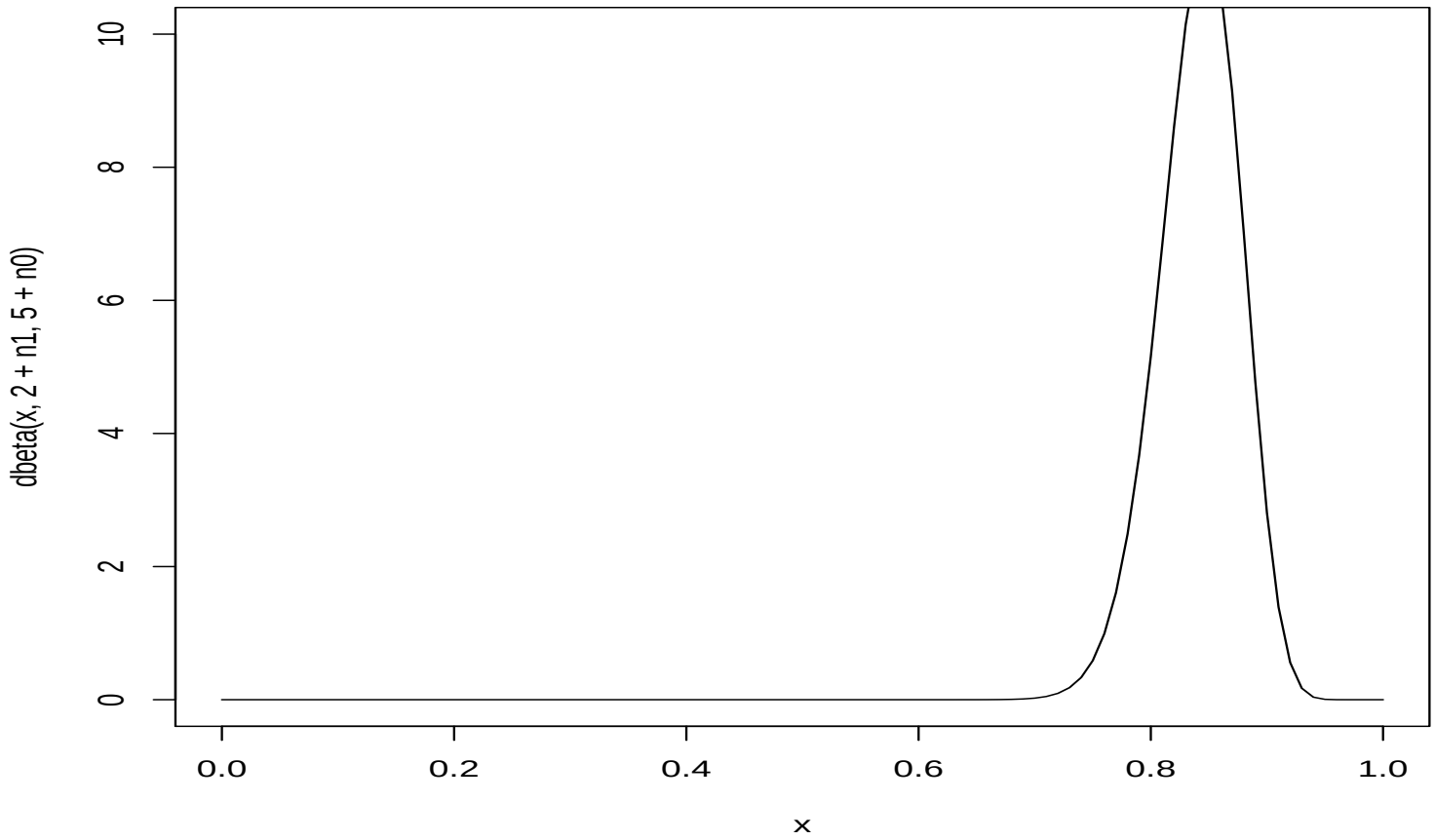
La courbe correspond à l'a posteriori Π_n avec $n = 60$. A priori : $\mathrm{Be}(2, 5)$;
Vrai modèle : Bernoulli $\mathcal{B}(0.8)$.

Convergence de la loi a posteriori



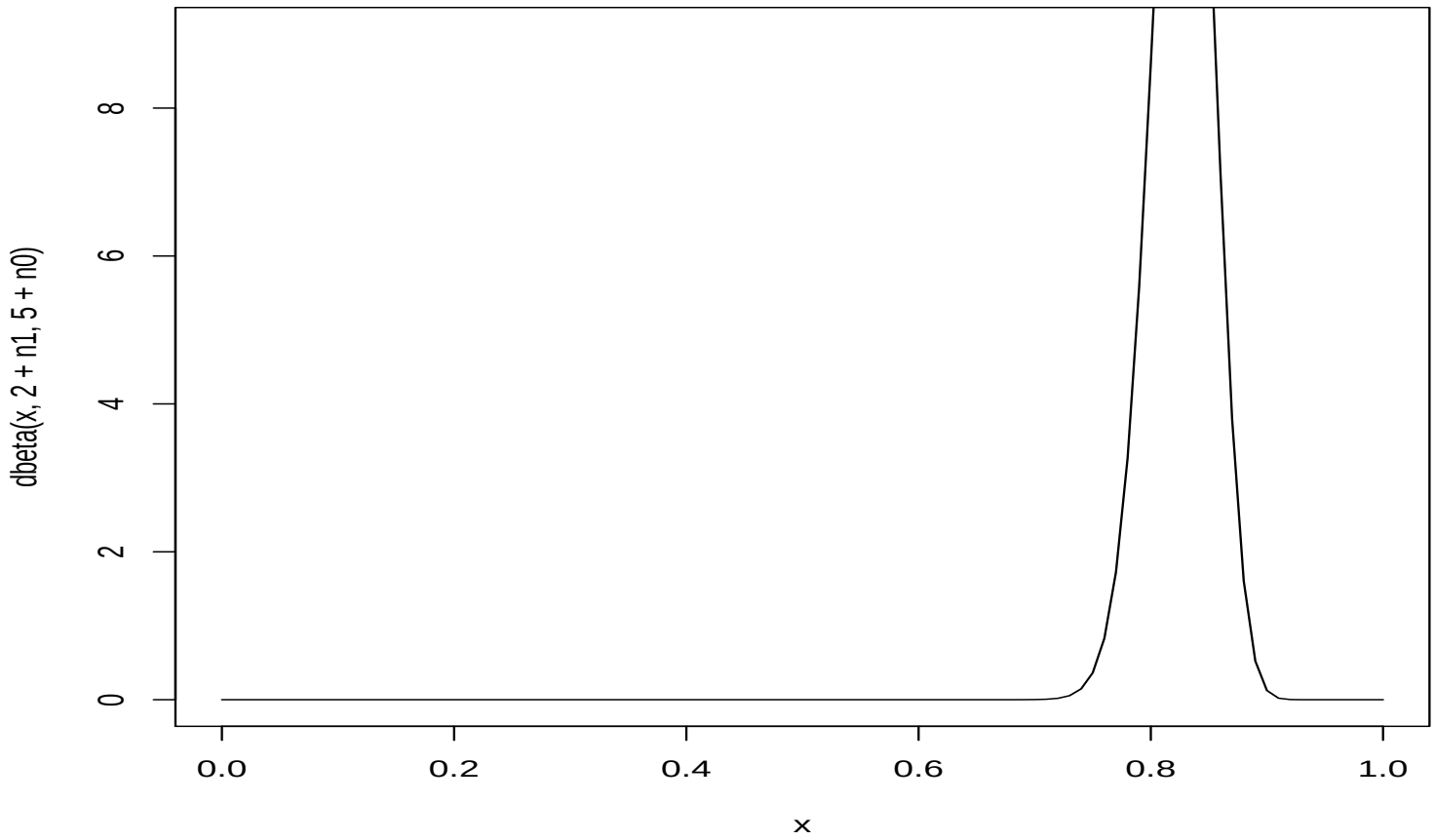
La courbe correspond à l'a posteriori Π_n avec $n = 80$. A priori : $\mathrm{Be}(2, 5)$;
Vrai modèle : Bernoulli $\mathcal{B}(0.8)$.

Convergence de la loi a posteriori



La courbe correspond à l'a posteriori Π_n avec $n = 100$. A priori : Be$(2, 5)$;
Vrai modèle : Bernoulli $\mathcal{B}(0.8)$.

Convergence de la loi a posteriori



La courbe correspond à l'a posteriori Π_n avec $n = 200$. A priori : Be$(2, 5)$;
Vrai modèle : Bernoulli $\mathcal{B}(0.8)$.

Il est même possible d'obtenir la convergence presque sûr de l'estimateur de la moyenne, $\overline{\theta}_n$, dans de nombreux cas.

Likelihood $p_\theta(x)$	Prior $\Pi(\theta)$	Posterior $\Pi_n(\theta)$	Estimator $\overline{\theta}_n$
$\mathcal{N}(\theta, 1)$	$\mathcal{N}(\mu, 1)$	$\mathcal{N}\left(\frac{\mu + n\overline{X}_n}{n+1}, \frac{1}{n+1}\right)$	$\frac{\mu + n\overline{X}_n}{n+1}$
$\mathcal{E}(\theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(n + \alpha, \beta + n\overline{X}_n)$	$\frac{n+\alpha}{\beta + n\overline{X}_n}$
$\mathcal{B}(\theta)$	$\mathrm{Be}(a, b)$	$\mathrm{Be}(a + n\overline{X}_n, b + n - n\overline{X}_n)$	$\frac{a + n\overline{X}_n}{a + b + n}$
$\mathcal{P}(\theta)$	$\Gamma(1, \beta)$	$\Gamma(1 + n\overline{X}_n, \beta + n)$	$\frac{1 + n\overline{X}_n}{\beta + n}$

Exercice : 1. Dans le cas où la vraisemblance est $\mathcal{N}(\theta, 1)$, déterminer un a priori bien choisi de façon à ne pas avoir la consistance de l'a posteriori en un certain θ.
2. Prouver qu'avec l'a priori $\mathcal{N}(\mu, 1)$, ce modèle est consistant en tout point de $\theta_0 \in \mathbb{R}$

Théorème de Bernstein-von Mises

Hypothèses :

- On considère un modèle $\mathbb{P}_\theta^{\otimes n}$, avec $\theta \in \Theta$ et un élément $\theta_0 \in \Theta$ étant le vrai paramètre. On suppose que ce modèle est dominé (densité par rapport à μ), différentiable en moyenne quadratique en θ_0 et l'information de Fisher $I(\theta_0)$ est inversible en θ_0.
- Soit Π un a priori possédant une densité par rapport à la mesure de Lebesgue et tel que : $\pi(\theta_0) > 0$ et π continue en θ_0.

Théorème

Sous les hypothèse ci-dessus, quand n tend vers $+\infty$, on

$$\left\| \Pi_n - \mathcal{N}(\hat{\theta}_n, \frac{I(\theta_0)^{-1}}{n}) \right\|_1 \to 0, \text{ en probabilité sous } \mathcal{P}_{\theta_0},$$

avec $\hat{\theta}_n$, l'estimateur du maximum de vraisemblance.

On rappelle que, sous des conditions de régularité, on a pour le maximum de vraisemblance le théorème asymptotique suivant :

$$\sqrt{n}(\hat{\theta}_n - \theta) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, I(\theta)^{-1}),$$

Le théorème de Bernstein-von Mises permet de rapprocher le comportement entre les lois limites bayésiennes et le maximum de vraisemblance.

Comportement asymptotique des régions de crédibilité

Contexte : La fonction de répartition a posteriori F_n de θ est considérée strictement croissante. Soit a_n et b_n définis par : $F_n(a_n) = \alpha/2$ et $F_n(b_n) = 1 - \alpha/2$, $\alpha \in [0, 1]$ de telle sorte que $[a_n, b_n]$ est l'intervalle de crédibilité de niveau α.

Théorème

Soit $q_{1-\alpha/2}$ le quantile de niveau $1 - \alpha/2$ d'une loi normale standard. Sous les conditions du théorème de Bernstein-von Mises, on a :

$$[a_n, b_n] = \left[\hat{\theta}_n \pm \frac{q_{1-\alpha/2}}{\sqrt{nI(\theta_0)}}(1 + o_{\mathbb{P}}(1)) \right],$$

où $\hat{\theta}_n$ est l'estimateur du maximum de vraisemblance.
On déduit de cela que $[a_n, b_n]$ est un intervalle de confiance asymptotique de niveau α : $\mathbb{P}_{\theta_0}[\theta_0 \in [a_n, b_n]] \underset{n \to +\infty}{\to} 1 - \alpha$.

Lemme : Soient P et Q deux mesures de probabilité sur l'espace mesurable $(X, \mathcal{A})$. On a :

$$\|P - Q\|_1 = 2 \sup_{A \in \mathcal{A}} |P(A) - Q(A)|.$$

Plan

Estimateurs de Bayes

Cadre bayésien et estimateurs : $\theta \sim \Pi$ et $X \sim \mathbb{P}_\theta$, le modèle possède une densité p_θ sur un espace E. On considère les estimateurs T, des applications mesurables de E dans Θ.

Fonction de perte et de risque :

- Une fonction de perte $l : \Theta \times \Theta \to \mathbb{R}^+$ mesurable vérifiant, $\forall \theta, \theta' \in \Theta$:
$\theta = \theta' \Leftrightarrow l(\theta, \theta') = 0$

Ex : la perte quadratique $l(\theta, \theta') = (\theta - \theta')^2$.

-La fonction de risque d'un estimateur $T(X)$ est une fonction $R : \Theta \to \mathbb{R}^+$ telle que :

$$R(\theta, T(X)) = \mathbb{E}_\theta(l(\theta, T(X))) = \int l(\theta, T(x)) \mathrm{d}\mathbb{P}_\theta(x).$$

• Le risque de Bayes pour un estimateur T et un a priori Π est :

$$R_b(\Pi, T) = \int R(\theta, T(X)) \, \mathrm{d}\Pi(\theta).$$

Un estimateur T_0 est dit de Bayes pour la loi a priori Π si il minimise le risque de Bayes parmi tous les estimateurs. L'infimum est appelé le risque de Bayes de Π :

$$R_b(\Pi) = R_b(\Pi, T_0) = \inf_T R_b(\Pi, T).$$

Cadre des tests bayésiens

Soit Θ_0 et Θ_1 deux sous-ensembles de Θ. Comme dans le cadre fréquentiste, on cherche à tester :

$$H_0 : \theta \in \Theta_0 \quad \text{contre} \quad H_1 : \theta \in \Theta_1.$$

Soit T le test valant 0 ou 1 pour recommander H_0 ou H_1, resp. La statistique de test T est une fonction mesurable de X_1^n.
La fonction de perte pondérée "0-1" est :

$$l(\theta, T) = \alpha_0 \mathbb{1}_{\{T=1, \theta \in \Theta_0\}} + \alpha_1 \mathbb{1}_{\{T=0, \theta \in \Theta_1\}},$$

avec $\alpha_i > 0$ Si $\alpha_0 = \alpha_1$, alors l'estimateur de Bayes de l est le même pour toute valeur α_0. Seul le ratio $\alpha_0/\alpha_1 = 1$ compte. Dans ce cas, l'estimateur-test de Bayes est :

$$T = \mathbb{1}_{\{\Pi_n(\theta_0) \leq \Pi_n(\theta_1)\}} = \mathbb{1}_{\left\{\frac{\Pi_n(\theta_0)}{\Pi_n(\theta_1)} \leq 1\right\}}$$

Exercice 1 : Déterminer la forme du test de Bayes pour des valeurs quelconques de α_0 et α_1.

Exercice 2 : On cherche à savoir si le paramètre générant un échantillon iid X_1^n de loi de Bernoulli $\mathcal{B}(\theta)$ excède 25%. On considère un a priori uniforme $\Pi \sim \mathcal{U}_{[0,1]}$ et une une fonction de perte pondérée "0-1" quelconque. Exprimer le test de Bayes pour une fonction de perte équilibré ($\alpha_0 = \alpha_1$) à l'aide de la fonction de répartition d'une loi Beta et de statistique de l'échantillon.

Facteur de Bayes

Le facteur de Bayes est une façon d'identifier le poids de l'a priori dans la décision.

$$B_n = \frac{\Pi_n(\theta_0)}{\Pi_n(\theta_1)} \frac{\Pi(\theta_1)}{\Pi(\theta_0)}.$$

B_n	Strength of evidence
$< 10^0$	Negative (support H_1)
10^0 to $10^{1/2} \approx 3.2$	Barely worth mentioning
$10^{1/2}$ to 10^1	Substancial
10^1 to $10^{3/2}$	Strong
$10^{3/2}$ to 10^2	Very strong
$> 10^2$	Decisive

TABLE – Echelle d'interprétation du facteur de Bayes d'après Jeffreys.

Il est le rapport des vraisemblances moyennes conditionnellement aux a priori restreints aux supports des hypothèses, $\Pi_0 = \Pi_{|\Theta_0}$ et $\Pi_1 = \Pi_{|\Theta_1}$:

$$B_n = \frac{\int_{\Theta_0} p_\theta^{\otimes n}(x_1^n)\, \mathrm{d}\Pi_0(\theta)}{\int_{\Theta_1} p_\theta^{\otimes n}(x_1^n)\, \mathrm{d}\Pi_1(\theta)}$$

Que vaut le rapport de Bayes dans le contexte de l'exercice 2 au slide précédent ?

Les tests de Bayes

On considère la fonction de perte pondérée "0-1" :

$$l(\theta, T) = \alpha_0 \mathbb{1}_{\{T=1, \theta \in \Theta_0\}} + \alpha_1 \mathbb{1}_{\{T=0, \theta \in \Theta_1\}},$$

avec $\alpha_i > 0$

Le tests de Bayes T pour cette fonction de perte et un a priori Π est :

$$T(X_1^n) = \mathbb{1}_{\left\{ B_n \leq \frac{\alpha_1 \Pi(\Theta_1)}{\alpha_0 \Pi \Theta_0} \right\}}.$$

Remarque : Si une hypothèse est ponctuelle, par exemple $H_0 : \Theta_0 = \{\theta_0\}$, l'a priori est généralement un mélange de loi continue et de Dirac en θ_0, e.g. :

$$\Pi = p\delta_{\theta_0} + (1-p)\mathcal{L}$$

où $\mathcal{L}$ est une loi de densité l par rapport à la mesure de Lebesgue sur Θ.

Plan

Méthode de Monte Carlo

On cherche à évaluer $I = \int_A f(x)g(x)\mathrm{d}x$, avec $A = \Theta$ (dénominateur de $\Pi_n(\theta)$) ou A un événement quelconque (numérateur de $\Pi_n(A)$). Dans la suite, on omet A.

Méthode de Monte Carlo : Soit g une densité et f une fonction intégrable par rapport à la mesure g. La loi des grand nombres implique :

$$I_n = \sum_{i=1}^{n} f(X_i) \underset{n\to+\infty}{\to} \mathbb{E}_g[f(X)] = I, \ g\text{-ps},$$

avec X_1^n un échantillon iid suivant la loi de densité g.

Remarques : Dans le cadre bayésien, plus la dimension du paramètre θ est élevée, plus cela devrait compliquer l'approximation de l'intégrale multiple et réclamer une exploration plus poussée de l'espace multidimensionnel. Mais pas avec la méthode de Monte-Carlo. D'après le TCL, on a :

$$\sqrt{n}(I_n - I) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Var_g[f(X)]).$$

La qualité de l'approximation dépend seulement de la variance de $f(X)$, où X suit la loi associée à g. Cependant, il faut être capable de tirer sous cette loi efficacement.

Méthode de Monte Carlo par Importance Sampling

Lorsque l'on ne sait pas tirer (facilement) sous la loi associée à g. Il est possible de tirer sous une loi h.
Méthode de Monte Carlo par Importance Sampling : Soit X_1^n un échantillon iid sous la loi associée h. Si fg/h est intégrable par rapport à la mesure h, on a :

$$I_n = \sum_{i=1}^{n} \frac{f(X_i)g(X_i)}{h(X_i)} \underset{n\to+\infty}{\to} \mathbb{E}_h\left[\frac{f(X)g(X)}{h(X)}\right] = \int f(x)g(x)\mathrm{d}x, \; h\text{-ps},$$

L'obtention d'un théorème central limite est possible si f^2g^2/h est intégrable par rapport à la mesure de Lebesgue (Si $f \in L^2(g)$, il suffit que h aient des queues plus lourdes que g).

Exercice : Soit $\theta \sim \mathcal{N}(0,1)$. Proposer deux solutions pour évaluer $\mathbb{P}(\theta > 4)$.

Propriété : Le choix optimal de la densité h pour la Méthode MCIS, est :

$$h_0 = \frac{|f|g}{\int |f|g}.$$

Il s'agit donc de ressembler au plus à la densité proportionnelle à $|f|g$.

Chaîne de Markov homogène

Une chaîne de Markov homogène d'espace d'états $\mathbb{R}$ est une suite $(X_i)_{i\in\mathbb{N}}$ telle que :

$$\mathcal{L}(X_{n+1} \mid X_1^n) = \mathcal{L}(X_{n+1} \mid X_n) = \mathcal{L}(X_1 \mid X_0).$$

La loi est donc caractérisée par la loi de X_0 et le noyau de transition (famille de densité conditionnelle valable pour tous i) :

$$K(x, y) = f_{X_i|X_{i-1}=x}(y).$$

Une loi de densité stationnaire f est une loi telle que si $X_i \sim \mathcal{L}(f)$ alors $X_{i+1} \sim \mathcal{L}(f)$, c'est à dire :

$$f(y) = \int f(x)p(x, y)\mathrm{d}x$$

La condition d'équilibre ponctuelle suivante implique que f est une densité stationnaire : $f(x)p(x, y) = f(y)p(y, x)$ (le démontrer).

Un résultat clé : Sous certaines conditions, si $(X_i)_i$ est un chaîne de Markov de densité stationnaire f alors :

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \underset{n\to+\infty}{\to} \int g(x)f(x)\mathrm{d}x.$$

Monte Carlos Markov Chain (MCMC, 1)

On cherche à évaluer $I = \int f(\theta)\pi_N(\theta)\mathrm{d}\theta$. On utilise des densités conditionnelles $k(y \mid x)$ sous lesquelles on sait générées facilement.

L'algorithme de Metropolis Hastings : Soit $\theta_0 \in \Theta$ quelconque. On génére la suite $(\theta_i)_i$. L'étape n s'effectue comme suit :

1. Générer $Z \sim \mathcal{L}(k(. \mid \theta_{n-1}))$,
2. Calculer $p = \min(\frac{k(\theta_{n-1} \mid Z)\pi_N(Z)}{k(Z \mid \theta_{n-1})\pi_N(\theta_{n-1})}, 1)$.
3. Générer $U \sim \mathcal{U}_{[0,1]}$ et $\theta_n = Z\mathbb{1}_{\{U \leq p\}} + \theta_{n-1}\mathbb{1}_{\{U > p\}}$.

Finalement, un estimateur de I est :

$$I_n = \sum_{i=1}^{n} f(\theta_i).$$

Exercice : Démontrer que π_N est une loi stationnaire de la suite $(\theta_i)_i$ ci-dessus. On utilisera la condition d'équilibre ponctuelle.
Remarques : 1) Dans les faits, la suite θ_i peut mettre du temps à converger vers la suite stationnaire et il préférable d'éliminer les premiers termes, on appelle cette phase le "burning".
2) Dans le calcul du terme de p, il est inutile d'avoir calculer π_N, $\pi \times \mathcal{L}_N$ suffit car les dénominateurs de l'a posteriori π_N s'éliminent.

Monte Carlos Markov Chain (MCMC, 2)

On cherche à générer un paramètre multidimensionnel $\theta = (\theta^1, \theta^2, \ldots, \theta^k) \sim \Pi$ (typiquement dans un modèle hiérarchique). On suppose que l'on sait générer sous les lois $\Pi(\theta^i \mid \theta^1, \ldots, \theta^{i-1}, \theta^{i+1}, \ldots, \theta^k)$, avec $1 \leq i \leq k$.

L'échantillonneur de Gibbs : Soit θ_0 une valeur initiale quelconque. On génére une suite $(\theta_i)_i$. À l'étape n, on procède comme suit :
• Pour tout $i \in \{1, \ldots, k\}$, générer $\theta_n^i \sim \Pi(\theta^i \mid \theta_{n-1}^1, \ldots, \theta_{n-1}^{i-1}, \theta_{n-1}^{i+1}, \ldots, \theta_{n-1}^k)$

La suite $(\theta_i)_i$ est une chaîne de Markov homogène de loi stationnaire Π. un tel algorithme fonctionne si seulement si il est facile de simuler une variable conditionnellement à toutes les autres.

Remarque : Les logiciels `stan`, `JAGS`, ou `Winbugs` propose des implémentations "clé en main" des algorithmes MCMC.