Méthodes de Monte Carlo et inférence bayésienne

Introduction

Pierre Gloaguen

pierre.gloaguen@agroparistech.fr

Déroulé du cours

- 7 séances de 3h;
- Chaque début de séance consacré aux questions et à la présentation du cours;
- TD en autonomie ensuite (avec réponse aux questions);
- Evaluation en contrôle continu (exercices à rendre);

Objectifs de cette 1ère séance

- Introduction et motivation du sujet du cours;
- Principe des méthodes de Monte Carlo;
- Premiers exercices et implémentation sous `R`.

Introduction et motivations

Modèlisation statistique

- Formulation probabiliste d'un problème:
  - *On suppose que les données sont les réalisations de variables aléatoires telles que . . .*;
- Quantification de l'incertitude:
  - *d'après le modèle posé, la probabilité que tel événement arrive est de . . .*;
- Recours permanent au **calcul de probabilités**;

Exemple immédiat

- Test d'hypothèse: On veut tester une hypothèse H_0 contre une hypothèse H_1
  - *Ex: Comparaison de moyennes de deux échantillons Gaussiens*;
- On construit une statistique de test T;
- Décision binaire, pour un risque de première espèce α, on définit un zone de rejet $\mathcal{R}$, telle que

$$\mathbb{P}_{H_0}\left(T \in \mathcal{R}\right) = \alpha$$

Exemple immédiat

- Test d'hypothèse: On veut tester une hypothèse H_0 contre une hypothèse H_1
  - *Ex: Comparaison de moyennes de deux échantillons Gaussiens*;
- On construit une statistique de test T;
- Décision binaire, pour un risque de première espèce α, on définit un zone de rejet $\mathcal{R}$, telle que

$$\mathbb{P}_{H_0}\left(T \in \mathcal{R}\right) = \alpha$$

Rappel:

$$\mathbb{P}_{H_0}\left(T \in \mathcal{R}\right) = \mathbb{E}_{H_0}\left[\mathbf{1}_{T \in \mathcal{R}}\right]$$

- On doit donc être capable d'évaluer une espérance (ici, la probabilité d'évènements) sous H_0 pour construire notre test.

Exemple de modèle statistique: Biomasse d'une population de poisson

Ce qu'on connaît

Captures C

Observations scientifiques Y

Quantité d'intérêt

Biomasse de poisson X

Modèle probabiliste d'observation de la dynamique de population

$$X_{t+1} = \left(X_t + rX_t \left(1 - \frac{X_t}{K} \right) - C_t \right) \exp(\varepsilon_{t+1}), \text{ Biomasse cachée}$$

$$Y_t | X_t = qX_t \exp(\nu_t), \ C_t \text{ Observations}$$

$$\varepsilon_t \overset{i.i.d}{\sim} \mathcal{N}\left(-\sigma^2/2, \sigma^2\right), \quad \nu_t \overset{i.i.d}{\sim} \mathcal{N}\left(-\sigma^2_{\text{obs}}/2, \sigma^2_{\text{obs}}\right).$$

- X_t: Biomasse à l'année t (non observée);
- Y_t: Abundance observée à l'année t;
- C_t: Captures à l'année t;
- K: Capacité d'accueil du milieu (paramètre);
- r: Taux de croissance de la population (paramètre);
- $\sigma, \sigma_{\text{obs}}$: Paramètres de l'aléa;
- q: Détectabilité (paramètre);

Questions classiques d'inférence

Questions classiques d'inférence

- Pour un tel modèle, étant donnée une population initiale X_0, quelle est la moyenne attendue du nombre de poissons au bout de 10 ans si on se fixe une quantité de captures?

Questions classiques d'inférence

- Pour un tel modèle, étant donnée une population initiale X_0, quelle est la moyenne attendue du nombre de poissons au bout de 10 ans si on se fixe une quantité de captures?
  - $\mathbb{E}[X_{10}|X_0]$?

Questions classiques d'inférence

- Pour un tel modèle, étant donnée une population initiale X_0, quelle est la moyenne attendue du nombre de poissons au bout de 10 ans si on se fixe une quantité de captures?
  - $\mathbb{E}[X_{10}|X_0]$?
- Etant données des observations sur 10 années, et en supposant tous les paramètres connus, que puis je dire sur la quantité de poissons qu'il y avait durant ces 10 ans?

Questions classiques d'inférence

- Pour un tel modèle, étant donnée une population initiale X_0, quelle est la moyenne attendue du nombre de poissons au bout de 10 ans si on se fixe une quantité de captures?
  - $\mathbb{E}[X_{10}|X_0]$?
- Etant données des observations sur 10 années, et en supposant tous les paramètres connus, que puis je dire sur la quantité de poissons qu'il y avait durant ces 10 ans?
  - $\mathbb{E}[X_{0:10}|Y_{0:10}]$?

Questions classiques d'inférence

- Pour un tel modèle, étant donnée une population initiale X_0, quelle est la moyenne attendue du nombre de poissons au bout de 10 ans si on se fixe une quantité de captures?
  - $\mathbb{E}[X_{10}|X_0]$?
- Etant données des observations sur 10 années, et en supposant tous les paramètres connus, que puis je dire sur la quantité de poissons qu'il y avait durant ces 10 ans?
  - $\mathbb{E}[X_{0:10}|Y_{0:10}]$?
- Etant données des observations, que puis je dire sur la valeur des paramètres de dynamique de population?

Questions classiques d'inférence

- Pour un tel modèle, étant donnée une population initiale X_0, quelle est la moyenne attendue du nombre de poissons au bout de 10 ans si on se fixe une quantité de captures?
  - $\mathbb{E}[X_{10}|X_0]$?
- Etant données des observations sur 10 années, et en supposant tous les paramètres connus, que puis je dire sur la quantité de poissons qu'il y avait durant ces 10 ans?
  - $\mathbb{E}[X_{0:10}|Y_{0:10}]$?
- Etant données des observations, que puis je dire sur la valeur des paramètres de dynamique de population?
  - Inférence des paramètres:
  - Méthode des moments (nécessite un calcul d'espérance);

Questions classiques d'inférence

- Pour un tel modèle, étant donnée une population initiale X_0, quelle est la moyenne attendue du nombre de poissons au bout de 10 ans si on se fixe une quantité de captures?
  - $\mathbb{E}[X_{10}|X_0]$?
- Etant données des observations sur 10 années, et en supposant tous les paramètres connus, que puis je dire sur la quantité de poissons qu'il y avait durant ces 10 ans?
  - $\mathbb{E}[X_{0:10}|Y_{0:10}]$?
- Etant données des observations, que puis je dire sur la valeur des paramètres de dynamique de population?
  - Inférence des paramètres:
  - Méthode des moments (nécessite un calcul d'espérance);
  - Méthode du maximum de vraisemblance (nécessite ici un calcul d'espérance);

Questions classiques d'inférence

- Pour un tel modèle, étant donnée une population initiale X_0, quelle est la moyenne attendue du nombre de poissons au bout de 10 ans si on se fixe une quantité de captures?
  - $\mathbb{E}[X_{10}|X_0]$?
- Etant données des observations sur 10 années, et en supposant tous les paramètres connus, que puis je dire sur la quantité de poissons qu'il y avait durant ces 10 ans?
  - $\mathbb{E}[X_{0:10}|Y_{0:10}]$?
- Etant données des observations, que puis je dire sur la valeur des paramètres de dynamique de population?
  - Inférence des paramètres:
  - Méthode des moments (nécessite un calcul d'espérance);
  - Méthode du maximum de vraisemblance (nécessite ici un calcul d'espérance);
  - Estimateur Bayésien: Nécessite un calcul d'espérance.

Questions classiques d'inférence

- Pour un tel modèle, étant donnée une population initiale X_0, quelle est la moyenne attendue du nombre de poissons au bout de 10 ans si on se fixe une quantité de captures?
  - $\mathbb{E}[X_{10}|X_0]$?
- Etant données des observations sur 10 années, et en supposant tous les paramètres connus, que puis je dire sur la quantité de poissons qu'il y avait durant ces 10 ans?
  - $\mathbb{E}[X_{0:10}|Y_{0:10}]$?
- Etant données des observations, que puis je dire sur la valeur des paramètres de dynamique de population?
  - Inférence des paramètres:
  - Méthode des moments (nécessite un calcul d'espérance);
  - Méthode du maximum de vraisemblance (nécessite ici un calcul d'espérance);
  - Estimateur Bayésien: Nécessite un calcul d'espérance.

Ces espérances n'ont, en général, pas d'expressions directes!

Méthodes de Monte Carlo

Méthodes de Monte Carlo

- **But:** Approcher des espérances (intégrales) en utilisant des simulations probabilistes;

Méthodes de Monte Carlo

- **But:** Approcher des espérances (intégrales) en utilisant des simulations probabilistes;
- **Idée:** La loi des grands nombres! La moyenne empirique d'une variable aléatoire va tendre, si on répète l'expérience, vers la moyenne théorique.

Exemple

On dispose d'un dé à 6 faces et seulement de ce dé. Comment peut on essayer de savoir s'il est biaisé?

- Si on lance le dé suffisament de fois, on obtient une information;

Exemple

On dispose d'un dé à 6 faces et seulement de ce dé. Comment peut on essayer de savoir s'il est biaisé?

- Si on lance le dé suffisament de fois, on obtient une information;
- Combien de fois faut il lancer le dé pour avoir une idée précise?

Exemple

On dispose d'un dé à 6 faces et seulement de ce dé. Comment peut on essayer de savoir s'il est biaisé?

- Si on lance le dé suffisament de fois, on obtient une information;
- Combien de fois faut il lancer le dé pour avoir une idée précise?
- À quel point peut être confiant en notre réponse?

Exemple

On dispose d'un dé à 6 faces et seulement de ce dé. Comment peut on essayer de savoir s'il est biaisé?

- Si on lance le dé suffisament de fois, on obtient une information;
- Combien de fois faut il lancer le dé pour avoir une idée précise?
- À quel point peut être confiant en notre réponse?
- Encore faut il savoir lancer le dé!

Programme du cours

- Présentation formelle des méthodes de Monte Carlo pour le calcul d'intégrales;
- Application directe en statistique classique (évaluation d'une probabilité, aide à la décision);

Programme du cours

- Présentation formelle des méthodes de Monte Carlo pour le calcul d'intégrales;
- Application directe en statistique classique (évaluation d'une probabilité, aide à la décision);
- Comment peut on simuler des variables aléatoires génériques (avec un ordinateur)?

Programme du cours

- Présentation formelle des méthodes de Monte Carlo pour le calcul d'intégrales;
- Application directe en statistique classique (évaluation d'une probabilité, aide à la décision);
- Comment peut on simuler des variables aléatoires génériques (avec un ordinateur)?
- Une méthode d'inférence dépendante de la simulation: l'inférence bayésienne;

Programme du cours

- Présentation formelle des méthodes de Monte Carlo pour le calcul d'intégrales;
- Application directe en statistique classique (évaluation d'une probabilité, aide à la décision);
- Comment peut on simuler des variables aléatoires génériques (avec un ordinateur)?
- Une méthode d'inférence dépendante de la simulation: l'inférence bayésienne;
- Une extension nécessaire, les Méthodes de Monte Carlo par chaîne de Markov (MCMC).

Prérequis

- Résulats statistiques asymptotiques:
  - Loi des grands nombres, théorème central limite, lemme de Slutsky, Delta méthode.
- Chaînes de Markov:
  - Loi de transition, irréductibilité, périodicité, mesure invariante ...
- Logiciel `R`
  - Logiciel `R` installé ainsi que l'IDE `Rstudio`.
  - Connaissance minimale du langage (boucles, fonctions, graphiques de base...).

Principe des méthodes de Monte Carlo

Pierre Gloaguen

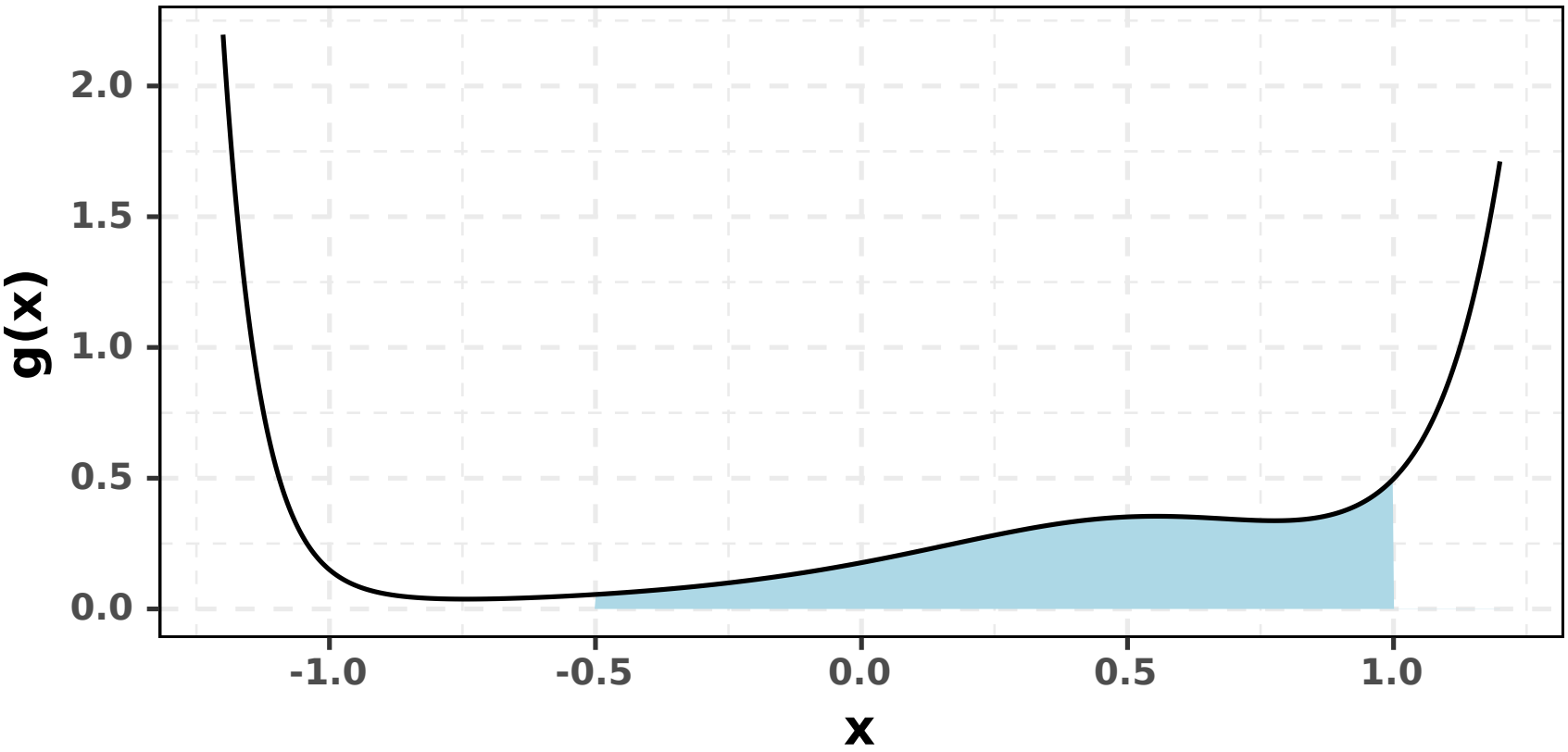
22/03/2020

Exemple introductif (1)

Soit g une fonction sur $\mathbb{R}$ et $a < b$ deux réels.

Supposons que l'on souhaite calculer une intégrale (finie) du type

$$I = \int_a^b g(x)\mathrm{d}x$$



Exemple introductif (2)

$$I = \int_a^b g(x)\mathrm{d}x$$

Exemple introductif (2)

$$
\begin{aligned}
I &= \int_a^b g(x)\mathrm{d}x \\
&= \int_{\mathbb{R}} \overbrace{(b-a)g(x)}^{:=\varphi(x)} \frac{\mathbf{1}_{a\leq x\leq b}}{b-a}\mathrm{d}x
\end{aligned}
$$

Exemple introductif (2)

$$
\begin{aligned}
I &= \int_a^b g(x)\mathrm{d}x \\
&= \int_{\mathbb{R}} \overbrace{(b-a)g(x)}^{:=\varphi(x)} \frac{\mathbf{1}_{a\leq x\leq b}}{b-a}\mathrm{d}x \\
&= \mathbb{E}[\varphi(X)], \text{ où } X \sim \mathcal{U}[a,b].
\end{aligned}
$$

Exemple introductif (2)

$$\begin{aligned} I &= \int_a^b g(x)\mathrm{d}x \\ &= \int_{\mathbb{R}} \overbrace{(b-a)g(x)}^{:=\varphi(x)} \frac{\mathbf{1}_{a\leq x\leq b}}{b-a}\mathrm{d}x \\ &= \mathbb{E}[\varphi(X)], \text{ où } X \sim \mathcal{U}[a,b]. \end{aligned}$$

Estimateur de Monte Carlo

On fixe un entier $M > 0$. On simule un échantillon $X_1, \ldots, X_M$ selon $X \sim \mathcal{U}[a,b]$, on pose alors l'estimateur:

$$\hat{I}_M = \frac{1}{M}\sum_{k=1}^{M} \varphi(X_k)$$

Exemple introductif (2)

$$\begin{aligned} I &= \int_a^b g(x)\mathrm{d}x \\ &= \int_{\mathbb{R}} \overbrace{(b-a)g(x)}^{:=\varphi(x)} \frac{\mathbf{1}_{a\leq x\leq b}}{b-a}\mathrm{d}x \\ &= \mathbb{E}[\varphi(X)], \text{ où } X \sim \mathcal{U}[a,b]. \end{aligned}$$

Estimateur de Monte Carlo

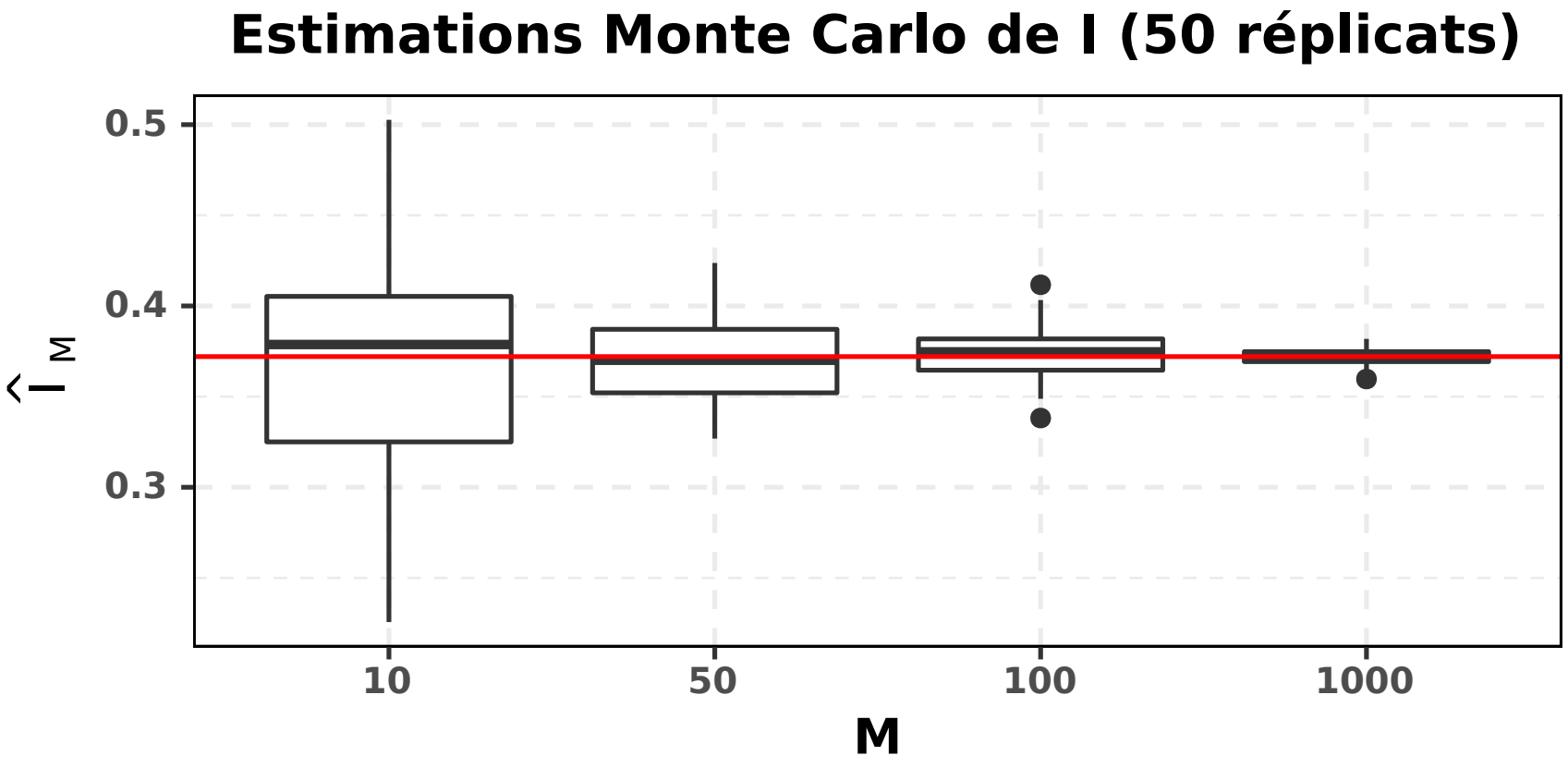
On fixe un entier $M > 0$. On simule un échantillon $X_1, \ldots, X_M$ selon $X \sim \mathcal{U}[a,b]$, on pose alors l'estimateur:

$$\hat{I}_M = \frac{1}{M}\sum_{k=1}^{M} \varphi(X_k)$$

Remarques

- M est appelé **effort de Monte Carlo**;
- On suppose pour le moment qu'on **sait simuler selon** $\mathcal{U}[a,b]$;
- L'estimateur de I est une **variable aléatoire**.

Exemple introductif (3)



Cas générique

On veut calculer une intégrale

$$I = \int_{\mathbb{R}^d} \varphi(x) f(x) \mathrm{d}x$$

où f est une fonction positive, telle que $\int_{\mathbb{R}^d} f(x)\mathrm{d}x = 1$, alors on se sert du fait que

$$I = \mathbb{E}[\varphi(X)]$$

où X est une variable aléatoire de densité f.

Estimateur de Monte Carlo

On fixe un entier $M > 0$. On simule un échantillon $X_1, \ldots, X_N$ selon $X \sim f(\cdot)$, on pose alors l'estimateur:

$$\hat{I}_N = \frac{1}{M} \sum_{k=1}^{M} \varphi(X_k)$$

Pourquoi?

Les $\varphi(X_1), \ldots, \varphi(X_N)$ sont des variables aléatoires i.i.d. d'espérance $\mathbb{E}[\varphi(X)]$ finie, avec $X \sim f(\cdot)$.

Pourquoi?

Les $\varphi(X_1), \ldots, \varphi(X_N)$ sont des variables aléatoires i.i.d. d'espérance $\mathbb{E}[\varphi(X)]$ finie, avec $X \sim f(\cdot)$.

Loi des grands nombres:

$$\frac{\varphi(X_1) + \cdots + \varphi(X_N)}{M} \xrightarrow[M \to \infty]{\text{p.s.}} \mathbb{E}[\varphi(X)]$$

Propriétés

Sans biais

$$\mathbb{E}[\hat{I}_N] = \frac{1}{M}\sum_{k=1}^{M}\mathbb{E}[\varphi(X_k)] \overset{\text{id. distrib}}{=} \mathbb{E}[\varphi(X)] = I$$

Propriétés

Sans biais

$$\mathbb{E}[\hat{I}_N] = \frac{1}{M}\sum_{k=1}^{M}\mathbb{E}[\varphi(X_k)] \overset{\text{id. distrib}}{=} \mathbb{E}[\varphi(X)] = I$$

Variance Si $\mathbb{V}[\varphi(X)] < \infty$

$$\mathbb{V}[\hat{I}_N] \overset{\text{ind.}}{=} \frac{1}{M^2}\sum_{k=1}^{M}\mathbb{V}[\varphi(X_k)] \overset{\text{id. distrib}}{=} \frac{1}{M}\mathbb{V}[\varphi(X)]$$

Propriétés

Sans biais

$$\mathbb{E}[\hat{I}_N] = \frac{1}{M}\sum_{k=1}^{M}\mathbb{E}[\varphi(X_k)] \overset{\text{id. distrib}}{=} \mathbb{E}[\varphi(X)] = I$$

Variance Si $\mathbb{V}[\varphi(X)] < \infty$

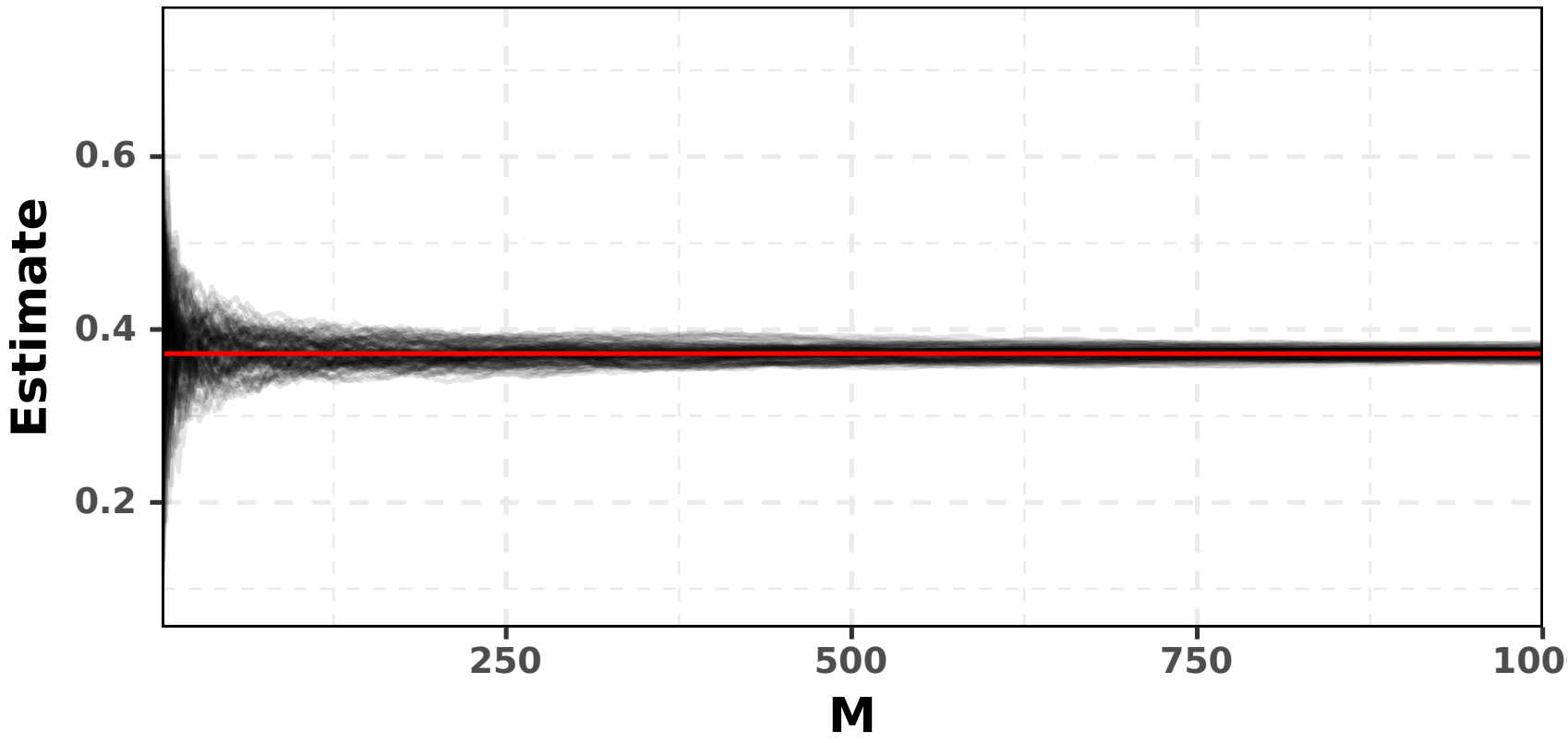
$$\mathbb{V}[\hat{I}_N] \overset{\text{ind.}}{=} \frac{1}{M^2}\sum_{k=1}^{M}\mathbb{V}[\varphi(X_k)] \overset{\text{id. distrib}}{=} \frac{1}{M}\mathbb{V}[\varphi(X)]$$

Loi La loi des grands nombres nous donne la loi asymptotique

$$\sqrt{M}\left(\hat{I}_N - I\right) \underset{M\to\infty}{\overset{\text{Loi}}{\longrightarrow}} \mathcal{N}(0, \mathbb{V}[\varphi(X)])$$

Loi de l'estimateur

- 100 réplicats d'échantillons Monte Carlo de taille $M = 1000$.



Intervalle de confiance

On note

$$\sigma^2 = \mathbb{V}[\varphi(X)].$$

Ainsi, en notant z_α le quantile d'ordre $\alpha \in]0,1[$ de la loi $\mathcal{N}(0,1)$, si on définit l'intervalle aléatoire

$$J_M = \left[\hat{I}_M - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{M}}; \hat{I}_M + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{M}}\right]$$

Alors,

$$\mathbb{P}(J_M \ni I) \underset{M\to\infty}{\longrightarrow} 1-\alpha$$

J_M est donc un intervalle de confiance asymptotique au niveau 1 - α pour la valeur de I.

Intervalle de confiance (2)

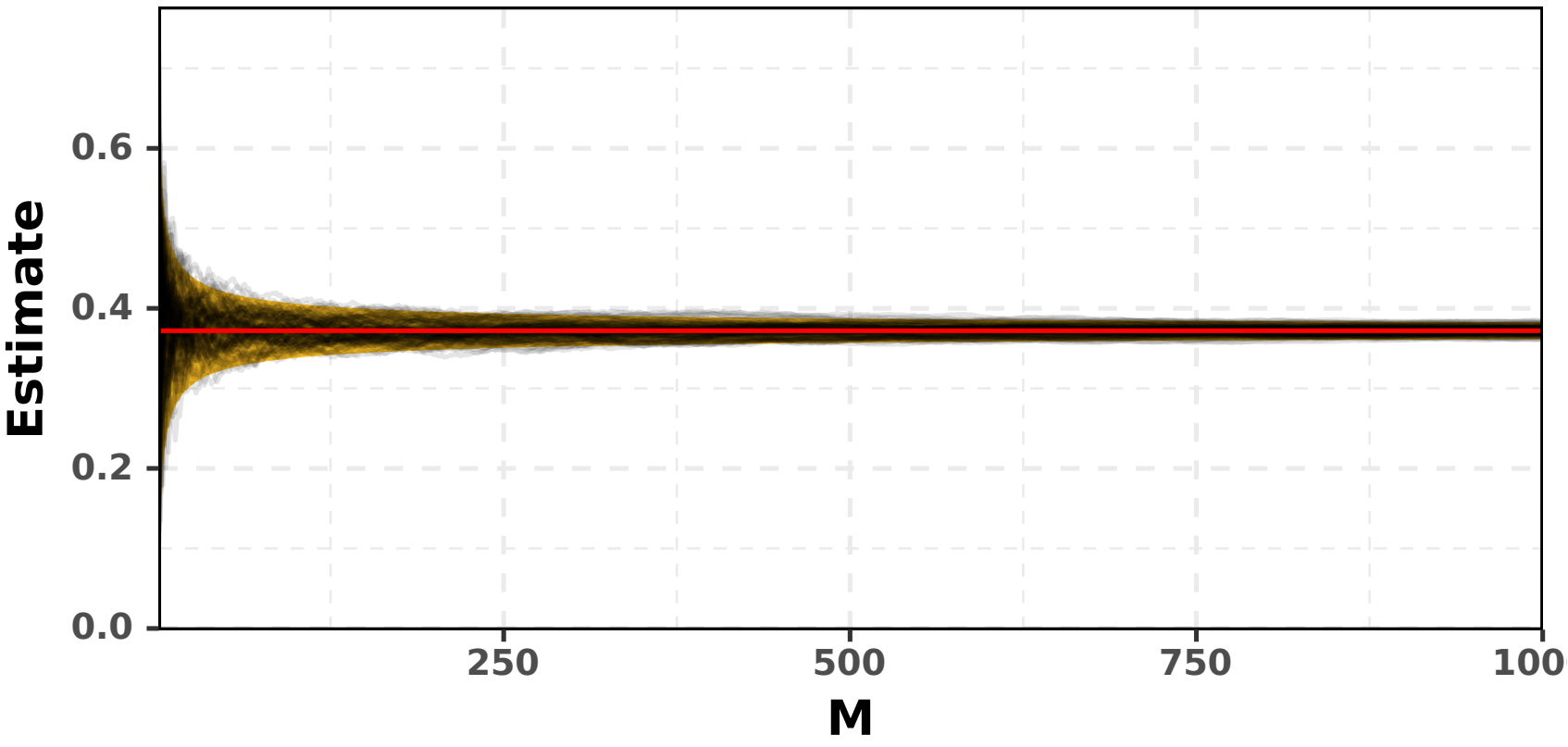
En pratique cependant, cet intervalle n'est pas calculable car σ^2 est inconnu

On dispose cependant d'un estimateur consistant de σ^2 donné par

$$\hat{\sigma}_M^2 = \frac{1}{M} \sum_{k=1}^{M} \left(\varphi(X_k) - \hat{I}_M \right)^2$$

- Dans l'expression précédente, on remplace σ^2 par son estimateur.
- Le lemme de Slutski nous assure que les propriétés de J_M restent vraies.

Intervalle de confiance (3)



Comparaison avec l'intégration numérique

L'objectif présenté ici est de calculer, en dimension d, une intégrale:

$$\int_{\mathbb{R}^d} g(x)\mathrm{d}x$$

- Calcul possible par méthode numérique (méthode des cubes).

Comparaison avec l'intégration numérique

L'objectif présenté ici est de calculer, en dimension d, une intégrale:

$$\int_{\mathbb{R}^d} g(x)\mathrm{d}x$$

- Calcul possible par méthode numérique (méthode des cubes).

- **Intégration numérique:** Pour une fonction g de classe C^s, l'erreur est de l'ordre $\frac{1}{M^{\frac{s}{d}}}$ (où M est le nombre d'évaluations de la fonction).
  - Il faut connaître la régularité de g!
  - L'erreur augmente avec la dimension.

Comparaison avec l'intégration numérique

L'objectif présenté ici est de calculer, en dimension d, une intégrale:

$$\int_{\mathbb{R}^d} g(x)\mathrm{d}x$$

- Calcul possible par méthode numérique (méthode des cubes).

- **Intégration numérique:** Pour une fonction g de classe C^s, l'erreur est de l'ordre $\frac{1}{M^{\frac{s}{d}}}$ (où M est le nombre d'évaluations de la fonction).
  - Il faut connaître la régularité de g!
  - L'erreur augmente avec la dimension.
- **Méthodes Monte Carlo:** Pour les méthodes de Monte Carlo, l'écart type de l'erreur est de l'ordre $\frac{1}{M^{\frac{1}{2}}}$.
  - Indépendamment de la régularité de g!
  - Indépendamment de la dimension!

Ainsi, ces méthodes deviennent vite avantageuses quand d est grand.

Echantillonnage préférentiel

Pierre Gloaguen

02/04/2020

Annonces:

- Premier rendu pour le 13 Avril (exo5 du TD1)
- À faire en binome (à compléter sur le lien envoyé par mail)
- Corrections des exos 1 à 4 du TD1 mis en ligne
- On regardera ces corrections et vos questions aujourd'hui

Rappel de l'épisode précédent

- Principes des méthodes de Monte Carlo;
- Approximation d'intégrales (espérances) par simulation;
- Fonctionne grâce à la loi des grands nombres, IC grâce au TCL.

Objectif du cours

- Présentation de l'échantillonnage préférentiel
  - Extension naturelle des méthodes de MC simples;
  - Améliore l'efficacité dans certains cas;
  - Utile quand on ne sait pas simuler selon une loi donnée;

Objectif du cours

- Présentation de l'échantillonnage préférentiel
  - Extension naturelle des méthodes de MC simples;
  - Améliore l'efficacité dans certains cas;
  - Utile quand on ne sait pas simuler selon une loi donnée;
- Motivation: cas de Monte Carlo problématiques;
- Définition;
- Propriétés: Analyse de la variance de ce nouvel estimateur;
- Illustration.

Echantillonnage préférentiel

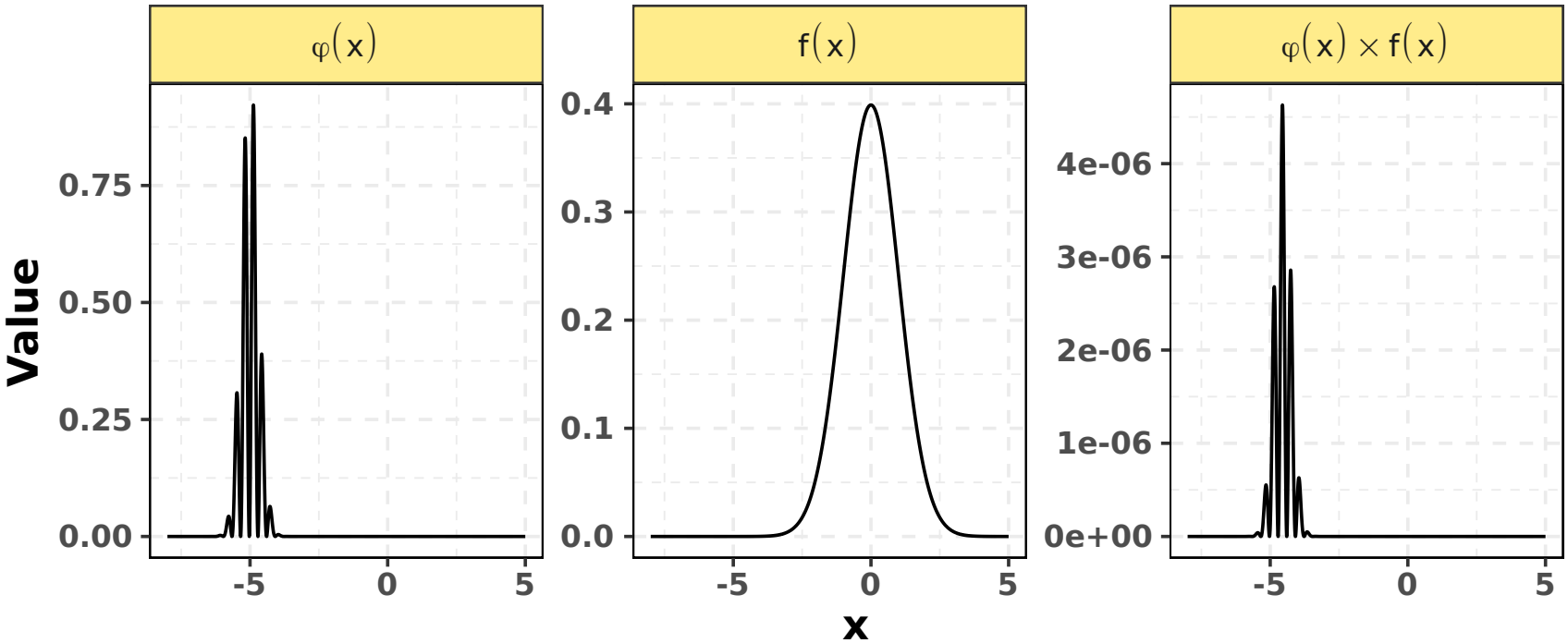
Exemple problématique:

On veut calculer

$$I = \mathbb{E}\left[\varphi(X)\right] = \int_{\mathbb{R}} \varphi(x) f(x) \mathrm{d}x$$

où $X \sim \mathcal{N}(0, 1)$ (de densité $f(x)$) et

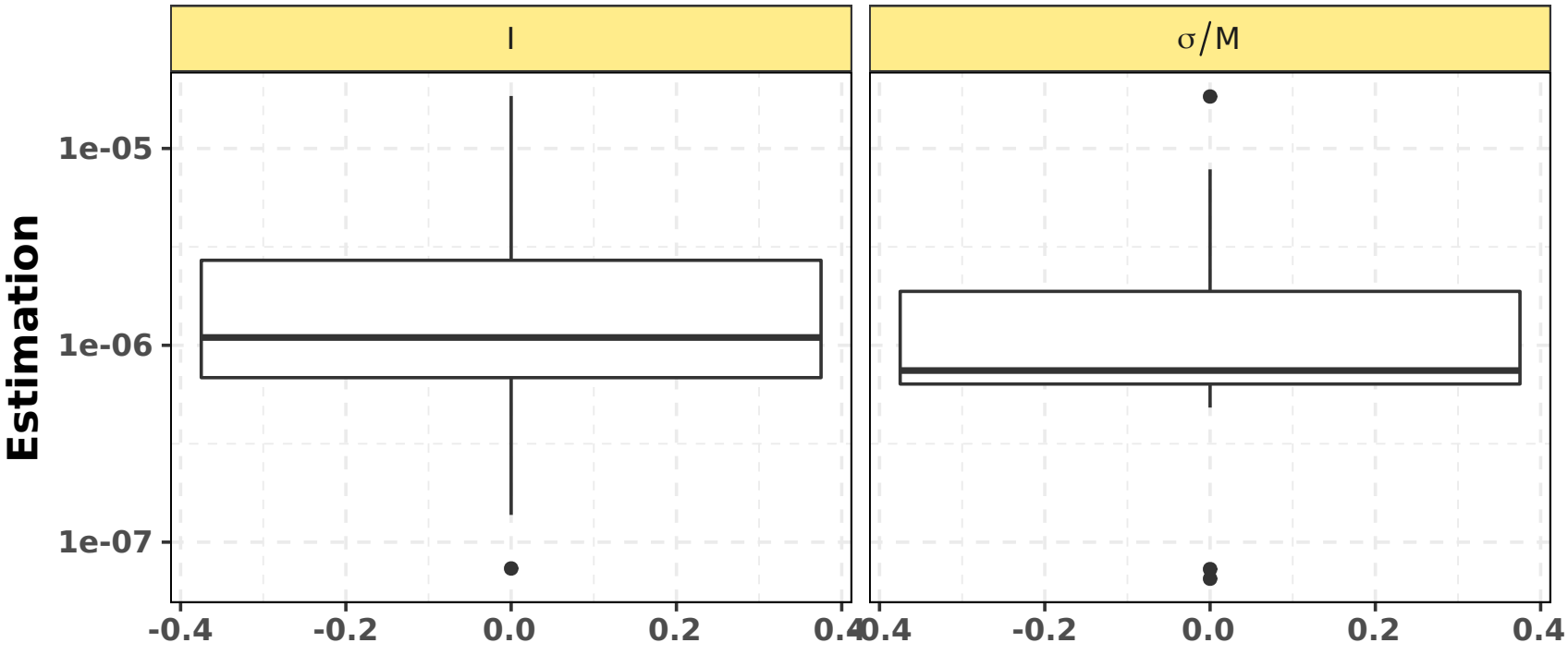
$$\varphi(x) = \sin^2(x) \mathrm{e}^{-5(x-5)^2}$$



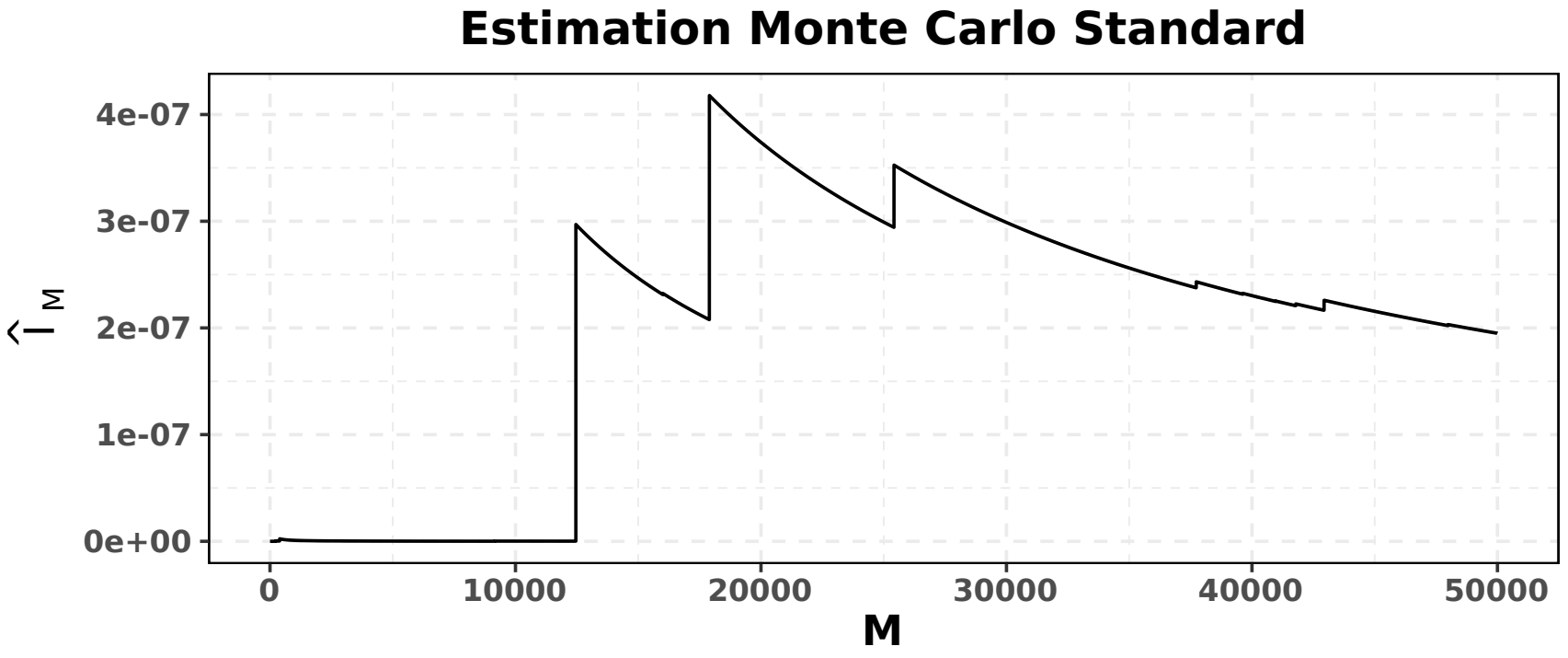
Estimateur Monte Carlo naturel

On tire $M = 50000$ points dans une loi $\mathcal{N}(0, 1)$, et on calcule la moyenne empirique.

Résultats obtenus:



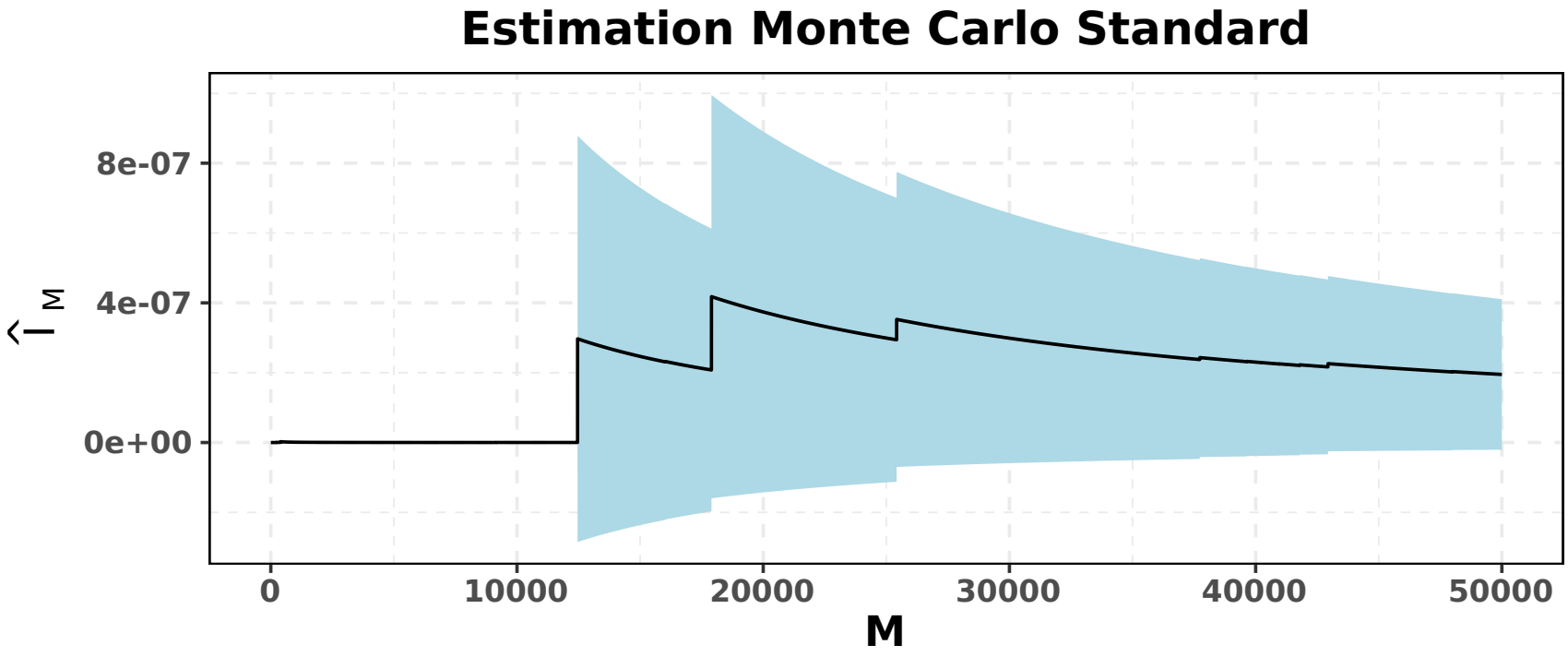
Une trajectoire d'estimation



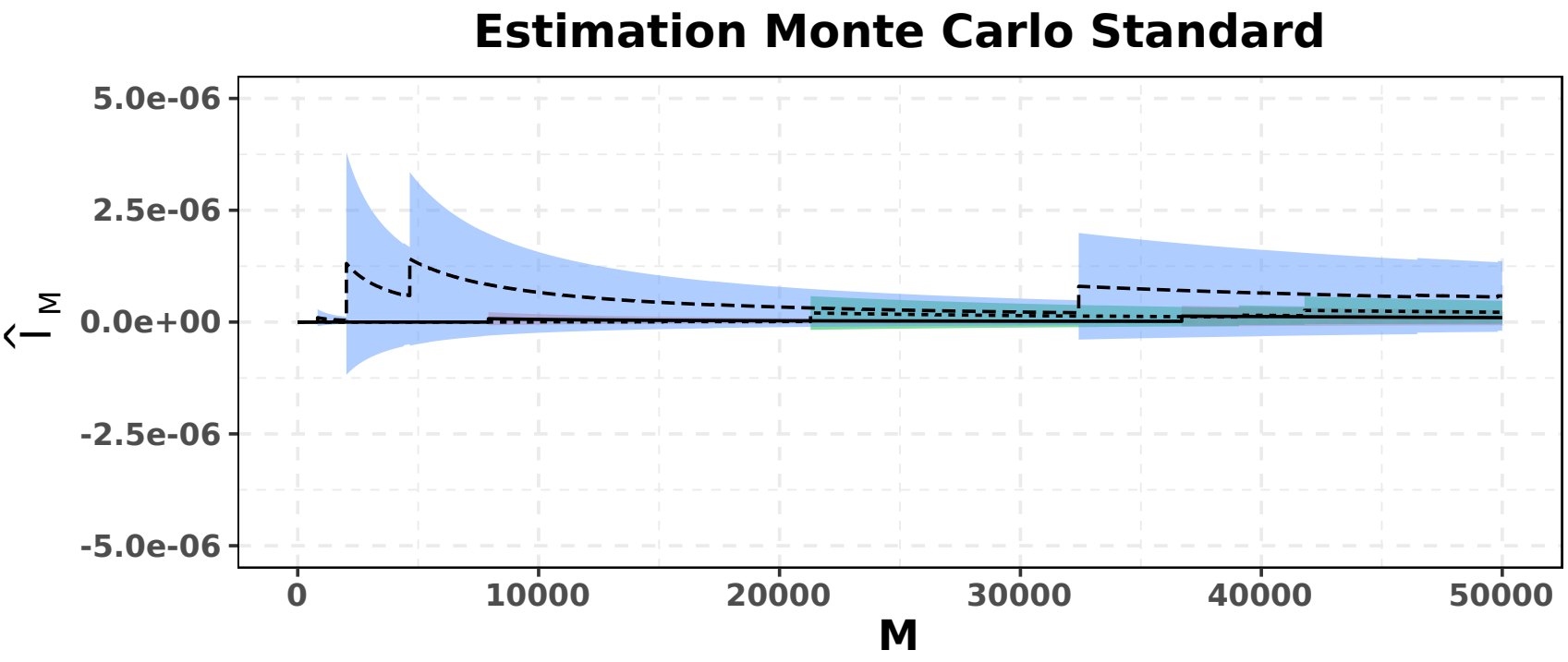
L'estimation est très instable.

Estimateur Monte Carlo naturel

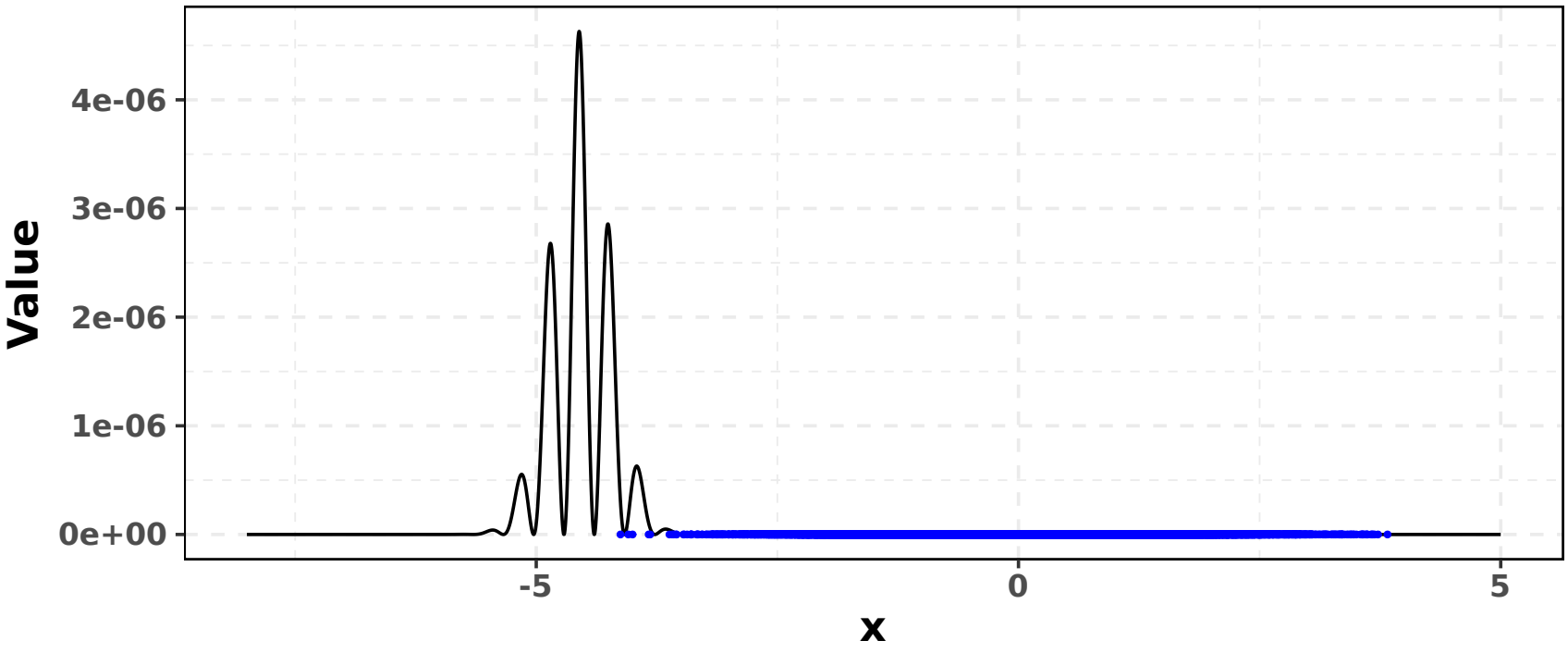
Estimation de la variance et intervalle de confiance asymptotique:



Manque de chance?



Origine du problème



On échantillonne loin des régions importantes!

Echantillonnage préférentiel

On cherche à estimer une intégrale du type:

$$I = \mathbb{E}_f[\varphi(X)] = \int_{\mathcal{D}_f} \varphi(x) f(x) \mathrm{d}x$$

où $\mathcal{D}_f \subset \mathbb{R}^d$, et f est une densité de probabilité sur $\mathcal{D}_f$ (donc $f(x) = 0$ pour $x \notin \mathcal{D}_f$) et X une v.a. de loi f.

Echantillonnage préférentiel

On cherche à estimer une intégrale du type:

$$I = \mathbb{E}_f[\varphi(X)] = \int_{\mathcal{D}_f} \varphi(x) f(x) \mathrm{d}x$$

où $\mathcal{D}_f \subset \mathbb{R}^d$, et f est une densité de probabilité sur $\mathcal{D}_f$ (donc $f(x) = 0$ pour $x \notin \mathcal{D}_f$) et X une v.a. de loi f.

Soit g une densité de probabilité sur $\mathcal{D}_g \supseteq \mathcal{D}_f$ telle que $x \in \mathcal{D}_f \Rightarrow g(x) > 0$ et Y une variable aléatoire de loi g, alors:

$$I = \int_{\mathcal{D}_f} \varphi(x) \frac{f(x)}{g(x)} g(x) \mathrm{d}x$$

Echantillonnage préférentiel

On cherche à estimer une intégrale du type:

$$I = \mathbb{E}_f[\varphi(X)] = \int_{\mathcal{D}_f} \varphi(x) f(x) \mathrm{d}x$$

où $\mathcal{D}_f \subset \mathbb{R}^d$, et f est une densité de probabilité sur $\mathcal{D}_f$ (donc $f(x) = 0$ pour $x \notin \mathcal{D}_f$) et X une v.a. de loi f.

Soit g une densité de probabilité sur $\mathcal{D}_g \supseteq \mathcal{D}_f$ telle que $x \in \mathcal{D}_f \Rightarrow g(x) > 0$ et Y une variable aléatoire de loi g, alors:

$$\begin{aligned} I &= \int_{\mathcal{D}_f} \varphi(x) \frac{f(x)}{g(x)} g(x) \mathrm{d}x \\ &= \int_{\mathcal{D}_g} \varphi(x) \frac{f(x)}{g(x)} g(x) \mathrm{d}x \qquad \text{as } x \notin \mathcal{D}_f \Rightarrow f(x) = 0 \end{aligned}$$

Echantillonnage préférentiel

On cherche à estimer une intégrale du type:

$$I = \mathbb{E}_f[\varphi(X)] = \int_{\mathcal{D}_f} \varphi(x) f(x) \mathrm{d}x$$

où $\mathcal{D}_f \subset \mathbb{R}^d$, et f est une densité de probabilité sur $\mathcal{D}_f$ (donc $f(x) = 0$ pour $x \notin \mathcal{D}_f$) et X une v.a. de loi f.

Soit g une densité de probabilité sur $\mathcal{D}_g \supseteq \mathcal{D}_f$ telle que $x \in \mathcal{D}_f \Rightarrow g(x) > 0$ et Y une variable aléatoire de loi g, alors:

$$\begin{aligned} I &= \int_{\mathcal{D}_f} \varphi(x) \frac{f(x)}{g(x)} g(x) \mathrm{d}x \\ &= \int_{\mathcal{D}_g} \varphi(x) \frac{f(x)}{g(x)} g(x) \mathrm{d}x \qquad \text{as } x \notin \mathcal{D}_f \Rightarrow f(x) = 0 \\ &= \mathbb{E}_g \left[\varphi(Y) \frac{f(Y)}{g(Y)} \right] \end{aligned}$$

Echantillonnage preferentiel

Comme estimateur de I, on peut ainsi proposer l'estimateur:

$$\hat{I}_M^{IS} = \frac{1}{M} \sum_{k=1}^{M} \frac{f(Y_k)}{g(Y_k)} \varphi(Y_k) = \frac{1}{M} \sum_{k=1}^{M} W(Y_k) \varphi(Y_k)$$

où $Y_1, \ldots, Y_M$ est un échantillon i.i.d. de variables aléatoires sur $\mathbb{R}^d$ de densité g.

Echantillonnage preferentiel

Comme estimateur de I, on peut ainsi proposer l'estimateur:

$$\hat{I}_M^{IS} = \frac{1}{M} \sum_{k=1}^{M} \frac{f(Y_k)}{g(Y_k)} \varphi(Y_k) = \frac{1}{M} \sum_{k=1}^{M} W(Y_k) \varphi(Y_k)$$

où $Y_1, \ldots, Y_M$ est un échantillon i.i.d. de variables aléatoires sur $\mathbb{R}^d$ de densité g.

Remarques:

- Comme $Y_k \sim g$, $g(Y_k)$ est p.s. $\neq 0$

Echantillonnage preferentiel

Comme estimateur de I, on peut ainsi proposer l'estimateur:

$$\hat{I}_M^{IS} = \frac{1}{M}\sum_{k=1}^{M}\frac{f(Y_k)}{g(Y_k)}\varphi(Y_k) = \frac{1}{M}\sum_{k=1}^{M} W(Y_k)\varphi(Y_k)$$

où $Y_1, \ldots, Y_M$ est un échantillon i.i.d. de variables aléatoires sur $\mathbb{R}^d$ de densité g.

Remarques:

- Comme $Y_k \sim g$, $g(Y_k)$ est p.s. $\neq 0$
- La variable aléatoire $W(Y_k) = \frac{f(Y_k)}{g(Y_k)}$ est appelée **poids d'importance** de Y_k.

Echantillonnage preferentiel

Comme estimateur de I, on peut ainsi proposer l'estimateur:

$$\hat{I}_M^{IS} = \frac{1}{M}\sum_{k=1}^{M}\frac{f(Y_k)}{g(Y_k)}\varphi(Y_k) = \frac{1}{M}\sum_{k=1}^{M} W(Y_k)\varphi(Y_k)$$

où $Y_1, \ldots, Y_M$ est un échantillon i.i.d. de variables aléatoires sur $\mathbb{R}^d$ de densité g.

Remarques:

- Comme $Y_k \sim g$, $g(Y_k)$ est p.s. $\neq 0$
- La variable aléatoire $W(Y_k) = \frac{f(Y_k)}{g(Y_k)}$ est appelée **poids d'importance** de Y_k.
- Quand $f = g$, on a l'estimateur MC standard (et chaque poids vaut 1).

Quel intérêt?

On peut choisir g afin d'échantillonner dans les zones d'importance!

Biais:

$$\mathbb{E}_g[\hat{I}_M^{IS}] = \frac{1}{M}\sum_{k=1}^{M}\int_{\mathcal{D}_g}\frac{f(x)}{g(x)}\varphi(x)g(x)\mathrm{d}x$$
$$= \int_{\mathcal{D}_f} f(x)\varphi(x)\mathrm{d}x = I$$

Donc, cet estimateur reste sans biais

Quel intérêt?

On peut choisir g afin d'échantillonner dans les zones d'importance!

Biais:

$$\mathbb{E}_g[\hat{I}_M^{IS}] = \frac{1}{M}\sum_{k=1}^{M}\int_{\mathcal{D}_g}\frac{f(x)}{g(x)}\varphi(x)g(x)\mathrm{d}x$$
$$= \int_{\mathcal{D}_f} f(x)\varphi(x)\mathrm{d}x = I$$

Donc, cet estimateur reste sans biais

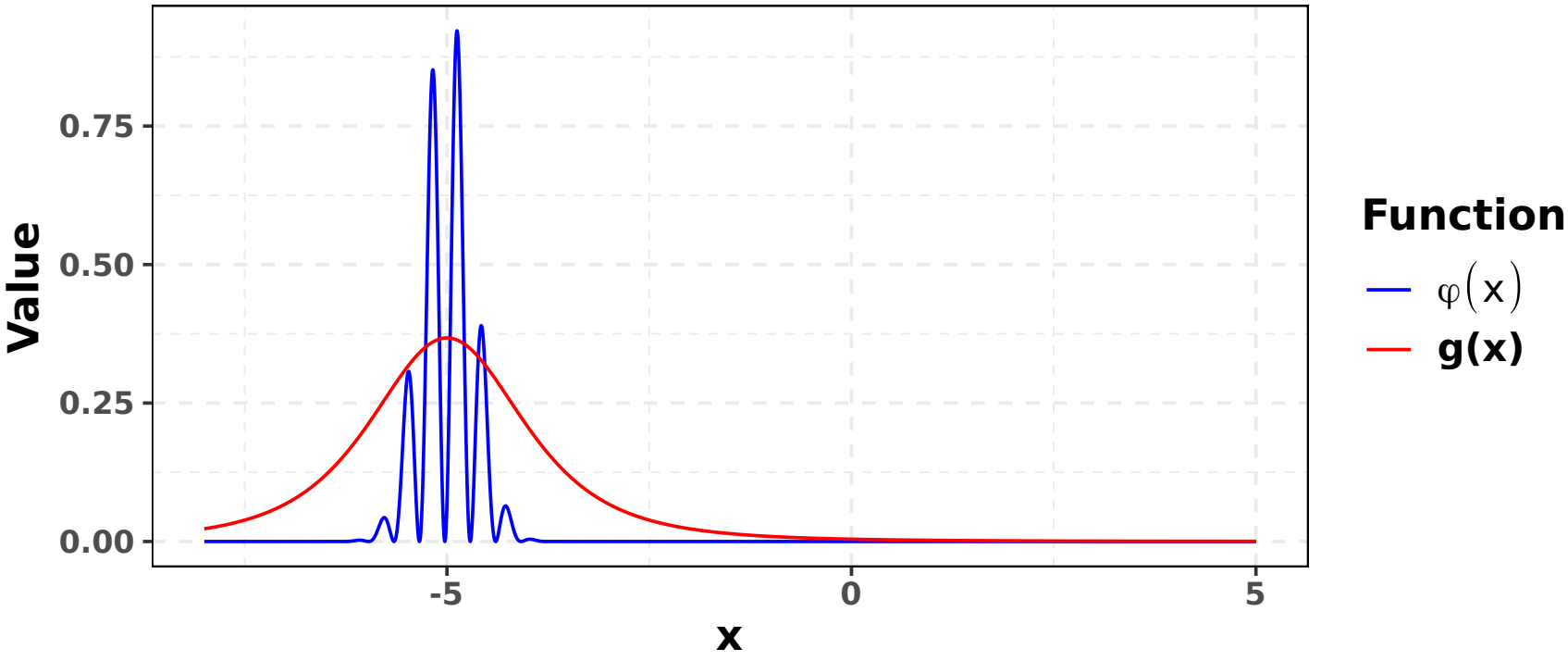
Variance:

$$\mathbb{V}_g[\hat{I}_1^{IS}] = \left[\mathbb{E}_g[(W(Y)\varphi(Y))^2] - I^2\right] = \int_{\mathcal{D}_g}\frac{(\varphi(y)f(y) - Ig(y))^2}{g(y)}\mathrm{d}y$$

- La variance peut être très réduite si $g(y) \approx \varphi(y)f(y)$!
- Ceci peut guider le choix de g!

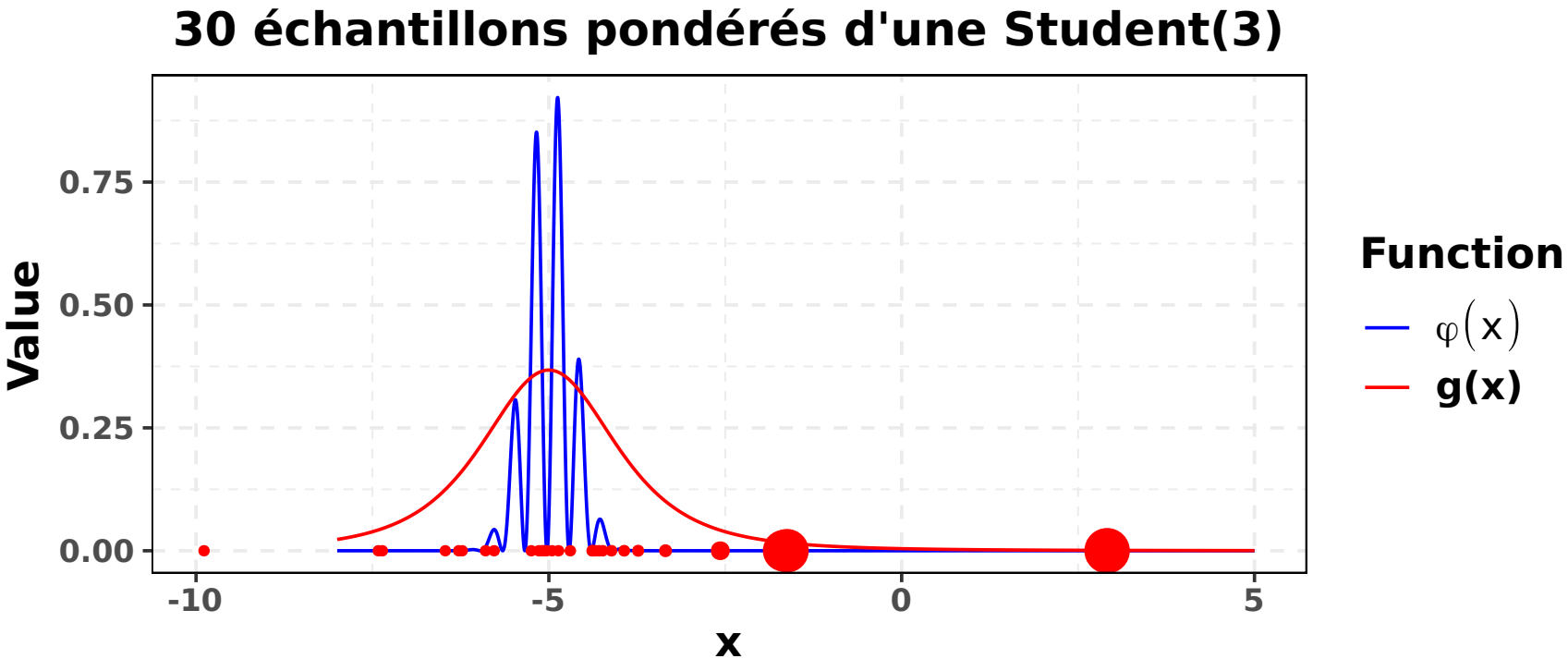
Exemple

$g(x)$ est la densité d'une loi de Student $\mathcal{T}(3)$.

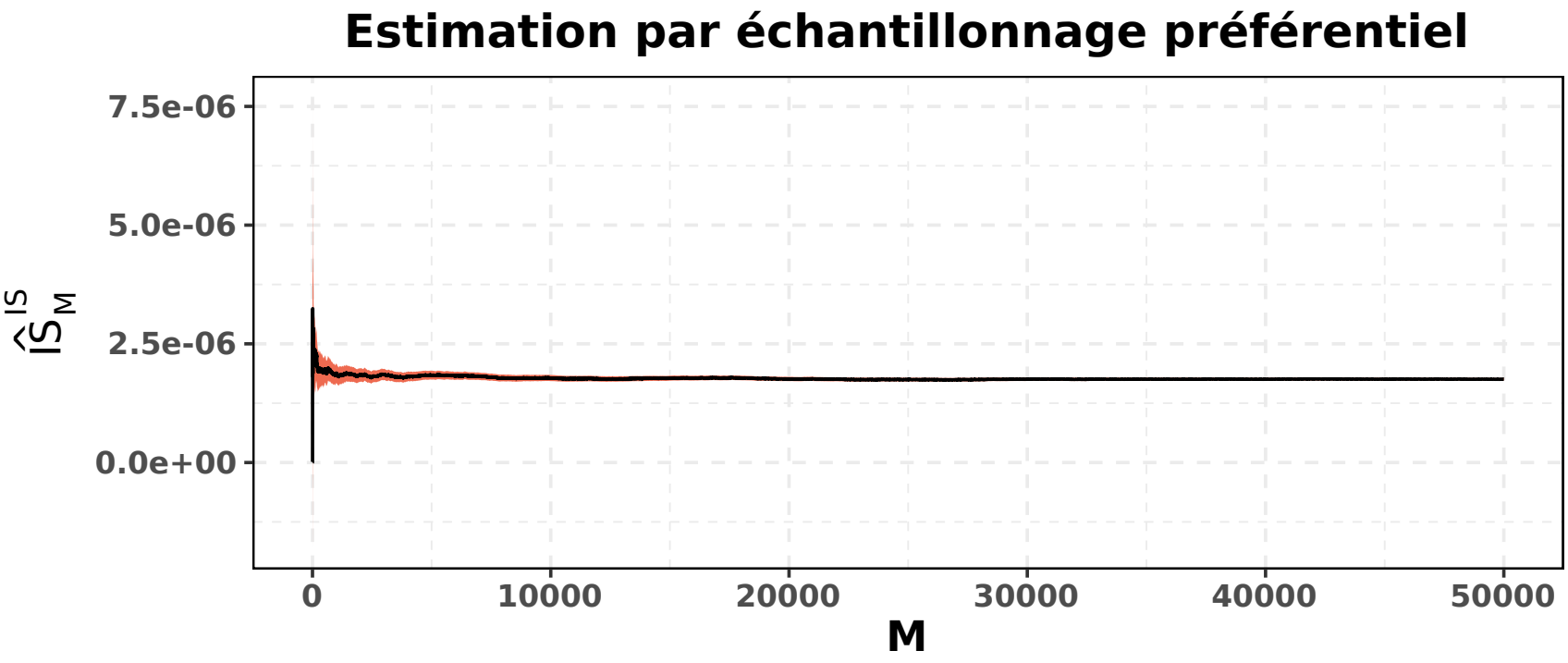


Exemple

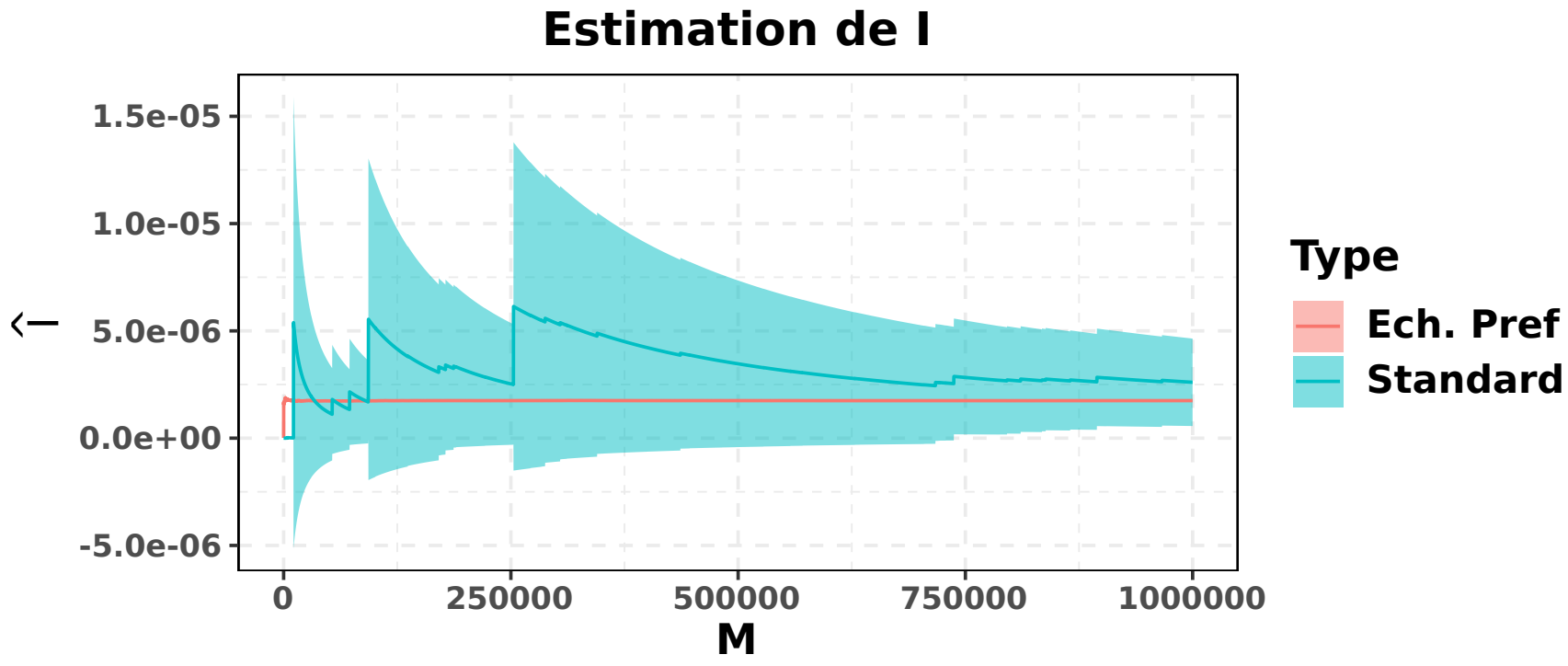
```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please
## "none")` instead.
```



Estimation de I (et IC asymptotique)



Comparaison sur cet exemple



Echantillonnage préférentiel

Avantages

- Très utile pour l'estimation de quantités petites (probabilités d'évènements rares).

Echantillonnage préférentiel

Avantages

- Très utile pour l'estimation de quantités petites (probabilités d'évènements rares).
- Peut amener à une forte réduction de variance

Echantillonnage préférentiel

Avantages

- Très utile pour l'estimation de quantités petites (probabilités d'évènements rares).
- Peut amener à une forte réduction de variance
- Peut aussi être utiliser quand on ne sait pas simuler selon f !

Echantillonnage préférentiel

Avantages

- Très utile pour l'estimation de quantités petites (probabilités d'évènements rares).
- Peut amener à une forte réduction de variance
- Peut aussi être utiliser quand on ne sait pas simuler selon f !

Attention!

- Nécessite le choix de g ! Pas toujours évident (notamment en grande dimension)!

Echantillonnage préférentiel

Avantages

- Très utile pour l'estimation de quantités petites (probabilités d'évènements rares).
- Peut amener à une forte réduction de variance
- Peut aussi être utiliser quand on ne sait pas simuler selon f !

Attention!

- Nécessite le choix de g! Pas toujours évident (notamment en grande dimension)!
- Un mauvais g peut amener à un estimateur de variance infinie! (voir TD).
- Un bon choix de g est souvent "problème dépendant", (ne conviendra que pour $\mathbb{E}[\varphi(X)]$ pour un φ spécifique).

Echantillonnage préférentiel normalisé

Problématique

Objectif, calculer:

$$I = \mathbb{E}[\varphi(X)] = \int_{\mathcal{D}_f} \varphi(x) f(x) d(x)$$

Supposons que f ne soit connue qu'à une constante près:

$$f(x) = \frac{\overbrace{f^{(u)}(x)}^{\text{Connu}}}{\underbrace{\int_{\mathcal{D}_f} f^{(u)}(z)\mathrm{d}z}_{\text{Inconnu}}}.$$

- Pour une densité de proposition g, on ne peut plus calculer le poids d'importance!

Problématique

Objectif, calculer:

$$I = \mathbb{E}[\varphi(X)] = \int_{\mathcal{D}_f} \varphi(x) f(x) d(x)$$

Supposons que f ne soit connue qu'à une constante près:

$$f(x) = \frac{\overbrace{f^{(u)}(x)}^{\text{Connu}}}{\underbrace{\int_{\mathcal{D}_f} f^{(u)}(z)\mathrm{d}z}_{\text{Inconnu}}}.$$

- Pour une densité de proposition g, on ne peut plus calculer le poids d'importance!
- Ce cas est en pratique très fréquent en inférence bayésienne!

Echantillonnage préférentiel normalisé

Si on dispose d'une loi de proposition g et de $Y_1, \ldots, Y_M$ tirés indépendemment selon g, alors l'estimateur:

$$\hat{I}_M^{IS,u} = \sum_{k=1}^{M} \frac{f^{(u)}(Y_k)/g(Y_k)}{\sum_{\ell=1}^{M} f^{(u)}(Y_\ell)/g(Y_\ell)} \varphi(Y_k)$$

est un estimateur consistant (convergence en proba.) de I.

Echantillonnage préférentiel normalisé

Si on dispose d'une loi de proposition g et de $Y_1, \ldots, Y_M$ tirés indépendemment selon g, alors l'estimateur:

$$\hat{I}_M^{IS,u} = \sum_{k=1}^{M} \frac{f^{(u)}(Y_k)/g(Y_k)}{\sum_{\ell=1}^{M} f^{(u)}(Y_\ell)/g(Y_\ell)} \varphi(Y_k)$$

est un estimateur consistant (convergence en proba.) de I.

- Estimateur biaisé pour M petit;
- Peut amener à un estimateur de variance plus faible que l'échantillonnage préférentiel classique.

Autres méthodes de réduction de variance

- Conditionnement
- Variables de contrôles
- Quasi Monte Carlo

Voir références dans le poly.

Simulations de variables aléatoires

Pierre Gloaguen

07/04/2020

Annonces:

- Premier rendu pour le 13 Avril à midi (exo5 du TD1)
- À faire en binome et à rendre sur Ecampus
- TD sur échantillonnage préférentiel en ligne sur ma page.
- Rendre l'exo 3 pour le 21 Avril au soir.

Rappel des précédents

- Nécessité en statistiques d'évaluer des espérances;
- Principes des méthodes de Monte Carlo:
  - Classiques et échantillonnage préférentiel.
- Approximation d'intégrales (espérances) par simulation de variables aléatoires.
- Fonctionne grâce à la loi des grands nombres, IC grâce au TCL.

Rappel des précédents

- Nécessité en statistiques d'évaluer des espérances;
- Principes des méthodes de Monte Carlo:
  - Classiques et échantillonnage préférentiel.
- Approximation d'intégrales (espérances) par simulation de variables aléatoires.
- Fonctionne grâce à la loi des grands nombres, IC grâce au TCL.
- Comment simule t'on ces lois?
  - Lois usuelles implémentées dans `R` (ou autre...)
  - Comment est ce fait?
  - Pour des lois non usuelles, comment faire?

Objectif du cours

- Comment simuler une loi uniforme continue avec un ordinateur?
  - Générateurs pseudo aléatoires;
- Comment simuler des lois génériques à partir de lois uniformes?
  - Méthode d'inversion;
- Comment simuler des lois non classique à partir de lois simulables?
  - Méthode d'acceptation rejet.

Générateurs pseudo aléatoires

Simulation de variables aléatoires par ordinateur

Objectif: Simuler une suite de variables aléatoires $X_1, \ldots, X_M$ telles qu'elles soient:

- Distribuées selon une même loi donnée:
- Mutuellement indépendantes.

Simulation de variables aléatoires par ordinateur

Objectif: Simuler une suite de variables aléatoires $X_1, \ldots, X_M$ telles qu'elles soient:

- Distribuées selon une même loi donnée:
- Mutuellement indépendantes.
- Une telle simulation est faite selon un algorithme **déterministe**;

Simulation de variables aléatoires par ordinateur

Objectif: Simuler une suite de variables aléatoires $X_1, \ldots, X_M$ telles qu'elles soient:

- Distribuées selon une même loi donnée:
- Mutuellement indépendantes.
- Une telle simulation est faite selon un algorithme **déterministe**;
- À l'heure actuelle, un ordinateur ne peut que **mimer l'aléa**;

Simulation de variables aléatoires par ordinateur

Objectif: Simuler une suite de variables aléatoires $X_1, \ldots, X_M$ telles qu'elles soient:

- Distribuées selon une même loi donnée:
- Mutuellement indépendantes.
- Une telle simulation est faite selon un algorithme **déterministe**;
- À l'heure actuelle, un ordinateur ne peut que **mimer l'aléa**;
- **Mimer l'aléa:** Pour un échantillon de loi donnée:
  - Passer les tests statistiques usuels d'adéquations (test du χ^2, test de Kolmogorov Smirnoff);
  - Passer les tests d'indépendances usuels (test du χ^2, test de corrélation linéaire, ...).

Générateur de la loi uniforme $\mathcal{U}[0,1]$

- En pratique, la loi “atomique” est la loi $\mathcal{U}[0,1]$;

Générateur de la loi uniforme $\mathcal{U}[0,1]$

- En pratique, la loi “atomique” est la loi $\mathcal{U}[0,1]$;
- Si on sait simuler selon cette loi, par différentes méthodes on pourra se ramener aux autres.

Générateur de la loi uniforme $\mathcal{U}[0,1]$

- En pratique, la loi "atomique" est la loi $\mathcal{U}[0,1]$;
- Si on sait simuler selon cette loi, par différentes méthodes on pourra se ramener aux autres.
- Nécessité d'un algorithme permettant de mimer un échantillon i.i.d. de loi $\mathcal{U}[0,1]$.

Algorithme de congruence linéaire

Algorithme basé sur 4 données initiales choisies par l'utilisateur:

- Un entier $m > 0$, appelé *module*;
- Un entier $0 < a < m$ appelé *multiplicateur*;
- Un entier $0 \leq c < m$ appelé *incrément*;
- Un entier $0 \leq x_0 < m$ appelé *graine*.

Algorithme de congruence linéaire

Algorithme basé sur 4 données initiales choisies par l'utilisateur:

- Un entier $m > 0$, appelé *module*;
- Un entier $0 < a < m$ appelé *multiplicateur*;
- Un entier $0 \leq c < m$ appelé *incrément*;
- Un entier $0 \leq x_0 < m$ appelé *graine*.

On créera alors une suite de nombres $x_1, \ldots x_n$ en utilisant la relation de récurrence

$$x_k = (ax_{k-1} + c) \text{ modulo } m.$$

Algorithme de congruence linéaire

Algorithme basé sur 4 données initiales choisies par l'utilisateur:

- Un entier $m > 0$, appelé *module*;
- Un entier $0 < a < m$ appelé *multiplicateur*;
- Un entier $0 \leq c < m$ appelé *incrément*;
- Un entier $0 \leq x_0 < m$ appelé *graine*.

On créera alors une suite de nombres $x_1, \ldots x_n$ en utilisant la relation de récurrence

$$x_k = (ax_{k-1} + c) \text{ modulo } m.$$

On définit enfin les nombres $u_1, \ldots, u_n$ dans l'intervalle $[0, 1]$:

$$u_k = \frac{x_k}{m}, \ 1 \leq k \leq n.$$

Algorithme de congruence linéaire

```r
mon_runif <- function(n , a, m, c, x0){
  echantillon <- rep(NA, n + 1)
  # %% est l'opérateur modulo
  x_vals[1] <- (x0 %% m) # Initialisation
  for(k in 2:(n + 1)){ # Iteration
    x_vals[k] <- (a * x_vals[k - 1] + c) %% m
  }
  u_vals <- x_vals / m # Mise entre 0 et 1
  return(u_vals[-1])
  # On ne retourne pas la graine
}
```

Choix de a, m, c, x0

- À x_0 fixé, la séquence obtenue est **toujours la même**.
  - En pratique, elle n'est pas demandée à l'utilisateur, mais obtenue en interne.
  - Exemple: nombre de millisecondes (modulo m) écoulé depuis le 1er Janvier 1970.

Choix de a, m, c, x0

- À x_0 fixé, la séquence obtenue est **toujours la même**.
  - En pratique, elle n'est pas demandée à l'utilisateur, mais obtenue en interne.
  - Exemple: nombre de millisecondes (modulo m) écoulé depuis le 1er Janvier 1970.
- Doit couvrir "tout" [0, 1]:
  - $\Rightarrow$ m grand;

Choix de a, m, c, x0

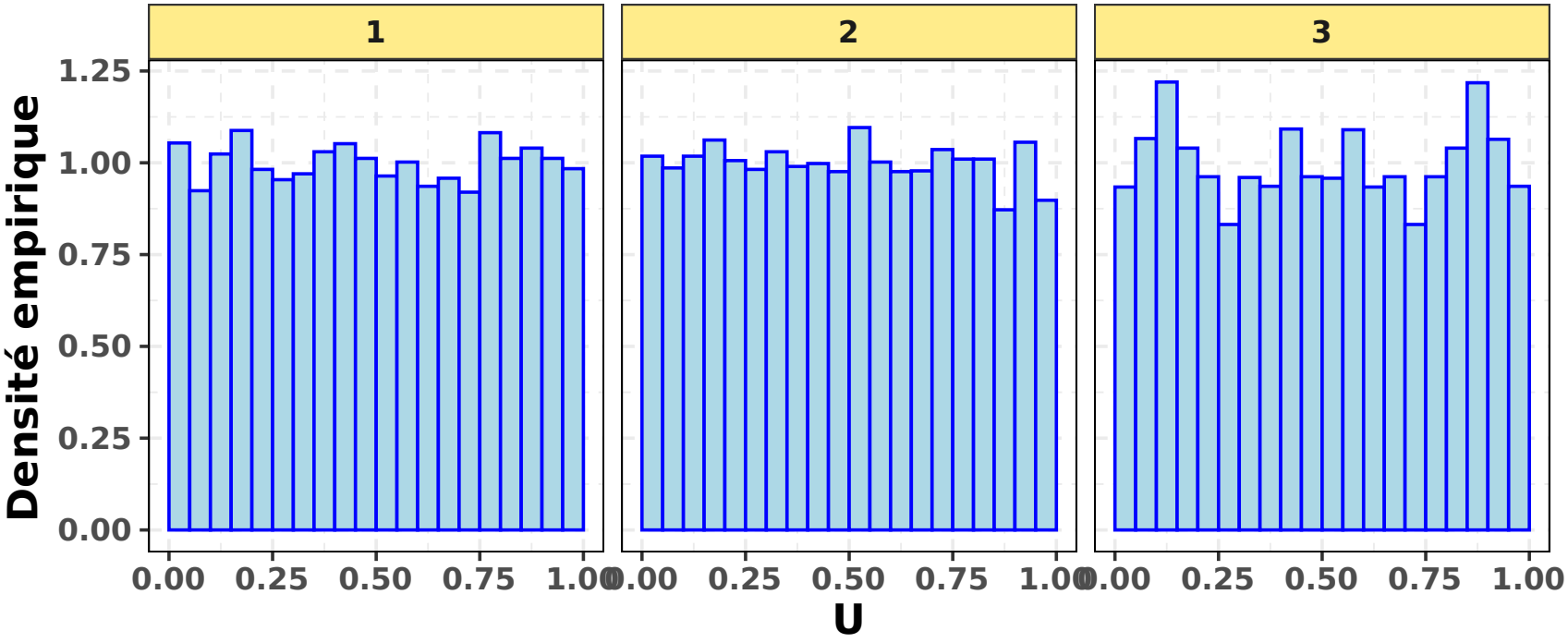
- À x_0 fixé, la séquence obtenue est **toujours la même**.
  - En pratique, elle n'est pas demandée à l'utilisateur, mais obtenue en interne.
  - Exemple: nombre de millisecondes (modulo m) écoulé depuis le 1er Janvier 1970.
- Doit couvrir "tout" [0, 1]:
  - $\Rightarrow$ m grand;
- La suite est nécessairement **périodique**!
  - On veut une période "invisible" (mimer l'indépendance);
  - $\Rightarrow$ a grand **et** relativement premier à m.

Choix de a, m, c, x0

- À x_0 fixé, la séquence obtenue est **toujours la même**.
  - En pratique, elle n'est pas demandée à l'utilisateur, mais obtenue en interne.
  - Exemple: nombre de millisecondes (modulo m) écoulé depuis le 1er Janvier 1970.
- Doit couvrir "tout" [0, 1]:
  - $\Rightarrow$ m grand;
- La suite est nécessairement **périodique**!
  - On veut une période "invisible" (mimer l'indépendance);
  - $\Rightarrow$ a grand **et** relativement premier à m.
- Considération algorithmiques sur le modulo.
- Voir références poly.

Choix important (distribution)

3 jeux de paramètres (voir TD), donnant 3 suites de 10000 valeurs entre 0 et 1



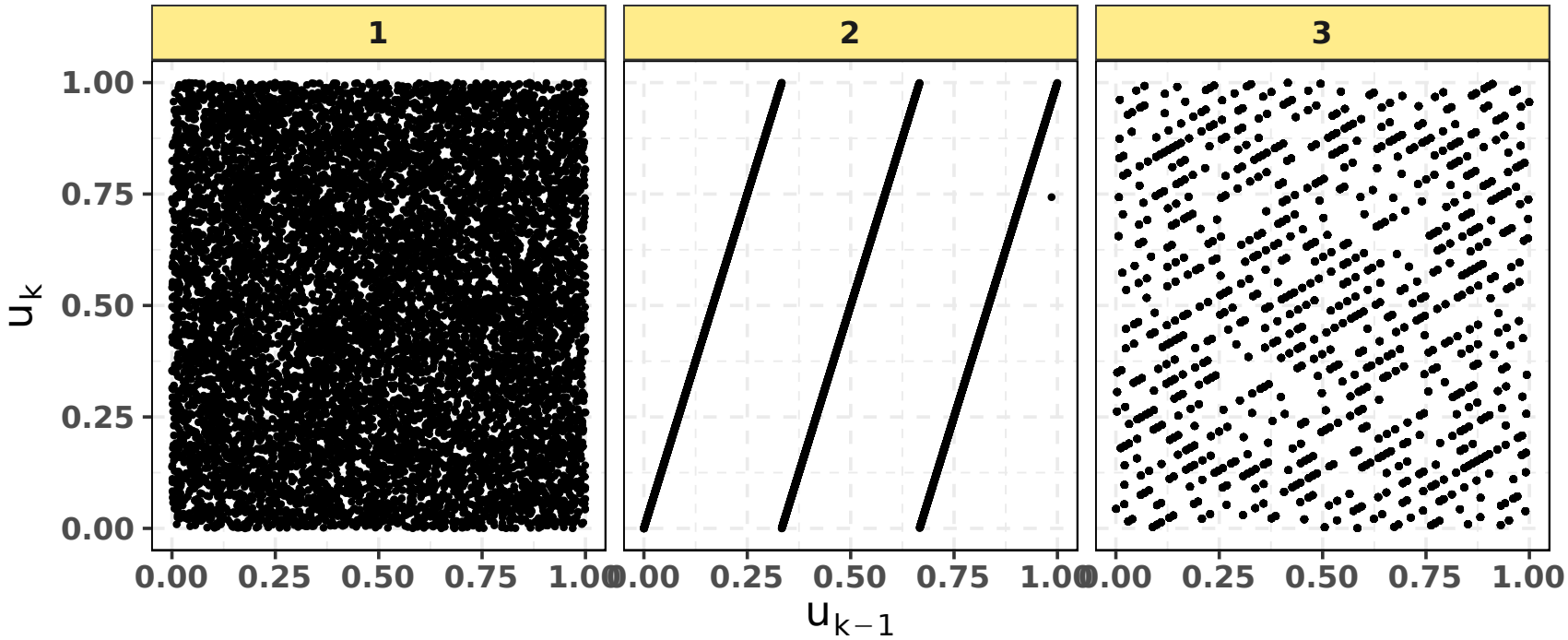
Au risque 5%, avec un test de Kolmogorov-Smirnoff, on rejette

- H_0 : Echantillon de loi uniforme $\mathcal{U}[0,1]$

seulement pour l'échantillon 3.

Choix important (indépendance)

On regarde, pour les 3 échantillons, la valeur de u_k en fonction de celle de u_{k-1}:



Il y a une forte autocorrélation empirique dans les deux derniers échantillons!

Loi uniforme générique

- Si on sait simuler $U \sim \mathcal{U}[0,1]$, alors, pour $a < b \in \mathbb{R}$

$$(b-a)U + a \sim \mathcal{U}[a,b]$$

- Dans `R`, la simulation d'une loi uniforme est faite avec `runif`;
- Dans la suite: on suppose qu'on sait simuler selon $\mathcal{U}[0,1]$.

Méthode d'inversion

Rappel: Fonction de répartition

Soit X une variable aléatoire à valeurs réelles. Pour tout réel x, on appelle fonction de répartition de X la fonction F_X:

$$\begin{array}{rcl} \mathbb{R} & \mapsto & [0,1] \\ x & \mapsto & F_X(x) = \mathbb{P}(X \leq x) \end{array}$$

Rappel: Fonction de répartition

Soit X une variable aléatoire à valeurs réelles. Pour tout réel x, on appelle fonction de répartition de X la fonction F_X:

$$\begin{array}{rcl} \mathbb{R} & \mapsto & [0,1] \\ x & \mapsto & F_X(x) = \mathbb{P}(X \leq x) \end{array}$$

Une fonction de répartition F_X est caractérisée par les propriétés suivantes:

1. F_X est partout continue à droite, i.e. pour tout $x \in \mathbb{R}$:

$$\lim_{\substack{h \to 0 \\ >0}} F(x+h) = F(x)$$

2. F_X est croissante.
3. $\lim_{x \to -\infty} F_X(x) = 0, \lim_{x \to +\infty} F_X(x) = 1$

Ainsi, toute fonction F sur $\mathbb{R}$ satisfaisant ces conditions est une fonction de répartition.

Exemple: Fonction de répartition

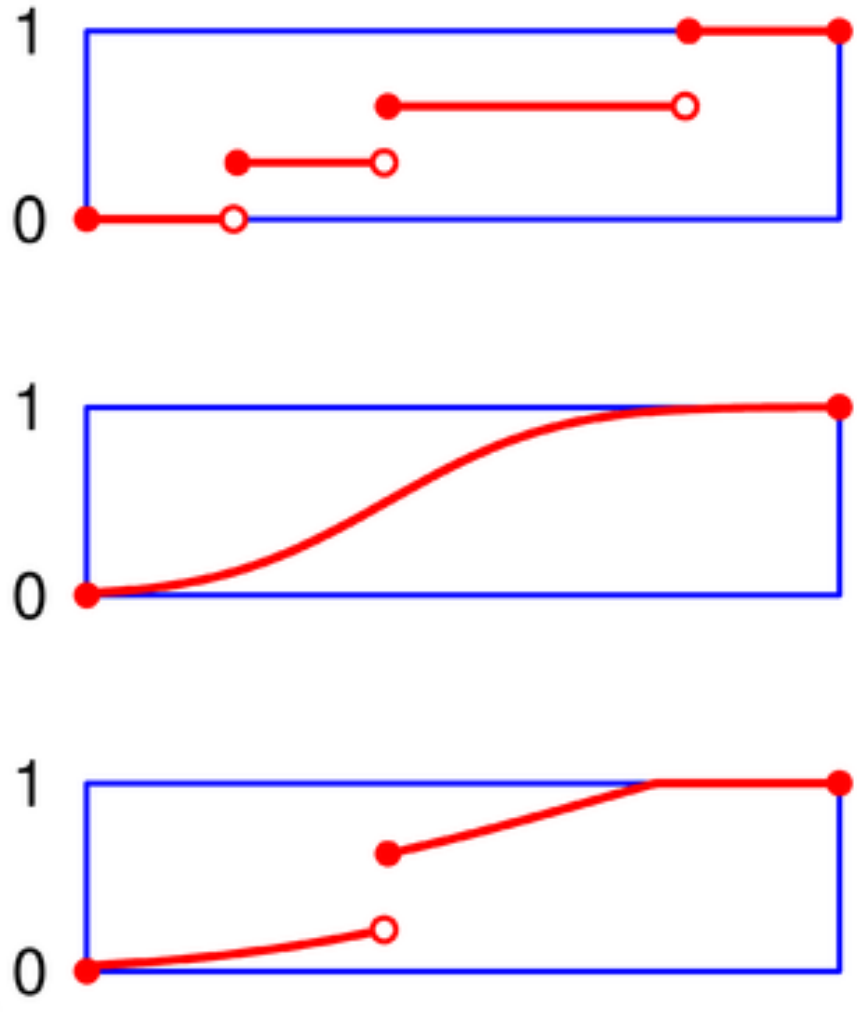


Figure 1: Exemples de fonction de répartition pour une variable aléatoire discrète (haut), continue (centre) ou avec atome (bas). Source *Wikipedia*.

Inverse généralisée de F (fonction quantile)

Soit F une fonction de répartition, on appelle inverse généralisée de F, notée, F^{-1} la fonction:

$$\begin{array}{rcl}]0,1[& \mapsto & \mathbb{R} \\ u & \mapsto & F^{-1}(u) = \inf\{z \in \mathbb{R} \text{ tel que } F(z) \geq u\} \end{array}$$

Pour une variable aléatoire X, la fonction F_X^{-1} est également appelée *fonction quantile* de la variable aléatoire X. On convient que $F_X^{-1}(0)$ et $F_X^{-1}(1)$ sont la plus petite et la plus grande des valeurs du support de X (éventuellement infinies).

Inverse généralisée

Remarque: Dans le cas d'une fonction de répartition F continue et strictement croissante sur $\mathbb{R}$, la fonction F^{-1} est simplement l'inverse de F.

Méthode d'inversion

Supposon qu'on ait une variable aléatoire X dont on connait la fonction de répartiion F, comment simuler X?

Méthode d'inversion

Supposon qu'on ait une variable aléatoire X dont on connait la fonction de répartiion F, comment simuler X?

Exemple: $X \sim \mathcal{E}xp(\lambda)$:

- **Densité:** $f(x) = \lambda \mathrm{e}^{-\lambda x} \mathbf{1}_{x \geq 0}$

Méthode d'inversion

Supposon qu'on ait une variable aléatoire X dont on connait la fonction de répartiion F, comment simuler X?

Exemple: $X \sim \mathcal{E}xp(\lambda)$:

- **Densité:** $f(x) = \lambda \mathrm{e}^{-\lambda x} \mathbf{1}_{x \geq 0}$
- **Fonction de répartition:**
  $F(x) = \int_{-\infty}^{x} f(z)\mathrm{d}z = (1 - \mathrm{e}^{-\lambda x})\mathbf{1}_{x \geq 0}$

Méthode d'inversion

Supposon qu'on ait une variable aléatoire X dont on connait la fonction de répartiion F, comment simuler X?

Exemple: $X \sim \mathcal{E}xp(\lambda)$:

- **Densité:** $f(x) = \lambda \mathrm{e}^{-\lambda x} \mathbf{1}_{x \geq 0}$
- **Fonction de répartition:** $F(x) = \int_{-\infty}^{x} f(z)\mathrm{d}z = (1 - \mathrm{e}^{-\lambda x})\mathbf{1}_{x \geq 0}$
- **Inverse généralisée:** $0 < u < 1,\ F^{-1}(u) = -\frac{\ln(1-u)}{\lambda}$

Méthode d'inversion

Supposon qu'on ait une variable aléatoire X dont on connait la fonction de répartiion F, comment simuler X?

Exemple: $X \sim \mathcal{E}xp(\lambda)$:

- **Densité:** $f(x) = \lambda \mathrm{e}^{-\lambda x} \mathbf{1}_{x \geq 0}$
- **Fonction de répartition:** $F(x) = \int_{-\infty}^{x} f(z)\mathrm{d}z = (1 - \mathrm{e}^{-\lambda x})\mathbf{1}_{x \geq 0}$
- **Inverse généralisée:** $0 < u < 1,\ F^{-1}(u) = -\frac{\ln(1-u)}{\lambda}$

Méthode d'inversion Soit F une fonction de répartition. Soit F^{-1} son inverse généralisée. Soit U une variable aléatoire de loi uniforme sur $[0, 1]$, alors la variable aléatoire

$$X := F^{-1}(U)$$

admet F comme fonction de répartition.

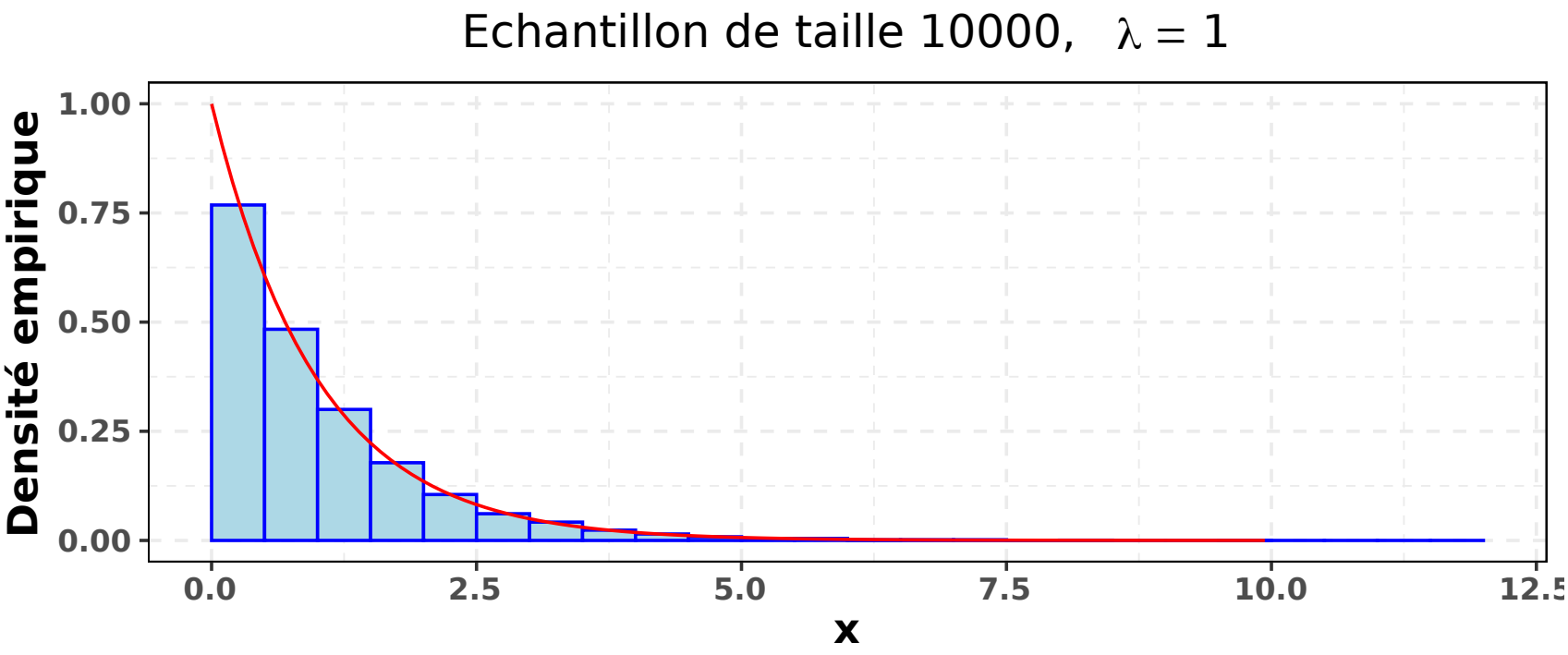
Exemple de méthode d'inversion

$$F^{-1}(u) = -\frac{\ln(1-u)}{\lambda}$$

```r
mon_rexp <- function(n, lambda){
  # On simule selon une loi uniforme
  us <- runif(n) # Echantillon IID U[0,1]
  # On applique la fonction quantile a l'échantillon
  - log(1 - us) / lambda
}
```

Exemple de méthode d'inversion

$$F^{-1}(u) = -\frac{\ln(1-u)}{\lambda}$$



Preuve de la méthode d'inversion

On veut montrer que, pour tout $x \in \mathbb{R}$, si $U \sim \mathcal{U}[0,1]$, alors

$$\mathbb{P}(F^{-1}(U) \leq x) = F(x).$$

Preuve de la méthode d'inversion

On veut montrer que, pour tout $x \in \mathbb{R}$, si $U \sim \mathcal{U}[0,1]$, alors

$$\mathbb{P}(F^{-1}(U) \leq x) = F(x).$$

Montrons tout d'abord que, pour tout $u \in]0,1[$

$$\forall x \in \mathbb{R}, F^{-1}(u) \leq x \Leftrightarrow u \leq F(x). \tag{1}$$

Preuve de la méthode d'inversion

On veut montrer que, pour tout $x \in \mathbb{R}$, si $U \sim \mathcal{U}[0,1]$, alors

$$\mathbb{P}(F^{-1}(U) \leq x) = F(x).$$

Montrons tout d'abord que, pour tout $u \in]0,1[$

$$\forall x \in \mathbb{R}, F^{-1}(u) \leq x \Leftrightarrow u \leq F(x). \tag{1}$$

En effet, si on y parvient, il restera à conclure en se servant de la définition d'une loi uniforme:

$$\mathbb{P}(F^{-1}(U) \leq x) \overset{\text{par } (1)}{=} \mathbb{P}(U \leq F(x)) \overset{\text{car } U \sim \mathcal{U}[0,1]}{=} F(x).$$

Preuve de la méthode d'inversion

Montrons d'abord que, pour tout $u \in]0,1[$

$$\forall x \in \mathbb{R}, F^{-1}(u) \leq x \Rightarrow u \leq F(x).$$

Preuve de la méthode d'inversion

Montrons d'abord que, pour tout $u \in]0,1[$

$$\forall x \in \mathbb{R}, F^{-1}(u) \leq x \Rightarrow u \leq F(x).$$

$\Rightarrow$ Soient $u \in]0,1[$ et $x \in \mathbb{R}$ tels que $F^{-1}(u) \leq x$.
Par croissance de F, on a donc:

$$F\left(F^{-1}(u)\right) \leq F(x)$$

Or, en se souvenant que par définition

$$F^{-1}(u) = \inf\left\{z \in \mathbb{R} \text{ tel que } F(z) \geq u\right\},$$

on a donc directement

$$u \leq F\left(F^{-1}(u)\right) \leq F(x)$$

Preuve de la méthode d'inversion

Montrons maintenant, pour tout $u \in]0,1[$

$$\forall x \in \mathbb{R}, F^{-1}(u) \leq x \Leftarrow u \leq F(x).$$

Preuve de la méthode d'inversion

Montrons maintenant, pour tout $u \in]0,1[$

$$\forall x \in \mathbb{R}, F^{-1}(u) \leq x \Leftarrow u \leq F(x).$$

$\Leftarrow$ Soient $u \in]0,1[$ et $x \in \mathbb{R}$ tels que $u \leq F(x)$.
Ainsi, $x \in \{z \in \mathbb{R} \text{ tel que } F(z) \geq u\}$, donc $F^{-1}(u) \leq x$.

Intérêt de la méthode d'inversion

Ainsi, pour une variable aléatoire à valeurs dans $\mathbb{R}$, on sait simuler si on connaît l'inverse généralisée de sa fonction de répartition.

Méthode d'acceptation rejet

Exemple motivant

On veut simuler selon la densité $f(x) = \sqrt{\frac{2}{\pi}} \mathrm{e}^{-\frac{x^2}{2}} \mathbf{1}_{x \geq 0}$.

La fonction quantile n'a pas d'expression analytique. La méthode d'inversion ne peut être appliquée.

Exemple motivant

On veut simuler selon la densité $f(x) = \sqrt{\frac{2}{\pi}} \mathrm{e}^{-\frac{x^2}{2}} \mathbf{1}_{x \geq 0}$.

La fonction quantile n'a pas d'expression analytique. La méthode d'inversion ne peut être appliquée.

Principe de la méthode d'acceptation rejet:

- Simuler des *candidats* selon une autre loi qu'on sait simuler.

Exemple motivant

On veut simuler selon la densité $f(x) = \sqrt{\frac{2}{\pi}} \mathrm{e}^{-\frac{x^2}{2}} \mathbf{1}_{x \geq 0}$.

La fonction quantile n'a pas d'expression analytique. La méthode d'inversion ne peut être appliquée.

Principe de la méthode d'acceptation rejet:

- Simuler des *candidats* selon une autre loi qu'on sait simuler.
- Choisir parmi les candidats grâce à la loi uniforme.

Méthode d'acceptation rejet (proposition)

Soit f et g deux densités sur $\mathbb{R}^d$. On suppose qu'il existe une constante M telle que

$$\forall x \in \mathbb{R}^d \quad f(x) \leq Mg(x)$$

On note

$$0 \leq r(x) := \frac{f(x)}{Mg(x)} \leq 1.$$

Méthode d'acceptation rejet (proposition)

Soit f et g deux densités sur $\mathbb{R}^d$. On suppose qu'il existe une constante M telle que

$$\forall x \in \mathbb{R}^d \quad f(x) \leq Mg(x)$$

On note

$$0 \leq r(x) := \frac{f(x)}{Mg(x)} \leq 1.$$

-Soient $(Y_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de densité g et $(U_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de loi uniforme sur $[0,1]$.

Méthode d'acceptation rejet (proposition)

Soit f et g deux densités sur $\mathbb{R}^d$. On suppose qu'il existe une constante M telle que

$$\forall x \in \mathbb{R}^d \quad f(x) \leq Mg(x)$$

On note

$$0 \leq r(x) := \frac{f(x)}{Mg(x)} \leq 1.$$

-Soient $(Y_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de densité g et $(U_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de loi uniforme sur $[0,1]$. - On note T la variable aléatoire (à valeurs dans $\mathbb{N}^*$):

$$T = \inf\{n, \text{ tel que } U_n \leq r(Y_n)\}.$$

.

Méthode d'acceptation rejet (proposition)

Soit f et g deux densités sur $\mathbb{R}^d$. On suppose qu'il existe une constante M telle que

$$\forall x \in \mathbb{R}^d \quad f(x) \leq Mg(x)$$

On note

$$0 \leq r(x) := \frac{f(x)}{Mg(x)} \leq 1.$$

-Soient $(Y_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de densité g et $(U_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de loi uniforme sur $[0,1]$. - On note T la variable aléatoire (à valeurs dans $\mathbb{N}^*$):

$$T = \inf\{n, \text{ tel que } U_n \leq r(Y_n)\}.$$

. - Alors, la variable aléatoire $X := Y_T$ (T-ième valeur de la suite $(Y_n)_{n\geq 1}$) a pour densité f.

Méthode d'acceptation rejet (algorithme)

On veut tirer un échantillon X de densité f. On ne sait simuler que selon la densité g. On suppose qu'il existe une constante M telle que

$$\forall x \in \mathbb{R}^d \quad f(x) \leq Mg(x)$$

Algorithme `Condition <- FALSE`

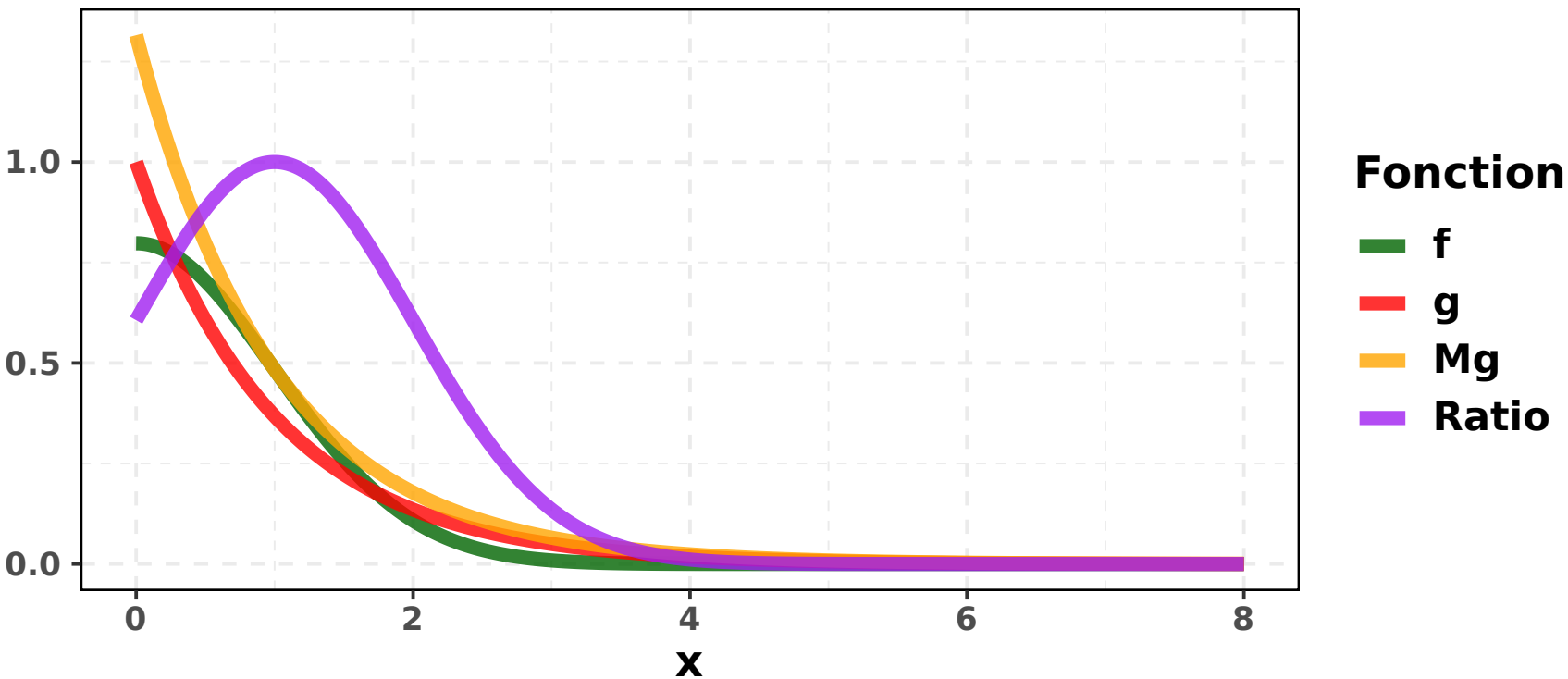
- Tant que `not Condition`:
  - Tirer $Y \sim g(y)$;
  - Tirer (indépendemment) $U \sim \mathcal{U}[0,1]$;
  - Si
  $$U \leq \frac{f(Y)}{Mg(Y)},$$
  alors on pose `Condition <- TRUE` et $X = Y$
  - Sinon `Condition <- FALSE`

En sortie, $X \sim f(x)$.

Méthode d'acceptation rejet: Exemple

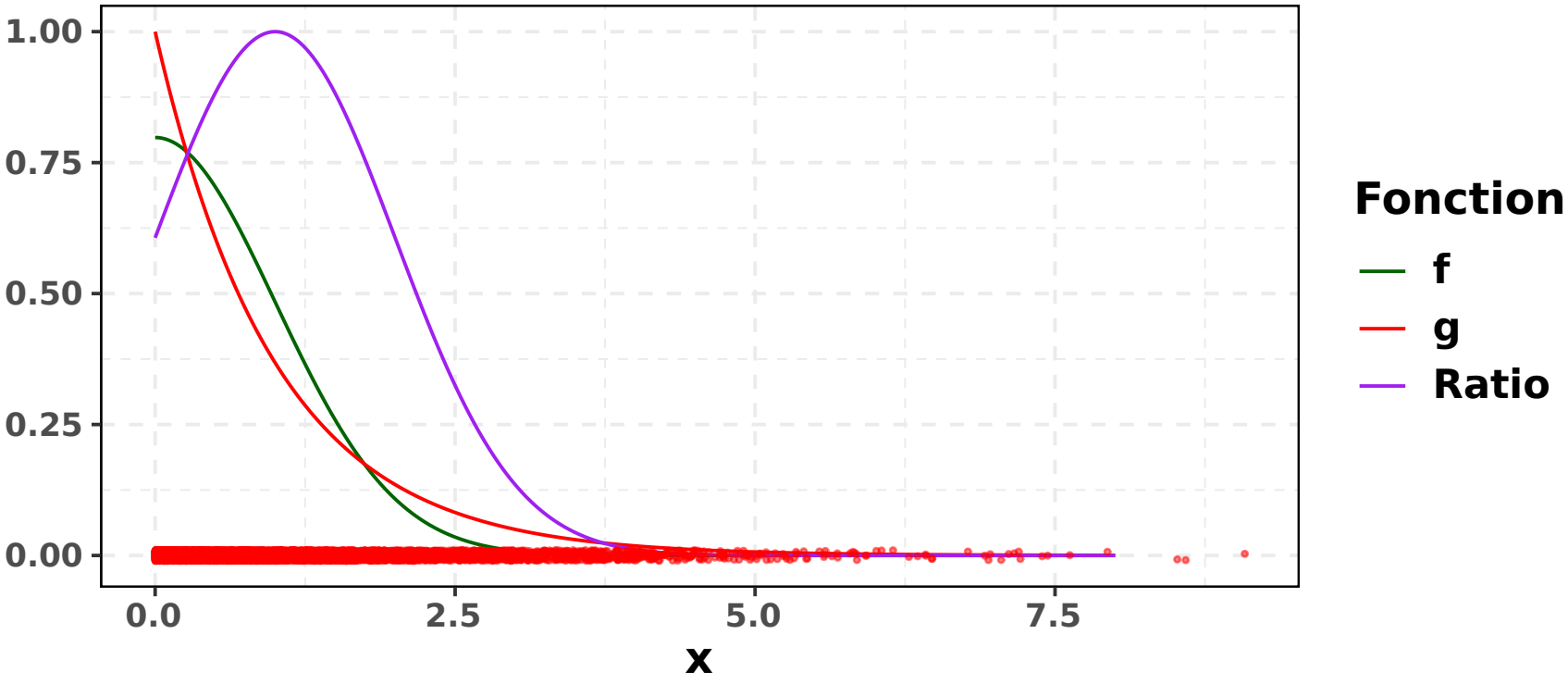
- On veut simuler selon $f(x) = \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}} \mathbf{1}_{x \geq 0}$
- On considère $g(x) = e^{-x} \mathbf{1}_{x \geq 0}$ (densité d'un $\mathcal{E}xp(\lambda = 1)$)
- On montre que

$$\forall x \in \mathbb{R},\ f(x) \leq \overbrace{\sqrt{\frac{2}{\pi}} e^{\frac{1}{2}}}^{M} g(x)$$



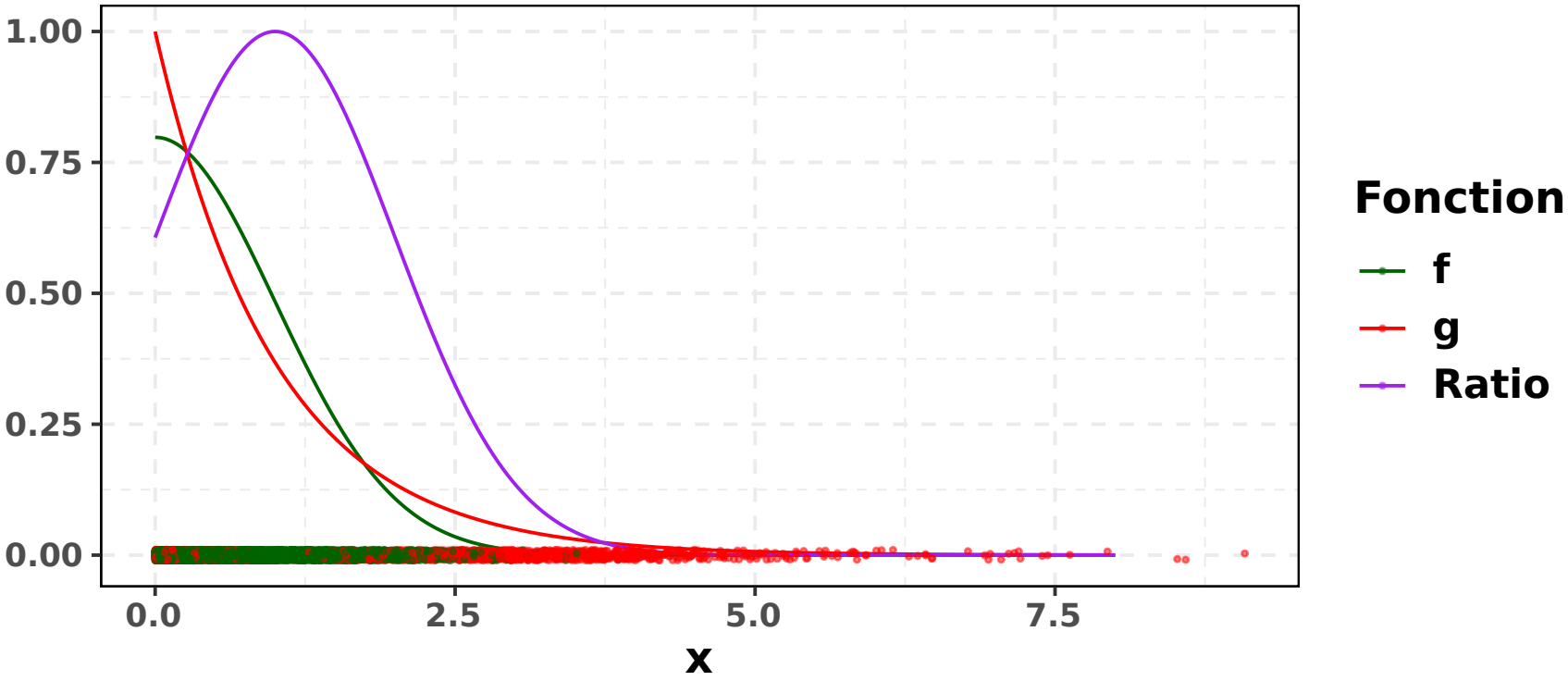
Méthode d'acceptation rejet: Exemple

- On simule 10000 points selon g



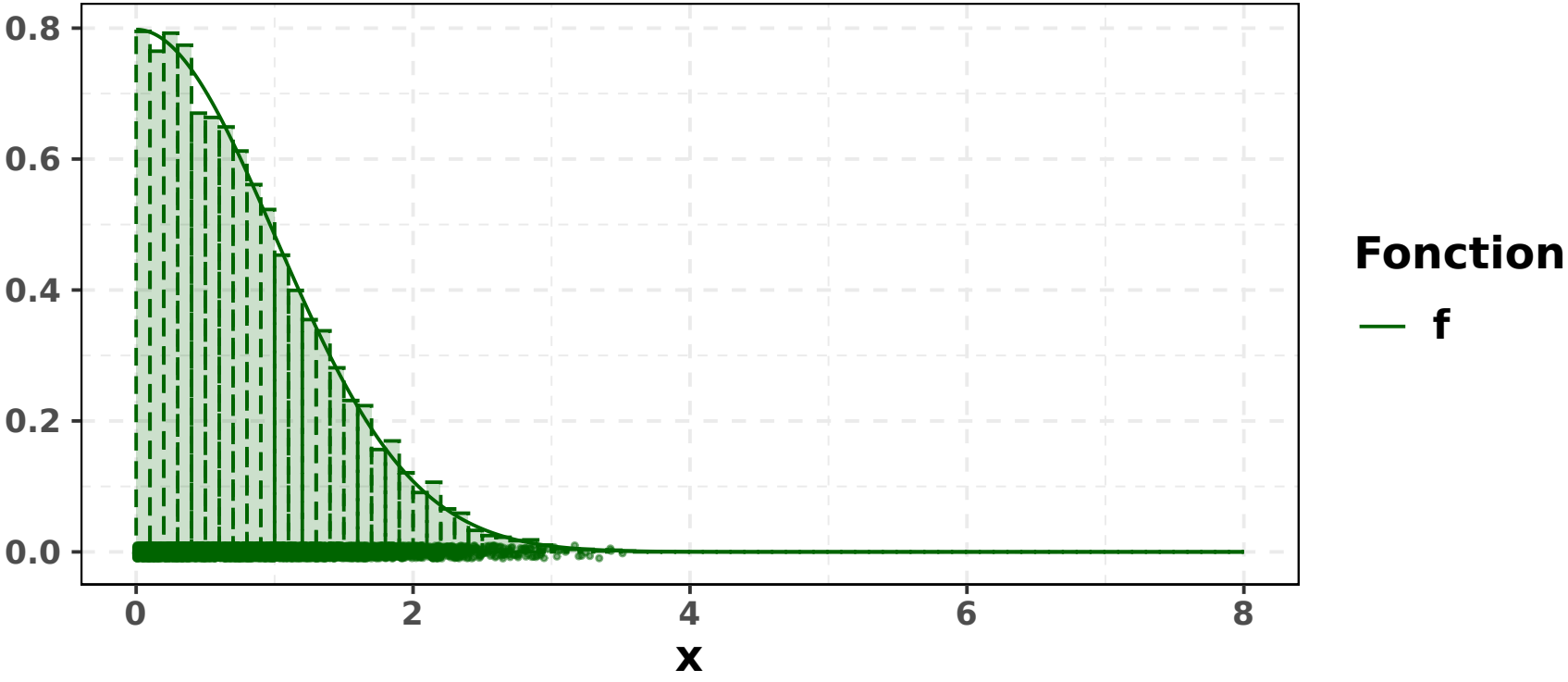
Méthode d'acceptation rejet: Exemple

- On simule 10000 points selon g
- On accepte avec une probabilité donnée par le ratio



Méthode d'acceptation rejet: Exemple

- On simule 10000 points selon g
- On accepte avec une probabilité donnée par le ratio
- Les points acceptés sont i.i.d. de densité f.



Remarque

Sur l'exemple précédent, au lieu de faire un "tant que", on a simuler 10000 points et on n'a retenu que les acceptés.

Remarque

Sur l'exemple précédent, au lieu de faire un "tant que", on a simuler 10000 points et on n'a retenu que les acceptés.

- Proportion empirique acceptée: 0.761
- D'un autre côté, on a $1/M = 0.76$

Preuve de la méthode d'acceptation rejet

Preuve à connaître!

- Voir le poly de cours.
- Analogue de la preuve sera demandée en devoir.

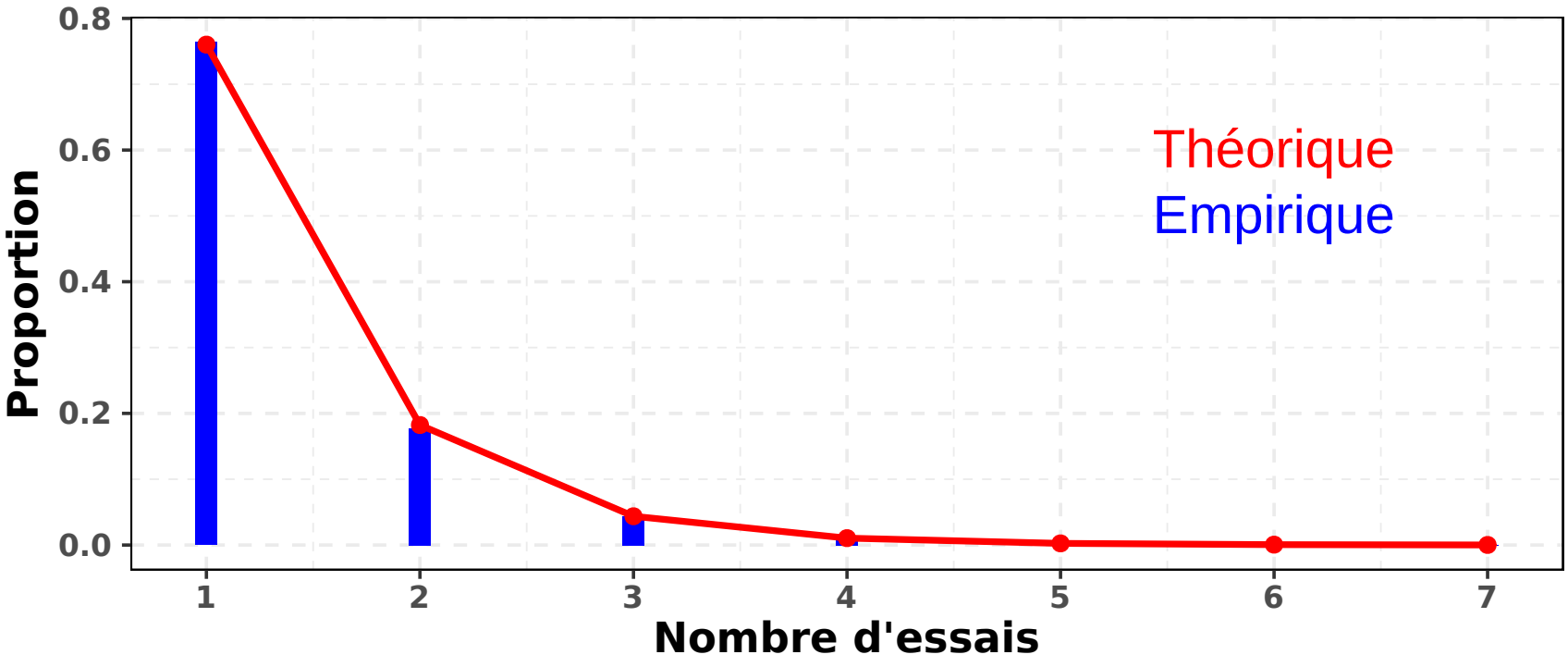
Code R

```r
get_one_sample <- function(){
  condition <- FALSE
  while(!condition){
    y <- simulate_g(...) # Simulation selon g
    u <- runif(1) # Uniform
    # On suppose que f, g, et M existent
    condition <- u <= f(y) / (M * g(y))
  }
  return(y)
}
```

Loi du temps d'attente

On s'arrête au premier temps tel qu'une uniforme est inférieure au ratio observée.

La loi du temps d'attente (voir preuve) est une **loi géométrique** sur $\mathbb{N}^*$ de paramètre $\frac{1}{M}$.



Vecteurs aléatoires

Pour simuler un vecteur aléatoire (X, Y), on pourra utiliser (voir poly et TD pour des exemples):

- Conditionnement;
- Changement de variables

Changements de variables pour densité

Soit un couple de variables aléatoires (U, V) de densité $f_{U,V}(u, v)$ définie sur $E_{UV} \subset \mathbb{R}^2$ et un couple de variables aléatoires (X, Y) à valeurs dans $E_{XY} \subset \mathbb{R}^2$. Supposons qu'il existe une application ϕ, C^1, inversible, et d'inverse C^1, tel que $(X, Y) = \phi(U, V)$, alors la densité jointe de (X, Y) est donnée par:

$$f_{X,Y}(x, y) = f_{U,V}(\phi^{-1}(x, y))|\det J_{\phi^{-1}}(x, y)|$$

où J_ϕ désigne la matrice jacobienne d'une application $\phi(u, v)$:

$$J_\phi(u, v) = \begin{pmatrix} \frac{\delta\phi_1}{\delta u}(u, v) & \frac{\delta\phi_1}{\delta v}(u, v) \\ \frac{\delta\phi_2}{\delta u}(u, v) & \frac{\delta\phi_2}{\delta v}(u, v) \end{pmatrix}$$

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:

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

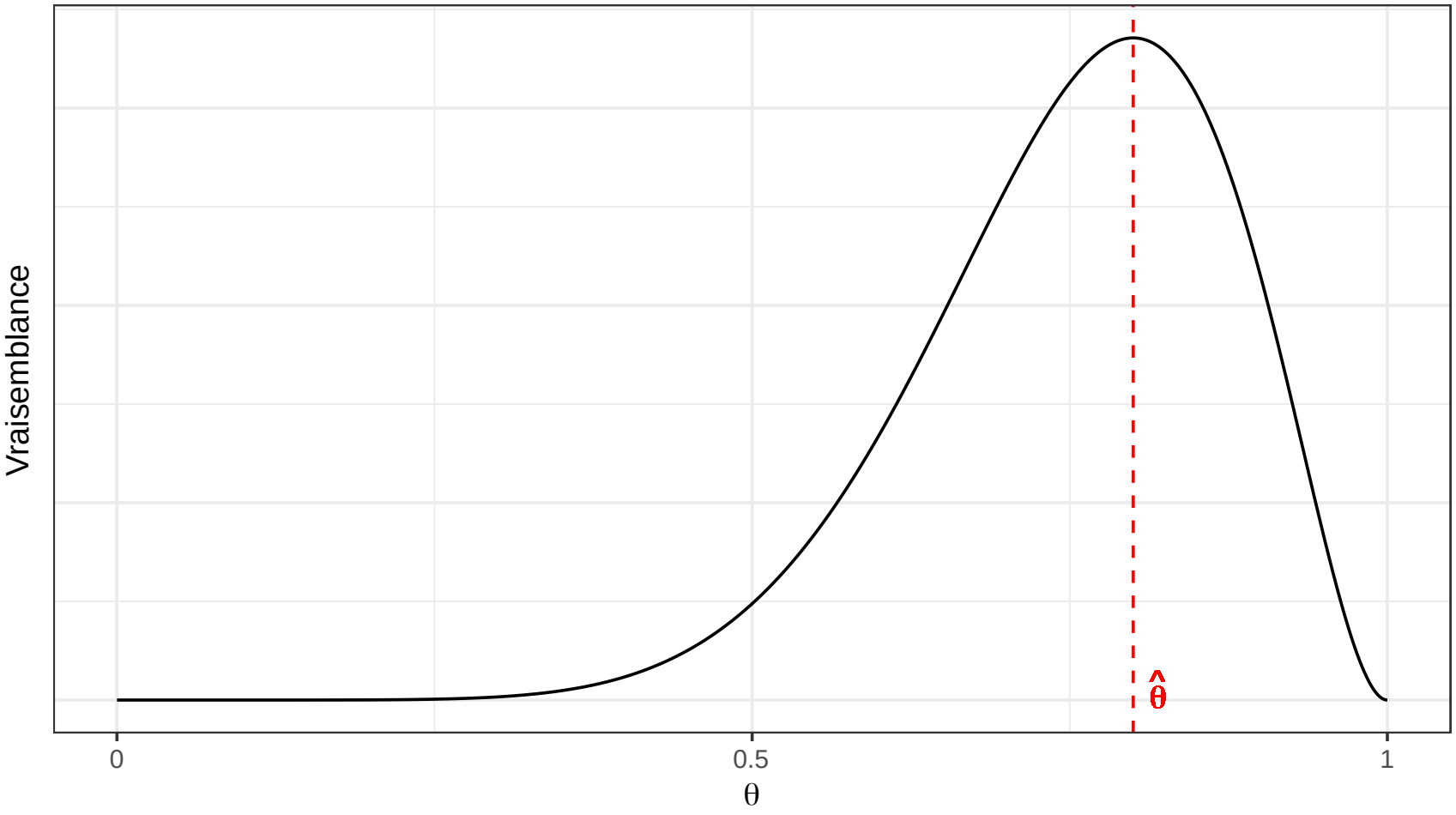
Maximum de vraisemblance

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**.

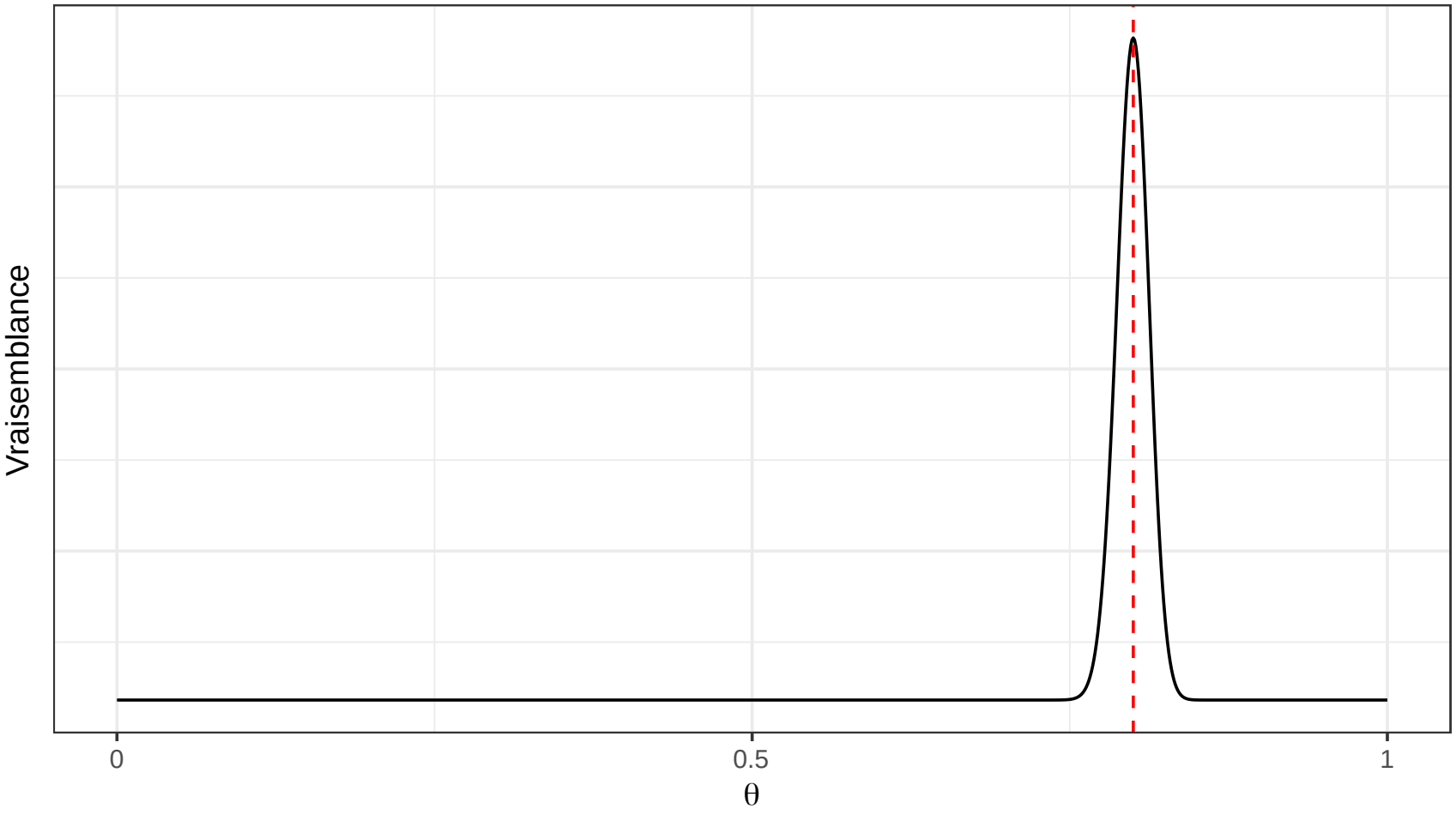
Incertitude sur $\hat{\theta}$

$\hat{\theta}$ 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.

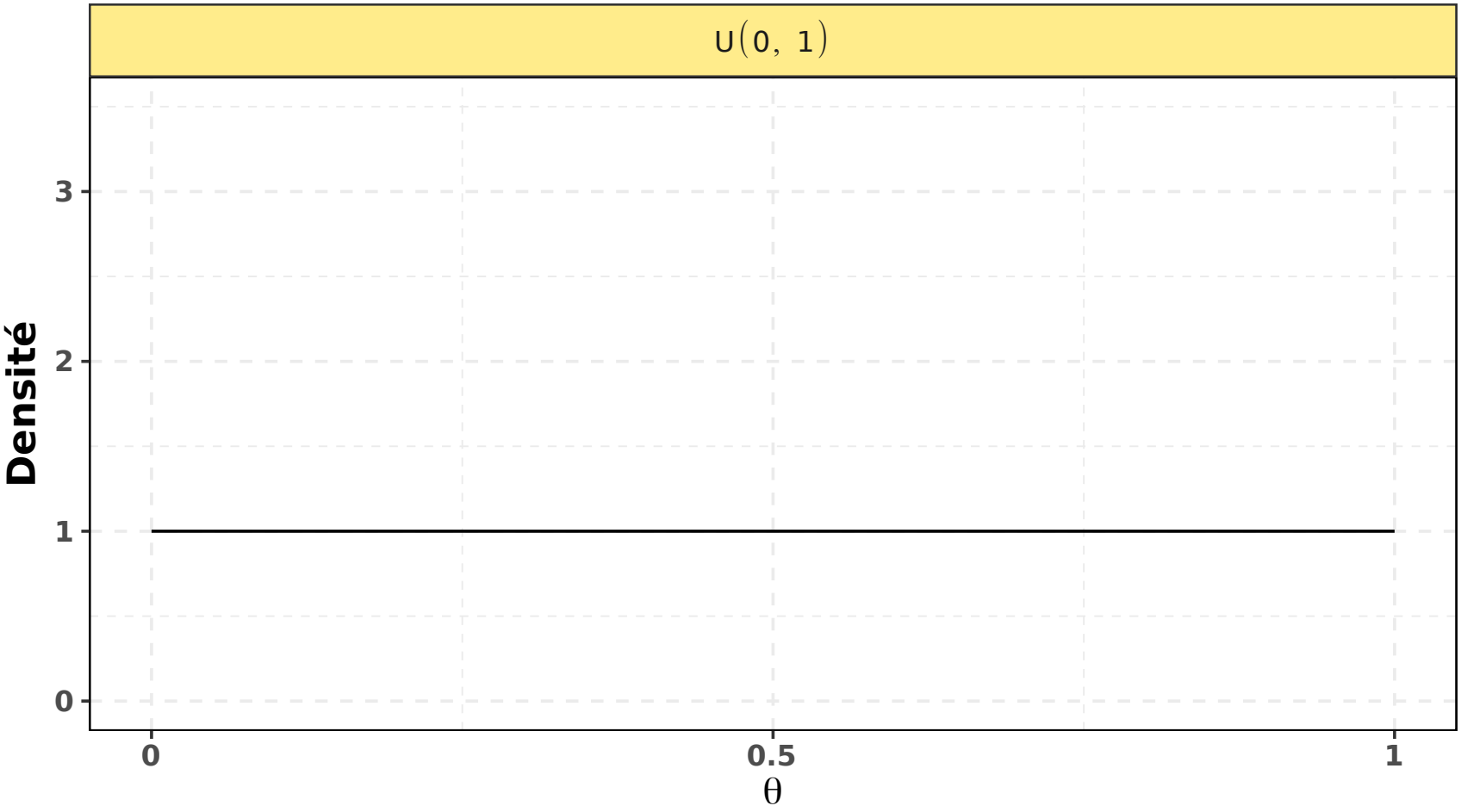
Inférence bayésienne

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

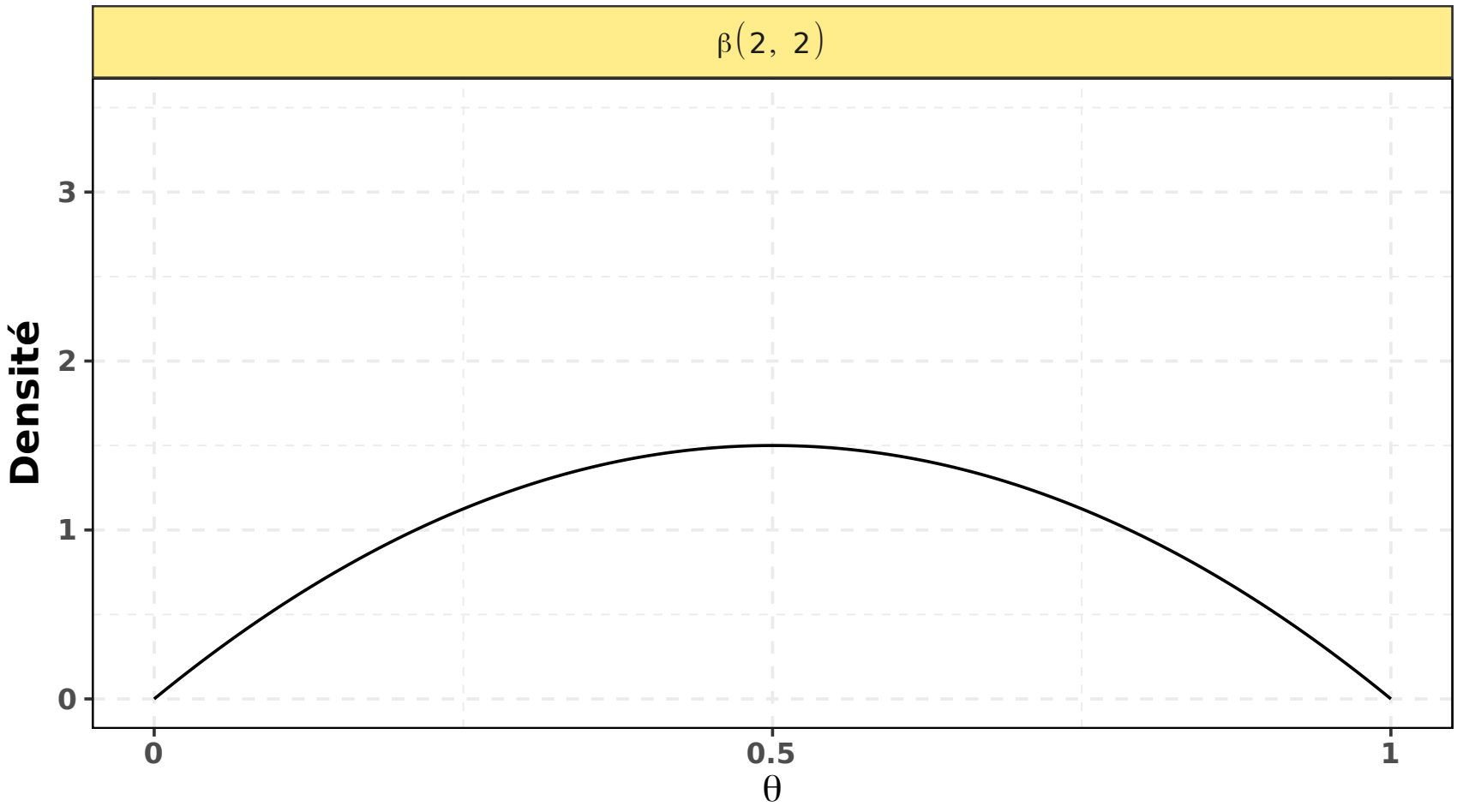
Aucune idée sur θ



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

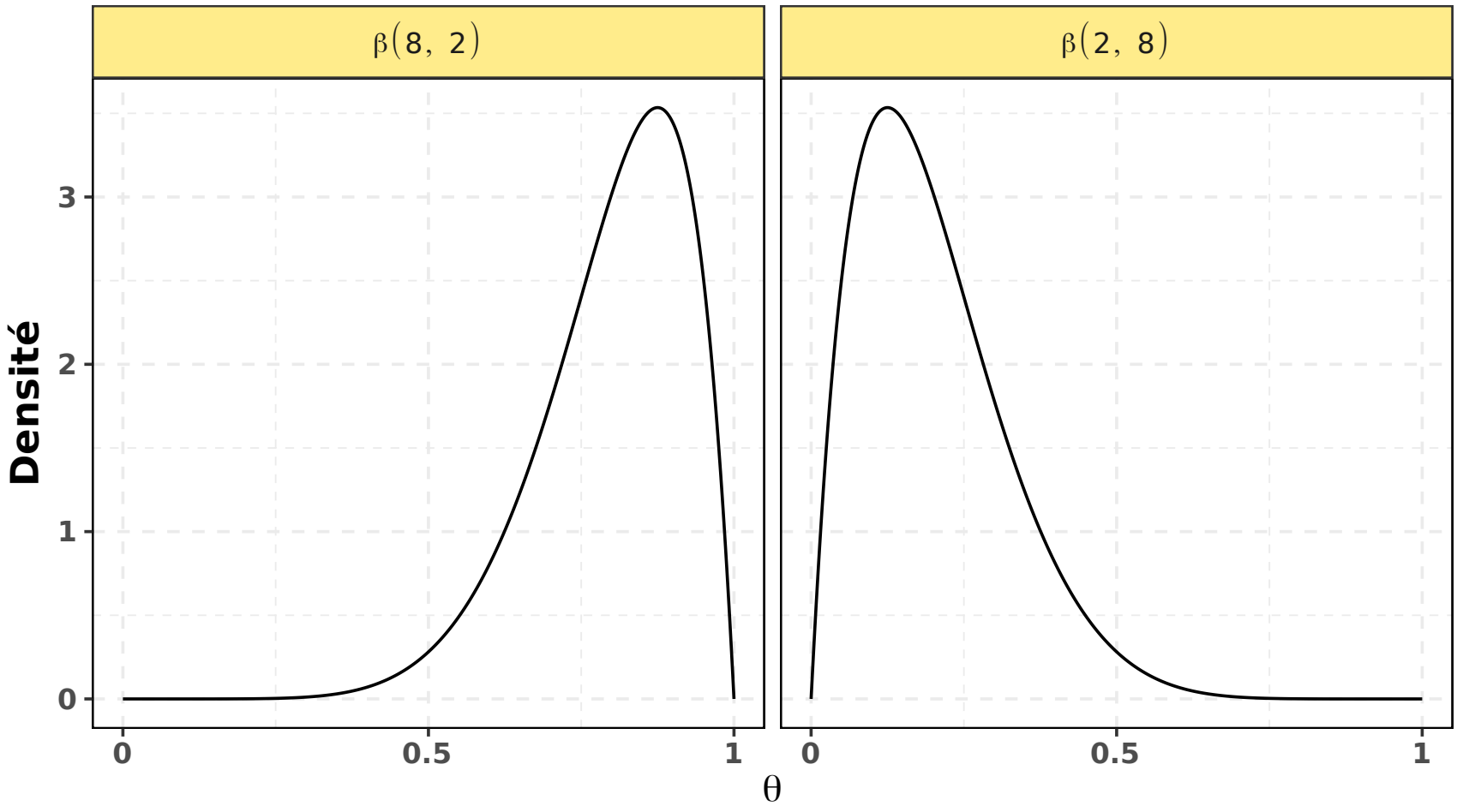
Exemples de loi a priori

A priori léger sur une pièce équitable



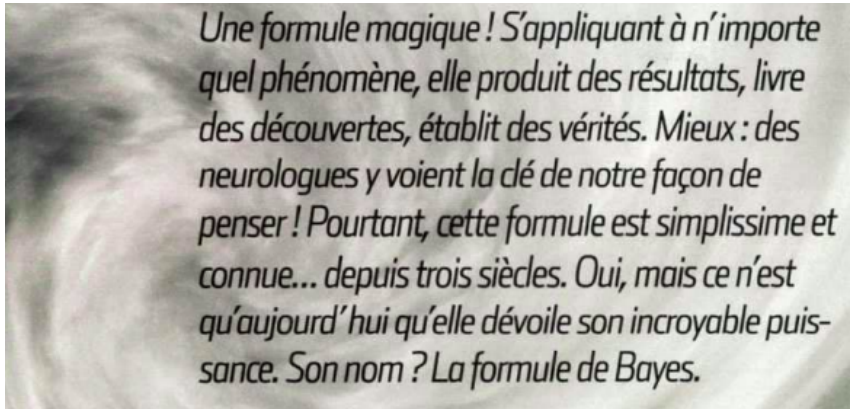
Exemples de loi a priori

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 $\theta|x_{1:n}$.

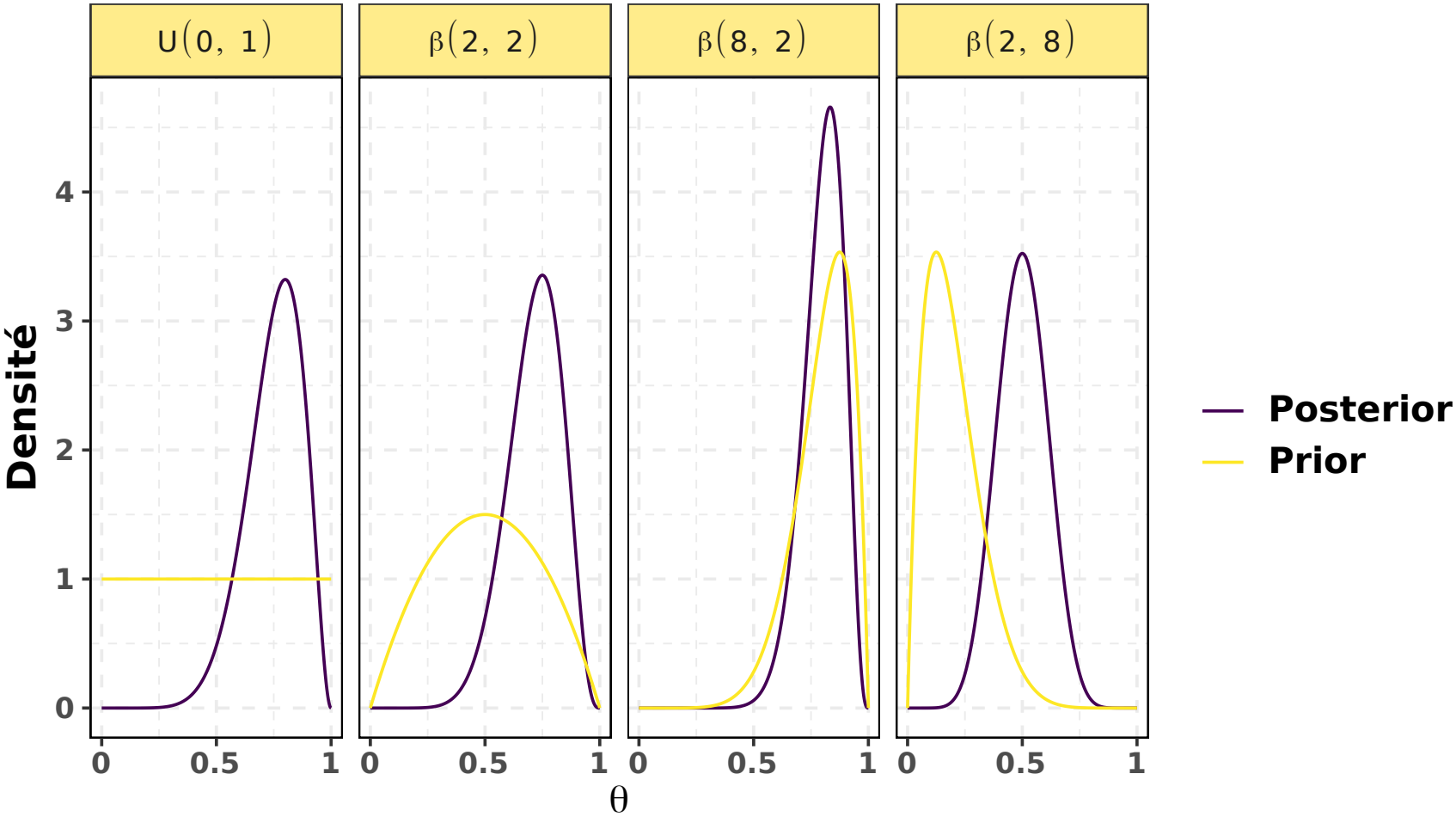
$$\begin{aligned}
\pi(\theta|x_{1:n}) &\propto L(x_{1:n}|\theta)\pi(\theta) \\
&\propto \theta^{\sum_{k=1}^n x_k} (1-\theta)^{n-\sum_{k=1}^n x_k}\, \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{0<\theta<1} \\
&\propto \theta^{a+\sum_{k=1}^n x_k-1}(1-\theta)^{b+n-\sum_{k=1}^n x_k-1}\mathbf{1}_{0<\theta<1}
\end{aligned}$$

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

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

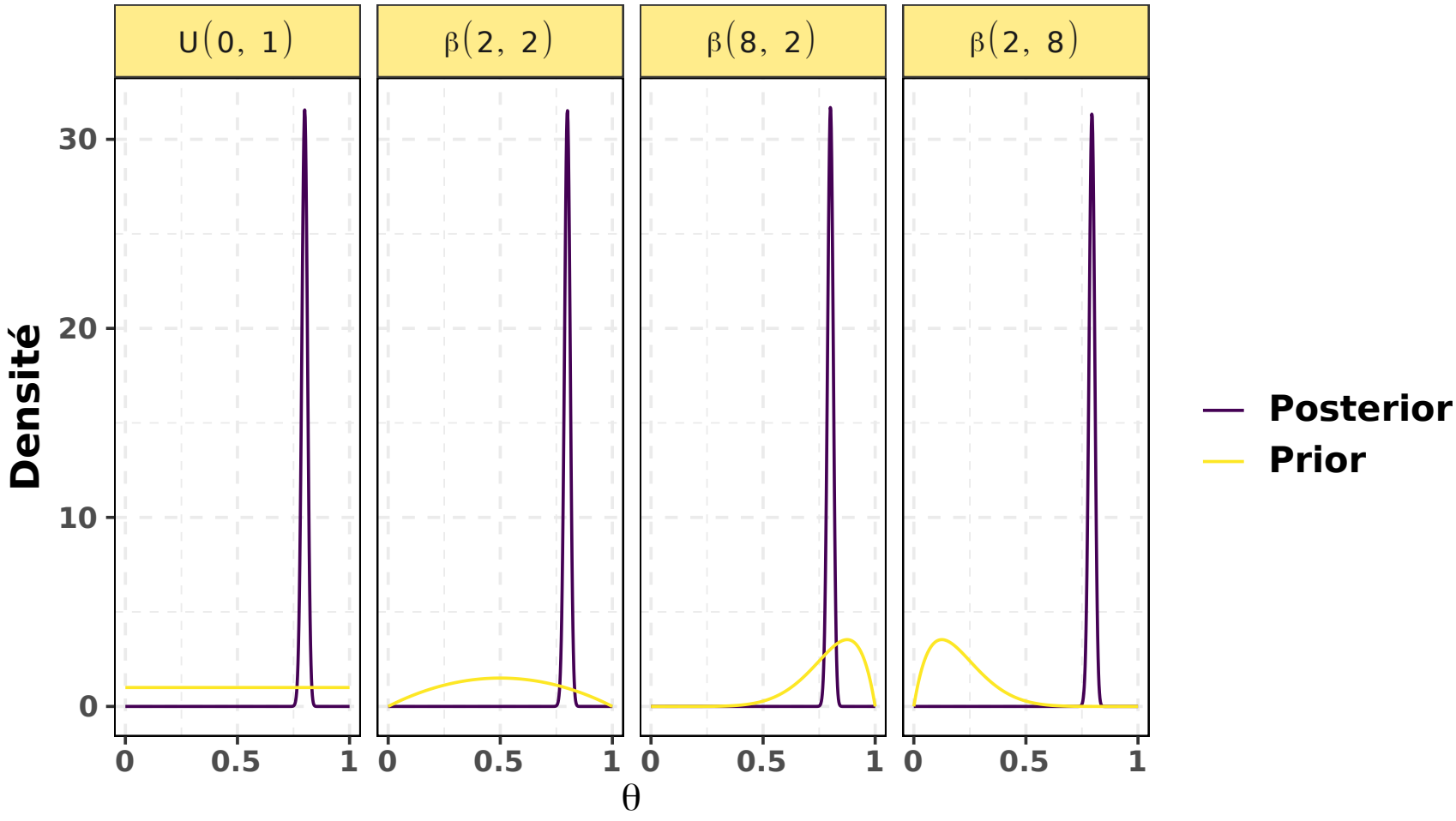
Cas $n = 10$ et 8 succès

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



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

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

Estimateurs Bayésiens

Intervalle de crédibilité

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

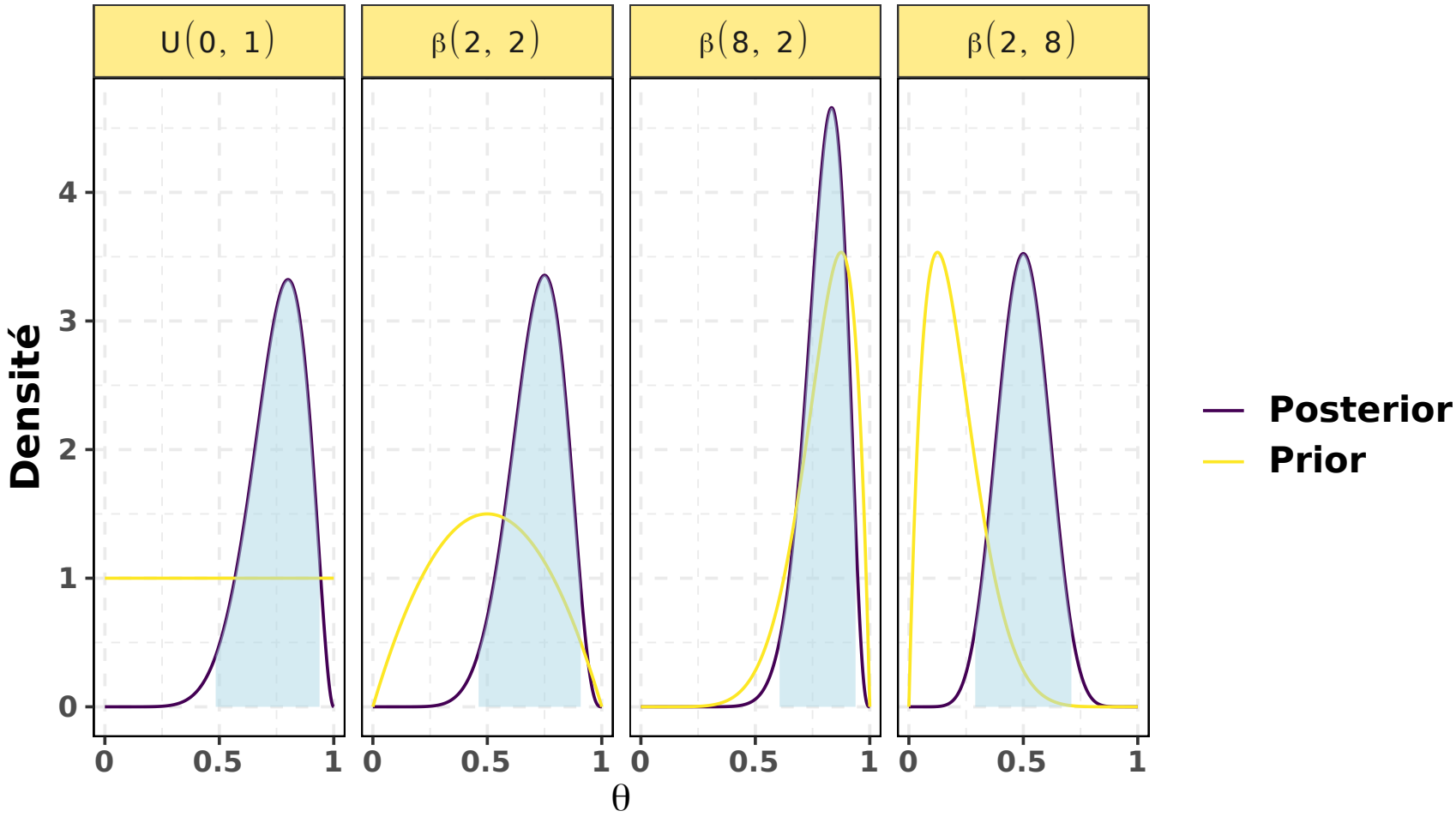
$$\mathbb{P}(\theta \in \mathcal{R} | \mathbf{X} = x_{1:n}) = \int_{\mathcal{R}} \pi(\theta | x_{1:n}) \mathrm{d}\theta$$

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

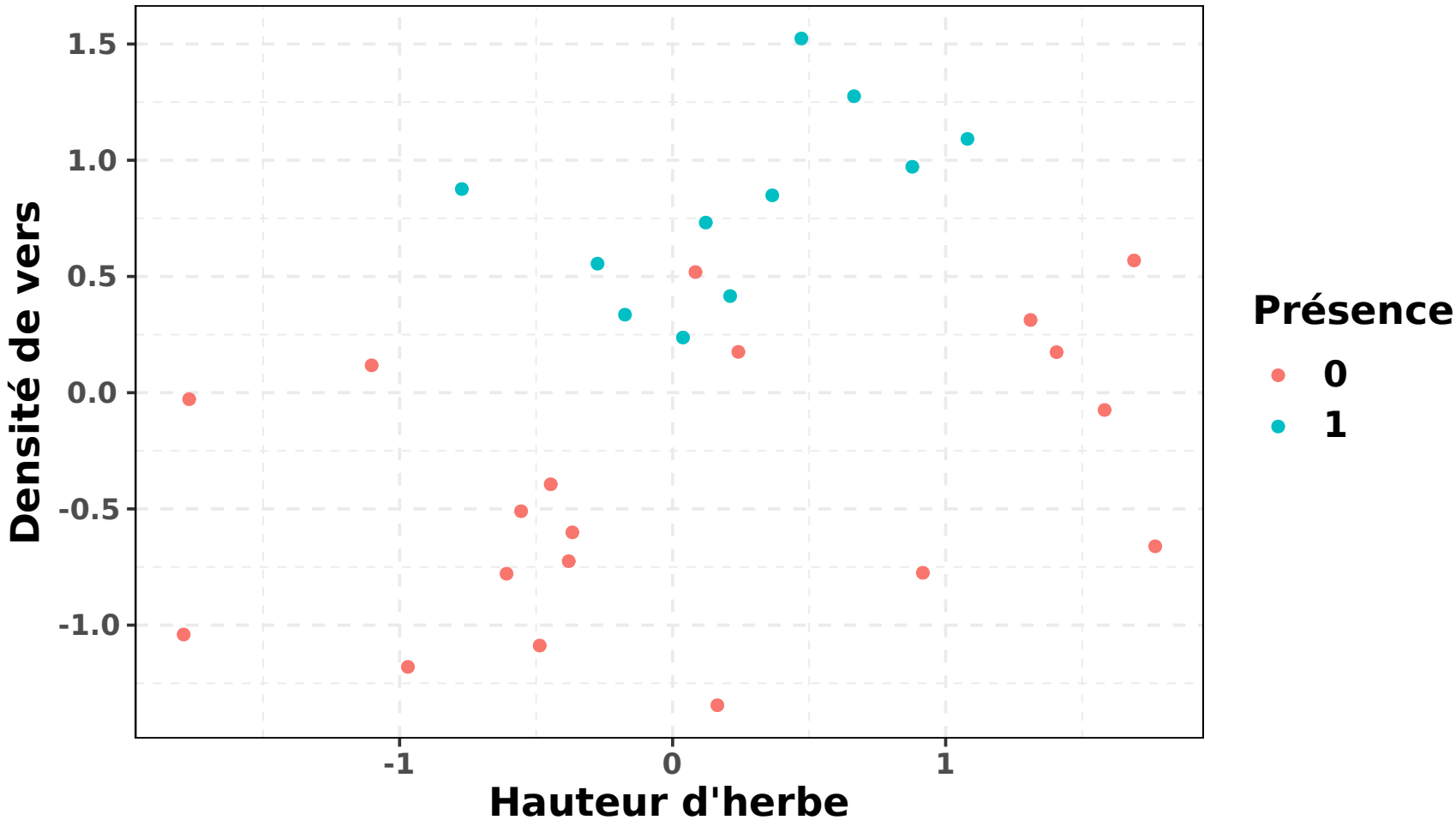
Une étude consiste en l'observation de la présence ou non de la linotte mélodieuse sur différents sites échantillonnés.

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

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

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 $\pi(\theta|y_{1:n})$

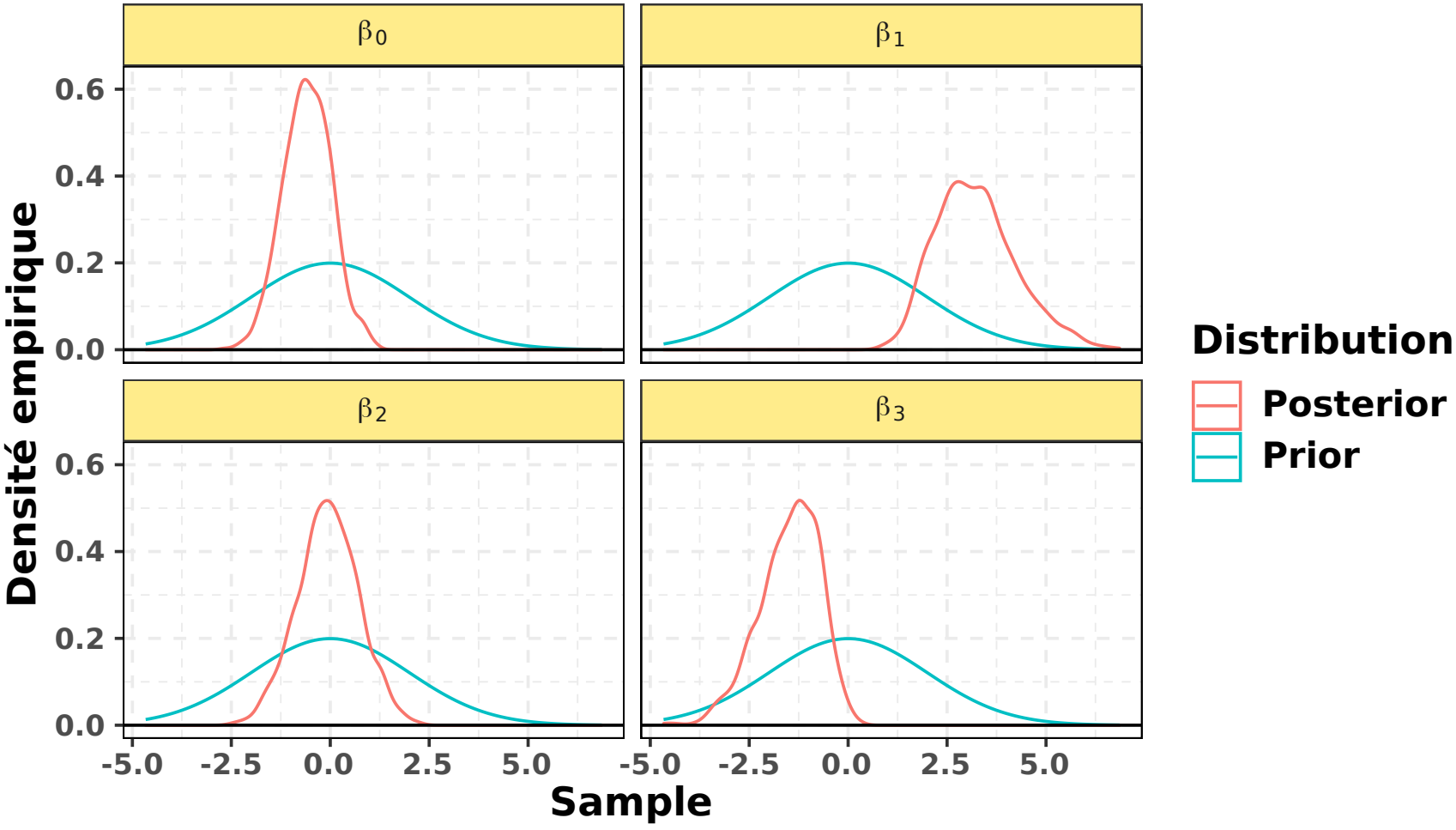
1. On tire $\theta_{cand} \sim \mathcal{N}(0, 4I)$
2. On tire (independamment) $U \sim \mathcal{U}[0,1]$
3. Si $U < \frac{L(y_{1:n}|\theta)}{M}$, on accepte θ_{cand}
4. Sinon on recommence

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

```
## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated
## numbers without specifying argument 'seed'. There is a risk that tho
## numbers are not statistically sound and the overall results might be
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel
## numbers are produced via the L'Ecuyer-CMRG method. To disable this c
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".
## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated
## numbers without specifying argument 'seed'. There is a risk that tho
## numbers are not statistically sound and the overall results might be
```

Echantillon du posterior, et loi a posteriori marginales

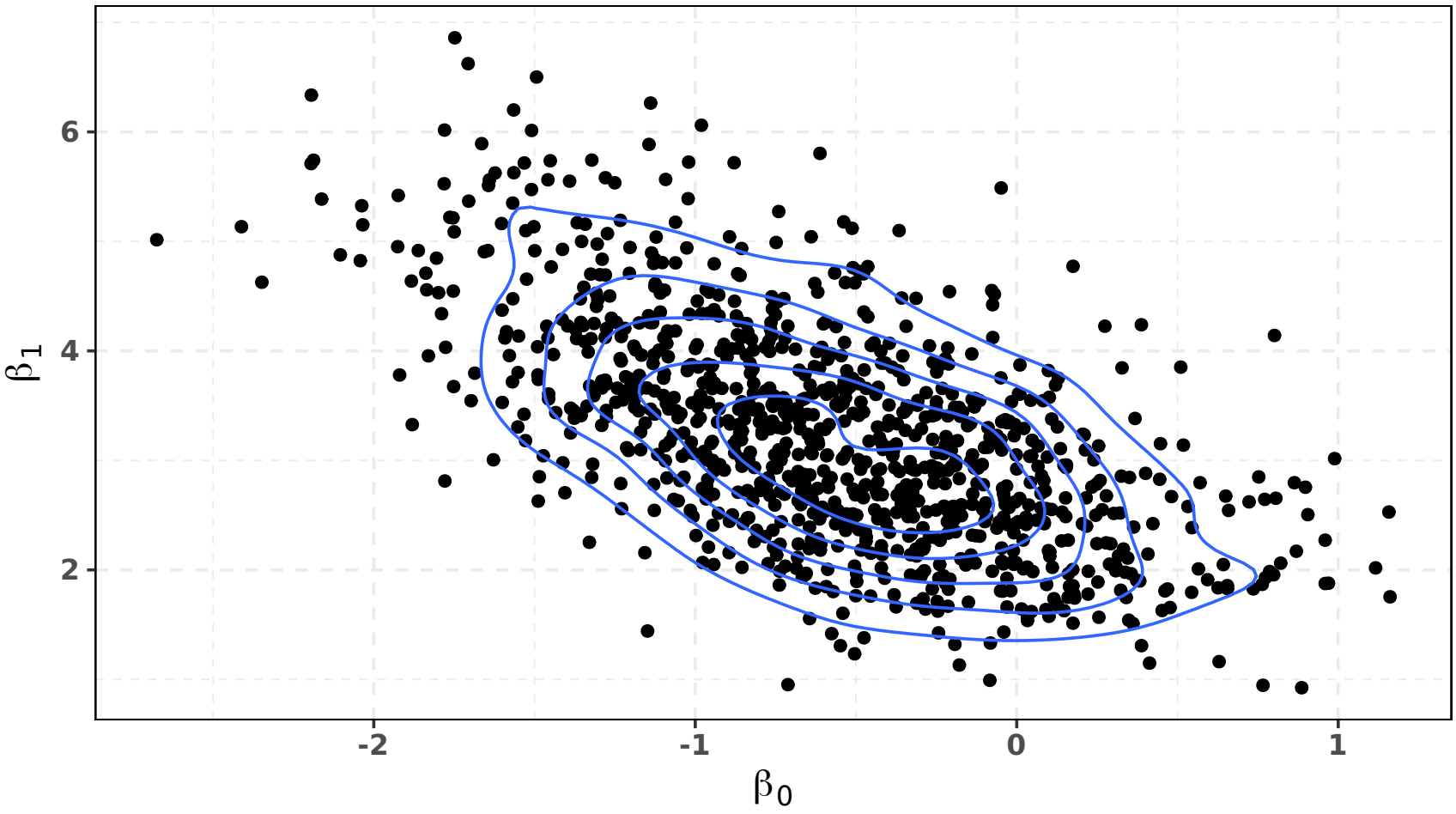
On effectue un tirage de taille $M = 1000$



- 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]	-0.576	-1.778926	0.6555852
beta[1]	3.231	1.603258	5.5504022
beta[2]	-0.055	-1.573656	1.4290713
beta[3]	-1.484	-3.194941	-0.2279059

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.

Méthodes de Monte Carlo par chaîne de Markov

Pierre Gloaguen

Avril 2020

Rappels des cours précédents

- Méthodes de Monte Carlo pour le calcul d'espérances
- Approche par simulation de vaariables aléatoires i.i.d.
- Méthodes de simulation de loi (échantillons i.i.d.)
- Inférence bayésienne, technique nécessitant des algos de simulations de lois

Modèle probit

On veut simuler selon une loi $\pi(\theta|y_{1:n}, \mathbf{x}_{1:n})$ telle que:

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

- Possible par acceptation rejet si n n'est pas trop grand;
- Ensuite, ne fonctionne plus **en pratique** (probabilité d'acceptation devient trop faible).
- Nécessité de définir un autre algorithme.

Objectif du cours

- Présentation des méthodes de Monte Carlo par chaîne de Markov
- Rappel sur les chaînes de Markov (définitions)
- Théorème ergodique
- Algorithme de Metropolis Hastings
- Algorithme de Gibbs

Rappel sur les chaînes de Markov

Chaîne de Markov (à espace d'états fini)

Soit X_0 une variable aléatoire sur $\{1, \ldots, K\}$ de loi π_0.

- La suite de variables aléatoires $(X_n)_{n \geq 0}$ à valeurs dans $\mathcal{K} = \{1, \ldots, K\}$ est une chaîne de Markov si pour tout $n \geq 1$ est pour tout suite $(k_0, \ldots, k_n)$ d'éléments de $\mathcal{K}$, on a :

$$\mathbb{P}\left(X_n = k_n | X_0 = k_0, \ldots, X_{n-1} = k_{n-1}\right) = \mathbb{P}(X_n = k_n | X_{n-1} = k_{n-1})$$

- Cette chaîne est *homogène* si, pour (i, j) dans $\mathcal{K} \times \mathcal{K}$:
  $\mathbb{P}(X_n = j | X_{n-1} = i) = \mathbb{P}(X_1 = j | X_0 = i) = P_{ij}$
- La matrice $P = (P_{ij})$ est la **matrice de transition** de la chaîne de Markov.
- Une chaîne de Markov homogène est entièrement caractérisée par π_0 et P.

Loi de la chaîne

Pour $n \geq 0$, on note π_n, la loi de l'état X_n, c'est à dire le vecteur ligne

$$\pi_n = (\pi_{n,1} = \mathbb{P}(X_n = 1), \ldots, \pi_{n,K} = \mathbb{P}(X_n = K)).$$

On a:

- $\mathbb{P}(X_1 = j) = \sum_{i=1}^{k} \mathbb{P}(X_0 = i) \times \mathbb{P}(X_1 = j | X_0 = i) = \sum_{i=1}^{k} \pi_{0,i} P_{ij}$ Cette relation est résumée par l'équation $\pi_1 = \pi_0 P$
- Par récurrence, on montre que

$$P_{ij}^{(n)} := \mathbb{P}(X_n = j | X_0 = i) = (P^n)_{ij}$$

  où P^n est la puissance n-ième de la matrice P.
- Ainsi:

$$\pi_n = \pi_0 P^n$$

Mesure invariante pour P

Soit π un vecteur (ligne) de probabilité sur $\mathcal{K}$.

- π est une **mesure invariante pour la chaîne de Markov de transition** P si:

$$\pi P = \pi$$

Mesure invariante pour P

Soit π un vecteur (ligne) de probabilité sur $\mathcal{K}$.

- π est une **mesure invariante pour la chaîne de Markov de transition** P si:

$$\pi P = \pi$$

- Si π_0 est une mesure invariante pour P, alors, pour tout n, $\pi_n = \pi_0$.
- Dans ce cas, les V.A. $X_0, \ldots, X_n$ sont identiquement distribuées (mais pas indépendantes!).

Irréductibilité

Une chaîne de Markov homogène sur $\mathcal{K}$, de transition P est **irréductible** si

$$\forall i, j \in \mathcal{K} \times \mathcal{K}, \ \exists \ n \text{ tel que } P_{i,j}^{(n)} > 0$$

- Pour deux états de la chaîne, il est possible d'accéder de l'un à l'autre en un temps fini.

Apériodicité

Soit $(X_n)_{n\geq 1}$ une chaîne de Markov homogène sur $\mathcal{K}$. Pour $k \in \mathcal{K}$,

- La *période* de l'état k, notée $d(k)$, est le P.G.C.D. de tous les entiers n tels que $P_{kk}^{(n)} > 0$ (avec la convention $pgcd(\emptyset) = +\infty$):

$$d(j) = pgcd\left\{n \geq 1, P_{kk}^{(n)} > 0\right\}$$

  Une chaîne est dite apériodique si pour tout k dans $\mathcal{K}$, $d(k) = 1$.

Apériodicité

Soit $(X_n)_{n\geq 1}$ une chaîne de Markov homogène sur $\mathcal{K}$. Pour $k \in \mathcal{K}$,

- La *période* de l'état k, notée $d(k)$, est le P.G.C.D. de tous les entiers n tels que $P_{kk}^{(n)} > 0$ (avec la convention $pgcd(\emptyset) = +\infty$):

$$d(j) = pgcd\left\{n \geq 1, P_{kk}^{(n)} > 0\right\}$$

  Une chaîne est dite apériodique si pour tout k dans $\mathcal{K}$, $d(k) = 1$.

Pour une chaîne irréductible, une condition suffisante pour être apériodique est qu'il existe un $k \in \mathcal{K}$ tel que $P_{kk} > 0$.

Théorème ergodique

Théorème ergodique

Soit $(X_n)_{n\geq 0}$ une chaîne de Markov sur $\mathcal{K}$ de loi initiale π_0 et de matrice de transition P. On suppose que cette chaîne est irréductible et apériodique. Alors:

Théorème ergodique

Soit $(X_n)_{n\geq 0}$ une chaîne de Markov sur $\mathcal{K}$ de loi initiale π_0 et de matrice de transition P. On suppose que cette chaîne est irréductible et apériodique. Alors:

1. Cette chaîne de Markov admet une unique mesure de probabilité invariante π.

Théorème ergodique

Soit $(X_n)_{n\geq 0}$ une chaîne de Markov sur $\mathcal{K}$ de loi initiale π_0 et de matrice de transition P. On suppose que cette chaîne est irréductible et apériodique. Alors:

1. Cette chaîne de Markov admet une unique mesure de probabilité invariante π.
2. $X_n \xrightarrow{loi} X$ où X est une v.a. de loi π.

Théorème ergodique

Soit $(X_n)_{n\geq 0}$ une chaîne de Markov sur $\mathcal{K}$ de loi initiale π_0 et de matrice de transition P. On suppose que cette chaîne est irréductible et apériodique. Alors:

1. Cette chaîne de Markov admet une unique mesure de probabilité invariante π.
2. $X_n \xrightarrow{loi} X$ où X est une v.a. de loi π.
3. Pour toute fonction φ intégrable par rapport à π, on a :

$$\frac{1}{M+1}\sum_{k=0}^{M}\varphi(X_k) \underset{M\to+\infty}{\xrightarrow{p.s.}} \mathbb{E}_\pi[\varphi(X)].$$

Théorème ergodique

Soit $(X_n)_{n\geq 0}$ une chaîne de Markov sur $\mathcal{K}$ de loi initiale π_0 et de matrice de transition P. On suppose que cette chaîne est irréductible et apériodique. Alors:

1. Cette chaîne de Markov admet une unique mesure de probabilité invariante π.
2. $X_n \xrightarrow{loi} X$ où X est une v.a. de loi π.
3. Pour toute fonction φ intégrable par rapport à π, on a :

$$\frac{1}{M+1}\sum_{k=0}^{M}\varphi(X_k) \underset{M\to+\infty}{\xrightarrow{p.s.}} \mathbb{E}_\pi[\varphi(X)].$$

4. Si $\varphi(X)$ admet un moment d'ordre supérieur à 2, on a

$$\sqrt{M}\left(\frac{1}{M+1}\sum_{k=0}^{n}\varphi(X_k) - \mathbb{E}_\pi[\varphi(X)]\right) \underset{M\to+\infty}{\xrightarrow{Loi}} \mathcal{N}(0,\sigma^2)$$

Théorème ergodique

Soit $(X_n)_{n\geq 0}$ une chaîne de Markov sur $\mathcal{K}$ de loi initiale π_0 et de matrice de transition P. On suppose que cette chaîne est irréductible et apériodique. Alors:

1. Cette chaîne de Markov admet une unique mesure de probabilité invariante π.
2. $X_n \xrightarrow{loi} X$ où X est une v.a. de loi π.
3. Pour toute fonction φ intégrable par rapport à π, on a :

$$\frac{1}{M+1}\sum_{k=0}^{M}\varphi(X_k) \underset{M\to+\infty}{\xrightarrow{p.s.}} \mathbb{E}_\pi[\varphi(X)].$$

4. Si $\varphi(X)$ admet un moment d'ordre supérieur à 2, on a

$$\sqrt{M}\left(\frac{1}{M+1}\sum_{k=0}^{n}\varphi(X_k) - \mathbb{E}_\pi[\varphi(X)]\right) \underset{M\to+\infty}{\xrightarrow{Loi}} \mathcal{N}(0,\sigma^2)$$

Une propriété analogue reste vraie quand la chaîne de Markov est à valeurs dans un ensemble continu (typiquement, $\mathbb{R}^d$).

Conséquence et intérêt pratique du théorème ergodique

- Pour estimer $\mathbb{E}_\pi[\varphi(X)]$, il suffit d'être capable de simuler une chaîne de Markov apériodique et irréductible de mesure de probabilité invariante π.

Conséquence et intérêt pratique du théorème ergodique

- Pour estimer $\mathbb{E}_\pi[\varphi(X)]$, il suffit d'être capable de simuler une chaîne de Markov apériodique et irréductible de mesure de probabilité invariante π.
- Il n'est pas nécessaire de savoir tirer selon π directement!

Conséquence et intérêt pratique du théorème ergodique

- Pour estimer $\mathbb{E}_\pi[\varphi(X)]$, il suffit d'être capable de simuler une chaîne de Markov apériodique et irréductible de mesure de probabilité invariante π.
- Il n'est pas nécessaire de savoir tirer selon π directement!
- Le point 2. dit qu'*au bout d'un certain temps*, les X_n simulés pourront être considérés comme de loi π (mais pas indépendants!)!

Conséquence et intérêt pratique du théorème ergodique

- Pour estimer $\mathbb{E}_\pi[\varphi(X)]$, il suffit d'être capable de simuler une chaîne de Markov apériodique et irréductible de mesure de probabilité invariante π.
- Il n'est pas nécessaire de savoir tirer selon π directement!
- Le point 2. dit qu'*au bout d'un certain temps*, les X_n simulés pourront être considérés comme de loi π (mais pas indépendants!)!
- Encore faut il être capable de construire une chaîne de Markov apériodique, irréductible, de loi invariante donnée par π!

Conséquence et intérêt pratique du théorème ergodique

- Pour estimer $\mathbb{E}_\pi[\varphi(X)]$, il suffit d'être capable de simuler une chaîne de Markov apériodique et irréductible de mesure de probabilité invariante π.
- Il n'est pas nécessaire de savoir tirer selon π directement!
- Le point 2. dit qu'*au bout d'un certain temps*, les X_n simulés pourront être considérés comme de loi π (mais pas indépendants!)!
- Encore faut il être capable de construire une chaîne de Markov apériodique, irréductible, de loi invariante donnée par π!
- $\longrightarrow$ Algorithme de Metropolis Hastings

Remarque sur le Théorème Central Limite

- Les V.A. dans l'estimateur Monte Carlo ne sont plus indépendantes.
- $\Rightarrow$ la variance σ^2 n'est absolument pas triviale (il ne s'agit pas de $\mathbb{V}[\varphi(X)]$)!
- Pas nécessairement facile à estimer!
- Ainsi, avoir un IC asymptotique sur $\mathbb{E}_\pi[\varphi(X)]$ n'est plus du tout immediat.

Algorithme de Metropolis Hastings

Réversibilité

Soit $\pi = (\pi_1, \ldots, \pi_K)$ une mesure de probabilité sur $\mathcal{K}$ et $(X_n)_{n \geq 0}$ une chaîne de Markov homogène de matrice de transition P et de loi initiale π_0.

Réversibilité

Soit $\pi = (\pi_1, \ldots, \pi_K)$ une mesure de probabilité sur $\mathcal{K}$ et $(X_n)_{n \geq 0}$ une chaîne de Markov homogène de matrice de transition P et de loi initiale π_0.

- π est **réversible** pour P si elle vérifie la condition d'équilibre:

$$\forall (i,j) \in \mathcal{K} \times \mathcal{K},\ \pi_i \times P_{ij} = \pi_j \times P_{ji}$$

- *Propriété:* Si π est réversible pour une chaîne de Markov de transition P, alors, π est une mesure de probabilité invariante pour P.

Réversibilité

Soit $\pi = (\pi_1, \ldots, \pi_K)$ une mesure de probabilité sur $\mathcal{K}$ et $(X_n)_{n\geq 0}$ une chaîne de Markov homogène de matrice de transition P et de loi initiale π_0.

- π est **réversible** pour P si elle vérifie la condition d'équilibre:

$$\forall (i,j) \in \mathcal{K} \times \mathcal{K},\ \pi_i \times P_{ij} = \pi_j \times P_{ji}$$

- *Propriété:* Si π est réversible pour une chaîne de Markov de transition P, alors, π est une mesure de probabilité invariante pour P.
- *Preuve* Soit π une mesure de probabilité réversible pour P. On a tout de suite que

$$\begin{aligned}
\forall j \in \mathcal{K} \quad (\pi P)_j &= \sum_{i=1}^{K} \pi_i P_{ij} \\
&= \sum_{i=1}^{K} \pi_j P_{ji} && \text{par réversibilité} \\
&= \pi_j && \text{par propriété de } P \\
\Rightarrow \pi P &= \pi
\end{aligned}$$

Objectif de l'algorithme

- On veut simuler selon la loi π.
- Construire une chaîne de Markov irréductible et apériodique, de loi initiale π_0 et de transition P, **réversible pour P**
- On va se servir pour ça d'une chaîne de Markov de transition Q (réversible et apériodique) **parcourant le même espace que** P (le support de π).

Algorithme de Metropolis Hastings (formulation discrete)

- Soit Q une matrice stochastique $K \times K$ satisfaisant la condition suivante:

$$\forall (i,j) \in \mathcal{K} \times \mathcal{K}, Q_{ij} > 0 \Leftrightarrow Q_{ji} > 0$$

- Soit $(X_n)_{n \geq 0}$ la suite de variables aléatoires construite ainsi:

Algorithme de Metropolis Hastings (formulation discrete)

- Soit Q une matrice stochastique $K \times K$ satisfaisant la condition suivante:

$$\forall (i,j) \in \mathcal{K} \times \mathcal{K}, Q_{ij} > 0 \Leftrightarrow Q_{ji} > 0$$

- Soit $(X_n)_{n \geq 0}$ la suite de variables aléatoires construite ainsi:

1. On simule X_0 selon π_0.

Algorithme de Metropolis Hastings (formulation discrete)

- Soit Q une matrice stochastique $K \times K$ satisfaisant la condition suivante:

$$\forall (i,j) \in \mathcal{K} \times \mathcal{K}, Q_{ij} > 0 \Leftrightarrow Q_{ji} > 0$$

- Soit $(X_n)_{n \geq 0}$ la suite de variables aléatoires construite ainsi:

1. On simule X_0 selon π_0.
2. Pour $n \geq 1$:

a. On tire Y_n selon la loi $Q_{X_{n-1}\bullet}$ (la ligne de Q donnée par X_{n-1}).

Algorithme de Metropolis Hastings (formulation discrete)

- Soit Q une matrice stochastique $K \times K$ satisfaisant la condition suivante:

$$\forall (i,j) \in \mathcal{K} \times \mathcal{K}, Q_{ij} > 0 \Leftrightarrow Q_{ji} > 0$$

- Soit $(X_n)_{n \geq 0}$ la suite de variables aléatoires construite ainsi:

1. On simule X_0 selon π_0.
2. Pour $n \geq 1$:

a. On tire Y_n selon la loi $Q_{X_{n-1}\bullet}$ (la ligne de Q donnée par X_{n-1}).
b. On tire une loi uniforme U indépendante de Y_n.

Algorithme de Metropolis Hastings (formulation discrete)

- Soit Q une matrice stochastique $K \times K$ satisfaisant la condition suivante:

$$\forall (i,j) \in \mathcal{K} \times \mathcal{K}, Q_{ij} > 0 \Leftrightarrow Q_{ji} > 0$$

- Soit $(X_n)_{n \geq 0}$ la suite de variables aléatoires construite ainsi:

1. On simule X_0 selon π_0.
2. Pour $n \geq 1$:

a. On tire Y_n selon la loi $Q_{X_{n-1}\bullet}$ (la ligne de Q donnée par X_{n-1}).
b. On tire une loi uniforme U indépendante de Y_n.
c. On calcule la quantité

$$\alpha(X_{n-1}, Y_n) = \min\left(1, \frac{\pi_{Y_n} Q_{Y_n X_{n-1}}}{\pi_{X_{n-1}} Q_{X_{n-1} Y_n}}\right)$$

Algorithme de Metropolis Hastings (formulation discrete)

- Soit Q une matrice stochastique $K \times K$ satisfaisant la condition suivante:

$$\forall (i,j) \in \mathcal{K} \times \mathcal{K}, Q_{ij} > 0 \Leftrightarrow Q_{ji} > 0$$

- Soit $(X_n)_{n \geq 0}$ la suite de variables aléatoires construite ainsi:

1. On simule X_0 selon π_0.
2. Pour $n \geq 1$:

a. On tire Y_n selon la loi $Q_{X_{n-1}\bullet}$ (la ligne de Q donnée par X_{n-1}).
b. On tire une loi uniforme U indépendante de Y_n.
c. On calcule la quantité

$$\alpha(X_{n-1}, Y_n) = \min\left(1, \frac{\pi_{Y_n} Q_{Y_n X_{n-1}}}{\pi_{X_{n-1}} Q_{X_{n-1} Y_n}}\right)$$

d. On pose:

$$X_n = \begin{cases} Y_n & \text{si } U \leq \alpha(X_{n-1}, Y_n) \\ X_{n-1} & \text{sinon} \end{cases}$$

Algorithme de Metropolis Hastings (formulation discrete)

- Soit Q une matrice stochastique $K \times K$ satisfaisant la condition suivante:

$$\forall (i,j) \in \mathcal{K} \times \mathcal{K}, Q_{ij} > 0 \Leftrightarrow Q_{ji} > 0$$

- Soit $(X_n)_{n \geq 0}$ la suite de variables aléatoires construite ainsi:

1. On simule X_0 selon π_0.
2. Pour $n \geq 1$:

a. On tire Y_n selon la loi $Q_{X_{n-1}\bullet}$ (la ligne de Q donnée par X_{n-1}).
b. On tire une loi uniforme U indépendante de Y_n.
c. On calcule la quantité

$$\alpha(X_{n-1}, Y_n) = \min\left(1, \frac{\pi_{Y_n} Q_{Y_n X_{n-1}}}{\pi_{X_{n-1}} Q_{X_{n-1} Y_n}}\right)$$

d. On pose:

$$X_n = \begin{cases} Y_n & \text{si } U \leq \alpha(X_{n-1}, Y_n) \\ X_{n-1} & \text{sinon} \end{cases}$$

- **Propriété 1:** $(X_n)_{n \geq 1}$ est une chaîne de Markov de transition P où

$$P_{ij} = Q_{ij}\alpha(i,j) \text{ si } i \neq j, \quad P_{jj} = 1 - \sum_{j \neq i} P_{ij}$$

- **Propriété 2:** De plus π **est invariante pour** P**.**

Preuve

Matrice de transition P

On veut montrer que:

$$P_{ij} = Q_{ij}\alpha(i,j) \text{ si } i \neq j, \quad P_{jj} = 1 - \sum_{j \neq i} P_{ij}$$

Preuve

Matrice de transition P

On veut montrer que:

$$P_{ij} = Q_{ij}\alpha(i,j) \text{ si } i \neq j, \quad P_{jj} = 1 - \sum_{j \neq i} P_{ij}$$

Soit $i \neq j$:

$$\begin{aligned}
\mathbb{P}(X_n = j | X_{n-1} = i) &= \mathbb{P}\left(Y_n = j, U \leq \alpha(X_{n-1}, Y_n) | X_{n-1} = i\right) \\
&= \mathbb{P}\left(Y_n = j, U \leq \alpha(i,j) | X_{n-1} = i\right) \\
&= \mathbb{P}\left(U \leq \alpha(i,j) | X_{n-1} = i, Y_n = j\right) \mathbb{P}\left(Y_n = j | X_{n-1} = i\right) \\
&= Q_{ij}\alpha(i,j)
\end{aligned}$$

Preuve que π est mesure invariante

Il suffit de montrer que π est réversible pour P.

Soient $i \neq j \in \mathcal{K}$:

Preuve que π est mesure invariante

Il suffit de montrer que π est réversible pour P.

Soient $i \neq j \in \mathcal{K}$:

$$\begin{aligned}
\pi_i P_{ij} &= \pi_i Q_{ij} \alpha(i,j) \\
&= \pi_i Q_{ij} \min\left(1, \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}\right) \\
&= \min\left(\pi_i Q_{ij}, \pi_j Q_{ji}\right) \\
&= \pi_j Q_{ji} \min\left(\frac{\pi_i Q_{ij}}{\pi_j Q_{ji}}, 1\right) \\
&= \pi_j Q_{ji} \alpha(j,i) \\
&= \pi_j P_{ji}
\end{aligned}$$

Algorithme dans le cas continu

Supposons qu'on veuille simuler dans $\mathbb{R}^d$ selon une densité π, éventuellement connue à une constante près, c'est à dire que

$$\forall x \in \mathbb{R}^d,\ \pi(x) = \frac{\tilde{\pi}(x)}{\int_{\mathbb{R}^d} \tilde{\pi}(z)\mathrm{d}z}$$

On remplace alors la matrice de transition par un *noyau de transition* sur $\mathbb{R}^d$, à savoir une fonction

$$\begin{array}{rccc} q: & \mathbb{R}^d \times \mathbb{R}^d & \mapsto & \mathbb{R}_+ \\ & (x,y) & \mapsto & q(x,y) \geq 0 \end{array}$$

telle que $\int_{\mathbb{R}^d} q(x,y)\mathrm{d}y = 1$ (typiquement, la loi d'une marche aléatoire centrée en x.

Algorithme dans le cas continu

Supposons qu'on veuille simuler dans $\mathbb{R}^d$ selon une densité π, éventuellement connue à une constante près, c'est à dire que

$$\forall x \in \mathbb{R}^d,\ \pi(x) = \frac{\tilde{\pi}(x)}{\int_{\mathbb{R}^d} \tilde{\pi}(z)\mathrm{d}z}$$

On remplace alors la matrice de transition par un *noyau de transition* sur $\mathbb{R}^d$, à savoir une fonction

$$\begin{array}{rccc} q: & \mathbb{R}^d \times \mathbb{R}^d & \mapsto & \mathbb{R}_+ \\ & (x,y) & \mapsto & q(x,y) \geq 0 \end{array}$$

telle que $\int_{\mathbb{R}^d} q(x,y)\mathrm{d}y = 1$ (typiquement, la loi d'une marche aléatoire centrée en x.

Si on sait simuler, pour x fixé, selon q, et qu'on a $q(x,y) > 0 \Leftrightarrow q(y,x) > 0$, alors, l'algorithme de Metropolis reste valide en remplaçant π par $\tilde{\pi}$ et Q par q.

Algorithme dans le cas continu

Supposons qu'on veuille simuler dans $\mathbb{R}^d$ selon une densité π, éventuellement connue à une constante près, c'est à dire que

$$\forall x \in \mathbb{R}^d,\ \pi(x) = \frac{\tilde{\pi}(x)}{\int_{\mathbb{R}^d} \tilde{\pi}(z)\mathrm{d}z}$$

On remplace alors la matrice de transition par un *noyau de transition* sur $\mathbb{R}^d$, à savoir une fonction

$$\begin{array}{rccc} q: & \mathbb{R}^d \times \mathbb{R}^d & \mapsto & \mathbb{R}_+ \\ & (x,y) & \mapsto & q(x,y) \geq 0 \end{array}$$

telle que $\int_{\mathbb{R}^d} q(x,y)\mathrm{d}y = 1$ (typiquement, la loi d'une marche aléatoire centrée en x.

Si on sait simuler, pour x fixé, selon q, et qu'on a $q(x,y) > 0 \Leftrightarrow q(y,x) > 0$, alors, l'algorithme de Metropolis reste valide en remplaçant π par $\tilde{\pi}$ et Q par q.

- Le ratio ne nécessite pas la constante de normalisation car

$$\frac{\tilde{\pi}(y)}{\tilde{\pi}(x)} = \frac{\pi(y)}{\pi(x)}$$

Exemple 2: cas non conjugué

Exemple: Prédiction de présence d'oiseaux

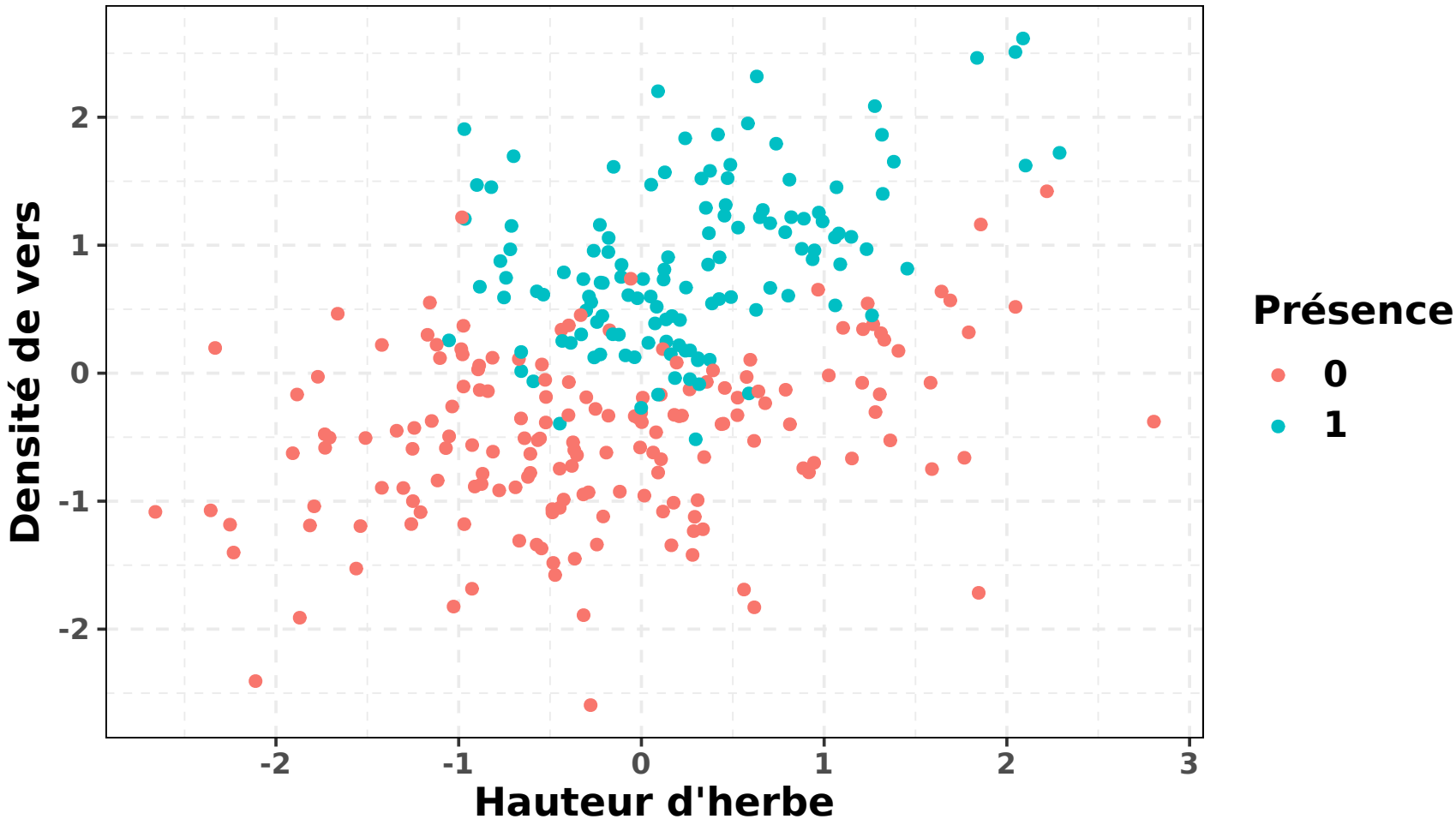
Une étude consiste en l'observation de la présence ou non de la linotte mélodieuse sur différents sites échantillonnés.

Caractéristiques des sites

Sur ces 300 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



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

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

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

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

Posterior

Le posterior est donc donné par:

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

Algorithme de Metropolis Hastings

La loi stationnaire cible est $\pi(\mathbf{y}|\theta)$. Pour $n = 300$, l'acceptation rejet vu au cours précédent fonctionnera très mal en pratique.

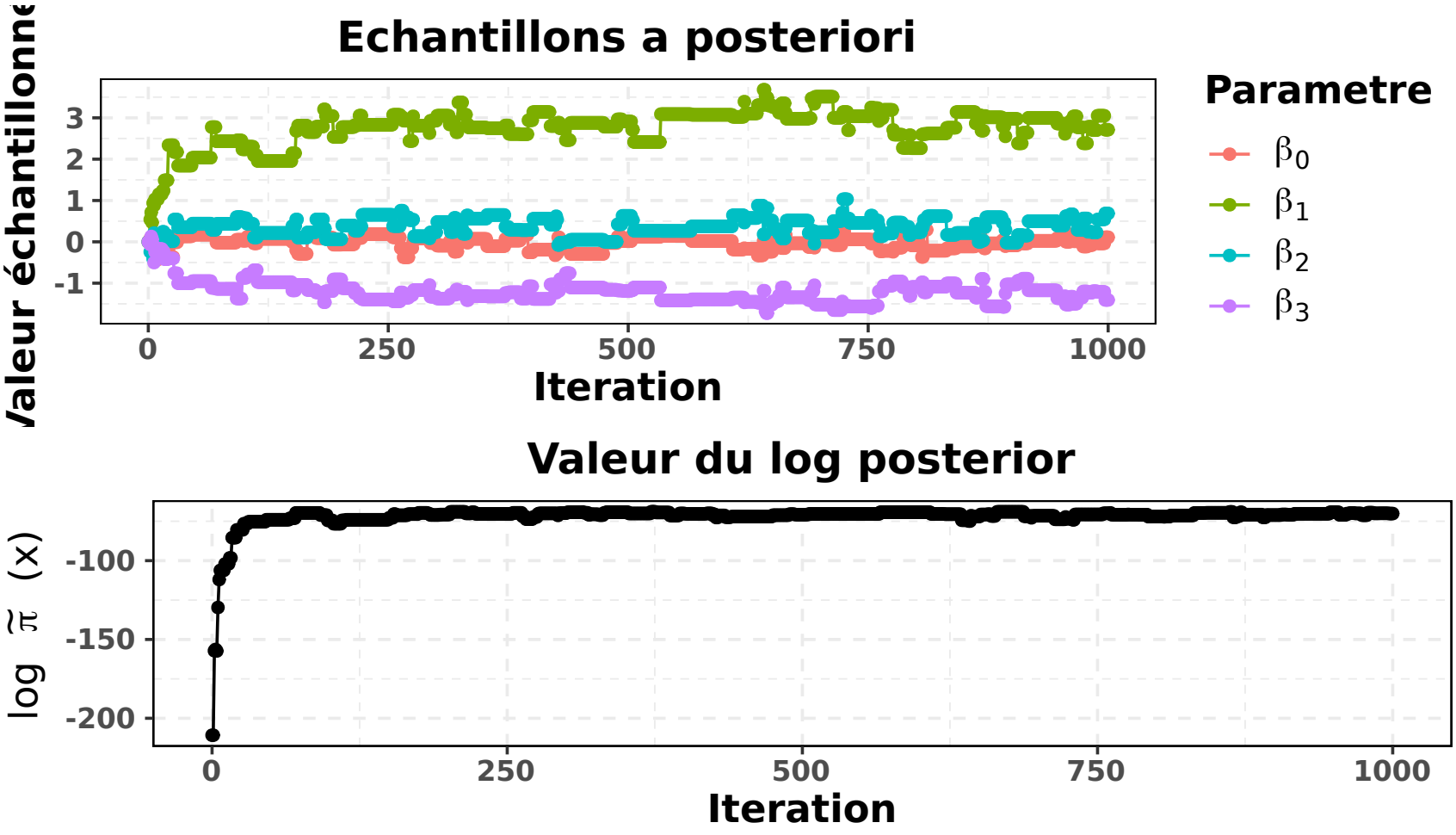
Algorithme de Metropolis Hastings

La loi stationnaire cible est $\pi(\mathbf{y}|\theta)$. Pour $n = 300$, l'acceptation rejet vu au cours précédent fonctionnera très mal en pratique.

On fait un algorithme de Metropolis Hastings avec comme loi de proposition une marche aléatoire dans $\mathbb{R}^4$, de matrice de covariance $\tau^2 \times I_4$.

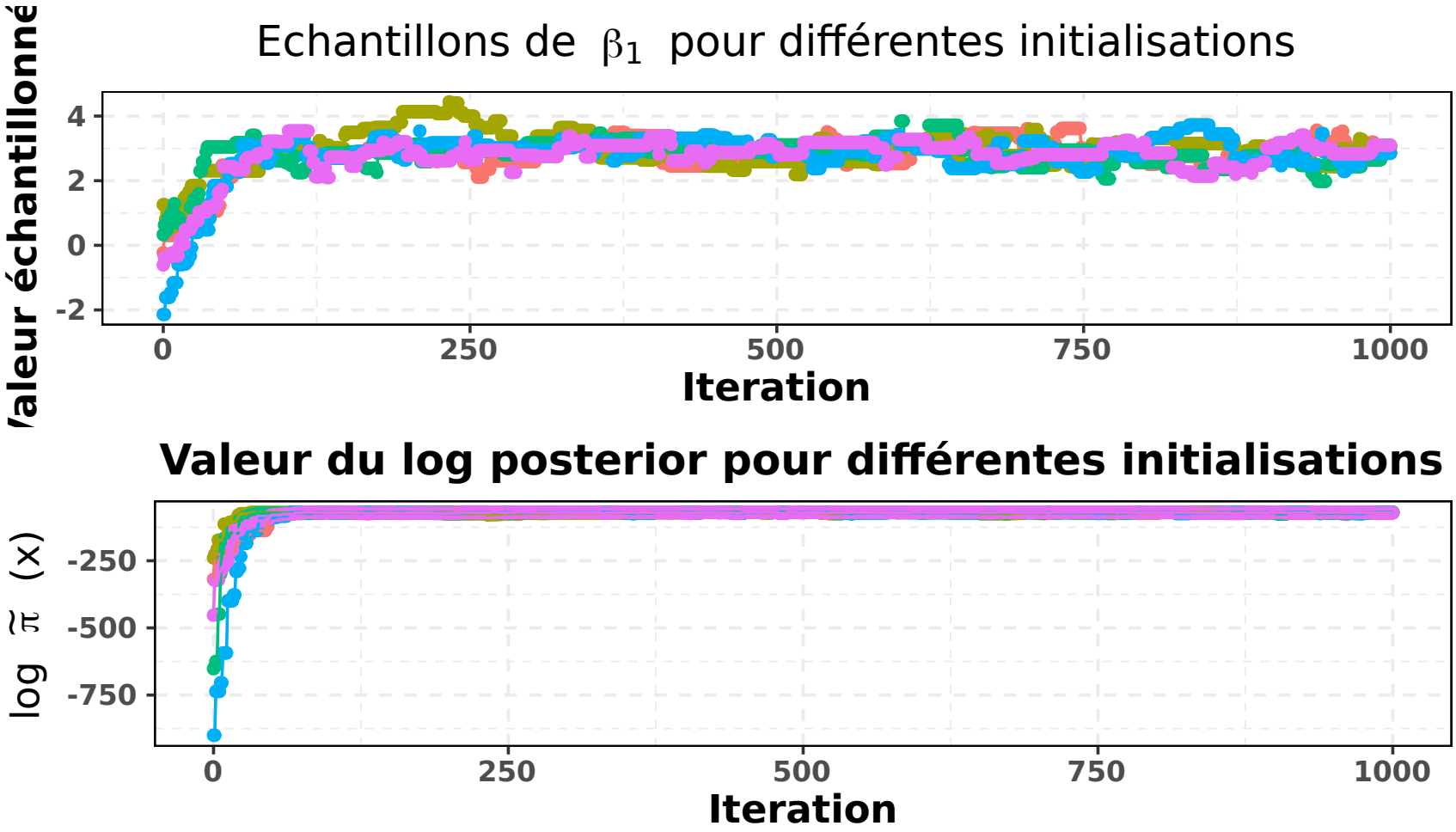
Résultat d'un algorithme lancé depuis un point de départ

- On choisit $\beta^{(0)} = (0, 0, 0, 0)$ et $\tau^2 = 0.1$, on lance 1000 itérations.

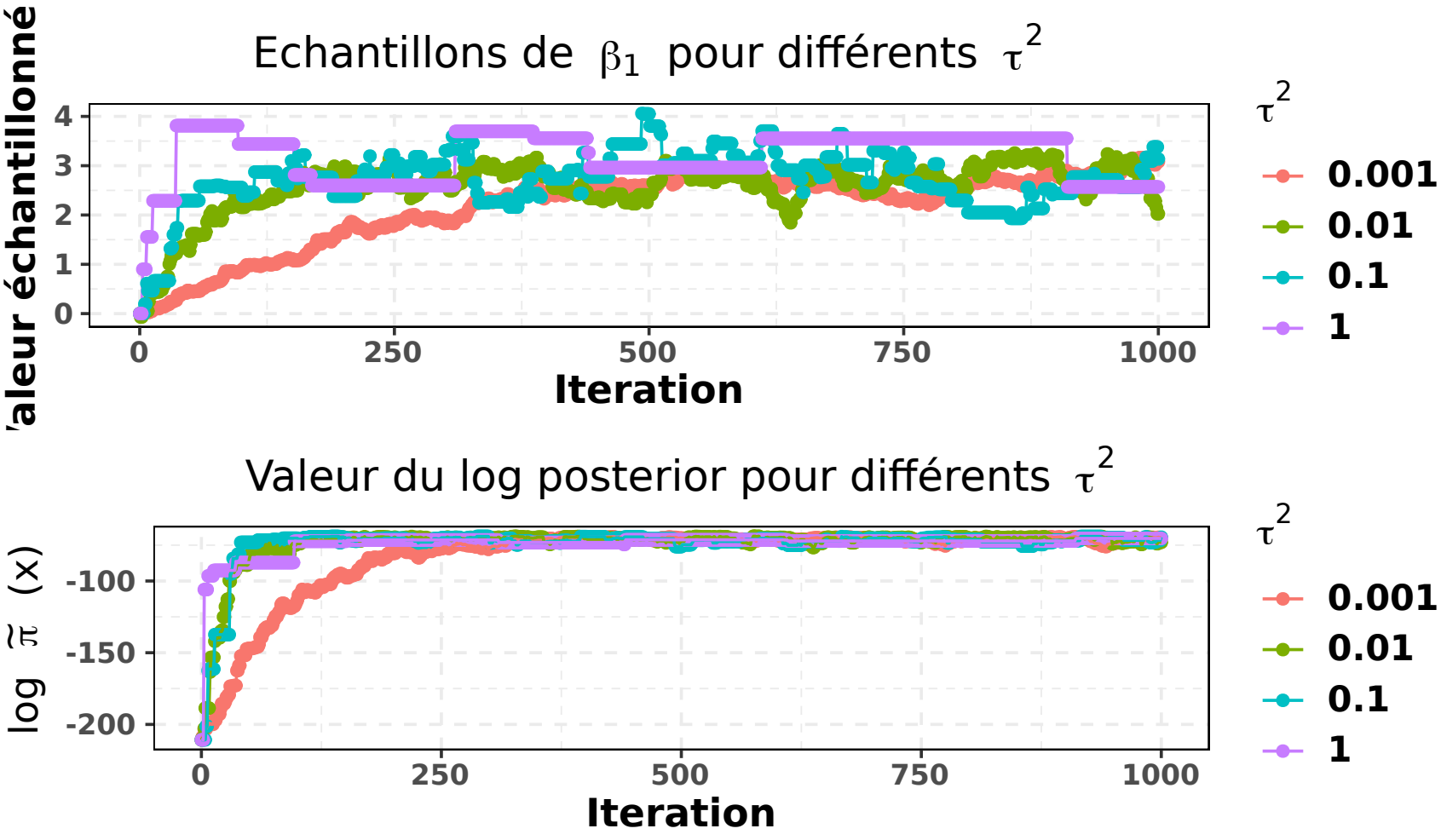


Sensibilité au point de départ

Il faut toujours vérifié la sensibilité au point de départ!



Influence de τ^2



Influence de τ^2

Taux d'acceptation dans l'algorithme

τ^2	Taux d'acceptation
0.001	0.781
0.010	0.515
0.100	0.147
1.000	0.013

Influence de τ^2

Taux d'acceptation dans l'algorithme

τ^2	Taux d'acceptation
0.001	0.781
0.010	0.515
0.100	0.147
1.000	0.013

Autocorrelation dans les chaînes

Correlation entre empirique entre β_1^n et $\beta_1^{(n+1)}$

τ^2	Autocorrelation
0.001	0.9993751
0.010	0.9917581
0.100	0.9835038
1.000	0.9864433

Reduction de l'autocorrelation

En pratique, on choisira une fracion des points. On appelle cela le **thinning**.

Reduction de l'autocorrelation

En pratique, on choisira une fracion des points. On appelle cela le **thinning**.

Autocorrélation en prenant un point sur 100.

τ^2	Autocorrelation
0.001	0.8092081
0.010	0.2395796
0.100	0.1502224
1.000	0.3861150

Estimation de la loi

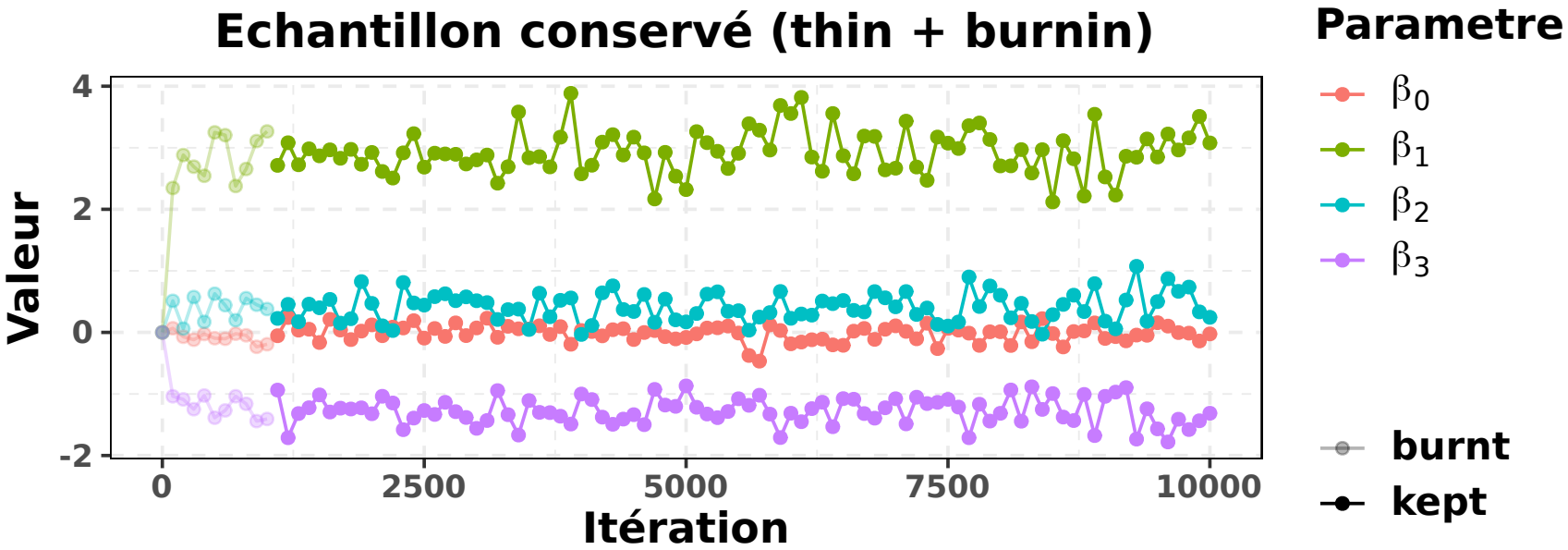
Les premières valeurs n'ont aucune raison d'être tirées selon la loi cible.

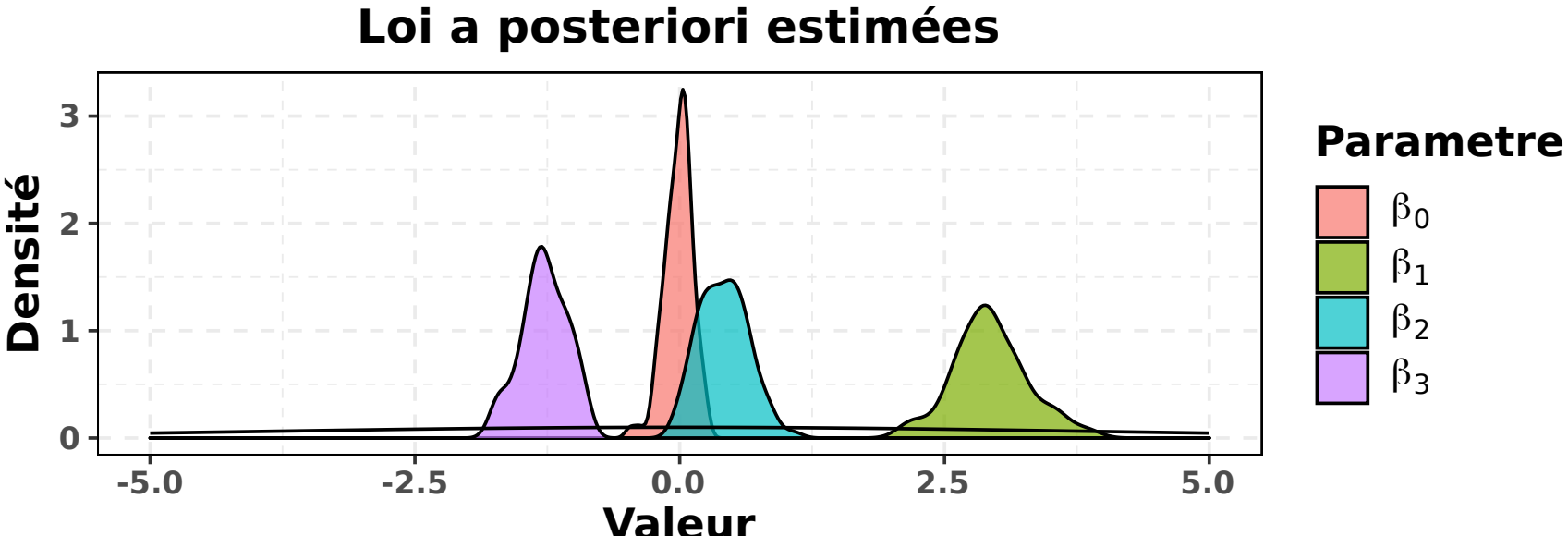
En pratique, on les supprimera. On appelle cela le **burn-in**.

Estimation de la loi

Les premières valeurs n'ont aucune raison d'être tirées selon la loi cible.

En pratique, on les supprimera. On appelle cela le **burn-in**.





Autres algorithmes MCMC

Echantillonneur de Gibbs

- Utile quand θ est en grande dimension;
- On suppose qu'on sait simuler selon les loi conditionnelles de θ

Echantillonneur de Gibbs

- Utile quand θ est en grande dimension;
- On suppose qu'on sait simuler selon les loi conditionnelles de θ
- Soit X un vecteur aléatoire en dimension d $X = (X^{(1)}, \ldots, X^{(d)})$.
- On note $X^{-(\ell)} = (X^{(1)}, \ldots, X^{(\ell-1)}, X^{(\ell+1)}, X^{(d)})$,
- Si on sait simuler la variable aléatoire $X^{(\ell)}|X^{(-\ell)}$, l'algo est le suivant:

1. Prendre $X_0 = (X_0^{(1)}, \ldots, X_0^{(d)})$ tiré selon une loi initiale.
2. Pour $k \geq 1$:

   2.1 Tirer ℓ uniformément dans $\{1, \ldots, d\}$;

   2.2 Simuler Y selon la loi $X^{(\ell)}|\{X^{(-\ell)} = X_{k-1}^{(-\ell)}\}$

   2.3 Poser $X_k = (X_{k-1}^{(1)}, \ldots, X_{k-1}^{(\ell-1)}, Y, X_{k-1}^{(\ell+1)}, X_{k-1}^{(d)})$

Propriété de l'échantillonneur de Gibbs

- L'échantillonneur de Gibbs est équivalent à un algorithme de Metropolis Hastings où la quantité α est toujours égale à 1,
- C'est à dire un Metropolis Hastings où on n'accepte tous les candidats!
- Algorithme utile dès que la simulation des lois conditionnelles est faisable.
- Si les lois conditionnelles induisent une matrice de transition (ou un noyau) de Markov irréductible et apériodique, alors le théorème ergodique s'applique.