Inference bayésienne

Pierre Gloaguen

01/05/2020

Rappel sur le maximum de vraisemblance

En statistique paramétrique, on suppose qu'un ensemble d'observations $\mathbf{X}$ est la réalisation d'une variable aléatoire dont la loi dépend d'un ensemble de paramètres θ inconnu et à valeurs dans un espace Θ. L'inférence statistique consiste en la définition d'un estimateur de θ.

Un estimateur générique commun est l'estimateur du maximum de vraisemblance.

Le modèle statistique posé permettant d'écrire la loi de $\mathbf{X}$ quand on connaît θ, que l'on note $L(\mathbf{X}, \theta)$. On choisit comme estimateur le paramètre $\hat{\theta}$ *le plus vraisemblable*, c'est à dire celui qui maximise (en θ) la fonction $L(\mathbf{X}, \theta)$.

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

Cet estimateur **est une variable aléatoire**. Sa loi dépend du modèle, mais asymptotiquement, un thérème central limite nous assure que sa distribution devient celle d'une loi Normale (dont l'expression de la variance est connue, au moins en théorie).

Dans ce contexte, le paramètre θ est donc une quantité fixe inconnue. Toute la connaissance sur sa valeur vient des données.

Inférence bayésienne

Connaissance a priori et définition du posterior

Dans le contexte de l'inférence bayésienne, on supposera que le paramètre θ est lui même aléatoire. On modélisera alors sa loi sous la forme d'une distribution. Cette distribution est indépendante des données et s'appelle la distribution *a priori*. Elle reflète la connaissance (et l'incertitude) que l'on a sur le paramètre. La loi a priori sur θ sera notée $\pi(\theta)$.

L'objectif de l'inférence bayésienne est d'actualiser cette connaissance (et son incertitude) grâce au données. La quantité d'intérêt, dans ce contexte est alors la loi de $\theta|\mathbf{X}$, quand appelle loi a posteriori (ou posterior). Cette quantité sera notée $\pi(\theta|\mathbf{X})$.

La formule de Bayes sur le conditionnement permet de lier cette loi a posteriori à la loi a priori et à la vraisemblance du modèle:

$$\pi(\theta|\mathbf{X}) = \frac{\pi(\theta)L(\mathbf{X}|\theta)}{\int_\Theta \pi(u)L(\mathbf{X}|u)}du$$

On remarque que le dénominateur ne dépend pas de θ, il s'agit d'une constante de normalisation. On écrira souvent cette relation

$$\pi(\theta|X) \propto \pi(\theta)L(\mathbf{X}|\theta)$$

Ce sont les caractéristiques de cette loi (ses quantiles, ses moments) que l'on cible dans le contexte de l'inférence bayésienne.

L'objectif de l'inférence bayésienne est donc la détermination de $\pi(\theta|\mathbf{X})$.

Choix du prior

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

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 cependant pas une loi de probabilité!

On peut cependant noter abusivement $\pi(\theta) \propto 1$. Dans ce cas, si $\frac{L(\mathbf{X})|\theta)}{\int_\Theta L(\mathbf{X}|\theta)\mathrm{d}\theta}$ définit une loi de probabilité en θ, alors le posterior $\pi(\theta|\mathbf{X})$ est bien défini. Le prior est alors dit **impropre**.

Estimateurs bayésiens

Les estimateurs bayésiens sont des quantités liées à la loi à posteriori.

On mentionnera:

- Le maximum a posteriori (MAP), correspondant à à la valeur de θ maximisant $\pi(\theta|\mathbf{X})$.
- L'espérance a posteriori
$$\mathbb{E}[\theta|\mathbf{X}] = \int_\theta \pi(\theta|\mathbf{X})\theta d\theta.$$
- Intervalles de crédibilités: Pour toute région $\mathcal{R} \subset \Theta$, on peut quantifier:
$$\mathbb{P}(\theta \in \mathcal{R}|\mathbf{X}) = \int_\mathcal{R} \pi(\theta|\mathbf{X})\mathrm{d}\theta$$
Pour $\alpha \in]0,1[$, une région de crédibilité de niveau $1-\alpha$ est une région $\mathcal{R} \subset \Theta$ telle que
$$\mathbb{P}(\theta \in \mathcal{R}|\mathbf{X} = \mathbf{x}) = 1 - \alpha$$
Cet intervalle n'est pas asymptotique, mais **dépend du prior**.

Détermination du posterior

Il existe deux cas différents en inférence bayésienne:

- Soit la loi a posteriori est dans une famille connue (loi normale, loi beta, etc,...), alors l'inférence est directe, et tous les estimateurs bayésiens peuvent être obtenus facilement.
- Soit la loi a posteriori n'appartient pas à une famille de loi connue. Dans ce cas, il faudra obtenir les quantités d'intérêt par méthode de Monte Carlo. Pour cela, il faudra souvent être capable d'obtenir un échantillon i.i.d. selon la loi a posteriori. Les méthodes vues jusqu'alors pourront être utilisées. On verra qu'elles ne suffiront pas toujours, et que d'autres méthodes, les méthodes de Monte Carlo par chaîne de Markov, aideront à s'en sortir.

Une manière astucieuse de se retrouver dans le cas 1 est d'utiliser les propriétés de conjugaisons de certaines lois. On parlera alors de priors conjugués au modèle.

Les deux sections suivantes décrivent chacune un exemple illustratif de ces cas.

Cas conjugué: modèle beta-binomial

Expérience et question

On suppose qu'on dispose d'une pièce, et l'on souhaite déterminée si elle est équilibrée. Pour cela, on effectue n tirages indépendant de pile ou face.

Modélisation

On note $\mathbf{x} = (x_1, \ldots, x_n)$ le résultat du lancer (0 si *face*, 1 si *pile*). On suppose que ces nombres sont les réalisations d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ où les $X_1, \ldots, X_n$ sont indépendantes et identiquements distribuées de loi $\mathcal{B}ern(\theta)$ où $\theta \in]0, 1[$ est la probabilité d'obtenir pile.

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

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

où $X \sim \mathcal{B}ern(\theta)$.

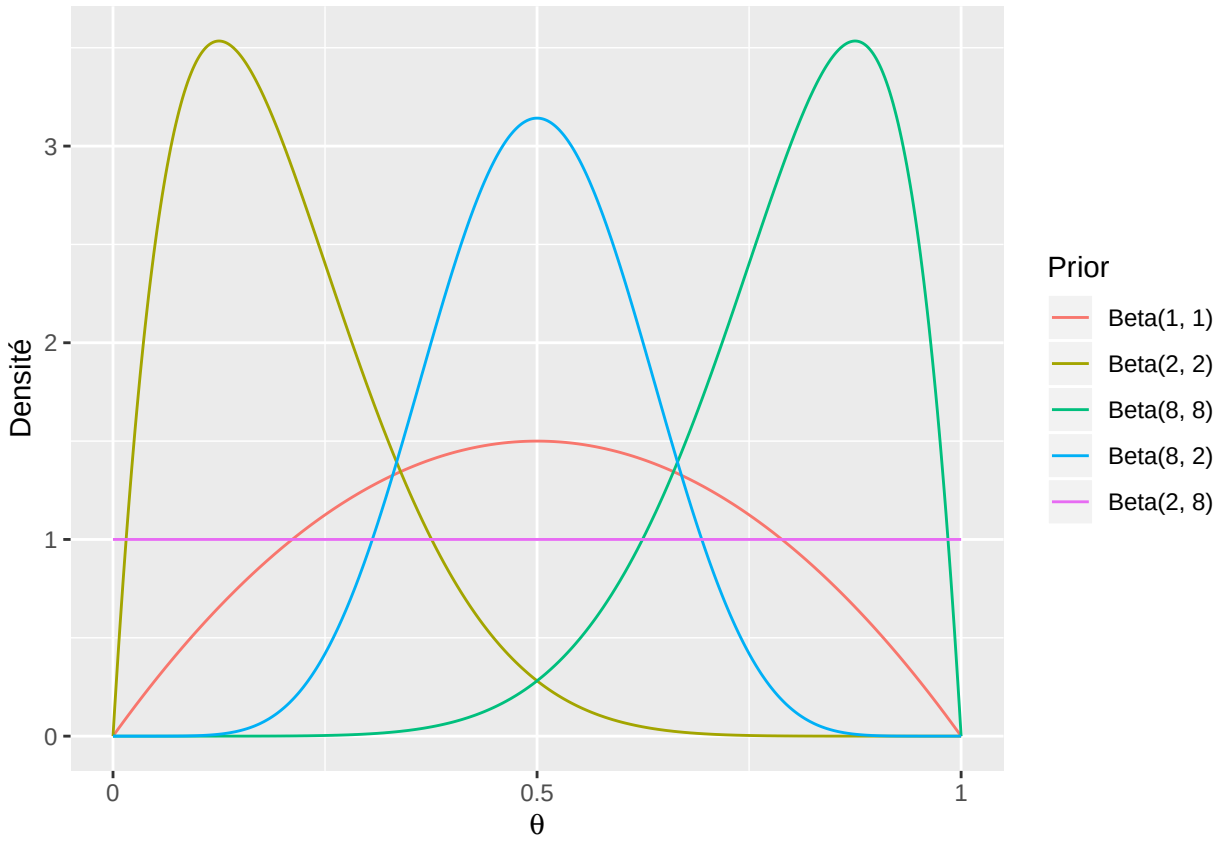
Prior

Pour l'inférence bayésienne, on pose comme *a priori* que $\theta \sim \mathcal{B}eta(a, b)$. Cette loi est censée illustrée notre connaissance indépendante des données sur θ. Le premier point trivial est que l'on sait que θ est entre 0 et 1, donc on a choisi une loi ayant ce support.

Ensuite, le choix des paramètres a et b déterminera notre a priori sur al pièce:

- Le cas $a = b = 1$, correspond à une loi uniforme. Cela traduira un a priori non informatif sur θ, chaque valeur entre $]0, 1[$ nous semble également vraisemblable.
- Le cas $a = b$ avec des valeurs supérieures à 1 traduira un a priori où la pièce est équilibrée. De grandes valeurs de a et b traduiront une plus grande certitude.
- Le cas $a > b$, traduira un a priori où la pièce est déséquilibrée en faveur de pile.
- Le cas $a < b$, traduira un a priori où la pièce est déséquilibrée en faveur de face.

La figure ci dessous illustre ces différents a priori:



L'expression analytique du prior est donc donnée par:

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

Loi a posteriori

On cherche la loi de $\theta|\mathbf{X}$.

On a directement que:

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

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

$$\theta|\mathbf{X} \sim \beta\left(a+\underbrace{\sum_{k=1}^n X_k}_{\text{Nb. piles}},\ b+\underbrace{n-\sum_{k=1}^n X_k}_{\text{Nb. faces}}\right)$$

Le fait que la *loi a posteriori* soit dans la même famille que la loi *a priori* est une propriété de conjugaison du modèle binomial avec le prior de loi beta. Ce prior est dit conjugué.

Estimateurs bayésiens

- **Maximum a posteriori (MAP)**

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

$$MAP(\theta|\mathbf{X}) = \frac{a+\sum_{k=1}^{n} X_k - 1}{a+b+n-2}$$

On remarque que pour $a=b=1$ (prior uniforme), il s'agit du maximum de vraisemblance, et que pour tout couple (a,b), cette quantité converge vers le maximum de vraisemblance quand n grandit.

- **Espérance a posteriori**

Par propriété de la loi β, on:

$$\mathbb{E}[\theta|\mathbf{X}] \overset{\text{loi } \beta}{=} \frac{a+\sum_{k=1}^{n} X_k}{a+b+n} = \underbrace{\frac{n}{a+b+n}}_{\text{Poids données}} \times \overbrace{\frac{\sum_{k=1}^{n} X_k}{n}}^{\text{Max. de vrais.}} + \underbrace{\frac{a+b}{a+b+n}}_{\text{Poids prior}} \times \overbrace{\frac{a}{a+b}}^{\mathbb{E} \text{ du prior}}$$

Encore une fois, la décomposition illustre le poids des données et le poids du prior. On remarque que pour n suffisament grand, tous les priors seront équivalents.

- **Régions de crédibilité**

Toute région de crédibilité peut facilement être obtenue à l'aide de la fonction quantile de la loi β, qui est implémentée dans tout logiciel de statistiques.

Posterior de loi inconnue: modèle de régression probit:

Prédiction de présence d'oiseaux

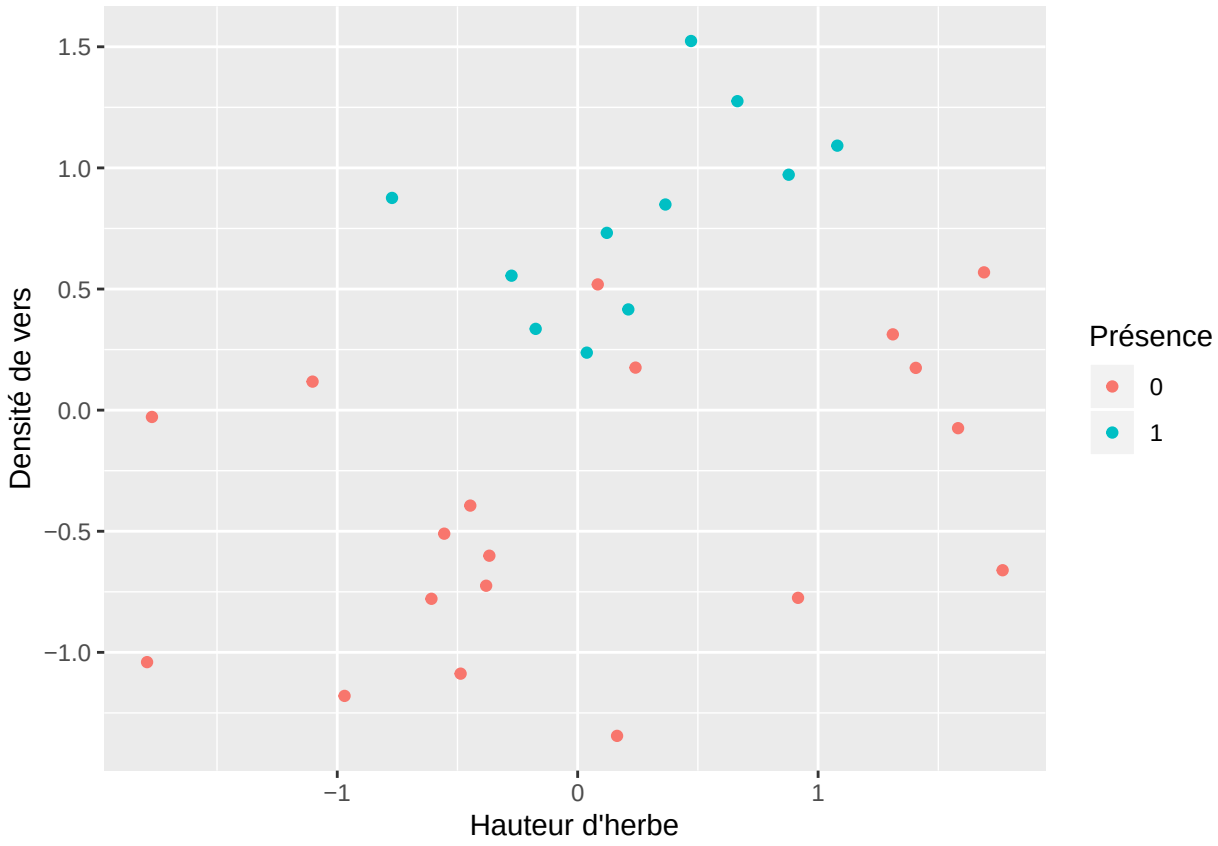
Une étude consiste en l'observation de la présence ou non de la linotte mélodieuse sur différents sites échantillonnés.

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).

On peut représenter la présence ou non d'oiseaux en fonctions des caractéristiques du site:

```
ggplot(donnees_presence, aes(x = haut_herbe, y = dens_vers)) +
  geom_point(aes(col = presence)) +
  labs(x = "Hauteur d'herbe",
       y = "Densité de vers",
       col = "Présence")
```



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

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

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.

Vraisemblance

Pour un vecteur d'observations $\mathbf{Y} = Y_{1:k}$, la vraisemblance

$$L(\mathbf{Y}|\theta) = \prod_{k=1}^{n} \underset{\text{Proba. présence}}{\phi(\mathbf{x}_k^T \theta)^{Y_k}} \times \underset{\text{Proba. absence}}{(1 - \phi(\mathbf{x}_k^T \theta))^{1-Y_k}}$$

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)

Posterior

Le posterior est donc donné par:

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

Cette densité n'est pas standard. Ainsi, on ne connaît pas ces caractéristiques associées et intéressantes (quantiles, espérance, variance). Ces différentes quantités peuvent cependant être approchées par méthode de Monte Carlo, si tant est qu'on soit capable de simuler selon cette loi.

Simulation d'échantillons a posteriori par acceptation rejet

Notre objectif est de simuler selon le posterior défini ci dessus. On va pour ce faire procéder par méthode d'acceptation rejet.

On remarque immédiatement qu'on ne peut pas utiliser l'acceptation rejet classique, car $\pi(\theta|\mathbf{Y})$ n'est connu qu'à une cnstante près!

Cependant, si on note

$$\tilde{\pi}(\theta|\mathbf{Y}) = \frac{1}{64\pi^2} \mathrm{e}^{-\frac{1}{8}\theta^T\theta} \prod_{k=1}^{n} \phi(\mathbf{x}_k^T\theta)^{Y_k}(1 - \phi(\mathbf{x}_k^T\theta))^{1-Y_k},$$

on peut utiliser la propriété vue en TD, qui dit qu'il suffit de connaître $\tilde{\pi}$ pour simuler selon π à partir d'acceptation rejet.

- **Choix de la densité de proposition**

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

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

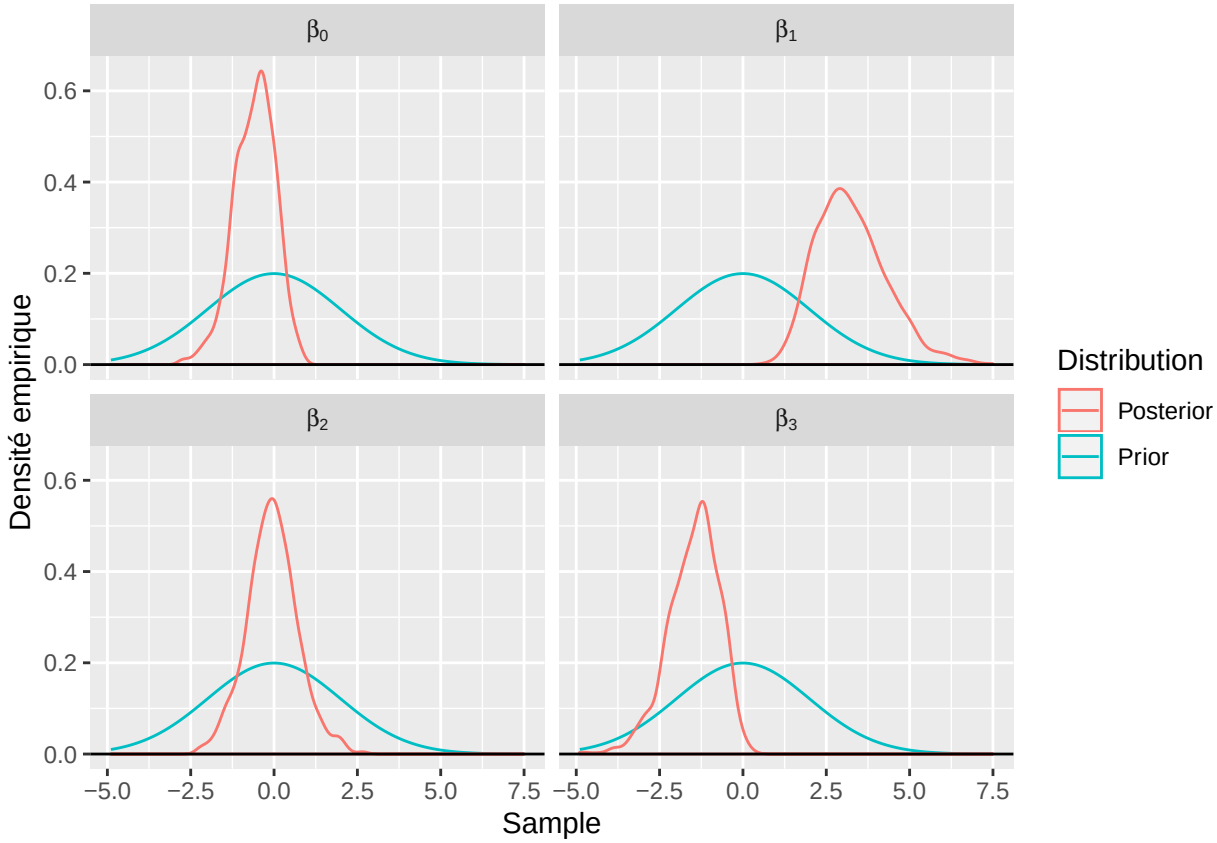
On a donc un majorant uniforme en θ et l'acceptation rejet suivant permet de tirer selon $\pi(\theta|\mathbf{Y})$:

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

Echantillon du posterior, loi a posteriori marginales et estimateurs bayésiens

- **Lois marginales**

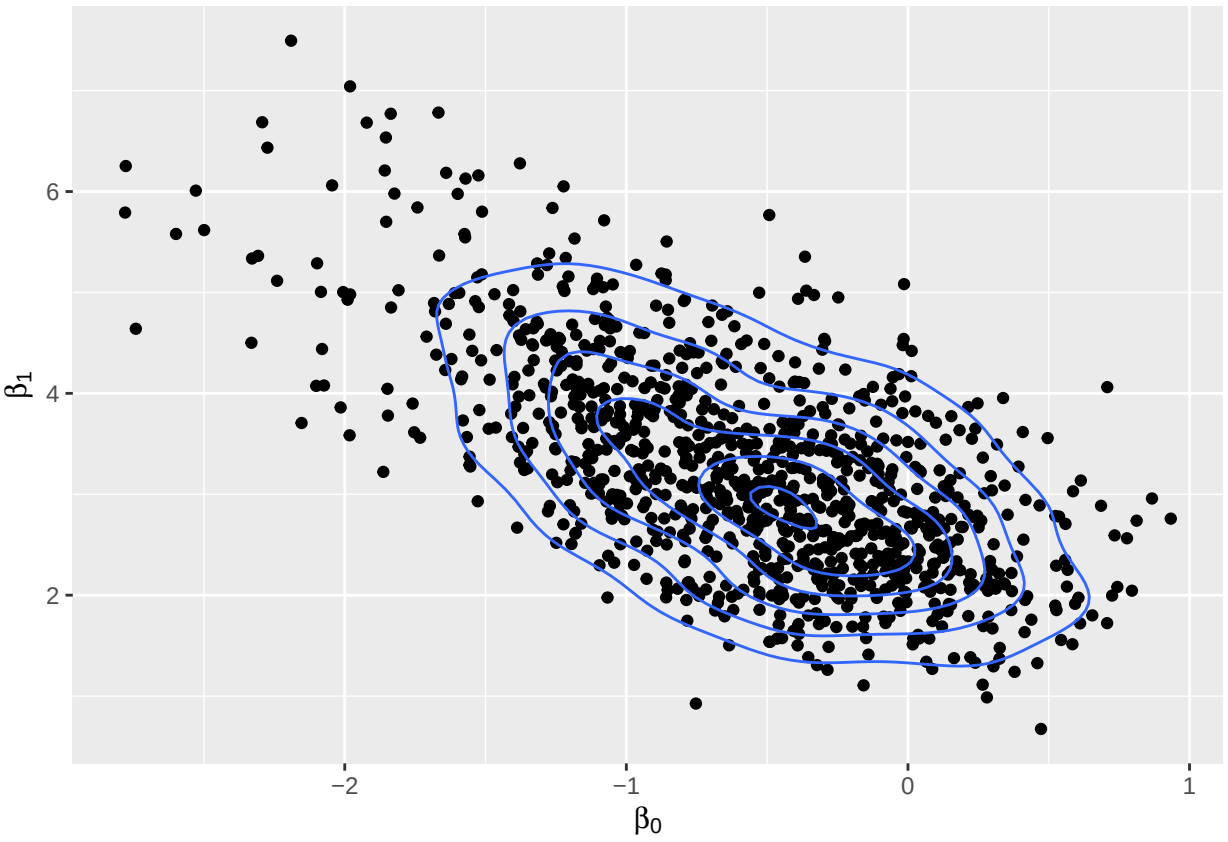
On effectue un tirage de taille $M = 1000$. On peut représenter la densité empirique de cet échantillon



On peut remarquer que les densités du posterior sont différentes du prior.

- **Loi jointe**

Il peut être intéressé de regarder la loi jointe obtenue par simulation. On pourra notamment remarquer qu'un prior indépendant sur les composantes n'entraîne pas une indépendance dans le posterior. Par exemple, on représente ici la loi du couple $(\beta_0, \beta_1|\mathbf{Y})$:



On voit qu'il existe une correlation linéaire négative entre ces deux variables aléatoires, qui étaient, a priori, supposées indépendantes.

- **Espérance a posteriori et intervalles de crédibilité**

Comme on sait simuler selon la loi cible, les espérances peuvent être approchées par méthode de Monte Carlo classique. On obtient ici une estimation del'espérance **a posteriori** ainsi qu'un intervalle de confiance a posterirori

Paramètre	Esperance a posteriori.	Quantile 2.5%	Quantile 97.5%
beta[0]	-0.586	-1.980890	0.5264341
beta[1]	3.250	1.538387	5.7005273
beta[2]	-0.051	-1.554202	1.5786397
beta[3]	-1.477	-3.171951	-0.2424187

Au delà de l'acceptation rejet

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

$$\frac{M}{\int L(\mathbf{Y}|\theta)\pi(\theta)\mathrm{d}\theta}$$

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.