Introduction à l'inférence bayesienne

Pierre Gloaguen

Avril 2020

Rappel des cours précédents

- Méthodes de Monte Carlo pour le calcul d'intégrales
- Echantillonnage préférentiel
- Méthodes de simulations de variables aléatoires

Rappel des cours précédents

- Méthodes de Monte Carlo pour le calcul d'intégrales
- Echantillonnage préférentiel
- Méthodes de simulations de variables aléatoires
- Intérêt statistique?
  - Permet l'approximation de probabilité (prise de décision)
  - Point clé de l'inférence bayésienne

Objectifs du cours

- Présentation du principe de l'inférence bayésienne;
- Deux exemples illustratifs;
- Définition des notions clés;

Objectifs du cours

- Présentation du principe de l'inférence bayésienne;
- Deux exemples illustratifs;
- Définition des notions clés;
- Lien avec le maximum de vraisemblance;
- Lien avec les premiers chapitres du cours;

Exemple introductif

Simple modèle paramètrique

Expérience et question

Supposons que l'on observe n $= 10$ tirages indépendant de pile ou face. On compte 8 observations de pile et 2 de face.

Quelle est la probabilité que la pièce tombe sur pile?

Modélisation

On note $x_1, \ldots, x_{10}$ le résultat du lancer (0 si *face*, 1 si *pile*). On suppose que ces nombres sont les réalisations de 10 V.A. $X_1, \ldots, X_{10}$ i.i.d. de loi $\mathcal{B}ern(\theta)$ où $\theta \in]0, 1[$ est la probabilité d'obtenir pile.

Donc, la loi jointe de $\mathbf{X} = (X_1, \ldots, X_n)$ est donnée par:

$$L(x_1, \ldots, x_n|\theta) = \prod_{k=1}^{n} \mathbb{P}_\theta(X = x_k) = \theta^{\sum_{k=1}^{n} x_k} (1-\theta)^{n-\sum_{k=1}^{n} x_k}$$

où $X \sim \mathcal{B}ern(\theta)$.

Inférence par maximum de vraisemblance

Pour un échantillon $\mathbf{X} = X_1, \ldots, X_n$, et pour un paramètre $\theta \in]0, 1[$, la *vraisemblance* de θ est:

$$L(x_{1:n}|\theta) = \prod_{k=1}^{n} \mathbb{P}_\theta(X = x_k) = \theta^{\sum_{k=1}^{n} x_k} (1-\theta)^{n - \sum_{k=1}^{n} x_k}$$

Inférence par maximum de vraisemblance

Pour un échantillon $\mathbf{X} = X_1, \ldots, X_n$, et pour un paramètre $\theta \in]0, 1[$, la *vraisemblance* de θ est:

$$L(x_{1:n}|\theta) = \prod_{k=1}^{n} \mathbb{P}_\theta(X = x_k) = \theta^{\sum_{k=1}^{n} x_k} (1-\theta)^{n - \sum_{k=1}^{n} x_k}$$

Maximum de vraisemblance

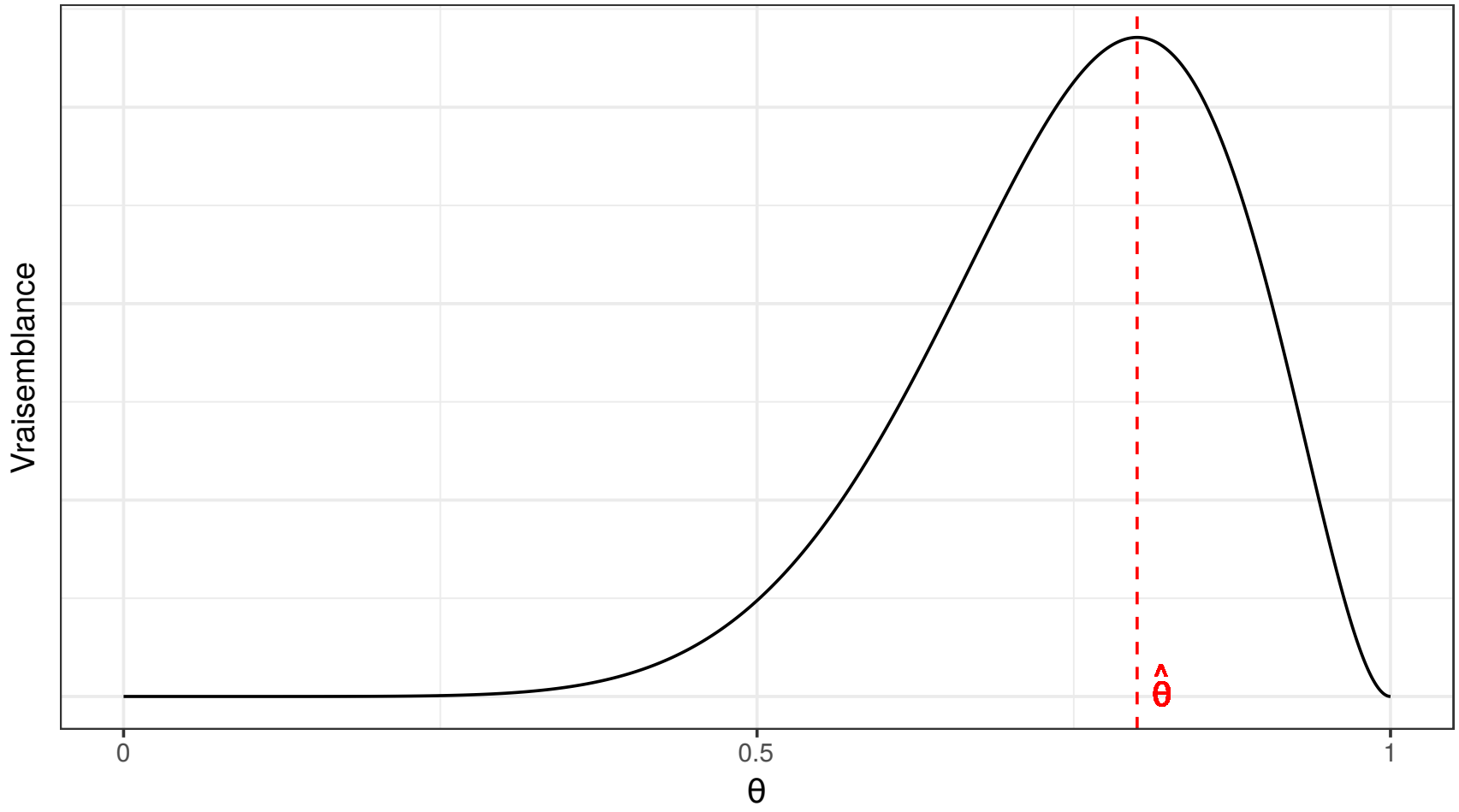
L'estimateur du maximum de vraisemblance pour $x_{1:n}$ est donné par $\hat{\theta} = \text{argmax}_\theta L(x_{1:n}|\theta) = \frac{\sum_{i=1}^{n} x_i}{n}$.

L'estimateur est **entièrement basé sur les données**.

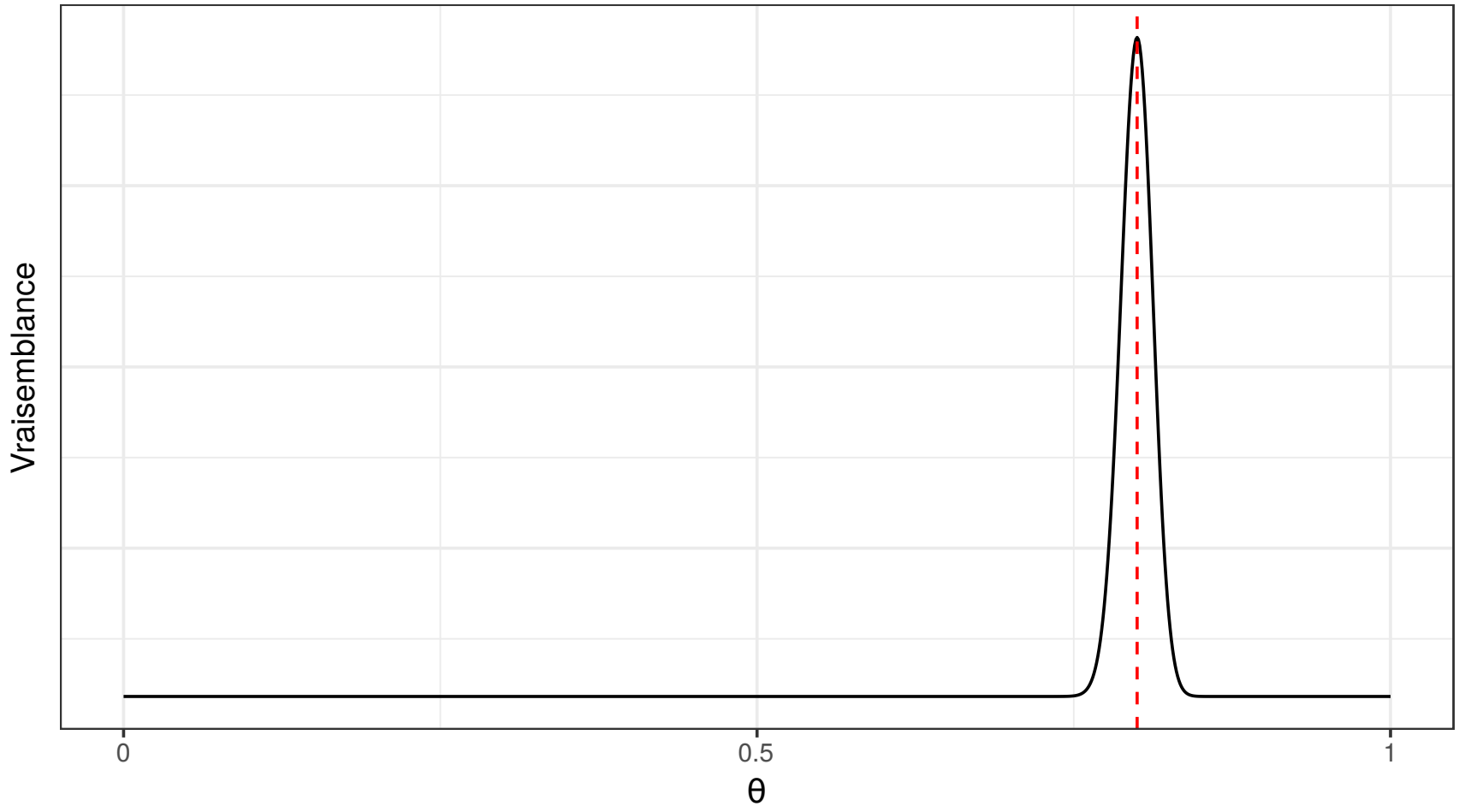
Incertitude sur $\hat{theta}$

$\hat{\theta}$ est une variable aléatoire. La théorie du MLE nous dit que cet estimateur admet un TCL. Ainsi, *asymptotiquement*, on a toujours un intervalle de confiance pour θ. Cet IC est aléatoire (mais pas θ!)).

Vraisemblance pour $n = 10$ et 8 succès



Vraisemblance pour $n = 1000$ et 800 succès



Inférence bayésienne

A priori sur θ

- On a potentiellement une connaissance *a priori* sur θ.

Inférence bayésienne

A priori sur θ

- On a potentiellement une connaissance *a priori* sur θ.
- On peut modéliser cet *a priori* sur le paramètre θ (savoir expert. . .) par une **variable aléatoire** de densité $\pi(\theta)$.

Inférence bayésienne

A priori sur θ

- On a potentiellement une connaissance *a priori* sur θ.
- On peut modéliser cet *a priori* sur le paramètre θ (savoir expert...) par une **variable aléatoire** de densité $\pi(\theta)$.
- Cette distribution est appelée **prior** sur θ.

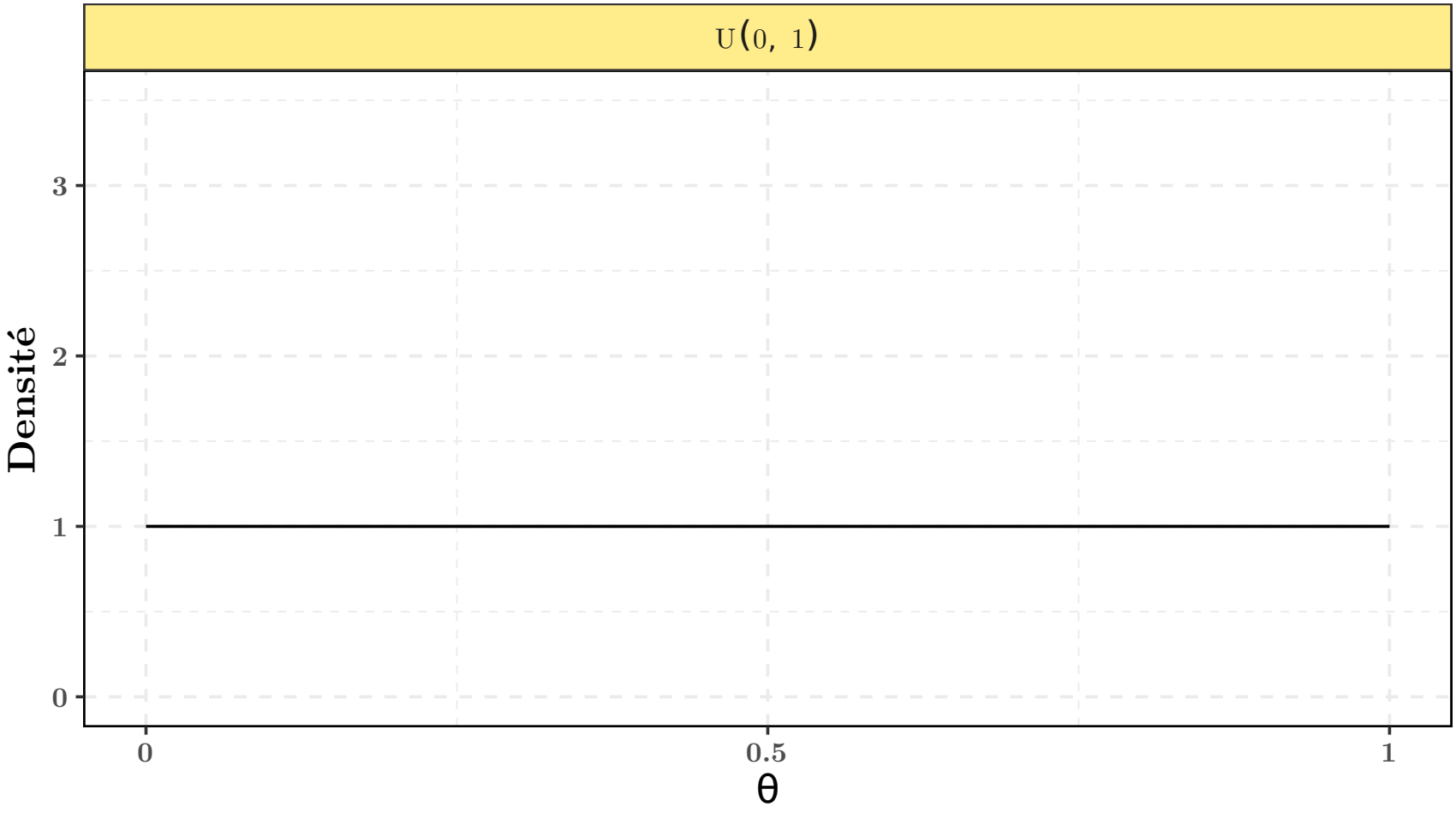
Inférence bayésienne

A priori sur θ

- On a potentiellement une connaissance *a priori* sur θ.
- On peut modéliser cet *a priori* sur le paramètre θ (savoir expert...) par une **variable aléatoire** de densité $\pi(\theta)$.
- Cette distribution est appelée **prior** sur θ.
- Dans ce contexte, θ est un variable aléatoire, on dispose d'un *a priori* sur sa loi.

Exemples de loi a priori

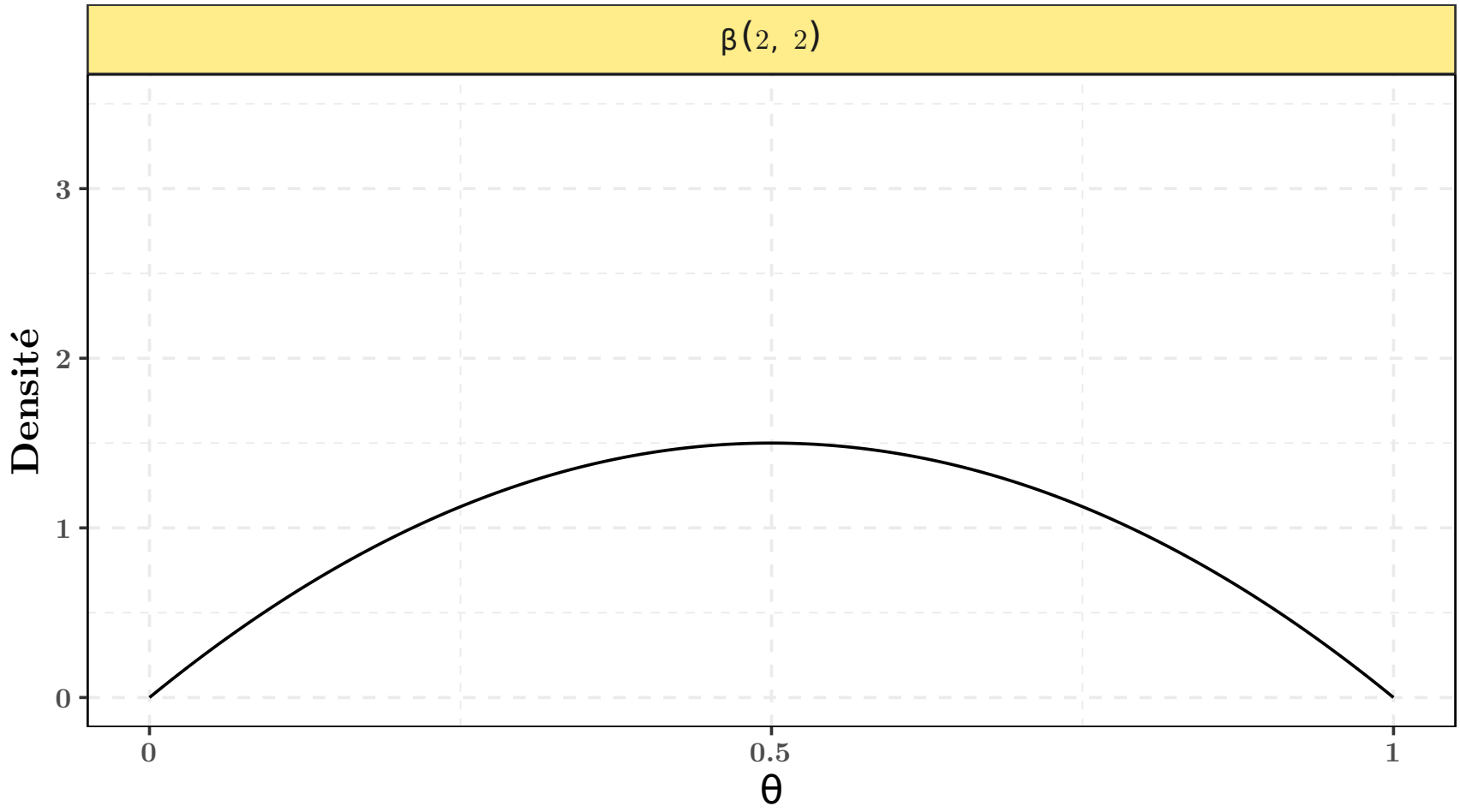
Aucune idée sur θ



Remarque une loi $\mathcal{U}[0,1]$ est **strictement** équivalente à une loi $\mathcal{B}eta(1,1)$.

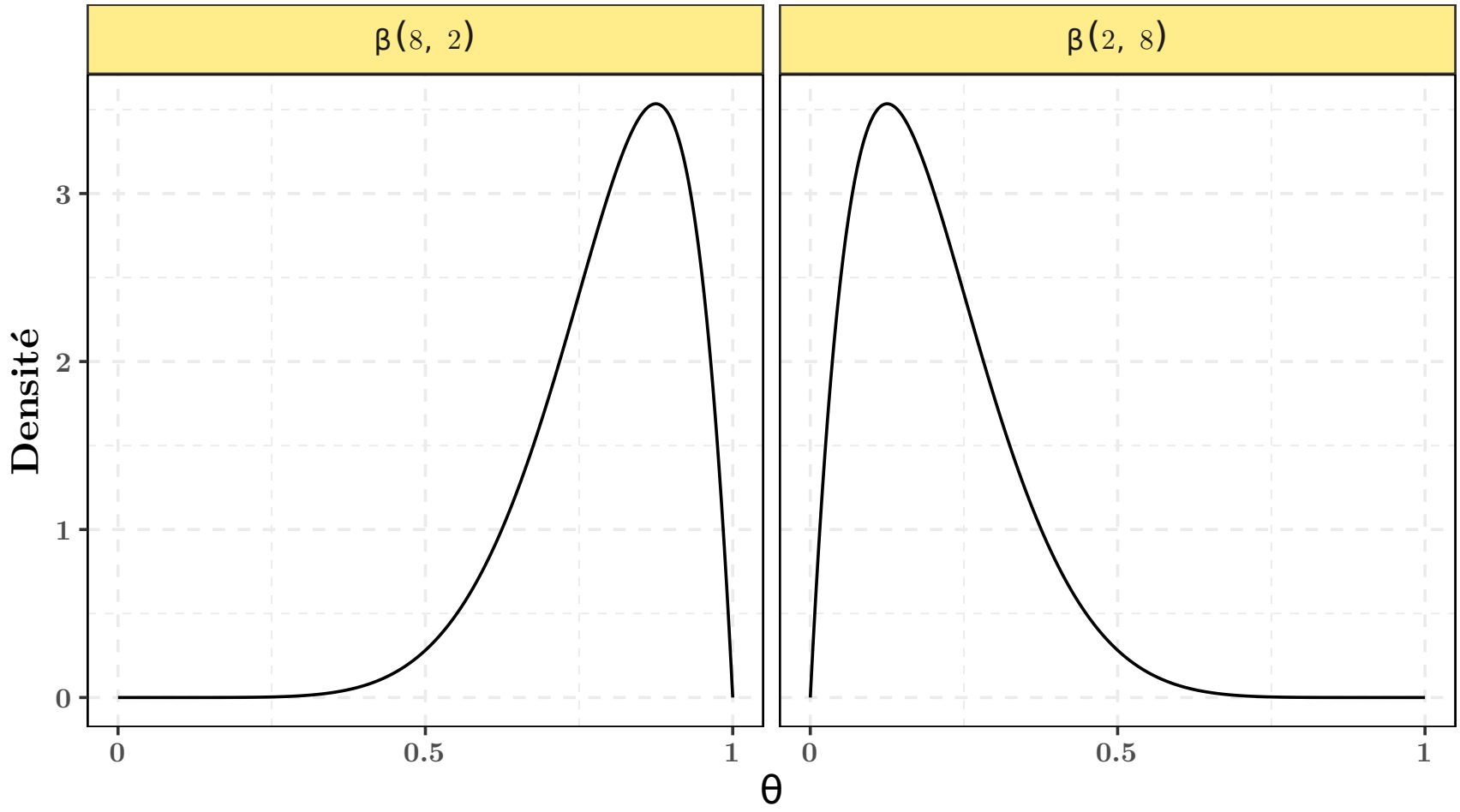
Exemples de loi a priori

A priori léger sur une pièce équitable



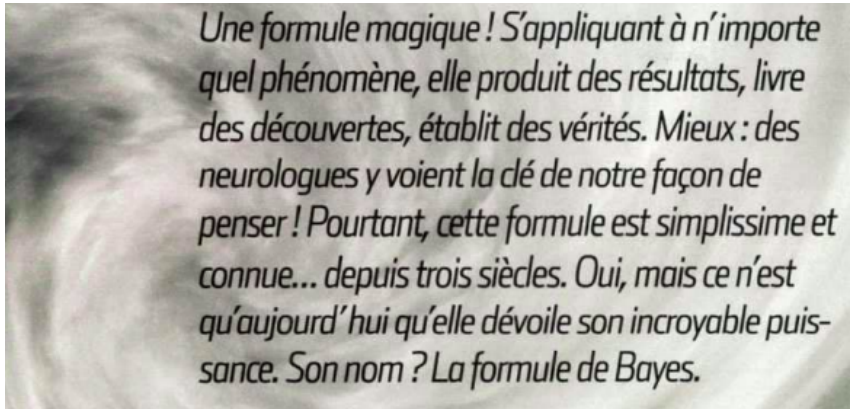
Exemples de loi a priori

A priori fort sur une pièce inéquitable



Inférence bayésienne

Inférence bayésienne



Inférence bayésienne

Influence des données, distribution a posteriori.

- L'objectif est de l'inférence est de **connaître la distribution de θ sachant les données**.

Inférence bayésienne

Influence des données, distribution a posteriori.

- L'objectif est de l'inférence est de **connaître la distribution de θ sachant les données**.
- La densité de cette distribution sur θ est notée $\pi(\theta|\mathbf{x})$, et est appelée **posterior** ou **loi a posteriori**.

Inférence bayésienne

Influence des données, distribution a posteriori.

- L'objectif est de l'inférence est de **connaître la distribution de θ sachant les données**.
- La densité de cette distribution sur θ est notée $\pi(\theta|\mathbf{x})$, et est appelée **posterior** ou **loi a posteriori**.
- On **actualise** notre connaissance sur θ grâce aux données.

Inférence bayésienne

Influence des données, distribution a posteriori.

- L'objectif est de l'inférence est de **connaître la distribution de θ sachant les données**.
- La densité de cette distribution sur θ est notée $\pi(\theta|\mathbf{x})$, et est appelée **posterior** ou **loi a posteriori**.
- On **actualise** notre connaissance sur θ grâce aux données.

Formule de Bayes

$$\mathbb{P}(B|A) = \frac{P(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

Inférence bayésienne

Influence des données, distribution a posteriori.

- L'objectif est de l'inférence est de **connaître la distribution de θ sachant les données**.
- La densité de cette distribution sur θ est notée $\pi(\theta|\mathbf{x})$, et est appelée **posterior** ou **loi a posteriori**.
- On **actualise** notre connaissance sur θ grâce aux données.

Formule de Bayes

$$\mathbb{P}(B|A) = \frac{P(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

Dans le cas avec des densités:

$$\pi(\theta|x_{1:n}) = \frac{p(x_{1:n}, \theta)}{p(x_{1:n})} = \frac{L(x_{1:n}|\theta)\pi(\theta)}{p(x_{1:n})}$$

où p est notation surchargée pour les densités.

Cette relation est résumée par:

$$\pi(\theta|\mathbf{x}) \propto L(x_{1:n}|\theta)\pi(\theta)$$

Objectif de l'inférence Bayésienne

$$\pi(\theta|\mathbf{x}) \propto L(x_{1:n}|\theta)\pi(\theta)$$

L'inférence Bayésienne a pour but la détermination (exacte, ou par simulation) du posterior $\pi(\theta|\mathbf{x})$.

Exemple 1: modèle avec prior conjugué

Posterior dans le modèle $\mathcal{B}eta$-Binomial

On revient au cas de pile ou on face où

$$L(x_{1:n}|\theta) = \prod_{k=1}^{n} \mathbb{P}_\theta(X = x_k) = \theta^{\sum_{k=1}^{n} x_k} (1-\theta)^{n-\sum_{k=1}^{n} x_k}$$

Posterior dans le modèle $\mathcal{B}eta$-Binomial

On revient au cas de pile ou on face où

$$L(x_{1:n}|\theta) = \prod_{k=1}^{n} \mathbb{P}_\theta(X = x_k) = \theta^{\sum_{k=1}^{n} x_k} (1-\theta)^{n-\sum_{k=1}^{n} x_k}$$

Pour l'inférence bayésienne, on pose comme *a priori* que $\theta \sim \mathcal{B}eta(a, b)$, ainsi:

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 u^{a-1}(1-u)^{b-1}\,du} \mathbf{1}_{0<\theta<1} \propto \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{0<\theta<1}$$

On cherche la loi de $\theta|x_{1:n}$.

Posterior dans le modèle $\mathcal{B}eta$-Binomial

On revient au cas de pile ou on face où

$$L(x_{1:n}|\theta) = \prod_{k=1}^{n} \mathbb{P}_\theta(X = x_k) = \theta^{\sum_{k=1}^n x_k} (1-\theta)^{n-\sum_{k=1}^n x_k}$$

Pour l'inférence bayésienne, on pose comme *a priori* que $\theta \sim \mathcal{B}eta(a, b)$, ainsi:

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 u^{a-1}(1-u)^{b-1}\,du}\mathbf{1}_{0<\theta<1} \propto \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{0<\theta<1}$$

On cherche la loi de $\theta|x_{1:n}$.

$$\begin{aligned}
\pi(\theta|x_{1:n}) &\propto L(x_{1:n}|\theta)\pi(\theta) \\
&\propto \theta^{\sum_{k=1}^n x_k} (1-\theta)^{n-\sum_{k=1}^n x_k}\, \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{0<\theta<1} \\
&\propto \theta^{a+\sum_{k=1}^n x_k-1}(1-\theta)^{b+n-\sum_{k=1}^n x_k-1}\mathbf{1}_{0<\theta<1}
\end{aligned}$$

Posterior dans le modèle $\mathcal{B}eta$-Binomial

On revient au cas de pile ou on face où

$$L(x_{1:n}|\theta) = \prod_{k=1}^{n} \mathbb{P}_\theta(X = x_k) = \theta^{\sum_{k=1}^n x_k} (1-\theta)^{n-\sum_{k=1}^n x_k}$$

Pour l'inférence bayésienne, on pose comme *a priori* que $\theta \sim \mathcal{B}eta(a, b)$, ainsi:

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 u^{a-1}(1-u)^{b-1}\,du}\mathbf{1}_{0<\theta<1} \propto \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{0<\theta<1}$$

On cherche la loi de $\theta|x_{1:n}$.

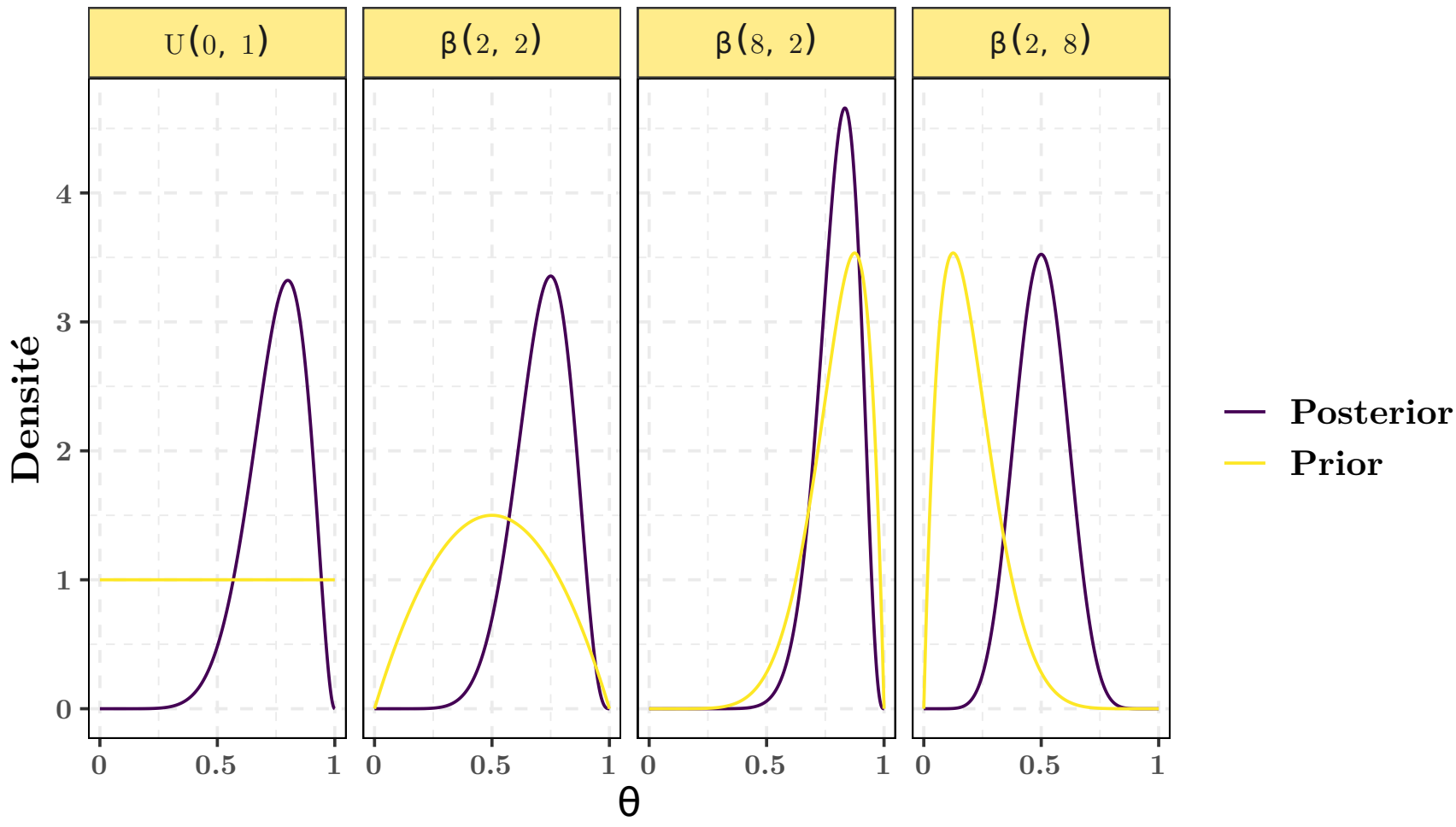
$$\begin{aligned}
\pi(\theta|x_{1:n}) &\propto L(x_{1:n}|\theta)\pi(\theta) \\
&\propto \theta^{\sum_{k=1}^n x_k} (1-\theta)^{n-\sum_{k=1}^n x_k}\, \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{0<\theta<1} \\
&\propto \theta^{a+\sum_{k=1}^n x_k-1}(1-\theta)^{b+n-\sum_{k=1}^n x_k-1}\mathbf{1}_{0<\theta<1}
\end{aligned}$$

On reconnaît que $\pi(\theta|\mathbf{x})$ est la densité d'une loi

$$\theta|x_{1:n} \sim \beta\left(a + \sum_{k=1}^n x_k,\, b + n - \sum_{k=1}^n x_k\right)$$

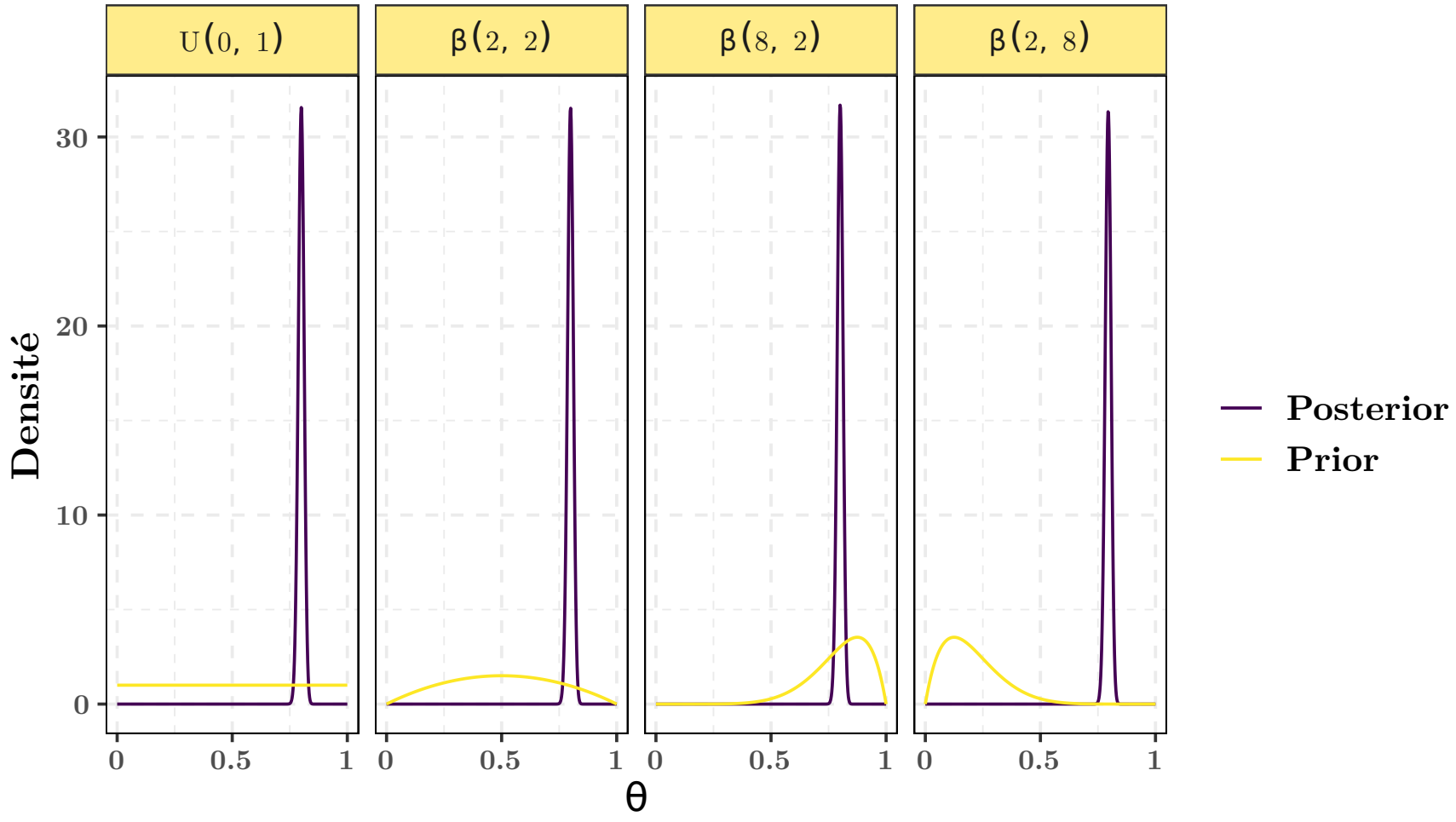
Cas $n = 10$ et 8 succès

$$\theta | x_{1:n} \sim \beta \left(a + \sum_i^n x_i, b + n - \sum_i^n x_i \right)$$



Cas $n = 1000$ et 800 succès

$$\theta | x_{1:n} \sim \beta \left(a + \sum_{i}^{n} x_i, b + n - \sum_{i}^{n} x_i \right)$$



Prior conjugué

Pour les modèles basés sur une vraisemblance “classique”, certains priors ont des priorités de conjugaison. Pour un modèle Bayésien, on appelle prior conjugué un prior $\pi(\theta)$ tel que le posterior $\pi(\mathbf{x}|\theta)$ est dans la même famille de loi que $\pi(\theta)$.

Exemples

- Modèle Bernouilli-Beta;
- Modèle Gaussien (prior: Normal Inverse Gamma);
- Modèle à densités dans la famille exponentielle.

Intérêt

L'inférence est directe!

Choix de prior et estimateurs Bayésiens

Influence et choix du prior

Pour un nombre de données limité, la **forme du prior** a un impact sur la forme du posterior.

Influence et choix du prior

Pour un nombre de données limité, la **forme du prior** a un impact sur la forme du posterior.

Choix du prior

La forme du prior peut être choisie en fonction du *savoir expert* (littérature existante, expériences passées).

ATTENTION: Le support du posterior sera toujours inclu dans le support du prior.

Influence et choix du prior

Pour un nombre de données limité, la **forme du prior** a un impact sur la forme du posterior.

Choix du prior

La forme du prior peut être choisie en fonction du *savoir expert* (littérature existante, expériences passées).

ATTENTION: Le support du posterior sera toujours inclu dans le support du prior.

Si le prior charge tout le support de manière égale, on dit qu'il est **non informatif**.

Prior impropre

Si le support de θ est sur $\mathbb{R}$, un prior non informatif est une "uniforme sur $\mathbb{R}$". Ceci n'est pas une loi.

Influence et choix du prior

Pour un nombre de données limité, la **forme du prior** a un impact sur la forme du posterior.

Choix du prior

La forme du prior peut être choisie en fonction du *savoir expert* (littérature existante, expériences passées).

ATTENTION: Le support du posterior sera toujours inclu dans le support du prior.

Si le prior charge tout le support de manière égale, on dit qu'il est **non informatif**.

Prior impropre

Si le support de θ est sur $\mathbb{R}$, un prior non informatif est une "uniforme sur $\mathbb{R}$". Ceci n'est pas une loi.

On peut cependant noter abusivement $\pi(\theta) \propto 1$. Dans ce cas, si $\frac{L(x_{1:n}|\theta)}{\int L(x_{1:n}|\theta)\mathrm{d}\theta}$ définit une loi de probabilité en θ, alors le posterior $\pi(\theta|\mathbf{x})$ est bien défini.

- Le prior est alors dit **impropre**.

Choix du prior

Exemple de prior impropre.

On suppose que **x** est issu d'un échantillon i.i.d. de taille n, de loi $\mathcal{N}(\mu, 1)$ où μ est inconnu. N'ayant aucune idée de la valeur de μ, on prend un prior non informatif. On a alors:

$$\begin{aligned}
\pi(\mu|x_{1:n}) &\propto L(x_{1:n}|\theta) \\
&\propto \mathrm{e}^{-\frac{1}{2}\sum_{k=1}^{n}(x_k-\mu)^2} \\
&\propto \mathrm{e}^{-\frac{1}{2}(n\mu^2-2\mu\sum_{k=1}^{n}x_k)} \\
&\propto \mathrm{e}^{-\frac{n}{2}(\mu-\frac{1}{n}\sum_{k=1}^{n}x_k)^2}
\end{aligned}$$

Ainsi,

$$\mu|x_{1:n} \sim \mathcal{N}\left(\frac{1}{n}\sum_{k=1}^{n}x_k, \frac{1}{n}\right)$$

Estimateurs Bayésiens

Maximum a posteriori (MAP)

Reprenant l'idée du MLE, il s'agit du mode de la distribution a posteriori:

$$MAP(\theta|x_{1:n}) = \text{argmax}_\theta \pi(\theta|x_{1:n})$$

Estimateurs Bayésiens

Maximum a posteriori (MAP)

Reprenant l'idée du MLE, il s'agit du mode de la distribution a posteriori:

$$MAP(\theta|x_{1:n}) = \text{argmax}_\theta \pi(\theta|x_{1:n})$$

Exemple sur la modèle Beta binomial

$$\theta|x_{1:n} \sim \beta\left(a + \sum_{k=1}^{n} x_k, b + n - \sum_{k=1}^{n} x_k\right)$$

On peut montrer que, pour $a + b + n > 2$ et $a + \sum_{k=1}^{n} x_k \geq 1$

$$MAP(\theta|x_{1:n}) = \frac{a + \sum_{k=1}^{n} x_k - 1}{a + b + n - 2}$$

Estimateurs Bayésiens

Maximum a posteriori (MAP)

Reprenant l'idée du MLE, il s'agit du mode de la distribution a posteriori:

$$MAP(\theta|x_{1:n}) = \text{argmax}_\theta \pi(\theta|x_{1:n})$$

Exemple sur la modèle Beta binomial

$$\theta|x_{1:n} \sim \beta\left(a + \sum_{k=1}^{n} x_k, b + n - \sum_{k=1}^{n} x_k\right)$$

On peut montrer que, pour $a + b + n > 2$ et $a + \sum_{k=1}^{n} x_k \geq 1$

$$MAP(\theta|x_{1:n}) = \frac{a + \sum_{k=1}^{n} x_k - 1}{a + b + n - 2}$$

On remarque que pour $a = b = 1$ (prior uniforme), il s'agit du maximum de vraisemblance, et que cela tend vers le MV quand n grandit.

Estimateurs Bayésiens

Espérance a posteriori

Soit un modèle Bayésien paramétré par une vraie valeur $\theta^* \in \Theta$ et de prior $\pi(\theta)$
Pour toute fonction φ, la variable aléatoire

$$\mathbb{E}[\varphi(\theta)|\mathbf{X}]$$

est un estimateur Bayésien de $\varphi(\theta^*)$.

Estimateurs Bayésiens

Espérance a posteriori

Soit un modèle Bayésien paramétré par une vraie valeur $\theta^* \in \Theta$ et de prior $\pi(\theta)$
Pour toute fonction φ, la variable aléatoire

$$\mathbb{E}[\varphi(\theta)|\mathbf{X}]$$

est un estimateur Bayésien de $\varphi(\theta^*)$.

Par exemple, pour un échantillon observé $\mathbf{x}$, une estimation bayésienne possible de θ^* est

$$\hat{\theta} = \mathbb{E}[\theta|\mathbf{X} = x_{1:n}] = \int_\Theta \theta\pi(\theta|x_{1:n})\mathrm{d}\theta$$

Exemple sur la modèle Beta-Binomial

Pour un prior $\beta(a, b)$, on a

$$\hat{\theta} \overset{\text{loi } \beta}{=} \frac{a + \sum_{i=1}^n x_i}{a + b + n} = \underbrace{\frac{n}{a + b + n}}_{\text{Poids données}} \times \overbrace{\frac{\sum_{i=1}^n x_i}{n}}^{\text{Max. de vrais.}} + \underbrace{\frac{a + b}{a + b + n}}_{\text{Poids prior}} \times \overbrace{\frac{a}{a + b}}^{\mathbb{E} \text{ du prior}}$$

Intervalle de crédibilité

Pour toute région $\mathcal{R} \subset \Theta$, on peut quantifier:

$$\mathbb{P}(\theta \in \mathcal{R} | \mathbf{X} = x_{1:n}) = \int_{\mathcal{R}} \pi(\theta | x_{1:n}) \mathrm{d}\theta$$

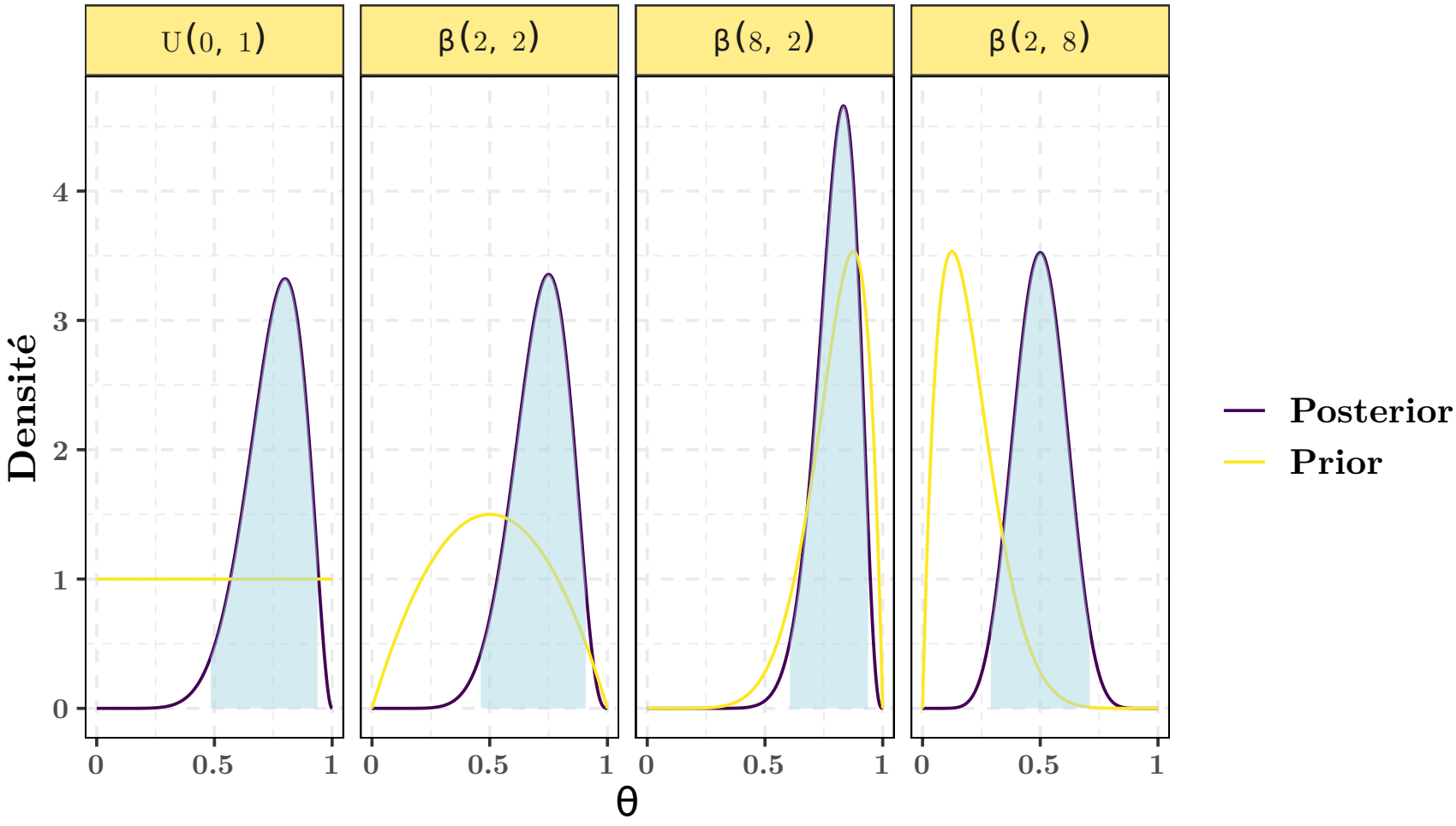
Pour $\alpha \in]0, 1[$, une région de crédibilité de niveau $1 - \alpha$ est une région $\mathcal{R} \subset \Theta$ telle que

$$\mathbb{P}(\theta \in \mathcal{R} | \mathbf{X} = x_{1:n}) = 1 - \alpha$$

Cet intervalle n'est pas asymptotique, mais **dépend du prior**.

Remarque, ici l'aléa est bien sur θ (contrairement à un intervalle de confiance).

Intervalles de crédibilités (centrés) à 95% dans le modèle Beta binomial



Exemple 2: cas non conjugué

Exemple: Prédiction de présence d'oiseaux

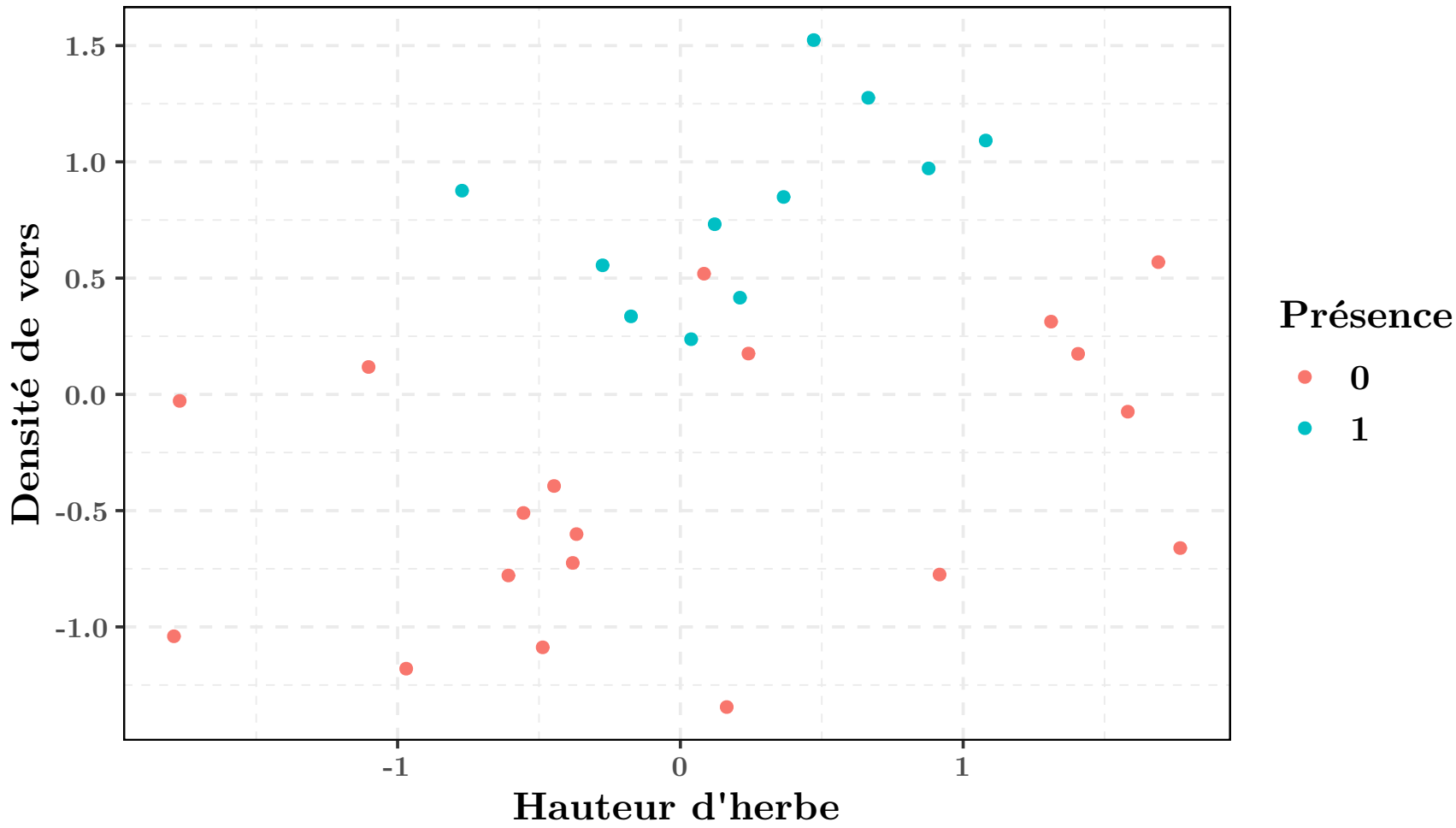
Une étude consiste en l'observation de la présence ou non de la linotte mélodieuse sur différents sites échantillonnés.

Caractéristiques des sites

Sur ces différents sites sont mesurées différentes caractéristiques:

- Le nombre de vers moyens sur une surface au sol de $1m^2$. (Covariable 1)
- La hauteur d'herbe moyenne sur une surface au sol de $1m^2$. (Covariable 2)
- On calcule cette hauteur d'herbe au carré. (Covariable 3).

Données


Densité de vers
1.5
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Hauteur d'herbe
Présence
0
1

Notations et modèle de régression probit

On note $y_1, \ldots, y_n$ les observations de présence (1 si on observe un oiseau, 0 sinon) sur les sites 1 à n.

On note

$$\mathbf{x}_k = \left(\overset{\text{Nb. vers}}{x_{k,1}} , \overset{\text{Haut. herbe}}{x_{k,2}} , \overset{\text{Haut. herbe}^2}{x_{k,3}} \right)^T$$

le vecteur des covariables sur le k-ème site ($1 \leq k \leq n$).

Notations et modèle de régression probit

On note $y_1, \ldots, y_n$ les observations de présence (1 si on observe un oiseau, 0 sinon) sur les sites 1 à n.

On note

$$\mathbf{x}_k = \left(\overset{\text{Nb. vers}}{x_{k,1}} , \overset{\text{Haut. herbe}}{x_{k,2}} , \overset{\text{Haut. herbe}^2}{x_{k,3}} \right)^T$$

le vecteur des covariables sur le k-ème site ($1 \leq k \leq n$).

On pose le modèle suivant:

$Y_k \sim \mathcal{B}ern(p_k)$ où

$$p_k = \phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) = \phi(\mathbf{x}_k^T \theta),$$

où

- ϕ est la fonction de répartition d'une $\mathcal{N}(0,1)$, i.e.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-\frac{u^2}{2}} \mathrm{d}u$$

- $\theta = \{\beta_0, \beta_1, \beta_2, \beta_3\}$ est le vecteur des paramètres à estimer.

Modèle Bayésien

Prior sur θ

Comme a priori sur θ, on choisit une normale avec une grande variance$\theta \overset{\text{prior}}{\sim} \mathcal{N}(0, 4I)$, donc

$$\pi(\theta) = \frac{1}{\sqrt{2\pi \times 4}^4} \mathrm{e}^{-\frac{1}{8}\theta^T\theta}$$

où I est la matrice Identité (ici 4×4)

Modèle Bayésien

Prior sur θ

Comme a priori sur θ, on choisit une normale avec une grande variance$\theta \overset{\text{prior}}{\sim} \mathcal{N}(0, 4I)$, donc

$$\pi(\theta) = \frac{1}{\sqrt{2\pi \times 4}^4} \mathrm{e}^{-\frac{1}{8}\theta^T\theta}$$

où I est la matrice Identité (ici 4×4)

Vraisemblance

Pour un vecteur d'observations $y_{1:k}$, la vraisemblance

$$L(y_{1:n}|\theta) = \prod_{k=1}^{n} \underbrace{\phi(\mathbf{x}_k^T\theta)^{y_k}}_{\text{Proba. présence}} \times \underbrace{(1 - \phi(\mathbf{x}_k^T\theta))^{1-y_k}}_{\text{Proba. absence}}$$

Modèle Bayésien

Prior sur θ

Comme a priori sur θ, on choisit une normale avec une grande variance$\theta \overset{\text{prior}}{\sim} \mathcal{N}(0, 4I)$, donc

$$\pi(\theta) = \frac{1}{\sqrt{2\pi \times 4}^4} \mathrm{e}^{-\frac{1}{8}\theta^T\theta}$$

où I est la matrice Identité (ici 4×4)

Vraisemblance

Pour un vecteur d'observations $y_{1:k}$, la vraisemblance

$$L(y_{1:n}|\theta) = \prod_{k=1}^{n} \underbrace{\phi(\mathbf{x}_k^T\theta)^{y_k}}_{\text{Proba. présence}} \times \underbrace{(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k}}_{\text{Proba. absence}}$$

Posterior

Le posterior est donc donné par:

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta)L(y_{1:n}|\theta) \propto \frac{1}{64\pi^2} \mathrm{e}^{-\frac{1}{8}\theta^T\theta} \prod_{k=1}^{n} \phi(\mathbf{x}_k^T\theta)^{y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k}$$

Posterior modèle Normal-Probit

$$\pi(\theta|y_{1:n}) \propto \pi(\theta)L(y_{1:n}|\theta) \propto \frac{1}{64\pi^2}\mathrm{e}^{-\frac{1}{8}\theta^T\theta}\prod_{k=1}^{n}\phi(\mathbf{x}_k^T\theta)^{y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k}$$

Cette densité n'est pas standard:

- On ne sait pas calculer des espérances associées (estimateurs bayésiens);
- On pourrait approcher ces espérances par méthodes de Monte Carlo

Posterior modèle Normal-Probit

$$\pi(\theta|y_{1:n}) \propto \pi(\theta)L(y_{1:n}|\theta) \propto \frac{1}{64\pi^2}\mathrm{e}^{-\frac{1}{8}\theta^T\theta}\prod_{k=1}^{n}\phi(\mathbf{x}_k^T\theta)^{y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k}$$

Cette densité n'est pas standard:

- On ne sait pas calculer des espérances associées (estimateurs bayésiens);
- On pourrait approcher ces espérances par méthodes de Monte Carlo
- Encore faut il savoir simuler!

Posterior modèle Normal-Probit

$$\pi(\theta|y_{1:n}) \propto \pi(\theta)L(y_{1:n}|\theta) \propto \frac{1}{64\pi^2}\mathrm{e}^{-\frac{1}{8}\theta^T\theta}\prod_{k=1}^{n}\phi(\mathbf{x}_k^T\theta)^{y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k}$$

Cette densité n'est pas standard:

- On ne sait pas calculer des espérances associées (estimateurs bayésiens);
- On pourrait approcher ces espérances par méthodes de Monte Carlo
- Encore faut il savoir simuler!
- Le cas où le posterior ne fait pas partie d'une famille connue est très fréquent.
- L'inférence bayésienne est une motivation énorme pour les algos de simulations de loi.

Simulation posterior modèle Normal-Probit

On veut simuler selon

$$\pi(\theta|y_{1:n}) \propto \frac{1}{64\pi^2} \mathrm{e}^{-\frac{1}{8}\theta^T\theta} \prod_{k=1}^{n} \phi(\mathbf{x}_k^T\theta)^{y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k}$$

Simulation posterior modèle Normal-Probit

On veut simuler selon

$$\pi(\theta|y_{1:n}) \propto \frac{1}{64\pi^2} \mathrm{e}^{-\frac{1}{8}\theta^T\theta} \prod_{k=1}^{n} \phi(\mathbf{x}_k^T\theta)^{y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k}$$

Simulation par acceptation rejet

On voudrait simuler selon $\pi(\theta|y_{1:n})$.

Simulation posterior modèle Normal-Probit

On veut simuler selon

$$\pi(\theta|y_{1:n}) \propto \frac{1}{64\pi^2} \mathrm{e}^{-\frac{1}{8}\theta^T\theta} \prod_{k=1}^{n} \phi(\mathbf{x}_k^T\theta)^{y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k}$$

Simulation par acceptation rejet

On voudrait simuler selon $\pi(\theta|y_{1:n})$.

- Idée 1: trouver une densité g selon laquelle on sait simuler et telle qu'il existe $M > 0$ tel que

$$\forall \theta \in \mathbb{R}^4, \ \frac{\pi(\theta|y_{1:n})}{g(\theta)} \leq M$$

Simulation posterior modèle Normal-Probit

On veut simuler selon

$$\pi(\theta|y_{1:n}) \propto \overbrace{\frac{1}{64\pi^2}\mathrm{e}^{-\frac{1}{8}\theta^T\theta}\prod_{k=1}^{n}\phi(\mathbf{x}_k^T\theta)^{y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k}}^{\tilde{\pi}(\theta|y_{1:n})}$$

Simulation par acceptation rejet

On voudrait simuler selon $\pi(\theta|y_{1:n})$.

- Idée 1: trouver une densité g selon laquelle on sait simuler et telle qu'il existe $M > 0$ tel que

$$\forall\theta \in \mathbb{R}^4,\ \frac{\pi(\theta|y_{1:n})}{g(\theta)} \leq M$$

  Mais $\pi(\theta|y_{1:n})$ n'est connu qu'a une constante près!

$$\pi(\theta|y_{1:n}) = \frac{\tilde{\pi}(\theta|y_{1:n})}{\int_{\mathbb{R}^4}\pi(u)L(y_{1:n}|u)\mathrm{d}u}$$

- **Rappel** L'acceptation rejet marche toujours si on ne connait la loi cible qu'à une constante près! (voir TD pour la preuve).

Simulation posterior modèle Normal-Probit

On veut simuler selon

$$\pi(\theta|y_{1:n}) \propto \overbrace{\frac{1}{64\pi^2}\mathrm{e}^{-\frac{1}{8}\theta^T\theta}\prod_{k=1}^{n}\phi(\mathbf{x}_k^T\theta)^{y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k}}^{\tilde{\pi}(\theta|y_{1:n})}$$

- Idée 2: trouver une densité g selon laquelle on sait simuler et telle qu'il existe $M > 0$ tel que

$$\forall\theta \in \mathbb{R}^4,\ \frac{\tilde{\pi}(\theta|y_{1:n})}{g(\theta)} \leq M$$

Implémentation de l'acceptation rejet

On peut par exemple prend pour g la densité correspondant au prior ($g(\theta) = \pi(\theta)$). On remarque que dans ce cas

$$\frac{\tilde{\pi}(\theta|y_{1:n})}{g(\theta)} = \frac{\pi(\theta)L(y_{1:n}|\theta)}{\pi(\theta)} = \prod_{k=1}^{n} \phi(\mathbf{x}_k^T\theta)^{y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k} \leq 1 =: M$$

Remarque: il existe un M optimal plus petit que 1.

Implémentation de l'acceptation rejet

On peut par exemple prend pour g la densité correspondant au prior ($g(\theta) = \pi(\theta)$). On remarque que dans ce cas

$$\frac{\tilde{\pi}(\theta|y_{1:n})}{g(\theta)} = \frac{\pi(\theta)L(y_{1:n}|\theta)}{\pi(\theta)} = \prod_{k=1}^{n} \phi(\mathbf{x}_k^T\theta)^{y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k} \leq 1 =: M$$

Remarque: il existe un M optimal plus petit que 1.

Algorithme de simulation selont $\pi(\theta|y_{1:n})$

1. On tire $\theta_{cand} \sim \mathcal{N}(0, 4I)$
2. On tire (independamment) $U \sim \mathcal{U}[0,1]$
3. Si $U < \frac{L(y_{1:n}|\theta)}{M}$, on accepte θ_{cand}
4. Sinon on recommence

Implémentation de l'acceptation rejet

On peut par exemple prend pour g la densité correspondant au prior ($g(\theta) = \pi(\theta)$). On remarque que dans ce cas

$$\frac{\tilde{\pi}(\theta|y_{1:n})}{g(\theta)} = \frac{\pi(\theta)L(y_{1:n}|\theta)}{\pi(\theta)} = \prod_{k=1}^{n} \phi(\mathbf{x}_k^T\theta)^{y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-y_k} \leq 1 =: M$$

Remarque: il existe un M optimal plus petit que 1.

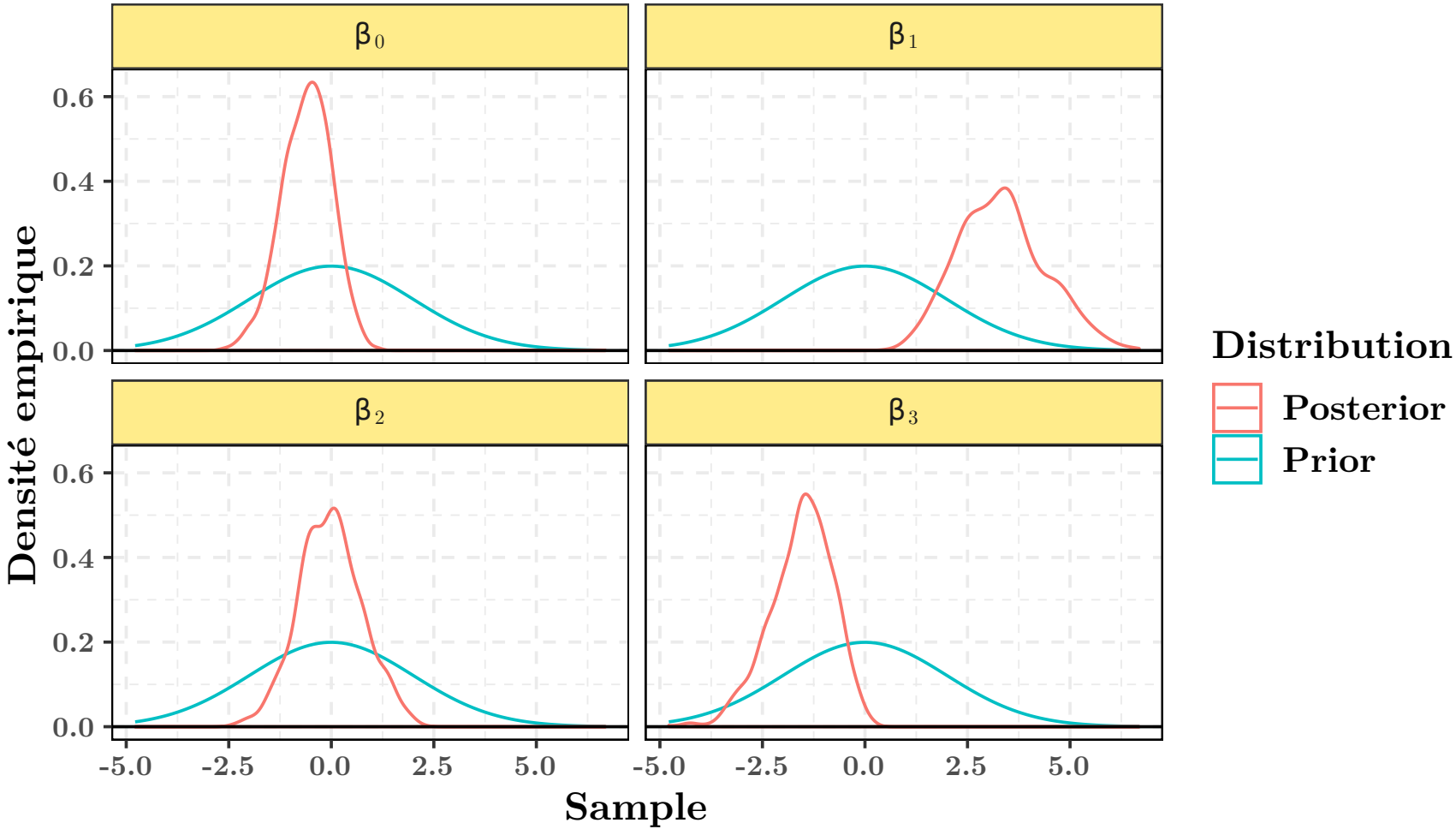
Algorithme de simulation selont $\pi(\theta|y_{1:n})$

1. On tire $\theta_{cand} \sim \mathcal{N}(0, 4I)$
2. On tire (independamment) $U \sim \mathcal{U}[0,1]$
3. Si $U < \frac{L(y_{1:n}|\theta)}{M}$, on accepte θ_{cand}
4. Sinon on recommence

Remarque, l'échantillon obtenu est tiré selon *la loi jointe* (on ne tire pas β_0 puis β_1, etc. . .)

Echantillon du posterior, et loi a posteriori marginales

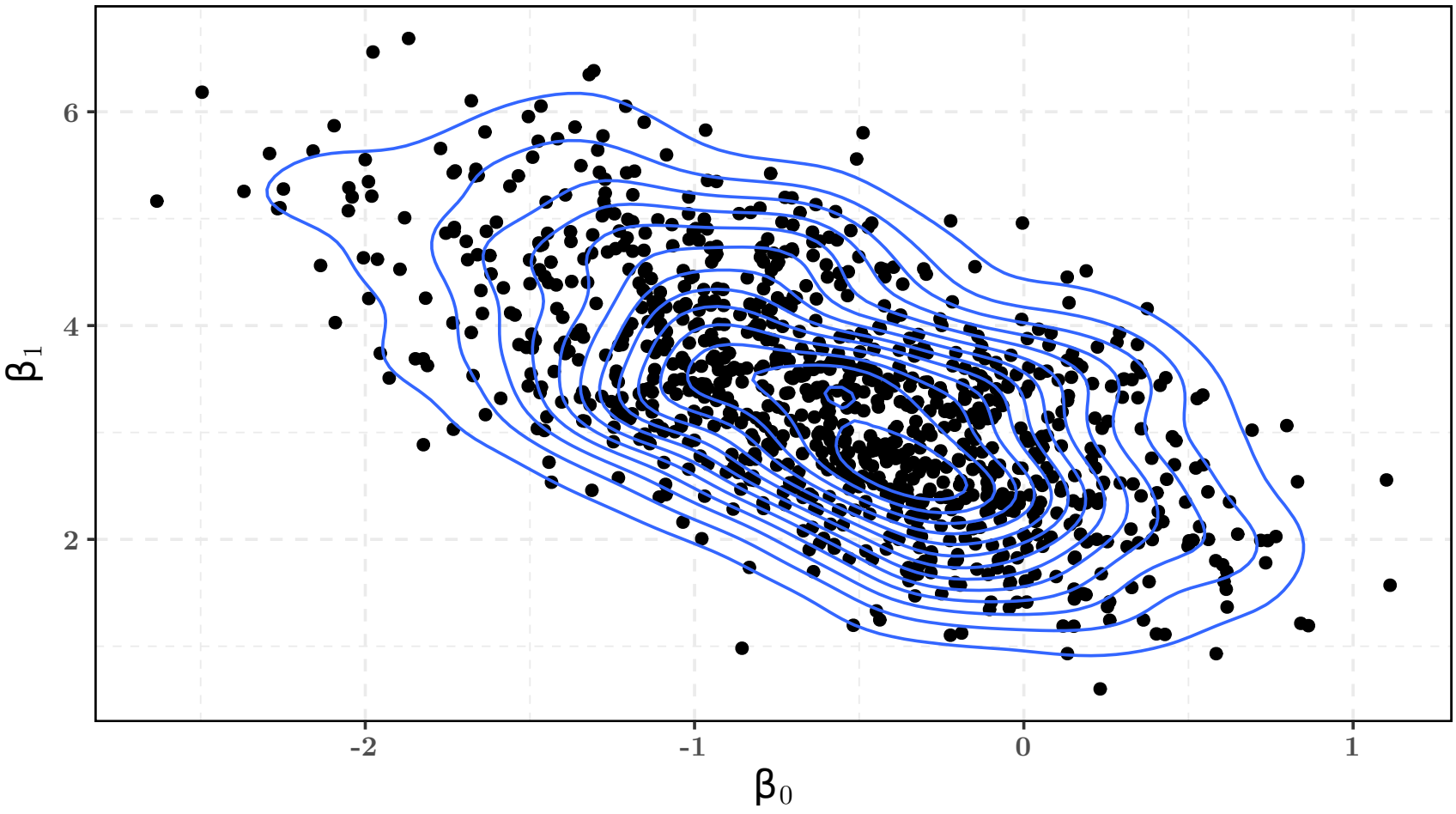
On effectue un tirage de taille $M = 1000$



- Les données ont bien actualisé la connaissance sur θ

Echantillon du posterior et loi jointe

On peut regarder la loi jointe de $(\beta_0, \beta_1 | y_{1:n})$:



Estimateurs bayésiens

On prend comme estimateur l'espérance **a posteriori**. De plus, on regarde l'estimation de l'intervalle

Parameter	Estimation	inf_IC95	sup_IC95
beta[0]	-0.595	-1.868800	0.5265336
beta[1]	3.329	1.441659	5.5525325
beta[2]	-0.019	-1.499485	1.5197882
beta[3]	-1.551	-3.222323	-0.2565622

Au delà de l'acceptation rejet

Dans le cas précédent, l'espérance du temps d'attente avant une acceptation est donnée par

$$\frac{M}{\int L(y_{1:n}|\theta)\pi(\theta)\mathrm{d}\theta}$$

Au delà de l'acceptation rejet

Dans le cas précédent, l'espérance du temps d'attente avant une acceptation est donnée par

$$\frac{M}{\int L(y_{1:n}|\theta)\pi(\theta)\mathrm{d}\theta}$$

Mécaniquement, cette quantité augmente quand n augmente, et l'acceptation rejet dvient prohibitif.

En pratique, l'inférence Bayésienne utilisera d'autres algorithmes de simulations de loi: les algorithmes de Monte Carlo par chaîne de Markov.