Régression Pénalisée

Julien Perrin, Marianne Sorba & Xingyuan Xuegre

ENSAE ParisTech

6 septembre 2017

Le modèle

On considère le modèle de régression pénalisée suivant :

- $Y_i = \beta^T X_i + U_i$ avec $U_i \sim \mathcal{N}(0, \sigma^2)$, $\beta = (\beta_1, ..., \beta_p)$ et $X_i = (X_1^i, ..., X_p^i)$. On note $Y = Vect((Y_i)_{i \leq n})$ et $X = Vect((X_i)_{i \leq n})$
- β est muni d'une loi à priori $\pi(\beta) \propto exp\{-\lambda \sum_{i=1}^{p} |\beta_i|^\kappa\}$ avec $\kappa \in [0, 2]$ et $\lambda > 0$

Objectifs :

- Approcher la densité de la loi à posteriori $\pi(\beta | Y, \kappa, \sigma, \lambda)$ selon différentes méthodes
- En déduire l'estimateur $\hat{\beta}_{Bayes} = \mathbb{E}^{\pi}_{Y,\kappa,\sigma,\lambda}(\beta | Y, \kappa, \sigma, \lambda)$ qui minimise le coût quadratique à posteriori

Résultats préliminaires

Pour $\kappa = 0$, la loi à posteriori $\beta|Y, \kappa, \sigma, \lambda$ suit une loi gaussienne

$$\mathcal{N}\left(m(X,Y), \Gamma(X)\right) \text{ avec } \begin{cases} m(X,Y) = \left(\sum_{i=1}^{n} X_i^T X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i^T Y_i\right) \\ \Gamma(X) = \sigma^2 (\sum_{i=1}^{n} X_i^T X_i)^{-1} \end{cases}$$

Preuve

$$\begin{aligned} \pi(\beta|Y, \kappa, \sigma, \lambda) &\propto \mathbb{P}(Y|\beta, \sigma)\pi(\beta|\kappa, \lambda) \\ &\propto \prod_{i=1}^{n} \mathbb{P}(Y_i|\beta, \sigma)\pi(\beta|\kappa, \lambda) \\ &\propto exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{Y_i - \beta^T X_i}{\sigma}\right)^2\right) exp\left(-\lambda p\right) \\ &\propto exp\left(-\frac{1}{2}(\beta - m(X,Y))^T)\Gamma(X)^{-1}(\beta - m(X,Y))\right) \end{aligned}$$

On reconnaît l'expression de la densité d'une loi $\mathcal{N}\left(m(X,Y), \Gamma(X)\right)$

Résultats préliminaires

Pour $\kappa = 2$, la loi à posteriori $\beta|Y, \kappa, \sigma, \lambda$ suit une loi gaussienne $\mathcal{N}(m(X,Y), \Gamma(X))$ avec $\begin{cases} m(X,Y) = \left(\sum_{i=1}^n X_i^T X_i + \lambda I_p\right)^{-1} \left(\sum_{i=1}^n X_i^T Y_i\right) \\ \Gamma(X) = \sigma^2 (\sum_{i=1}^n X_i X_i^T + \lambda I_p)^{-1} \end{cases}$

Preuve

$$\begin{aligned}
\pi(\beta|Y, \kappa, \sigma, \lambda) &\propto \mathbb{P}(Y|\beta, \sigma)\pi(\beta|\kappa, \lambda) \\
&\propto \prod_{i=1}^n \mathbb{P}(Y_i|\beta, \sigma)\pi(\beta|\kappa, \lambda) \\
&\propto exp\left(-\frac{1}{2}\sum_{i=1}^n \left(\frac{Y_i - \beta^T X_i}{\sigma}\right)^2\right) exp\left(-\lambda\|\beta\|_2^2\right) \\
&\propto exp\left(-\frac{1}{2}(\beta - m(X,Y))^T \Gamma(X)^{-1}(\beta - m(X,Y))\right)
\end{aligned}$$

On reconnait l'expression de la densité d'une loi $\mathcal{N}(m(X,Y), \Gamma(X))$

$\kappa \in]0,2[$: Une loi à posteriori difficilement simulable

Pour $\kappa \in]0,2[$, la densité de la loi à posteriori ne correspond pas à une loi connu : $\pi(\beta|Y,\kappa,\sigma,\lambda) \propto exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{Y_i-\beta^T X_i}{\sigma}\right)^2 - \lambda\sum_{i=1}^{p}|b_i|^\kappa\right)$

Notre travail : approcher cette loi selon trois méthodes de Monte-Carlo :

1. Échantillonnage préférentiel (Importance Sampling)
2. Algorithme de Metropolis à marche aléatoire
3. Échantillonnage de Gibbs (Gibbs sampling)

Importance Sampling : Le principe

On souhaite estimer $\hat{\beta}_{Bayes} = \mathbb{E}^{\pi}_{Y,\kappa,\sigma,\lambda}(\beta|Y,\kappa,\sigma,\lambda)$ sauf qu'on ne sait pas simuler selon $\beta|Y,\kappa,\sigma,\lambda$.

$\Longrightarrow$ On se ramène à une loi de densité g que l'on sait simuler, et on calcule

$$\hat{\beta}_{Bayes} = \mathbb{E}^{\pi}_{Y,\kappa,\sigma,\lambda}(\beta|Y,\kappa,\sigma,\lambda) = \mathbb{E}^{g}[\beta \times \frac{\pi(\beta|Y,\kappa,\sigma,\lambda)}{g(\beta)}] = \mathbb{E}^{g}[\beta \times w(\beta)]$$

avec w la fonction de poids.

Importance sampling

1. Calculer $\begin{cases} \Gamma = \sigma^2 \left(\sum_{i=1}^n X_i^T X_i\right)^{-1} \\ m = \left(\sum_{i=1}^n X_i^T X_i\right)^{-1} \left(\sum_{i=1}^n X_i^T Y_i\right) \end{cases}$ les paramètres de la loi à posteriori pour $\kappa = 0$ (On peut aussi prendre $\kappa = 2$)

2. simuler $(\beta_1, ..., \beta_N)$ selon une normale $\mathcal{N}(m, \Gamma)$ de densité g

3. $\hat{\beta}_{AIS} = \hat{\mathbb{E}}^g[\beta \times w(\beta)] = \frac{\sum_{i=1}^N \beta_i w(\beta_i)}{\sum_{i=1}^N w(\beta_i)} = \frac{\sum_{i=1}^N \beta_i exp\{-\lambda \sum_{j=1}^p |\beta_{i,j}|^\kappa\}}{\sum_{i=1}^N exp\{-\lambda \sum_{j=1}^p |\beta_{i,j}|^\kappa\}}$

4. Évaluer l'importance sampling selon le critère de taille effectif de l'échantillon : $ESS = \frac{\left(\sum_{i=1}^n w(\beta_i)\right)^2}{\sum_{i=1}^n w(\beta_i)^2}$

Algorithme de Metropolis-Hastings

On souhaite engendrer une chaîne de Markov $(\beta^{(t)})$ dont la loi stationnaire est la loi cible de densité $\pi(\beta|Y,\kappa,\sigma,\lambda)$:

1. Initialiser $\beta^{(0)}$ selon une $\mathcal{U}_{[-1,1]}$
2. Engendrer $\epsilon^{(t)}$ selon une loi de proposition q préalablement choisi
3. Calculer le rapport d'acceptation
   $\rho(\beta^{(t)},\epsilon^{(t)}) = min\left\{1, \frac{\pi(\beta^{(t)}|Y,\kappa,\sigma,\lambda)}{\pi(\beta^{(t)}+\epsilon^{(t)}|Y,\kappa,\sigma,\lambda)}\right\}$
4. Tirer une loi $U \sim \mathcal{U}_{[0,1]}$: $\begin{cases} \beta^{(t+1)} = \beta^{(t)} + \epsilon^{(t)} \text{ si } U \leq \rho(\beta^{(t)},\epsilon^{(t)}) \\ \beta^{(t+1)} = \beta^{(t)} \text{ sinon} \end{cases}$

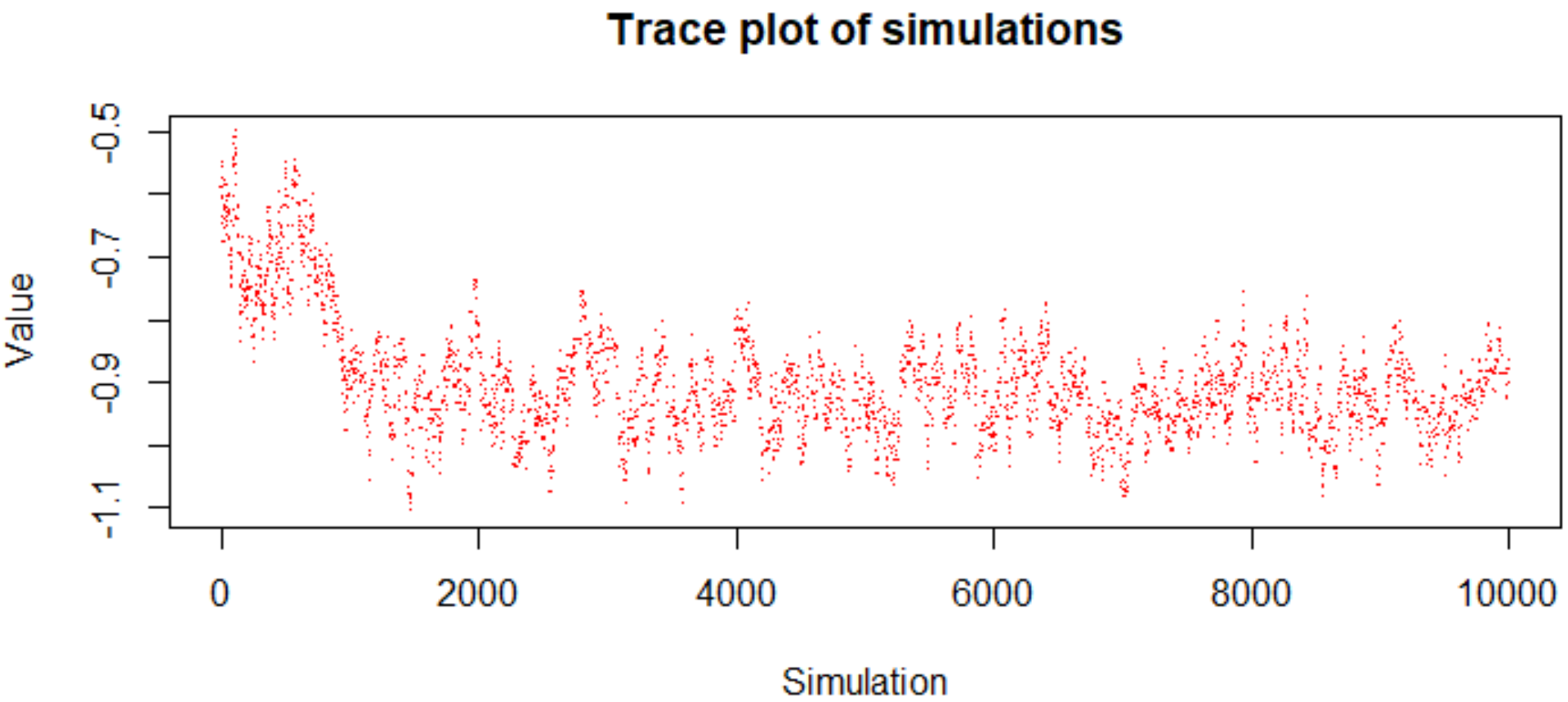
Les $\beta^{(t)}$ convergent vers la loi cible $\Longrightarrow$ On considère qu'à partir de T_0 suffisamment grand, les $\beta^{(t)}$ suivent la loi cible. On calcule ensuite $\hat{\beta}_{Bayes} = \hat{\mathbb{E}}^{\pi}_{Y,\kappa,\sigma,\lambda}(\beta|Y,\kappa,\sigma,\lambda)$ par LGN.

Calibration de l'algorithme de Metropolis

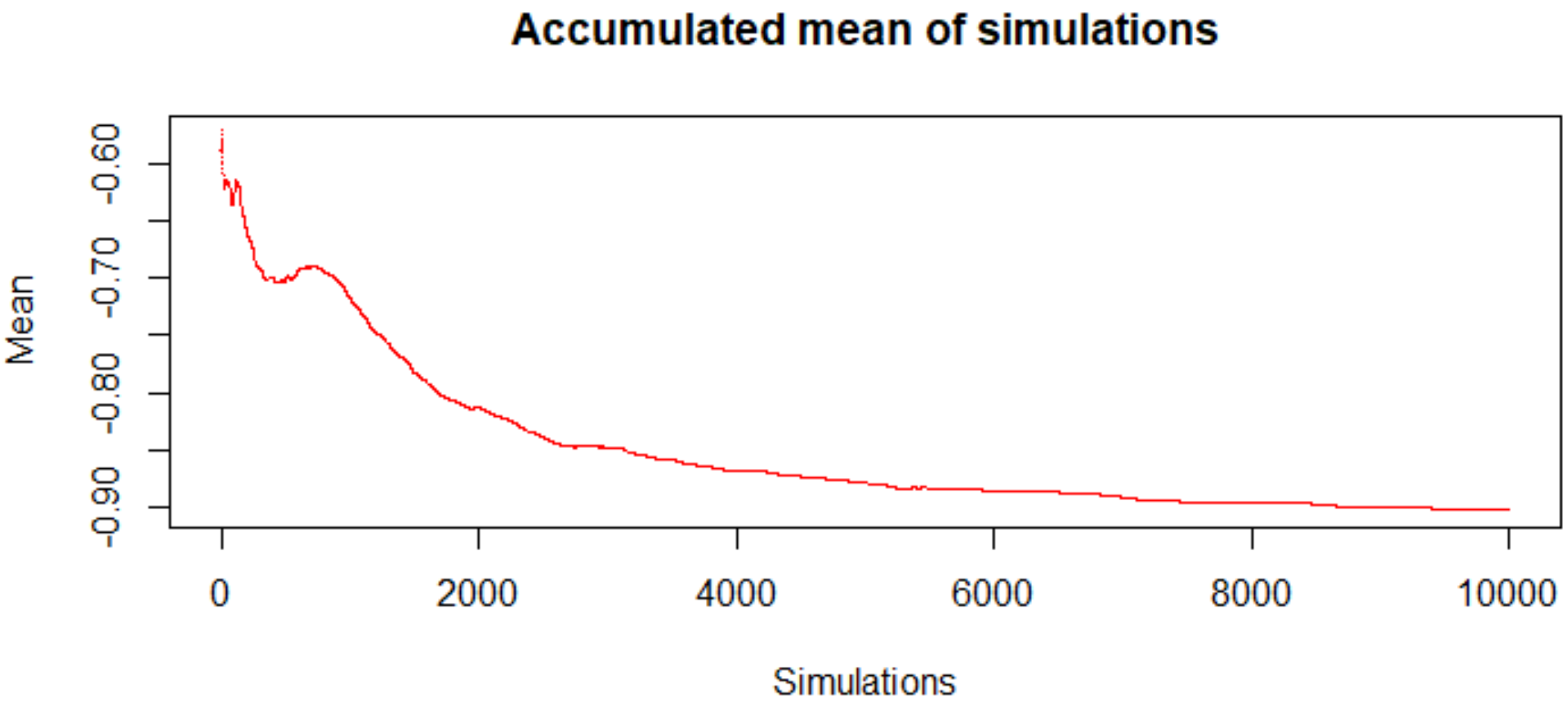
On regarde les quatre indices ci-dessous pour calibrer l'algorithme :

- Le taux d'acceptation (moyenne des $\rho(\beta^{(t)}, y_t)$) : On le veux le plus proche possible de 0.25.
- Le graphique de la trajectoire(Trace plot) : Quand la chaîne converge, cette trajectoire est stable sans tendance ni période.
- Les moyennes des sommes accumulées (Ergodic mean). Quand la chaîne converge, la moyenne empirique tend vers une constante.
- L'autocorrélogramme : grande autocorrélation $\rightarrow$ vitesse de convergence lente. Si l'autocorrélation ne diminue pas avec les lags, on doit reparamétrer la loi de proposition.

Exemple d'un graphique de la trajectoire



Exemple d'un graphique des moyennes accumulées

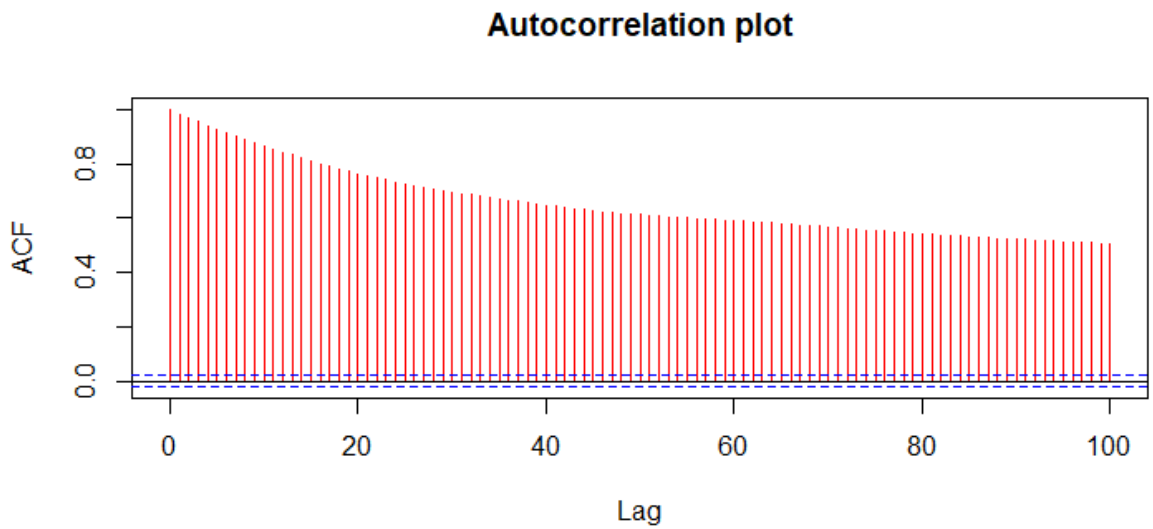


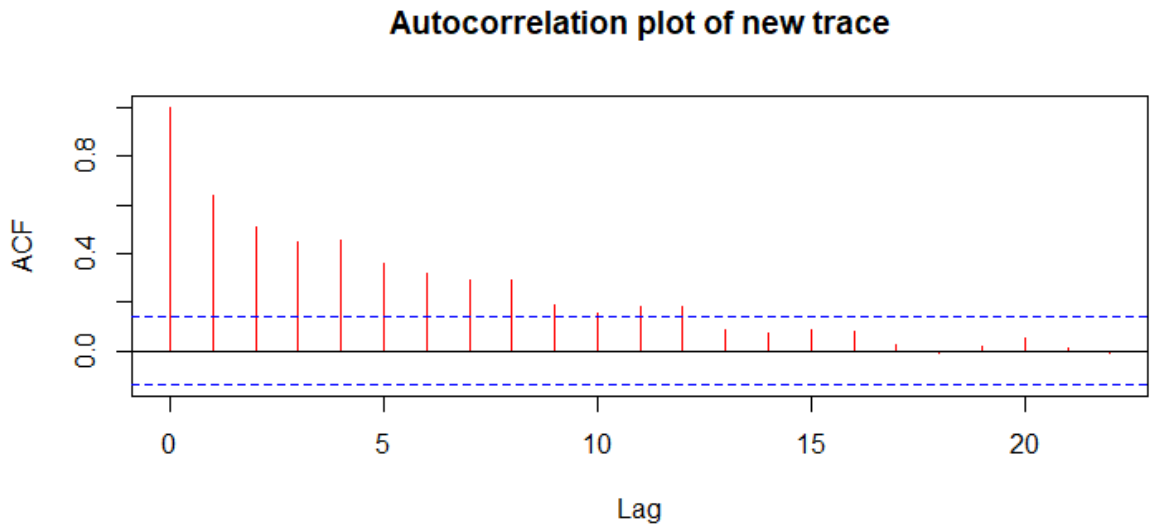
Échantillonnage de Metropolis

1. On prend un T_0 suffisamment grand et on prend les simulations engendrées après cet instant.

2. Afin d'éviter l'autocorrélation élevée qui peut empêcher la convergence vers la loi cible, on saute dans l'échantillonnage avec une largeur de saut w (Jumping width), c'est-à-dire que l'on prend $\beta_{T_0}, \beta_{T_0+w}, \beta_{T_0+2w}, \ldots$.

Autocorrélogramme avant/après échantillonnage





Comparaison avec l'Importance Sampling

Nombre de simulations	σ	κ	λ
10000	1	0.5	0.5

Table 1 – Coefficents du modèle

En utilisant la moyenne empirique obtenue par la méthode d'Importance Sampling avec un nombre de simulations égal à 1,000,000 qui est considéré comme le vrai β ici, on calcule l'erreur quadratique de nos estimateurs par les deux méthodes.

Comparaison sans effets croisés

Méthode	Durée	Erreur	Taux d'acceptation
Importance Sampling	1.83s	0.006	-
Metropolis avec loi uniforme	2.52min	2.34	0.30
Metropolis avec loi gaussienne	2.55min	2.50	0.42

Table 2 – Comparaison des deux méthodes

Conclusion de la comparaison

- La méthode d'Importance Sampling est plus rapide que la méthode de Metropolis pour atteindre un même niveau de précision.
- La méthode de Metropolis nécessite des calculs plus compliqués et est donc plus coûteuse en temps.
- La méthode de Metropolis présente des autocorrélations élevées $\Rightarrow$ vitesse de convergence lente.
- La calibration de la méthode de Metropolis est délicate car si le paramètre n'est pas optimal, la vitesse de convergence diminue

Echantillonage de Gibbs : Le principe

- on fixe $\kappa = 1$ et on considère une approche bayésienne de la régression Lasso.

- On souhaite estimer $\hat{\beta}_{Bayes} = \hat{\mathbb{E}}^{\pi}_{Y,\kappa,\sigma,\lambda}(\beta|Y,\sigma,\lambda)$ sauf qu'on ne sait pas simuler selon $\beta|Y,\sigma,\lambda$.

  $\Longrightarrow$ Comme pour l'algorithme de Metropolis, On souhaite engendrer une chaîne de Markov dont la loi stationnaire est la loi cible, c'est-à-dire $\pi(\beta|Y,\sigma,\lambda)$.

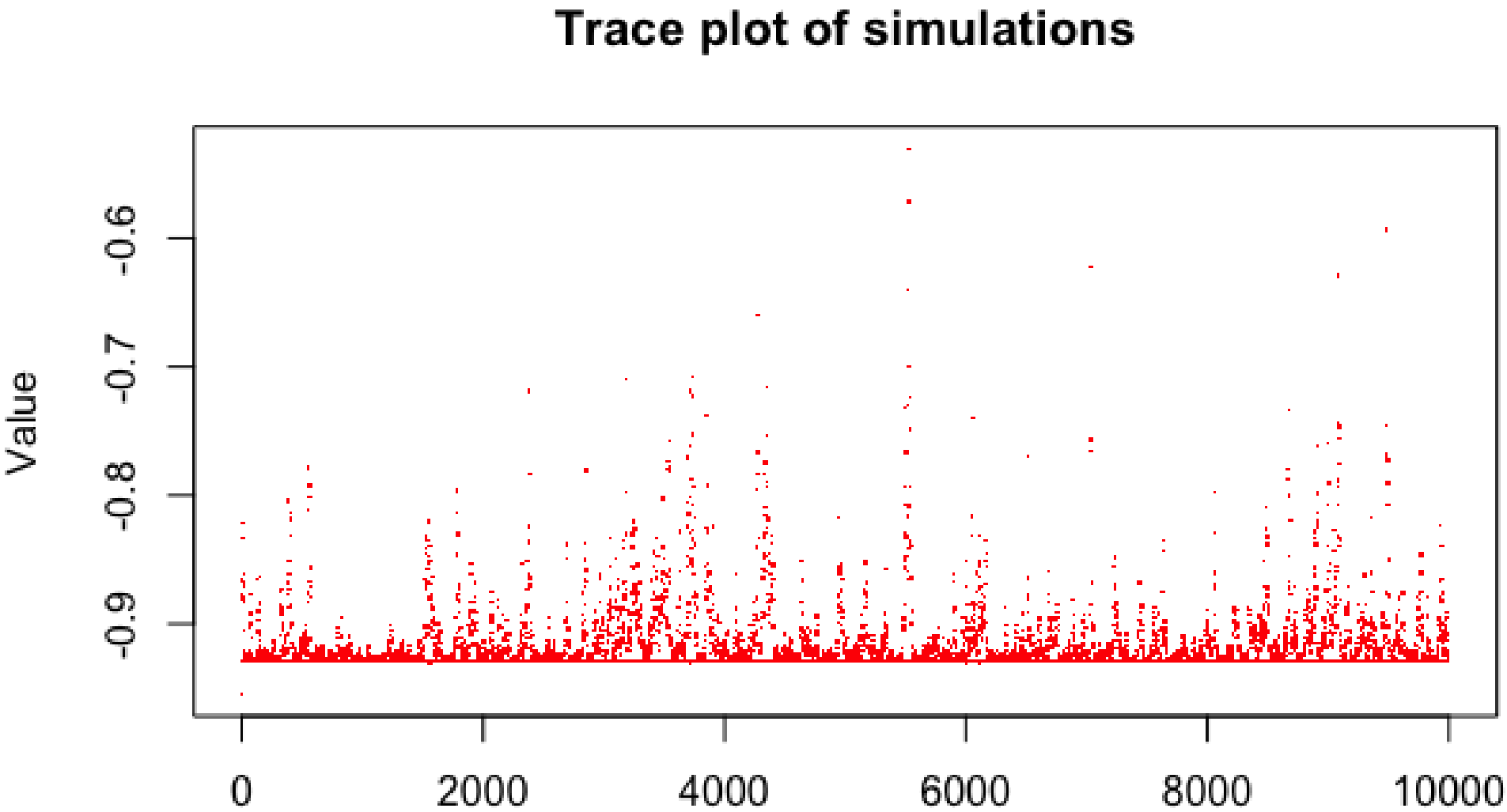
- Comme la densité de la loi $\pi(\beta|Y,\sigma,\lambda)$ est difficile à simuler, on utilise les densités conditionnelles.

Echantillonage de Gibbs : Lois conditionnelles

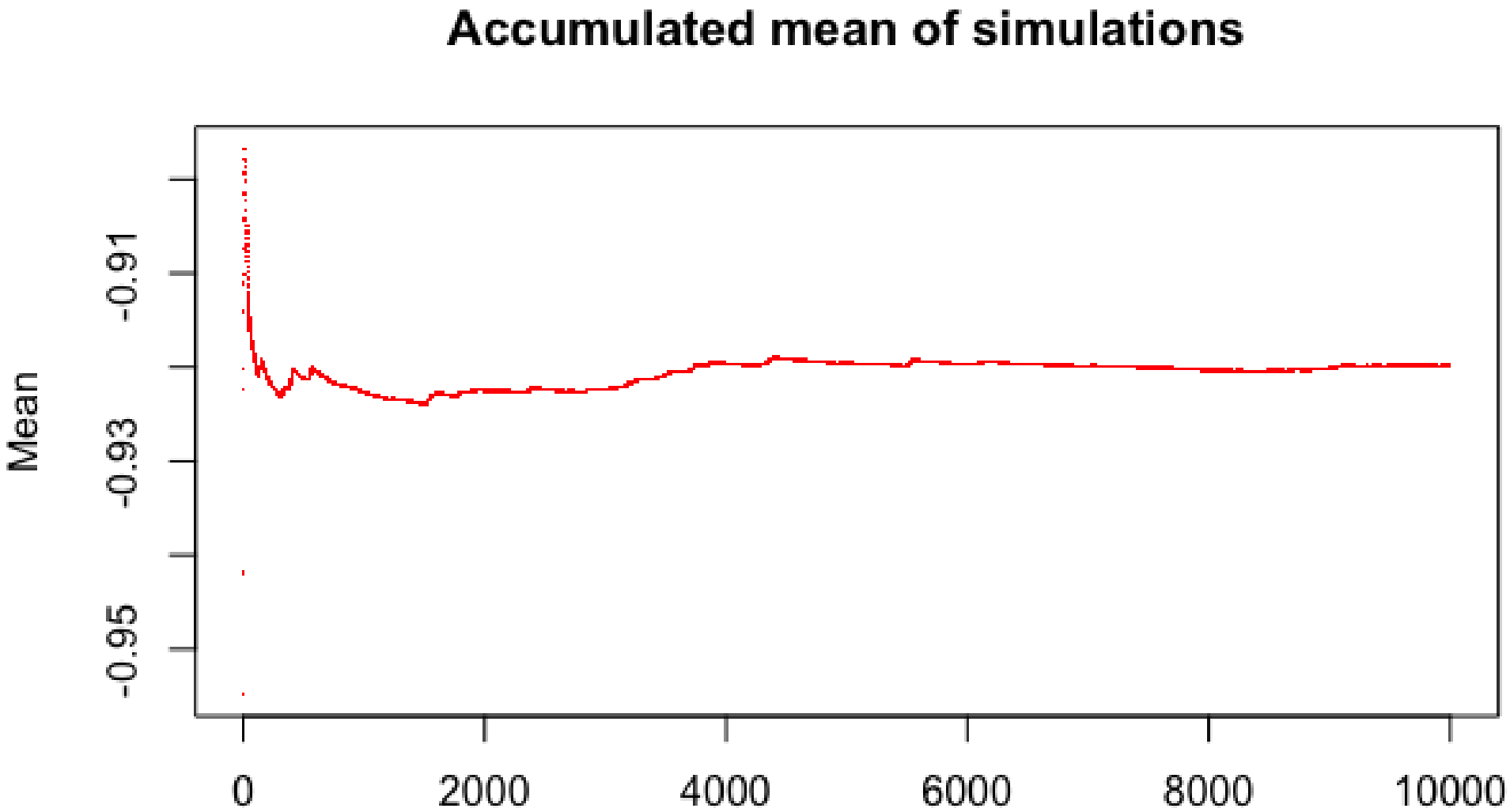
- $\beta \mid \tilde{y}, \tau_1^2, \cdots, \tau_p^2, \sigma^2 \sim \mathcal{N}\left(A^{-1} X^T \tilde{y}, \sigma^2 A^{-1}\right)$
  - $A = X^T X - D_\tau^{-1}$
  - $\tilde{y}$ le vecteur de réponse centré
- $\sigma^2 \mid \tilde{y}, \tau_1^2, \cdots, \tau_p^2, \beta \sim \Gamma^{-1}\left(\frac{n-1}{2} + \frac{p}{2}, \frac{(\tilde{y}-X\beta)^T(\tilde{y}-X\beta)}{2} + \frac{\beta^T D_\tau^{-1} \beta}{2}\right)$
- $\tau_1^2, \cdots, \tau_p^2$ iid

  avec $1/\tau_j^2 \mid \beta, \sigma^2, \tilde{y} \sim IG\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda^2\right)$

Exemple d'un graphique de la trajectoire



Exemple d'un graphique des moyennes accumulées

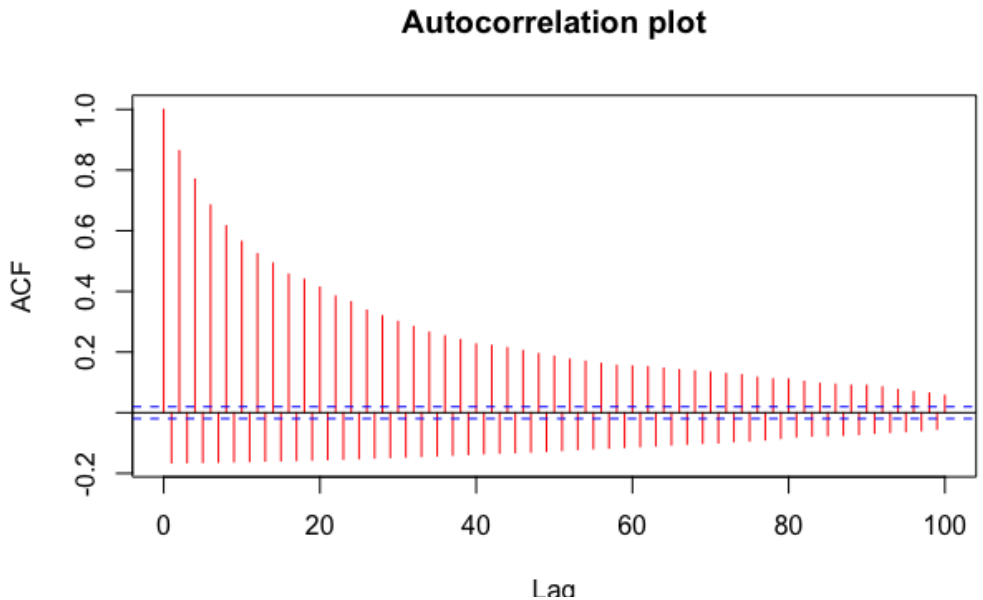


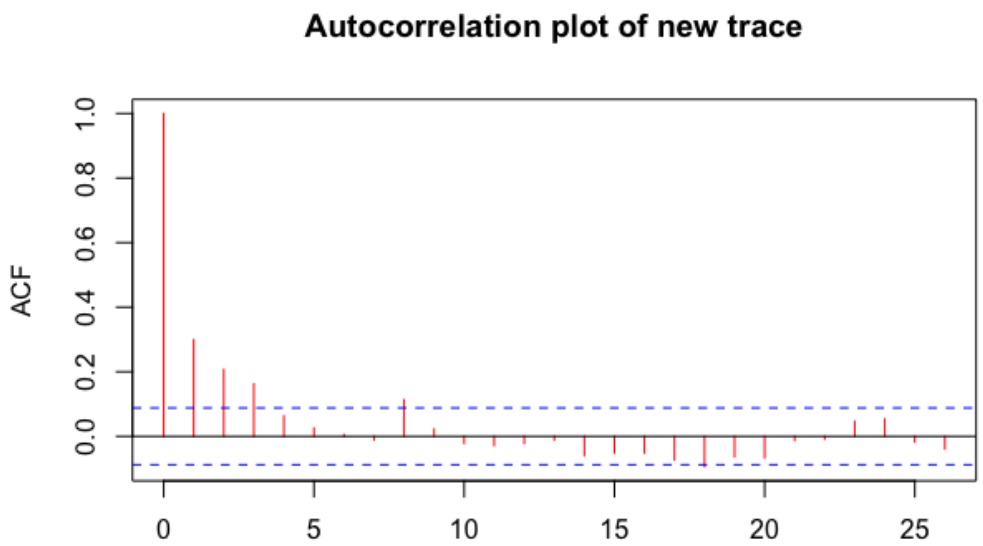
Échantillonnage de Gibbs

1. On prend un T_0 suffisamment grand et on prend les simulations engendrées après cet instant.

2. Afin d'éviter l'autocorrélation élevée qui peut empêcher la convergence vers la loi cible, on saute dans l'échantillonnage avec une largeur de saut w (Jumping width), c'est-à-dire que l'on prend $\beta_{T_0}, \beta_{T_0+w}, \beta_{T_0+2w}, \ldots$.

Autocorrélogramme avant/après échantillonnage





Comparaison avec l'Importance Sampling

Nombre de simulations	σ	κ	λ
10000	1	1	0.5

Table 3 – Coefficents du modèle

En utilisant la moyenne empirique obtenue par la méthode d'Importance Sampling avec un nombre de simulations égal à 1,000,000 qui est considéré comme le vrai β ici, on calcule l'erreur quadratique de nos estimateurs par les deux méthodes.

Comparaison sans effets croisés

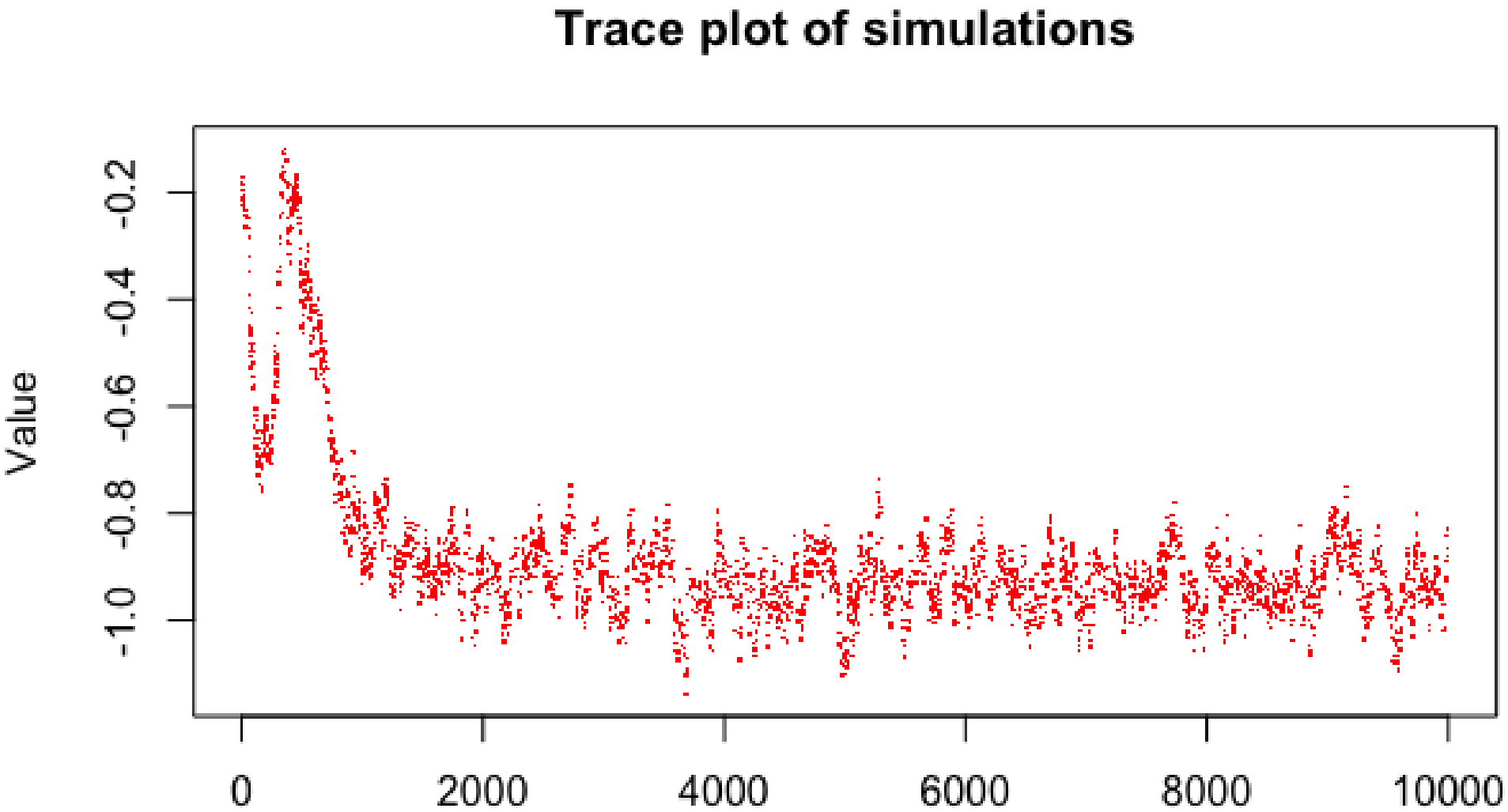
Méthode	Durée	Erreur	Taux d'acceptation
Importance Sampling	1.1s	0.09	-
Metropolis avec loi uniforme	1.2min	10	0.30
Metropolis avec loi gaussienne	1.3min	10,3	0.42
Gibbs sampler	4.20s	19.4	-

Table 4 – Comparaison des trois méthodes

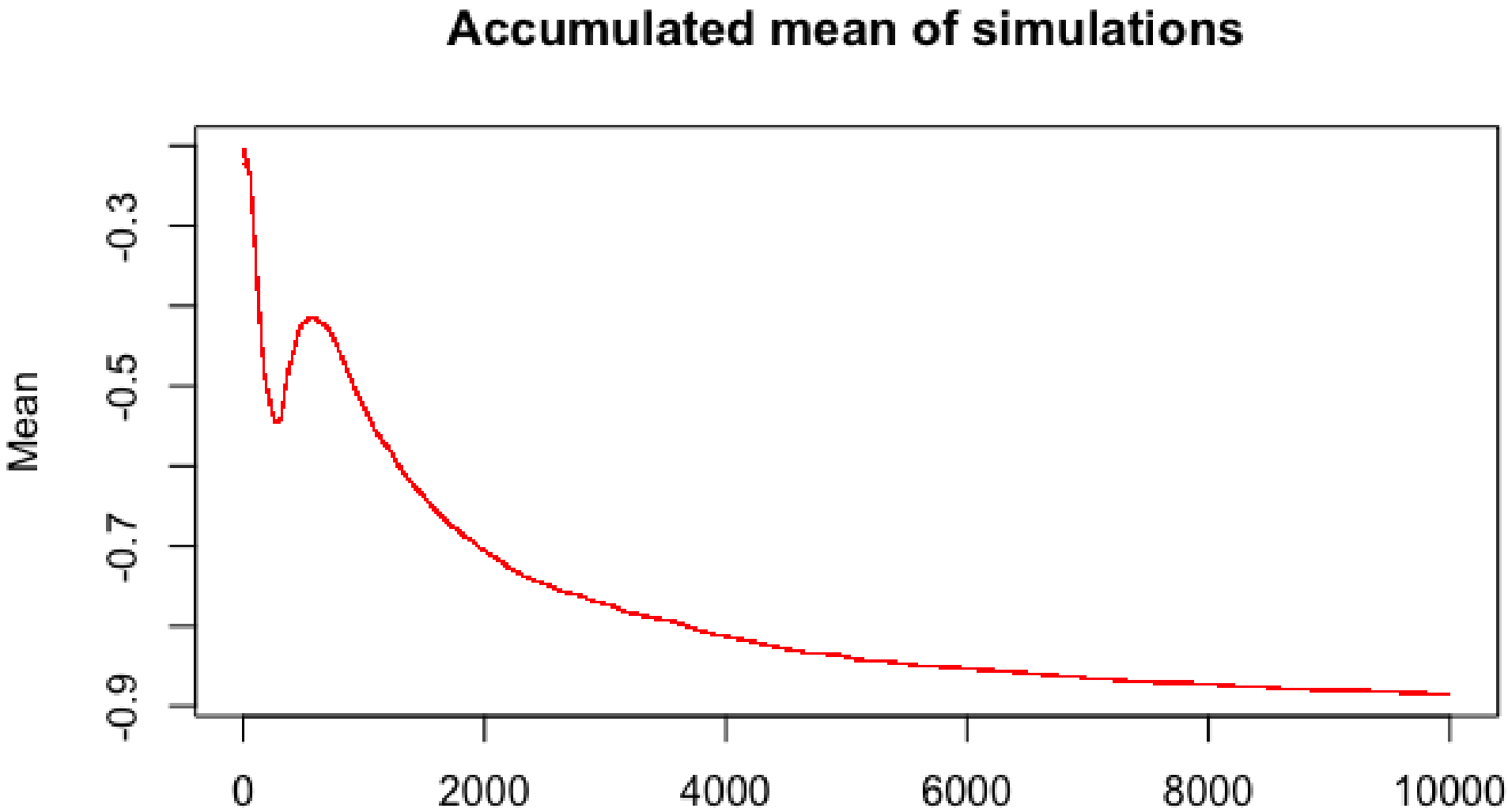
Conclusion de la comparaison

- La méthode d'Importance Sampling est plus rapide que la méthode de Metropolis pour atteindre un même niveau de précision.
- La méthode de Metropolis nécessite des calculs plus compliqués et est donc plus coûteuse en temps.
- La méthode de Metropolis présente des autocorrélations élevées $\Rightarrow$ vitesse de convergence lente.
- La calibration de la méthode de Metropolis est délicate car si le paramètre n'est pas optimal, la vitesse de convergence diminue
- L'échantillonnage de Gibbs est plus rapide et converge plus vite que la méthode Metropolis mais a une erreur plus élevée.

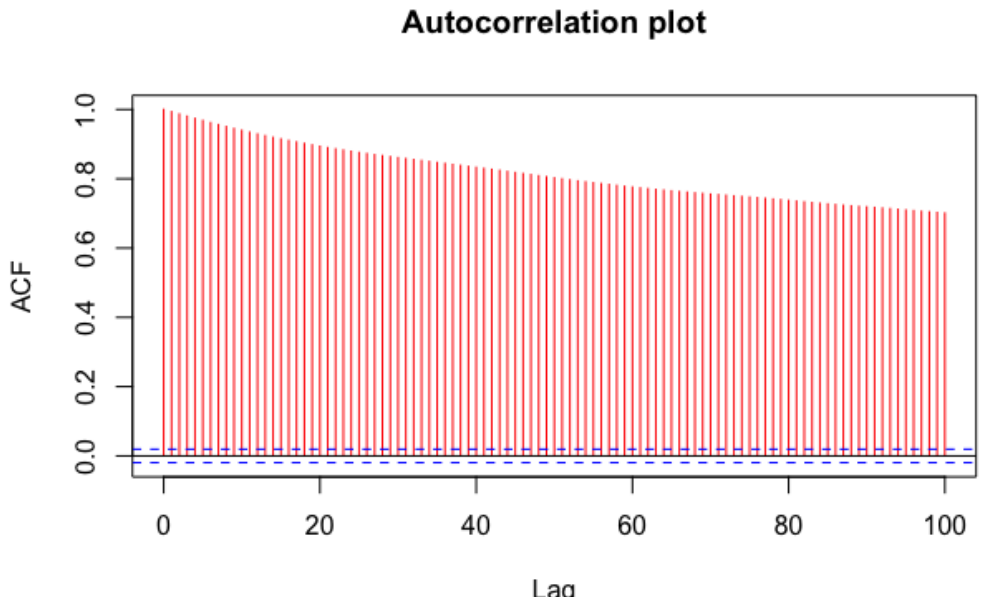
Exemple d'un graphique de la trajectoire

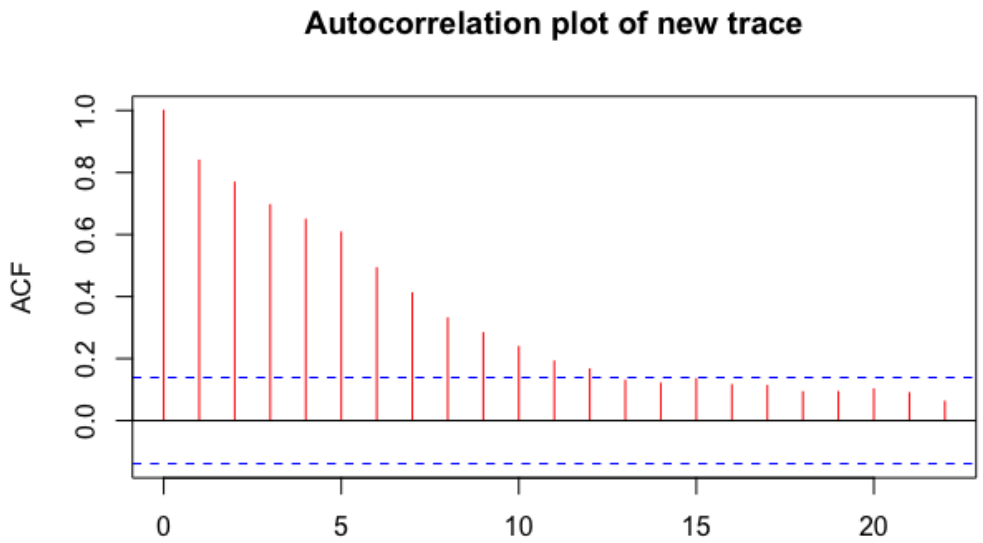


Exemple d'un graphique des moyennes accumulées



Autocorrélogramme avant/après échantillonnage





Echantillonage de Gibbs : Modèle Hiérarchisé

- $Y \mid X, \beta, \sigma^2 \sim \mathcal{N}_n\left(X\beta, \sigma^2 I_n\right)$
- $\beta \mid \tilde{y}, \tau_1^2, \cdots, \tau_p^2, \sigma^2 \sim \mathcal{N}_p\left(0_p, \sigma^2 D_\tau\right)$
- $D_\tau = \begin{pmatrix} \tau_1^2 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \tau_p^2 \end{pmatrix}$
- $\sigma^2, \tau_1^2, \cdots, \tau_p^2 \sim \pi(\sigma^2) d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2,$

  $\sigma^2, \tau_1^2, \cdots, \tau_p^2 > 0$