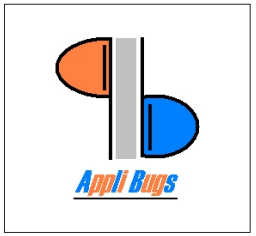
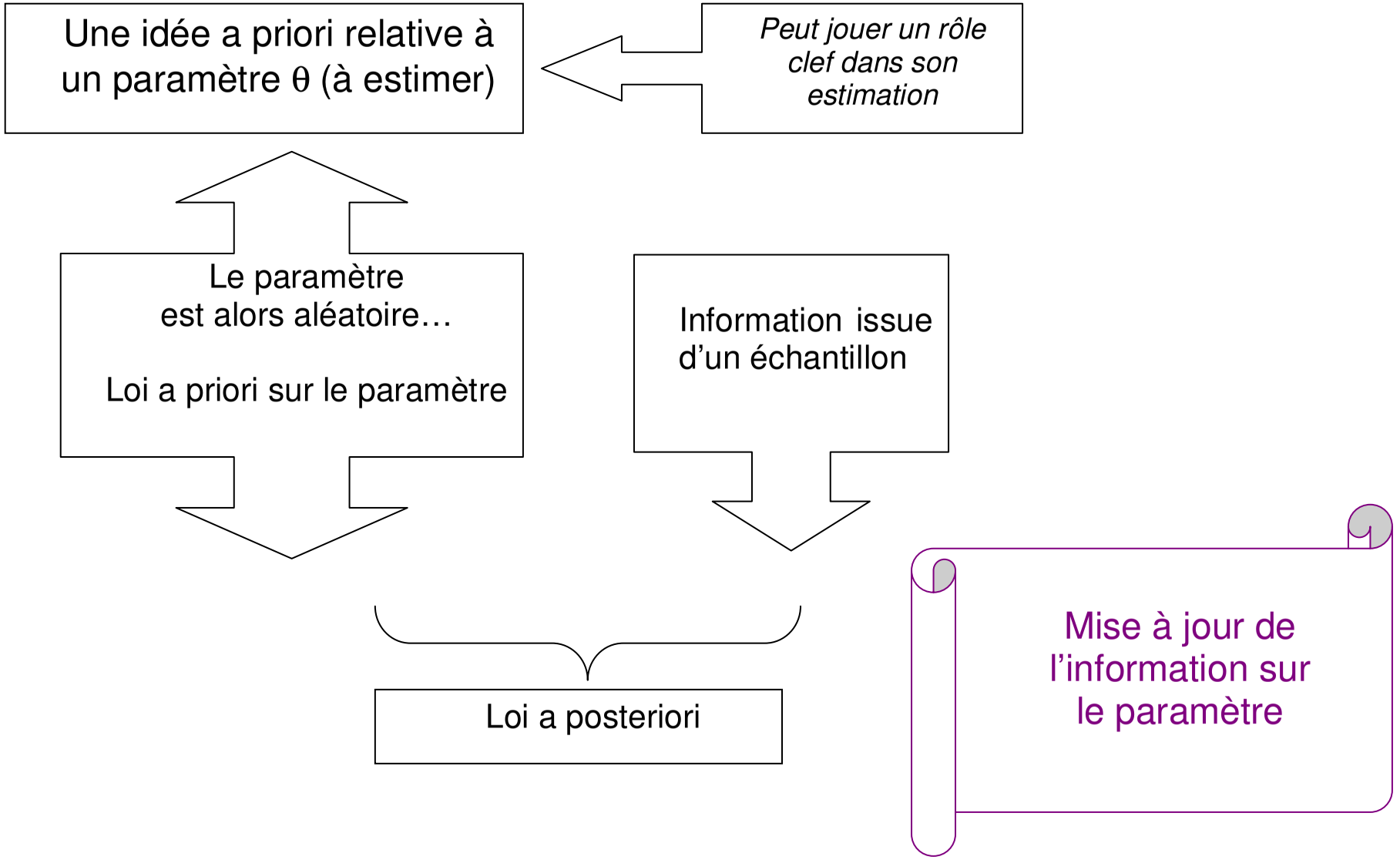


Introduction à la statistique bayésienne

et au logiciel Winbugs

APPROCHE BAYESIENNE



La loi *a posteriori* est le résultat de la combinaison entre la loi *a priori* et l'information des données

Pourquoi faire de la modélisation bayésienne ?

Plusieurs catégories de réponses

- Information a priori : c'est fait pour !
- volonté de donner des résultats conditionnellement à l'observation …
- vraisemblance « complexe»
- …

Modélisation fine

- quantités cachées : modèle de Markov caché, modèle de mélanges
- spécificité individuelle, corrélation : modèle mixte (gaussien, logistique, Poisson …)
- dimension inconnue : modèle de mélanges (nbre de composantes inconnu)
- modèles hiérarchiques : décomposition conditionnelle de la variabilité

Utilisation de plus en plus fréquente des modélisations bayésiennes ?

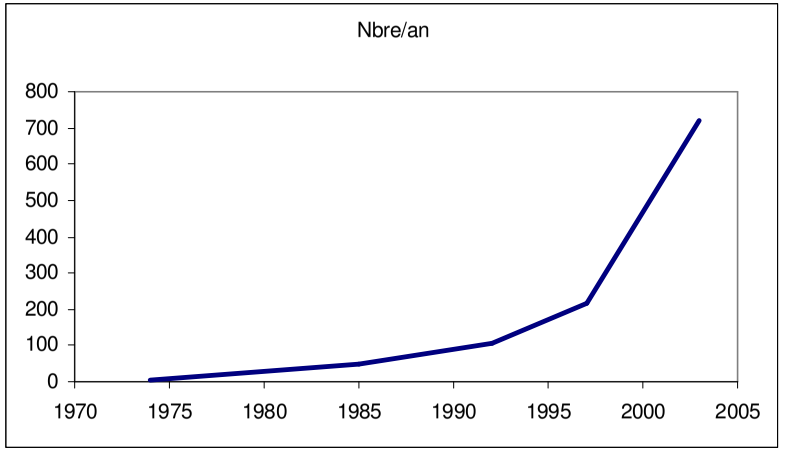
Oui, par exemple, 2 indices :

Enseignement

Publications

Par exemple, sous PubMed, avec Bayesian comme mot clé

période	Nbre	Nbre/an
1968-80	80	6,15
1981-1990	474	47,40
1991-1995	532	106,40
1995-2000	1080	216,00
2001-2006	4315	719,17



Pourquoi surtout depuis les années 90 ?

Modélisation fine (déjà dit …)

Problèmes calculatoires : algorithmes stochastiques MCMC

Logiciels (Winbugs 1989)

Avantages et désavantages des logiciels ?

Avantages

Convivialité

Généralité

Accessibles à tous

Déjà testés

Mise à jour ….

Désavantages

Lenteur

Limitation dans les modèles …

La démarche générale

et

pourquoi des problèmes calculatoires ?

La démarche générale

De la loi a priori vers la loi a posteriori

Soit θ le paramètre dont la loi *a priori* est notée π :

$$[\,\theta\,] \sim \pi(\theta)$$

Soit X les données dont la loi conditionnelle est notée : $[X \mid \theta] = V$ (vraisemblance).

(a) **La loi jointe** de (θ, X) : $[\,\theta, X\,] = [\,\theta\,][\,X \mid \theta\,] = \pi(\theta)\, V$

(b) **La loi marginale** de X est : $m(x) = \int [\theta, X]\, d\theta = \int [\theta]\, [X \mid \theta]\, d\theta$

Donc calcul d'une intégrale multidimensionnelle

$\Rightarrow$ **La loi *a posteriori*** de θ est : $[\,\theta \mid X\,] = [\,\theta\,][\,X \mid \theta\,] / m(x)$

$$\propto [\,\theta\,][\,X \mid \theta\,]$$

Idée de mise à jour

Par exemple,

On observe des premières données X_1,

on obtient la loi a posteriori $[\theta \mid X_1] \propto [\theta][X_1 \mid \theta]$

Plus tard, on a à nouveau des données X_2 (que nous supposons indépendantes de X_1)

par l'indépendance, $[X_1, X_2 \mid \theta] = [X_1 \mid \theta][X_2 \mid \theta]$

et donc

$$[\theta \mid X_1, X_2] \propto [\theta][X_1, X_2 \mid \theta] = \underbrace{[\theta][X_1|\theta]}\ \underbrace{[X_2|\theta]}$$

La première loi a posteriori est utilisée comme loi a priori pour la seconde loi a posteriori

Exemple Simple $X \sim \mathcal{B}in(5,\theta)$, θ est la probabilité pour un joueur de Basket de manquer son tir pendant un match où il fait 5 tirs

Observation : 2 tirs ratés parmi les 5.

Pas d'idée a priori $\theta \sim \mathcal{U}[0\ ,\ 1]$ $\Rightarrow$ $[\ \theta\ ] = \mathbf{1}_{[0\ ,\ 1]}(\theta)$

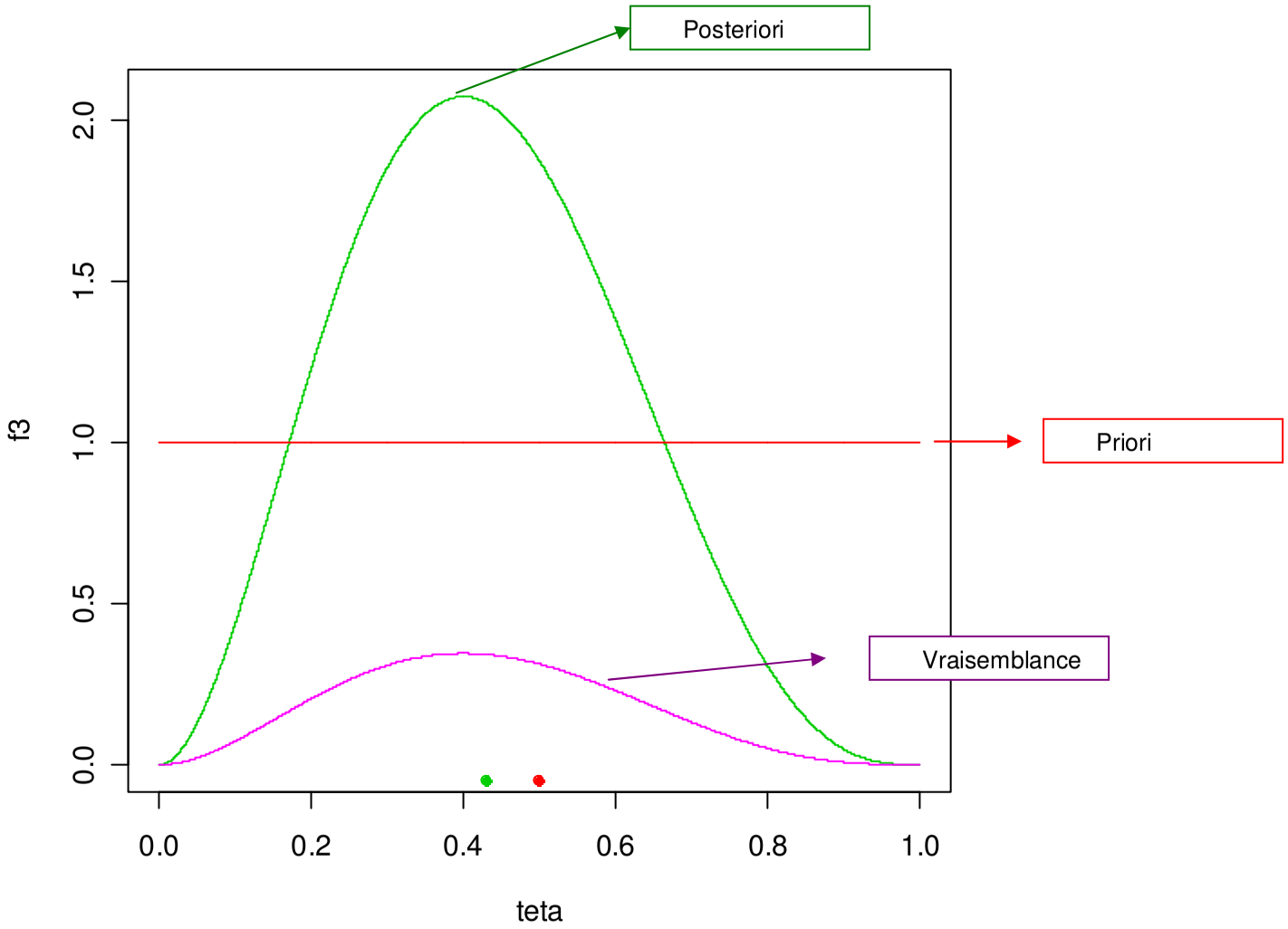
$[\ X \mid \theta\ ] \propto \theta^2\,(1-\theta)^{5-2}$ donc $[\ \theta \mid X\ ] \propto [\ X \mid \theta\ ]\,[\ \theta\ ] = \theta^2\,(1-\theta)^3\ \mathbf{1}_{[0\ ,\ 1]}(\theta)$

$\Rightarrow$ $\theta \mid X \sim \mathcal{B}e(2+1\ ;\ 3+1) = \mathcal{B}e(\alpha\ ;\ \beta)$ $\alpha = 3$ et $\beta = 4$

	A priori	A posteriori
Espérance	0,5	$\alpha / (\alpha+\beta) = 0{,}43$
Variance	0,08	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0{,}03$
Valeur modale	-	$\frac{\alpha-1}{\alpha+\beta-2} = 0{,}4$

Rappel : $Z \sim \mathcal{B}e\ (\alpha,\ \beta)$ $\alpha>0$ $\beta>0$ alors $f(z) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha-1}(1-z)^{\beta-1}$ si $z \in\ ]\,0;1\,[$

(remarque : la loi uniforme sur [0,1] est la loi Beta(1,1))


Posteriori
Priori
Vraisemblance
f3
teta
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0

Une information a priori sur ce joueur est disponible (suivi de quelques matchs sans détail) où le mode et l'espérance de θ étaient de 0.3 et 0.335 respectivement et la variance de 0.03

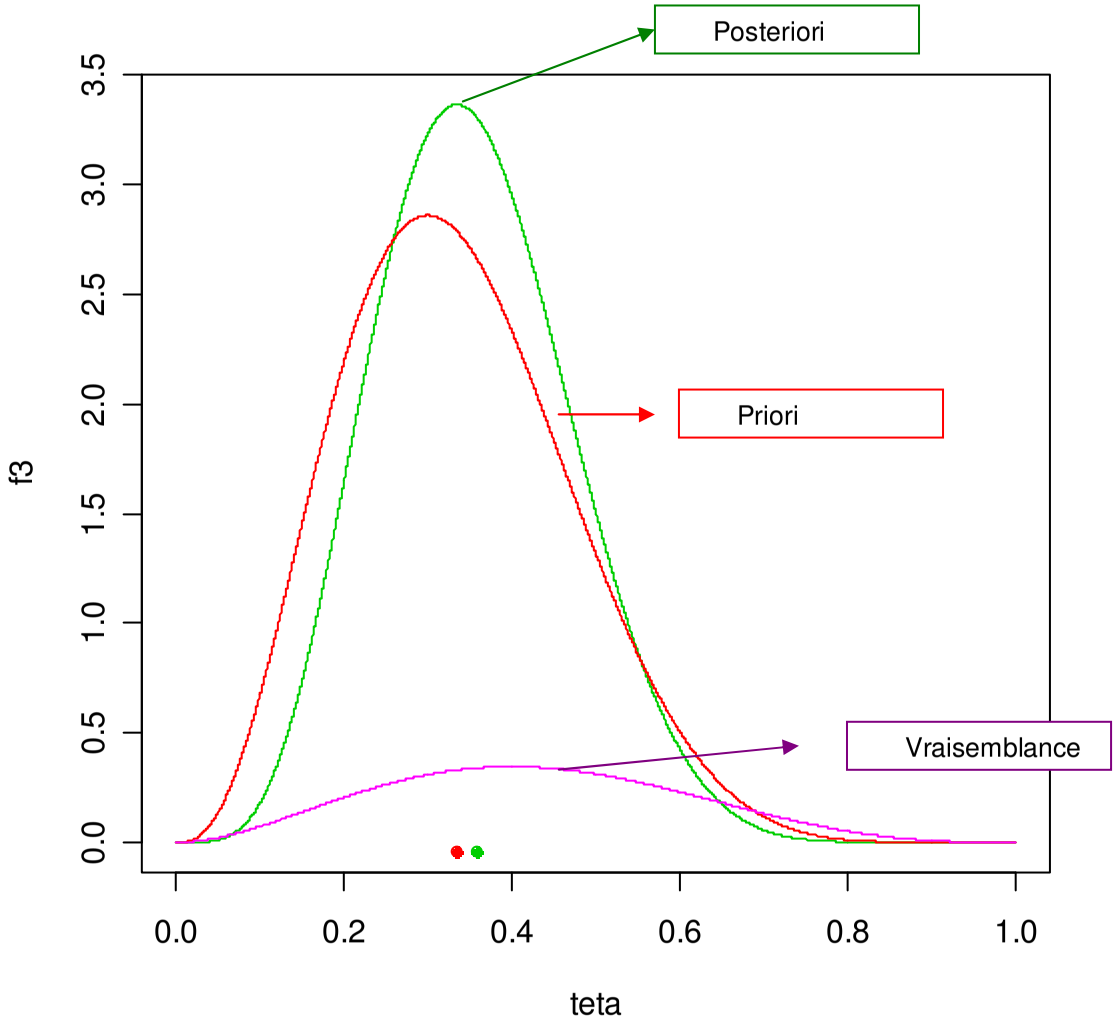
Quelle loi a priori choisir ?

Par exemple, une loi Beta qui respecte le **mode** et la **moyenne**

$\theta \sim \mathcal{B}e(3{,}83\ ;\ 7{,}6) \quad \Rightarrow \quad [\theta] \propto \theta^{2{,}83}\,(1-\theta)^{6{,}6}$

$[X \mid \theta] \propto \theta^{2}\,(1-\theta)^{5-2}$ donc $[\theta \mid X] \propto \theta^{2+2{,}83}\,(1-\theta)^{6{,}6+3} \quad \Rightarrow \quad \theta \mid X \sim \mathcal{B}e(5{,}83\ ;\ 10{,}6)$

	A priori	A posteriori
Espérance	0,335	0,36
Variance	0,02	0,013
Valeur modale	0,3	0,34

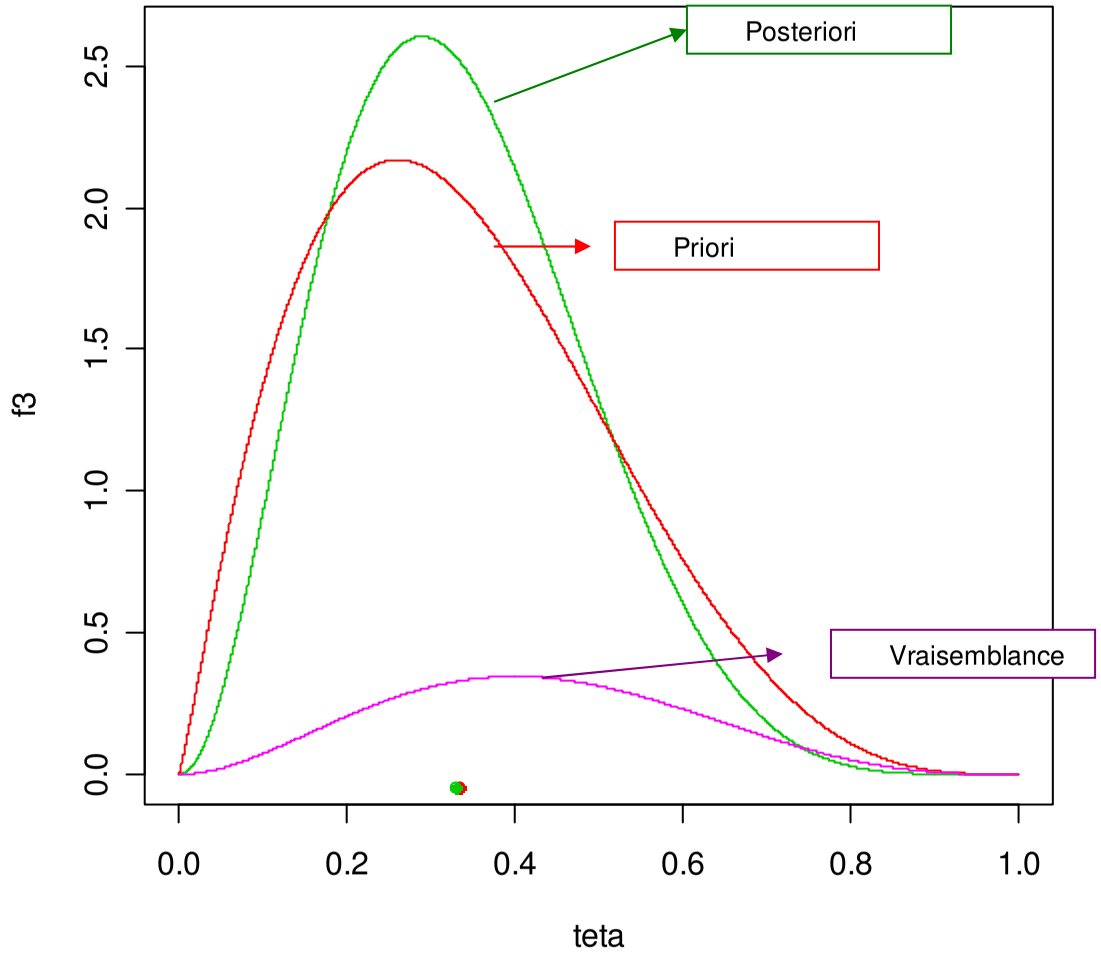

Posteriori
Priori
Vraisemblance
f3
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
teta

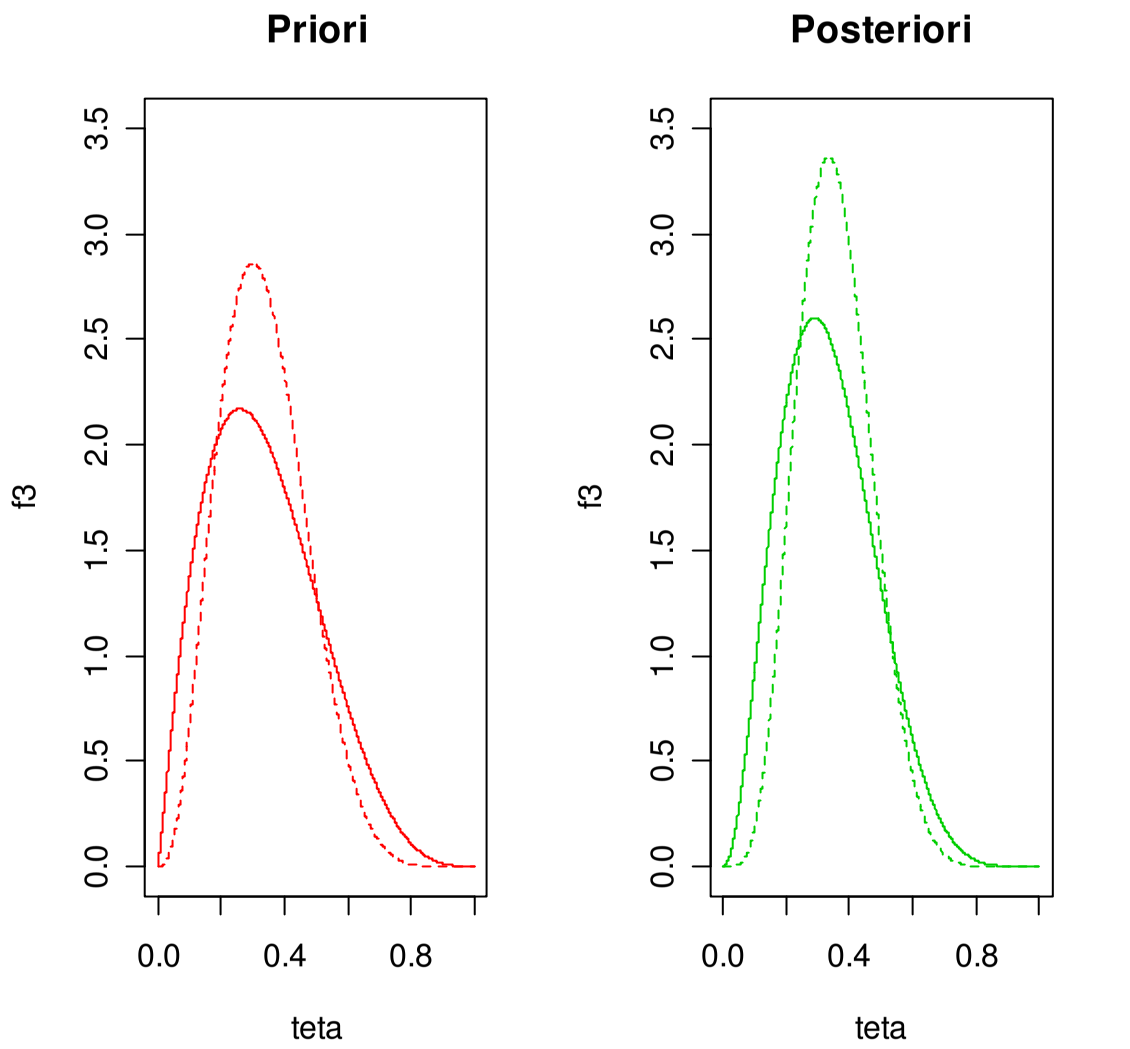
Autre choix de loi a priori ? Une loi Beta qui respecte la **moyenne** et la **variance**

$$\theta \sim \mathcal{B}e(2{,}15\ ;\ 4{,}27) \quad \Rightarrow \quad [\theta] \propto \theta^{1,15}\,(1-\theta)^{3,27}$$

$$[\,X \mid \theta\,] \propto \theta^{2}\,(1-\theta)^{5-2} \text{ donc } \quad [\,\theta \mid X\,] \propto \theta^{2+1,15}\,(1-\theta)^{3+3,27} \quad \Rightarrow \quad \theta \mid X \sim \mathcal{B}e(3{,}15\ ;\ 6{,}27)$$

	A priori	A posteriori
Espérance	0,335	0,33
Variance	0,03	0,021
Valeur modale	0,26	0,29


Posteriori
Priori
Vraisemblance
f3
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
teta


Priori
Posteriori
f3
f3
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.4
0.8
teta
teta

Choix des lois a priori ????

Si information a priori sur le paramètre

=> loi a priori contient cette information (par exemple sur le support)

Comment passer d'une information en loi a priori ?

Si pas d'information a priori sur le paramètre

=> Lois a priori peu informatives ??? On y reviendra …

Différentes catégories de lois a priori peuvent être définies :

Lois conjuguées (famille déterminée par la vraisemblance)

Loi a priori et loi a posteriori
de la même famille
Pratique car évite le calcul de m(x) mais est-ce utile aujourd'hui ?

Lois uniformes pas toujours une bonne solution

Lois impropres souvent limites de « lois propres » comme $\mathcal{U}$[-A,+A] et A $\rightarrow+\infty$

$$\int f(x)dx=\infty$$

Lois a priori conjuguées en θ usuelles

$V = f(x \mid \theta)$	$\pi(\theta)$	$\pi(\theta \mid x)$
Normale $\mathcal{N}(\theta , \sigma^2)$	Normale $\mathcal{N}(\mu,\tau^2)$	$\mathcal{N}(\rho(\sigma^2\mu+\tau^2\Sigma x_i),\rho\sigma^2\tau^2)$ avec $\rho^{-1} = \sigma^2 + (m\,\tau^2)$
Normale $\mathcal{N}(\mu , 1/\theta)$	Gamma $\mathcal{G}(\alpha , \beta)$	$\mathcal{G}(\alpha + m/2 ; \beta + \Sigma (\mu - x_i)^2/2)$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha , \beta)$	$\mathcal{G}(\alpha + \Sigma x_i ; \beta + m)$
Gamma $\mathcal{G}(\mu , \theta)$	Gamma $\mathcal{G}(\alpha , \beta)$	$\mathcal{G}(\alpha + m\,\mu ; \beta + \Sigma x_i)$
Binomiale $\mathcal{B}(n,\theta)$	Beta $\mathcal{B}e(\alpha , \beta)$	$\mathcal{B}e(\alpha + \Sigma x_i ; \beta + nm - \Sigma x_i)$

Rem : on observe ici un m-échantillon $(x_1, \ldots, x_m)$ de X

Pourquoi la loi uniforme n'est pas toujours une bonne solution ?

Si on veut la loi uniforme comme principe, cela ne respecte pas la **reparamétrisation**

Si $\pi[\theta] = 1$ pour $0 \leq \theta \leq 1$, on *reparamétrise* en $\phi = e^{\theta}$ alors $\pi[\phi] = 1/\phi$ pour $1 \leq \phi \leq e$

ou de manière générale

Si $\pi[\theta] = c$ pour $a \leq \theta \leq b$, on reparamétrise en $\phi = g(\theta)$ alors $\pi[\phi] = (g^{-1}(\theta))'$ pour $g(a) \leq \phi \leq g(b)$

Est-ce un problème ?

Certains disent, « on sait quel paramètre est d'intérêt »
Oui, mais même, si on s'intéresse à θ, on aimerait en déduire son espérance et aussi, son moment d'ordre 2 = espérance de θ^2 !

Autre soucis :
si $X \sim N(\mu ; \sigma^2)$ et $\pi[\mu] \propto 1$, $\pi[\sigma] \propto 1$ alors $\pi[\mu , \sigma |x] \propto \sigma^{-1} \exp(-\sigma^{-2} (x-\mu)^2/2)$

Donc, la loi marginale a posteriori de σ est
$\pi[\sigma |x] \propto \int \sigma^{-1} \exp(-\sigma^{-2} (x-\mu)^2/2) \, d\mu =$ **cste**

Une famille de lois peu informatives :

les lois de Jeffreys

définies par La distribution a priori correspondante $\pi(\theta) \propto [I(\theta)]^{1/2}$

où $I(\theta)$ est l'information de Fisher (cas uni-dimensionnel) :

$$I(\theta) = E_\theta \left(\frac{\partial \log [f(x|\theta)]}{\partial \theta} \right)^2$$

ou (sous certaines conditions),

$$I(\theta) = - E_\theta \left(\frac{\partial^2 \log [f(x|\theta)]}{\partial \theta^2} \right)$$

Remarque :

Si θ est multidimensionnel alors la loi a priori de Jeffreys est :

$$\pi(\theta) \; \alpha \; (\mathbf{Det}(I(\theta)))^{1/2}$$

Exemple sur Beta-Binomiale : $X \sim \mathcal{B}(n,\theta)$

$$V(\theta) = [\, x \mid \theta \,] = \binom{n}{x} \theta^{x} (1-\theta)^{n-x} \quad \Rightarrow \quad LV(\theta) = \text{cste} + x \log(\theta) + (n-x) \log(1-\theta)$$

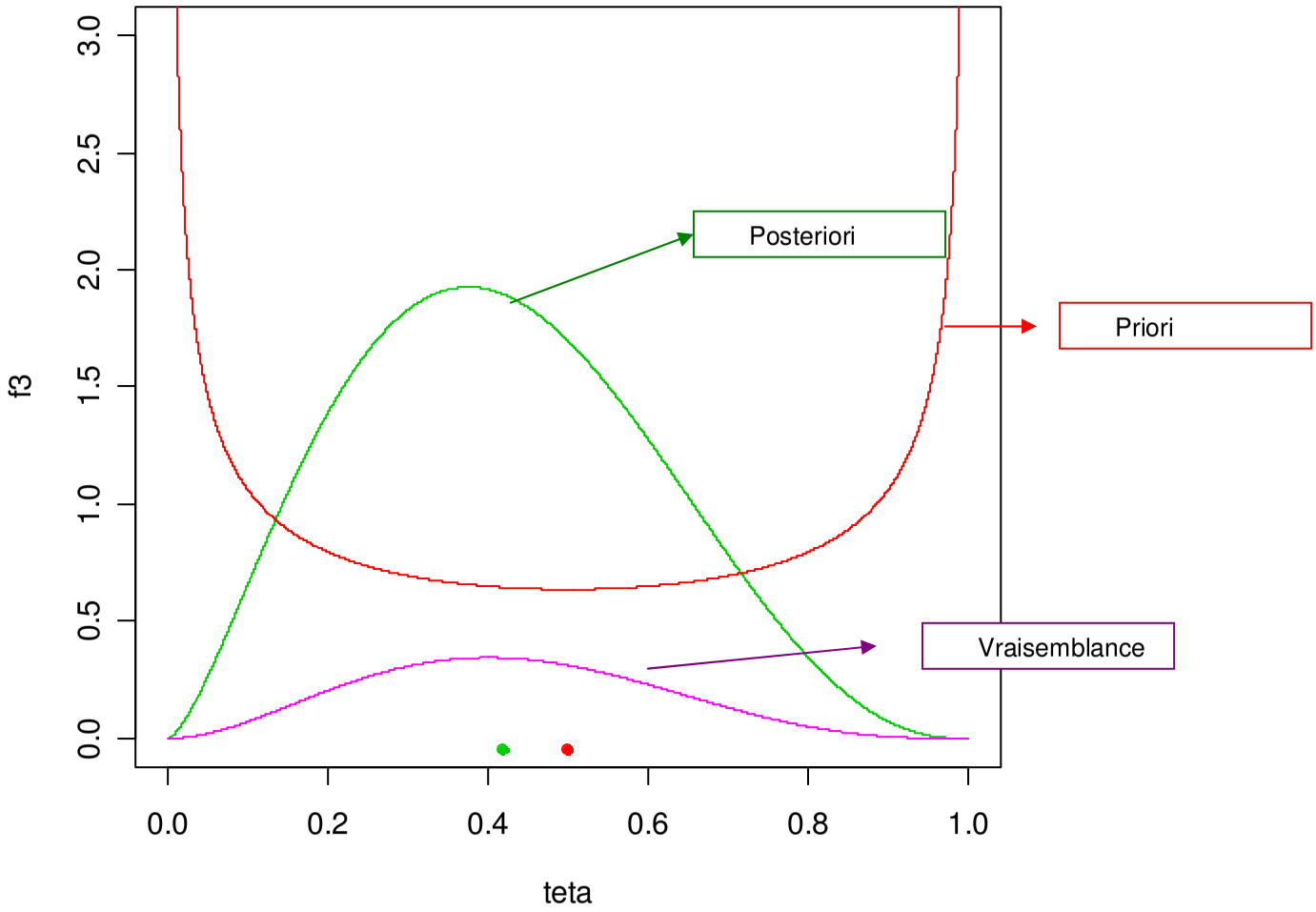
$$\frac{\partial^2 LV(\theta)}{\partial \theta^2} = -\left(\frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2} \right) \quad \Rightarrow \quad I(\theta) = -E_\theta \left(\frac{\partial^2 LV(\theta)}{\partial \theta^2} \right) = n \left(\frac{1}{\theta} + \frac{1}{1-\theta} \right) = \frac{n}{\theta\,(1-\theta)}$$

La loi a priori de Jeffreys pour ce modèle est :

$$\pi(\theta) \propto [\, I(\theta) \,]^{1/2} \propto [\, \theta\,(1-\theta) \,]^{-1/2} \quad \Leftarrow \quad \mathcal{B}e\,(1/2\,,\,1/2)$$

et donc *la loi a posteriori* est (pour n=5 et x=2) :

$$\mathcal{B}e\,(1/2 + 2\,;\; 1/2 + (5 - 2)) = \mathcal{B}e\,(\,2{,}5\,;\,3{,}5\,)$$



Espérance a priori : 0,5 Variance a priori : 0,125

Espérance a posteriori : 0,42 Variance a posteriori : 0,035

Le résultat d'une modélisation bayésienne est donc une loi a posteriori **[θ | X]**

Inférences bayésiennes

Fonction de perte

θ = le paramètre
d = la « décision » suite à l'inférence statistique (par exemple, d peut être directement l'estimation de θ)

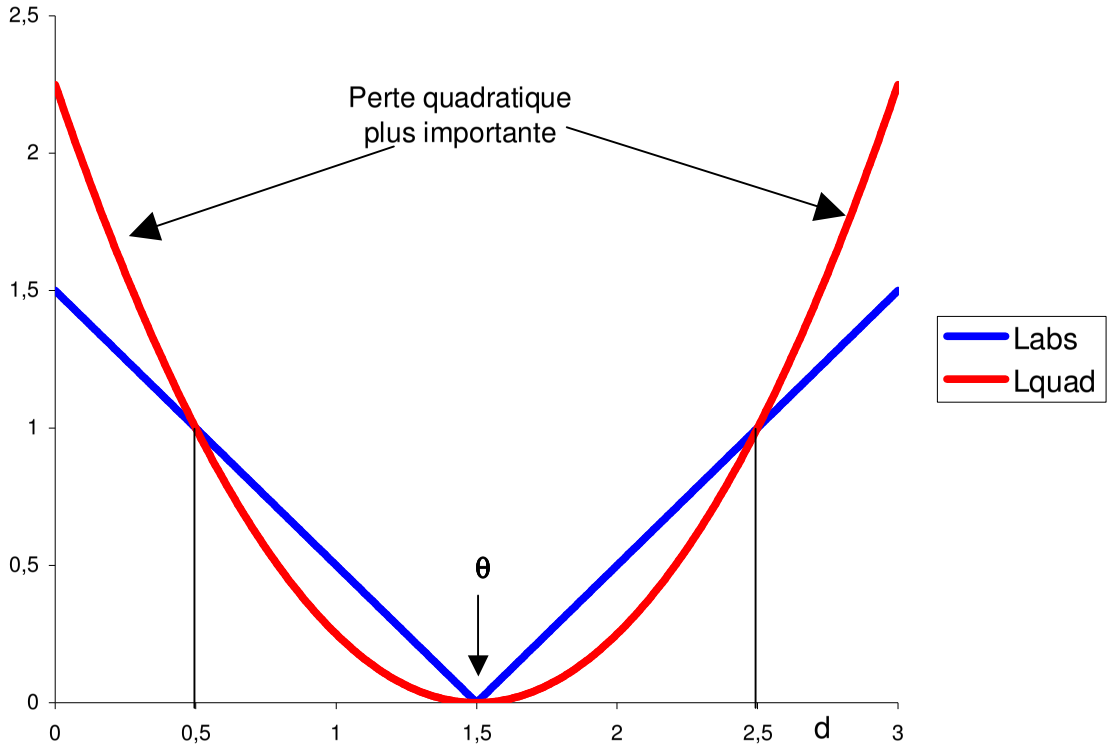
La fonction de perte L(θ,d) *(à valeurs dans [0, +∞))*

⇒ évalue la perte due à la prise de décision d quand le paramètre prend la valeur θ.

- Choix arbitraire de L
  - Exemples de fonctions L(θ,d)

Perte quadratique : $L(\theta,d) = (d-\theta)^2$

Perte absolue : $L(\theta,d) = | d - \theta |$



Perte a posteriori :

La perte a posteriori est

$$\rho_x(\pi,d) = \int_\theta L(\theta,d)\ \pi(\theta|x)\ d\theta\ = E_{[\theta|x]}\ [L(\theta,d)\ |x].$$

Elle correspond à la moyenne a posteriori de la perte (ou de l'erreur) ***conditionnellement à l'observation x*** *selon la loi a posteriori de θ.*

Risque intégré :

On construit le risque intégré

$$\mathbf{r(\pi,d) = \int_x\ \rho_x(\pi,d)\ m(x)\ dx}$$

perte a posteriori moyenne ***selon toutes les observations de x*** *probables (m(x) étant la loi marginale de X).*

Estimateur basé sur la fonction de perte:

Un estimateur bayésien d_L du paramètre θ pour une fonction de perte L est celui qui réalise **le minimum en d de $r(\pi,d)$** et donc, pour chaque donnée x, $d_L(x)$ est celui qui réalise

le minimum en d de $\rho_x(\pi,d)$ ($= \int_\theta L(\theta,d)\, \pi(\theta|x)\, d\theta$)

La valeur de $r(\pi,d_L)$ est appelée le risque bayésien.

On montre que …

1) si la fonction de perte L = perte quadratique — d_L est $E_\pi[\theta|x]$

2) si la fonction de perte L = perte absolue — d_L est Médiane$[\theta|x]$

3) si la fonction de perte L = $1_{|d-\theta|>\varepsilon}$ pour ε "petit" fixé — d_L est Mode de $[\theta|x]$

Par exemple, pour L = perte quadratique , d_L est $E_\pi[\theta|x]$?

On doit minimiser, en d, $A(d) = \int_\theta \mathbf{(d-\theta)^2}\, \pi(\theta|x)\, d\theta$

$$A(d) = d^2 \int_\theta \pi(\theta|x)\, d\theta - 2d \int_\theta \theta\, \pi(\theta|x)\, d\theta + \text{cste}$$

$$= d^2 - 2d\, E_\pi[\theta|x] + \text{cste}$$

$$A'(d) = 2d - 2\, E_\pi[\theta|x]$$

$$\mathbf{A'(d_L) = 0 \iff d_L = E_\pi[\theta|x]} \qquad \mathbf{(et\ A''(d_L) > 0)}$$

Exemple de la Beta-Binomiale :

$X \sim Bin(n,\theta)$ et $[\theta] \sim Beta(\alpha, \beta)$ donc $[\theta|X] \sim Beta(\alpha+x, \beta+n-x)$

on choisit $L(\theta,d) = (d-\theta)^2$

l'estimateur bayésien: $d^*(x) = E_\pi[\theta|x] = (x + \alpha) / (n + \alpha + \beta)$

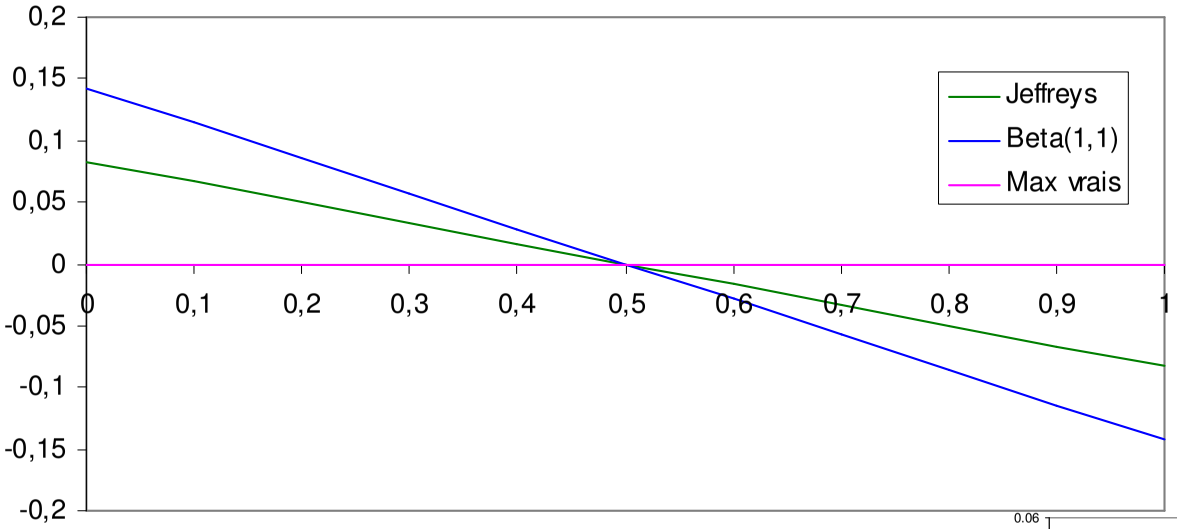
$$\mathbf{r(\pi,d) = \int_x \rho_x(\pi,d)\, m(x)\, dx}$$

Priori	alpha	beta	Risque bayésien (d*)	Risque bayésien (x/n)
Jeffreys	0,5	0,5	0,021	0,025
Uniforme	1	1	0,024	0,033
Informative 1	3,83	7,6	0,012	0,041
Informative 2	2,15	4,27	0,017	0,039

Si on choisit la fonction de perte quadratique, l'estimateur minimise cette erreur

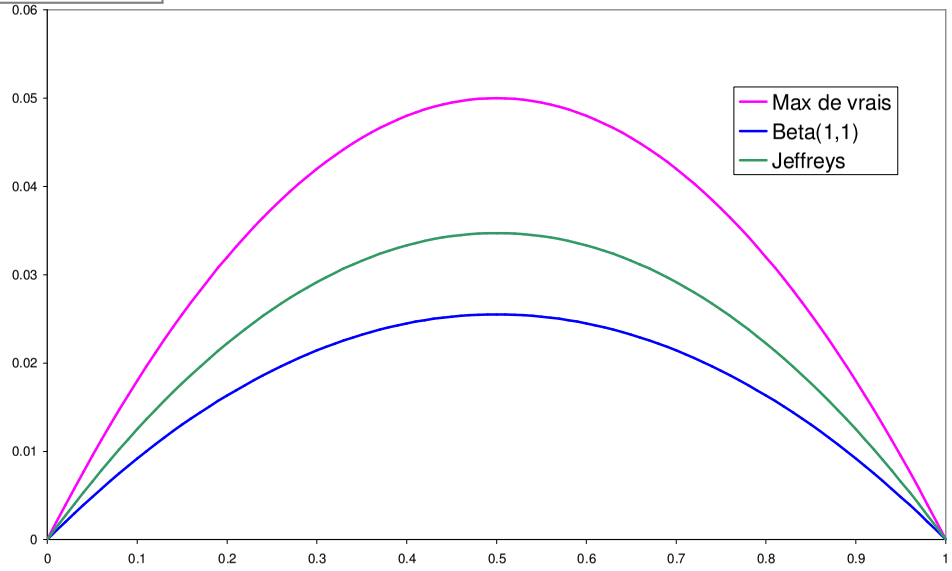
Et le biais alors ?? EQ = VAR + Biais^2

L'estimateur du max de vrais. => Estimateur sans biais (au moins asymp.)
L'estimateur bayésien => Estimateur avec biais



Biais

Variance



Estimateur bayésien par région :

Régions dites "de plus haute densité a posteriori" (en anglais, HPD, Highest Posterior Density) :

$$C_{x,\alpha}^{\pi} = \{ \theta ; [\theta \mid x] \geq k \} \text{ tq } P_{\pi}(\theta \in C_{x,\alpha}^{\pi} \mid x) \geq 1-\alpha$$

(On se fixe une confiance (1-α) et en déduire k)

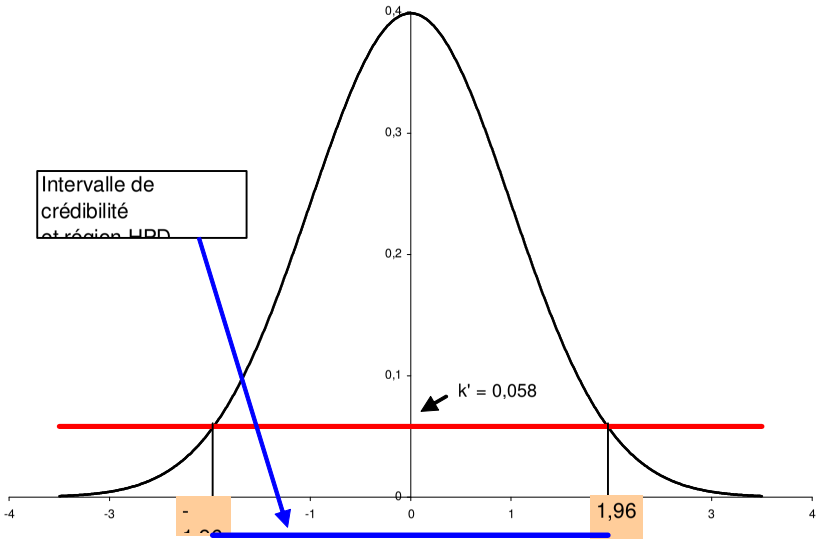
Intervalles de crédibilité au risque α (passage en unidimensionnel)

si $\theta=(\theta_1, ..., \theta_L)$ alors l'intervalle de crédibilité de θ_j est [A,B] tq :

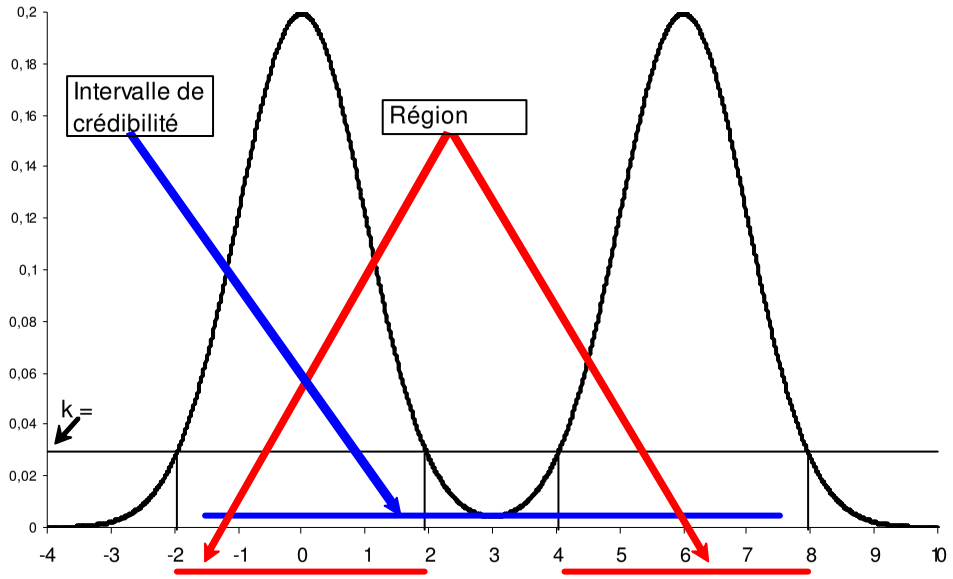
$$P_{[\theta j \mid x]}(A < \theta_j < B) \geq 1-\alpha$$

où $[\theta_j \mid x]$ est la loi marginale en θ_j de $[\theta \mid x]$.

[θ | x] ~ N(0,1)



[θ | x] ~ 0.5 N(0,1) + 0.5 N(6,1)



Pourquoi des problèmes calculatoires ?

Car beaucoup de calculs d'intégrales …

Le résultat d'une modélisation bayésienne est une loi a posteriori multidimensionnelle ! **$[\theta \mid X]$ avec $\theta = (\theta_1, \ldots, \theta_p)$**

- Pour avoir la loi a posteriori, calcul de **$m(x) = \int\int\ldots\int \pi(\theta, X)\, d\theta$**

- Difficile de « raisonner » en multidimensionnel

  Passage à l'unidimensionnel $\Rightarrow$ lois marginales

  $$\pi(\theta_i \mid X) = \int\int\ldots\int \pi(\theta \mid X)\, d\theta_{-i}$$

  avec $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$

- Résumés ponctuels des lois marginales

  $$E_\pi[\, g(\theta) \mid x] = \int_\theta g(\theta)\, \pi(\theta \mid x)\, d\theta$$

- Intervalles de crédibilité à $(1-\alpha)$

$$[A,B] \text{ tq } \int_A^B \pi(\theta_i \mid \mathbf{X})\, d\theta_i \geq (1-\alpha)$$

- Comparaison via le facteur de Bayes

$$H_0 : \theta \in \Theta_{H0} \quad \text{versus} \quad H_1 : \theta \in \Theta_{H1}$$

$$BF = \frac{\int_{\Theta_{H0}} p(X|\theta,H_0)\, \pi(\theta|H_0)\, d\theta}{\int_{\Theta_{H1}} p(X|\theta,H_1)\, \pi(\theta|H_1)\, d\theta} \ \ldots$$

Algorithmes MCMC
Gibbs sampling
et Hasting Metropolis

Avant, pour comprendre pourquoi cela marche, on a besoin d'un peu de Monte Carlo et d'un peu de chaînes de Markov…

Monte Carlo :

Principe : On désire calculer l'espérance de g(θ) sous la loi π(θ |X).

$$\int_{\Theta} g(\theta)\, \pi(\theta|\, x)\, d\theta \quad = \quad E_{\pi}[\, g(\theta)\,]$$

Une approximation consiste à simuler n réalisations $\theta_1, \ldots \theta_n$ selon cette loi π et d'approcher $E_{\pi}[g(\theta)]$ *par leur moyenne empirique* ***(loi des grands nombres)***

Si $\theta_1, \ldots \theta_n \sim \pi(\theta|\, x)$ **iid** **alors** $\lim_{n\to\infty} \Sigma\, g(\theta_i)/n = \int_{\Theta} g(\theta)\, \pi(\theta|\, x)\, d\theta$

approchée par $\overline{g}_n = \Sigma\, g(\theta_i)/n$

On a de plus un contrôle de convergence par :

$\mathrm{Var}(\overline{g}_n) = [\, \int_{\Theta} g^2(\theta)\, \pi(\theta|\, x)\, d\theta \; - E_{\pi}^2[\, g(\theta)\,]\,]$ *approchée par* $\Sigma\, g^2(\theta_i)/n^2 - (\overline{g}_n)^2/n^2$

Problème = on ne sait pas toujours simuler selon une loi multidimensionnelle de manière indépendante ...

Chaîne de Markov :

Une suite (θ_t) de v. a. forme une chaîne de Markov si la loi conditionnelle de θ_t sachant θ_{t-1}, θ_{t-2}, .. est la même que la loi de θ_t sachant θ_{t-1}.

Exemple AR(1) : $\theta_n = \alpha\,\theta_{n-1} + \varepsilon_n$, $\varepsilon_n \sim \mathcal{N}(0,\sigma^2)$ *iid et* $\alpha \in R$ *, alors*

$$[\theta_n \mid \theta_{n-1}, \theta_{n-2}, ..] = [\theta_n \mid \theta_{n-1}] \sim \mathcal{N}(\alpha\,\theta_{n-1},\sigma^2)$$

Définition de la loi invariante :

Une mesure π *est* ***invariante*** *pour le noyau K si* $\pi(B) = \int_\Theta K(\theta,B)\,\pi(\theta)\,d\theta$ *où K est la loi de transition de* θ *vers un ensemble B*

Exemple AR(1) : $K(\theta_{n-1}, \theta) = N(\alpha\,\theta_{n-1}; \sigma^2)$

La chaîne AR(1) admet une loi stationnaire pour $1 > \alpha^2$
La loi π, $N(\mu; \tau^2)$ est stationnaire pour la chaîne AR(1) avec $\mu=0$ et $\tau^2 = \sigma^2/(1-\alpha^2)$

Idée : une CM possède une loi invariante si elle fait communiquer tous les états entre eux et le nombre de visites est infini (notion de récurrence)….

Théorème : ergodicité

Si $\theta_1, .., \theta_n$ est une réalisation d'une chaîne de Markov qui possède une loi invariante π alors

$$1/N \ \Sigma_{n=1, N} \{ g(\theta_n) \} \rightarrow E_\pi [g(\theta)] \qquad ps.$$

De plus, quand la chaîne est réversible (cad K symétrique) alors

$$1/\sqrt{N} \ [\Sigma_{n=1, N} \{ g(\theta_n) \} - E_\pi[g]] \rightarrow N(0, \sigma_h^2) \ \text{(cv en loi).}$$

Ceci permet d'avoir les mêmes propriétés (loi des grands nombres ...) dans le cas markovien que dans le cas indépendant !!!

Ceci est la base des algorithmes MCMC (Monte Carlo par Chaîne de Markov) !

Objectifs

On a k paramètres $\theta=(\theta_1, \ldots \theta_k)$ et on veut simuler un échantillon de θ selon la loi ***multidimensionnelle*** a posteriori $[\theta|x]$ <= loi d'intéret, difficile à simuler

⇨ une itération t de l'algorithme va fournir une réalisation θ^t de la loi $[\theta|x]$ de dimension k, $\theta^t=(\theta_1^t, \ldots, \theta_k^t)$

⇨ l'ensemble des réalisations $\{\theta^t\}$ est une chaîne de Markov (le temps est ici l'itération) *(rem : échantillon avec dépendance markovienne).*

⇨ Si on obtient beaucoup d'itérations, alors on pourra avoir une bonne estimation de la loi a posteriori (ainsi que des caractéristiques des distributions a posteriori marginales de chaque paramètre comme la moyenne, variance, quantiles a posteriori ...), ***ceci d'après le théorème ergodique.***

Les deux algorithmes les plus connus sont les suivants :

⇨ **Hasting Metropolis** : *Méthode générale mais choix d'une loi instrumentale (« proposal ») dont on simule une réalisation puis on garde ou on rejette la valeur proposée selon la loi a posteriori recherchée*

⇨ **Echantillonneur de Gibbs** : *Il faut savoir échantillonner selon les lois conditionnelles complètes des paramètres. Une itération de l'algorithme consiste en la simulation de chaque paramètre selon leur loi conditionnelle complète (sous Winbugs)*

Hasting Metropolis

On dispose d'une loi instrumentale $q(y|\theta)$:

1) on donne des valeurs de départ $(\theta_1^0, \ldots \theta_p^0)$ « cohérentes » avec lois a priori

2) itération t+1 :

tirage d'un candidat y selon la loi $q[y\ |(\theta^t)]$

θ^{t+1} <- y avec la probabilité $\alpha(\theta^t,y)$
θ^{t+1} <- θ^t avec la probabilité 1- $\alpha(\theta^t,y)$

3) si t+1<T retour en 2) pour l'itération t+2

où $\alpha(\theta^t,y) = \min (1 ; \pi[y\ |x]\ q[\theta^t\ |\ y] / (\pi[\theta^t\ |x]\ q[y\ |\theta^t]))$

Est-ce que la loi stationnaire est $\pi(\theta \mid X)$?

oui mais cela dépend du choix de q …

Quelques catégories de lois instrumentales

- Cas indépendant **$q(. \mid \theta^{(t)})$ indépendant de $\theta^{(t)} = q(.)$**

  par ex, $q(y) \sim \mathcal{N}(0,\sigma^2)$ alors attention ! choix de q tq

  $\exists M$ tq $\pi(y|x) \leq M\, q(y)$

  proba moyenne d'acceptation est ≥1/M

- Cas marche aléatoire **$y = \theta^{(t)} + \varepsilon$ où $\varepsilon \sim$f indép. de $\theta^{(t)}$**

  par ex, si $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ alors

  $q(y|\theta^{(t)}) \sim \mathcal{N}(\theta^{(t)},\sigma^2)$ et $q(\theta^{(t)}|y) \sim \mathcal{N}(y,\sigma^2)$ *<= f symétrique*

si f est symétrique alors $\alpha(\theta^t,y) = \min\left(1 ; \dfrac{\pi[y \mid x]}{\pi[\theta^t \mid x]}\right)$

Exemple sur la loi a posteriori est $\pi(\mid X) \sim \mathcal{N}(0,25)$
et choix instrumental est $y = \theta^t + \delta\,\varepsilon$ où $\varepsilon \sim$ Unif[-1 ; 1]

Exemple de déroulement de l'algorithme (avec δ=20) :

A t=100 => $\theta^{(100)} = 2.1$
Que se passe-t-il en t=101 ?

- Simulation de y ~ Unif[2.1-20 ; 2.1+20] = Unif[-17.9 ; 22.1]

  $\Rightarrow$ y=-3

- Probabilité d'acceptation $\alpha(\theta^{100},y) = \min\left(1 ; \exp\left(\frac{(\theta^{100})^2 - y^2}{2}\right)\right) = 10\%$

- Tirage de u ~ Unif[0,1]

  si u <10% alors $\theta^{101}=y=-3$ sinon $\theta^{101}=\theta^{100}=2.1$

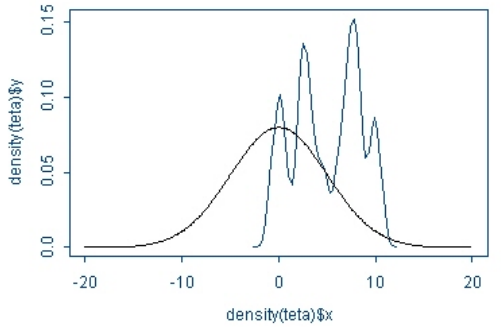
  $\Rightarrow$ u=0.15

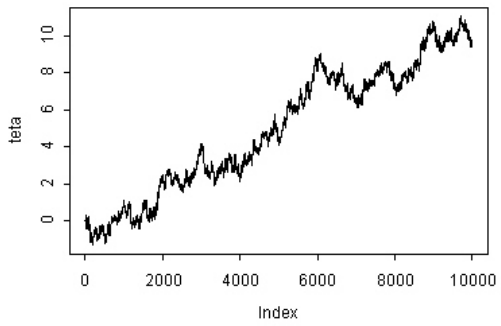
  $\Rightarrow$ $\theta^{101}=2.1$

Exemple sur la loi a posteriori est $\pi(\mid X) \sim \mathcal{N}(0,25)$

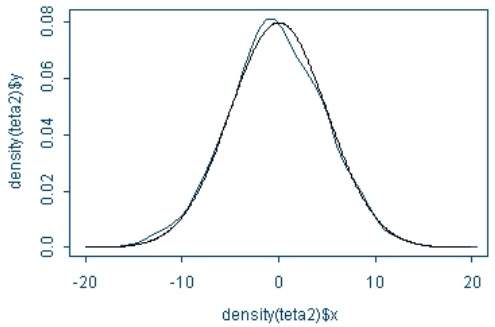
et choix instrumental est $y = \theta^t + \delta\,\varepsilon$ où $\varepsilon \sim$ Unif[-1 ; 1]

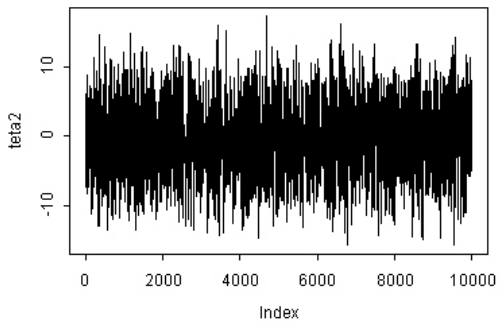
$\delta = 0.1$ (accept = 63%)





$\delta = 20$ (accept = 39%)





Remarque en pratique :

⇨ si le nombre d'acceptations est trop élevé

- Queues de distribution de $\pi(\mid X)$ peu visitées "trop local"

⇨ si le nombre d'acceptations est trop faible

- Mauvaise approximation car trop à "l'extérieur" du support de π

⇨ % d'acceptation de l'ordre de 30% ou 40%

l'échantillonneur de Gibbs

Si on désire faire T simulations, la valeur de T étant fixée :

1) on donne des valeurs de départ $(\theta_1^{(0)}, \ldots \theta_p^{(0)})$ « cohérentes » avec lois a priori

2) itération t+1 :

tirage de $\theta_1^{(t+1)}$ selon la loi de $[. \,|(\theta_2^{(t)}, \ldots, \theta_p^{(t)})]$

tirage de $\theta_2^{(t+1)}$ selon la loi de $[. \,|(\theta_1^{(t+1)}, \theta_3^{(t)} \ldots, \theta_p^{(t)})]$

...

tirage de $\theta_p^{(t+1)}$ selon la loi de $[.|(\theta_1^{(t+1)}, \ldots, \theta_{p-1}^{(t+1)})]$

3) si t+1<T retour en 2) pour l'itération t+2

Exemple La régression cste (*Obs X* : n-échantillon $\{x_i\}$)

$$X_i = \beta + \varepsilon_i \text{ où } \varepsilon \sim N(0,\sigma^2)$$

Lois a priori : $[\sigma^{-2}] \sim \text{Gamma}(a ; b)$ par ex., a=b=0.001 ($E(\sigma^{-2}) = a / b$ et $Var(\sigma^{-2}) = a / b^2$)

$[\beta] \sim N(0 ; \tau^2)$ par ex., τ^2=10000

Loi a posteriori : $\pi[\beta ,\sigma^{-2} \mid D] \propto$ priori . vrais ?

$$\propto (\sigma^{-2})^{a-1}\exp(-b\,\sigma^{-2})\ \exp(-\beta^2\tau^{-2}/2)\ (\sigma^{-2})^{n/2}\exp(-\sigma^{-2}\,\Sigma(x_i-\beta)^2/2)$$

$$= (\sigma^{-2})^{n/2+a-1}\exp(-\sigma^{-2}\{\,b + \Sigma(x_i-\beta)^2/2\,\})\ \exp(-\beta^2\tau^{-2}/2) \quad \text{ou}$$

$$= (\sigma^{-2})^{n/2+a-1}\exp(-b\,\sigma^{-2})\exp(-0.5\{\,\sigma^{-2}\Sigma x_i^2 - 2\,\beta\,\sigma^{-2}\Sigma x_i + \beta^2\,(n\sigma^{-2}+\tau^{-2})\,\})$$

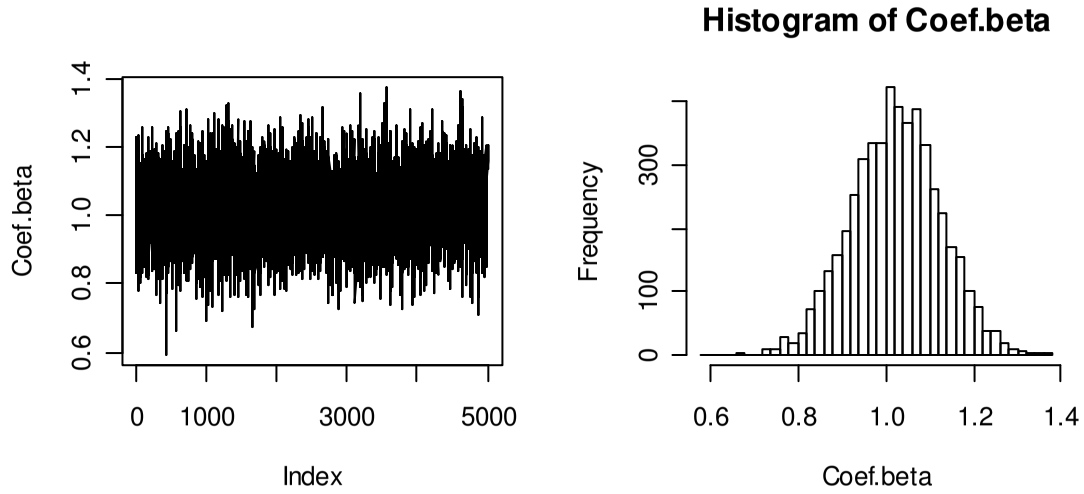
Algorithme :Itération t -> t+1

$$(\sigma^{-2})^{(t+1)} \sim \text{Gamma}(n/2+a\ ,\ b+\Sigma(x_i-\beta^{(t)})^2/2)$$

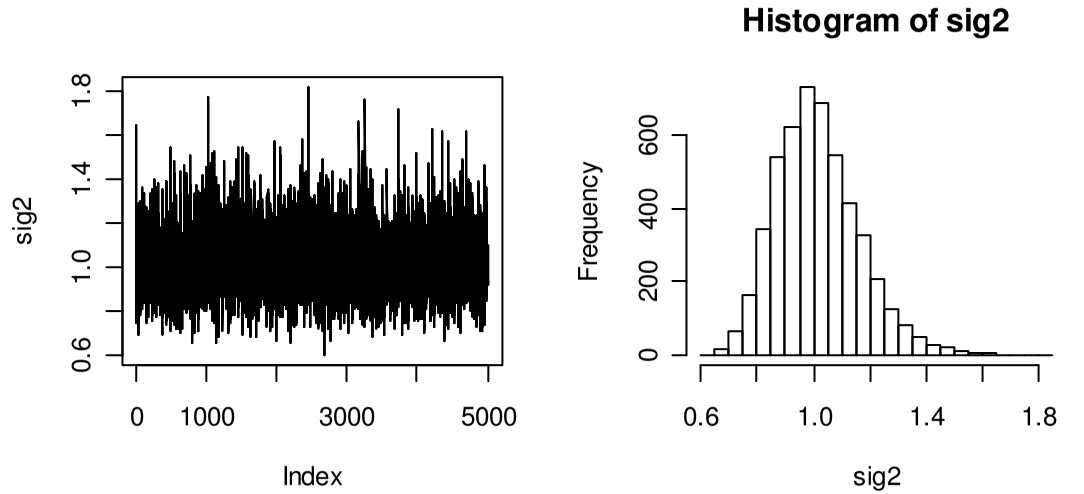
$$\beta^{(t+1)} \sim \mathcal{N}((\sigma^{-2})^{(t+1)}\Sigma(x_i)/\{n(\sigma^{-2})^{(t+1)}+\tau^{-2}\};[n(\sigma^{-2})^{(t+1)}+\tau^{-2}]^{-1})$$

	Moyenne a posteriori	Ecart type a posteriori	IC(95%)
β	1.02	0.10	[0.83 ; 1.22]
σ^2	1.02	0.15	[0.77 ; 1.35]

β



σ^2



Exemple sous WinBugs … B*ayesien inference* ***U****sing* ***G****ibbs* ***S****ampling*

- *Les développeurs* sont : “MRC Biostatistics, Cambridge” et “Imperial College School of Medicine at St Mary's, London”.

- *Compatibilité* avec R, S+, stata…

Freeware : à télécharger : http://www.mrc-bsu.cam.ac.uk/bugs/
(une demande de clé…)

- *Des modules additionnels* : GEOBUGS, CODA, WBDIFF…

- *Des versions sans clé* : OpenBugs, BRugs

Et d’autres logiciels …

BACC : Bayesian Analysis, Computation and Communication (2003, économétrie)
BATS : Pour les séries chronologiques (pas de MAJ depuis 1997)
BayesX : Pour régression et modèle de survie, modèle mixte
BIPS : Bayesian Inference for the Physical Sciences
JAGS : Just Another Gibbs Sampler (sous unix) **…**

Problèmes et caractéristiques :

- ⇨ Méthodes itératives (peut être long...)
- ⇨ Savoir simuler des lois : Gibbs + parfois Metropolis (marche aléatoire Normale) …
- ⇨ Choix du nombre d'itérations : ce nombre doit être assez grand pour l'oubli des points de départ (burn in ou temps de chauffe) et la bonne couverture de la loi a posteriori.

  En pratique, on se fixe une valeur M itérations (temps de chauffe -> loi stationnaire) et T itérations (T très grand) afin d'obtenir T réalisations de la chaîne de Markov sous la loi a posteriori recherchée

- ⇨ L'algorithme comporte donc (M+T) itérations dont les M premières ne sont pas utilisées dans les inférences.
- ⇨ Diagnostics de convergence (pb encore actuel) ... M et T assez grands ?

Quelques conseils

Voici quelques conseils pratiques, facilement accessibles sous WinBugs.

⇨ Le graphe des réalisations de chaque paramètre en fonction des itérations peut permettre de détecter un problème de lenteur de mélange (« slow mixing ») et dans ce cas, il est indispensable de faire un plus grand nombre d'itérations.

⇨ La vérification de la stabilité (par rapport aux itérations) de quelques statistiques comme la moyenne, la variance et surtout, les quantiles d'ordre α (pour α petit ou grand) est importante car ces statistiques correspondent, en général, aux résumés utilisés pour caractérisées la loi marginale a posteriori de chaque paramètre.

⇨ La comparaison des lois a priori et des lois marginales a posteriori permet de voir si les données ont apporté une information.

⇨ Le graphe des autocorrélations de chaque paramètre par rapport aux itérations permet de détecter une bonne « mélangeance »

⇨ Vérification pour chaque paramètre que la "MC error" est inférieure à 5% de l'écart-type a posteriori du paramètre

⇨ Etude de la sensibilité aux lois a priori

Quelques diagnostics

Les diagnostics de convergence (ou plutôt de « non divergence ») sont nombreux. Deux diagnostics assez populaires et accessibles sous WinBugs (par l'intermédiaire du logiciel CODA qui lui est associé) sont :

⇨ Raftery et Lewis : Il donne des estimations du temps de chauffe M et du nombre d'itérations T. Mais le nombre estimé T est souvent très conservateur, donnant un nombre d'itérations gigantesque !

⇨ Gelman et Rubin : Ce diagnostic nécéssite de lancer plusieurs fois l'algorithme avec des valeurs initiales différentes et donc d'obtenir plusieurs réalisations des chaînes de Markov. Il vérifie que chaque algorithme amène bien à un échantillon provenant de la même loi en comparant les variances intra-chaîne et inter-chaîne (Var totale/Var within proche de 1).

Deviance Information Criterion DIC

"Bayesian Measures of Model Complexity and Fit (with Discussion)", Journal of the Royal Statistical Society, Series B, 2003 64(4):583-616.

$$DIC(M) = \bar{D}(\theta) + 2c$$

où c mesure la complexité (*proche de la notion de nombre de paramètres*)

et $D(\bar{\theta})$ est « l'éloignement » du modèle aux données (*déviance cad –2 logV en la moyenne a posteriori*)

$\Rightarrow$ Une estimation de $c = \bar{D}(M) - D(\bar{\theta})$ où $\bar{D}(M) \approx$ moyenne a posteriori de la déviance (calculée à chaque itération)

Directement accessible sous WinBugs !

Attention :

Le critère DIC suppose que la moyenne a posteriori

est un bon estimateur des paramètres aléatoires

Sous WinBugs

Voyons la Beta-binomiale simple (n=5 et x=2) avec [θ]~Beta(1,1) …*(modèle 0)*

On connaît les résultats analytiques : $\theta \mid X \sim \mathcal{B}e(3 ; 4)$

$E(\theta \mid X)=0.43$, $\sigma^2(\theta \mid X)=0.03$, $\sigma(\theta \mid X)=0.175$, Méd=0.43 et $IC_{95}=[0.12 – 0.78]$

Nous suivons ce joueur pendant 39 matchs (donc 40 matchs au total)

Fichier basket.txt donne les n et x

Regardons 3 modèles :

1) Béta-binomiale avec θ constant ***(modèle 1)***

2) la probabilité θ varie d'un match à l'autre indépendamment ***(modèle 2)***

3) la probabilité du match dépend de celle du match précédent ***(modèle 3)***

BUGS directement sous R …

Au préalable, *Sous R, charger le package BRugs*
Puis taper library(BRugs)

Fonction BrugsFit :

```
BRugsFit(modelFile="nom.txt",data="data.txt",inits="init.txt",numChains=1,para=c("t
heta"),nBurnin=1000,nIter=10000,DIC=TRUE,working.directory="C:/…")
```

Que peut-on récupérer ? Tout. **Par exemple,**

```
# les stats sur  theta          Stat.complet<- samplesStats("theta")
                                Stat.moyenne<- samplesStats("theta")$mean
                                dicStats()

# les valeurs de theta          theta1<-samplesSample("theta")

# les graphes de theta          samplesDensity("theta")
                                samplesHistory("theta")
```

Par exemple, si vos fichiers (programme du modèle 1 (prog-bb40.txt), données (basket.txt) et valeurs initiales (init-bb40.txt)) se trouvent dans le répertoire C:\applibugsCGJ, la commande suivante fait tourner le programme 10000 itérations dont 1000 de chauffe :

```
modbb40<-BRugsFit(modelFile="prog-bb40.txt",data="basket.txt",inits="init-
bb40.txt",numChains=1,para=c("theta","sigma","thetam"),nBurnin=1000,nIter=10000,DIC=TR
UE,working.directory="C:/applibugsCGJ")
```

 # les stats

```
samplesStats(c("theta","sigma","thetam"))        ou   modbb40$Stats
samplesStats("theta")                            ou   modbb40$Stats[2 :41,]
samplesStats("theta")$mean                       ou   modbb40$Stats$mean[2 :41]
```

 # le DIC modbb40$DIC *ou* modbb40$DIC$DIC *ou …*

 # les valeurs de theta theta1<-samplesSample("theta[1]")

 # les graphes de theta samplesDensity("theta[1]",beg=1000,end=10000)

samplesDensity("theta[1 :4]",mfrow=c(2,2))

samplesHistory("theta[1]")

…

Quelques fonctions sous WinBugs …

abs(e)	$\|e\|$
cos(e)	cosinus(e)
equals(e1, e2)	1 if $e1 = e2$; 0 otherwise
exp(e)	exp(e)
inprod(v1, v2)	$\Sigma v_{1i}\, v_{2i}$
inverse(v)	V^{-1} for symmetric positive-definite matrix v
log(e)	ln(e)
logdet(v)	ln(det(v)) for symmetric positive-definite v
logfact(e)	ln(e!)
loggam(e)	ln($\Gamma(e)$)
logit(e)	ln(e / (1-e))
max(e1, e2)	$e1$ if $e1 > e2$; $e2$ otherwise
mean(v[1:n])	moyenne(v)
min(e1, e2)	$e1$ if $e1 < e2$; $e2$ otherwise
phi(e)	standard normal cdf
pow(a, b)	a^b
sin(e)	sinus(e)
sqrt(e)	$e^{1/2}$
rank(v, s)	number of components of v less than or equal to v_s
ranked(v, s)	the s^{th} smallest component of v
round(e)	nearest integer to e
sd(v[1:n])	standard deviation of components of v (n-1 in denominator)
step(e)	1 if $e >= 0$; 0 otherwise
sum(v[1:n])	Σv_i
trunc(e)	greatest integer less than or equal to e