Rapport du Projet : Model Aggregation

Réalisé par :
Henri ALAM (3670259)
Redwan MEKRAMI (3105180)

Dans le cadre du cours :
Modèle Linéaire et Grande Dimension

Travail présenté à
Etienne ROQUAIN

Année scolaire : 2020-2021

Table des matières

1 Introduction

Une branche assez importante des Statistiques repose sur la reconstitution d'un signal quelconque. Ce signal peut être bruité, parcimonieux etc. Mais nous pouvons tout de même le retrouver à l'aide de différentes astuces.
En cas de parcimonie, une technique classique consiste à construire l'estimateur des moindres carrés $\hat{f}$ sur un sous-ensemble $m \subset \{1, \ldots, p\}$ des colonnes de la matrice X de dimension $n \times p$. On choisira alors $\hat{m}$ tel que :

$$\hat{m} \in \arg\min_{m \in \mathcal{M}} \left\{ \|Y - \hat{f}_m\| + \sigma^2 pen(m) \right\} \tag{1}$$

On procède donc ici à une sélection de modèles qui potentiellement peut être extrêmement coûteux lorsque $|m|$ croît.

Pour essayer d'améliorer notre modèle, nous allons présenter l'agrégation de modèles, une véritable alternative à la technique vu juste au-dessus. À travers ce projet, nous allons donc étudier ces deux techniques, les aborder dans un cadre spécifique, et ensuite les mettre en pratique dans un modèle simpliste.

2 Présentation du problème / Notations

Tout au cours de notre partie pratique, nous appliquerons nos connaissances sur un modèle assez spécifique et avec des notations précises que voici :

Notre modèle est de la forme :

$$Y = f^* + \varepsilon = X\beta^* + \varepsilon \tag{2}$$

avec $\varepsilon = (\varepsilon_i)_{i=1,\dots,n}$ tel que $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$.
De plus, nous allons considérer le cas où le signal est parcimonieux, ou "sparse" : β^* n'aura que certaines coordonnées non nulles.
Nous allons noter $\mathcal{M} = \mathcal{P}\left(\{1,\dots,p\}\right)$ l'ensemble des sous-ensembles de $\{1,\dots,p\}$.
D'autre part, nous définissons les collections de modèles $\mathcal{S}_m = Vect\,(X_j, j \in m)$, avec $m \in \mathcal{M}$ et tel que $d_m = dim(\mathcal{S}_m)$.
Nous allons noter $\Pi_{\mathcal{S}}$ l'opérateur de projection orthogonale sur un ensemble $\mathcal{S}$.
Enfin, nous écrirons $\hat{f}_m = \Pi_{\mathcal{S}_m} Y$.

Maintenant que nous avons bien défini notre cadre de travail, rentrons dans le vif du sujet !

3 Agrégation de modèles

Nous allons maintenant présenter en quoi consiste la technique d'agrégation de modèles.

L'estimateur $\hat{f}_m$ vu précédemment est assez efficace, mais le fait d'ignorer complètement les autres $\hat{f}_{m'}$ est assez triste car ils pourraient améliorer le risque !
Ainsi, on pourrait créer un estimateur qui est une combinaison convexe des $\hat{f}_m$ et ainsi réduire l'erreur : c'est l'agrégation de modèles !

Plus formellement, on définit l'estimateur par agrégation $\hat{f}$ tel que :

$$\hat{f} = \sum_{m \in \mathcal{M}} w_m \hat{f}_m, \tag{3}$$

avec, $w_m \geq 0\ \forall m \in \mathcal{M}$ et tel que $\sum_{m \in \mathcal{M}} w_m = 1$.

On attribue ainsi à chaque estimateur un certain poids, ce qui rendra, l'on espère, un modèle plus stable et plus performant.

3.1 Mélange de Gibbs

Soit $\widehat{r}_m = \left\|Y - \widehat{f}_m\right\|^2 + 2d_m\sigma^2 - n\sigma^2$, un estimateur non-biaisé de $r_m = \mathbb{E}\left[\left\|f^* - \widehat{f}_m\right\|^2\right]$.

Soit $\beta > 0$ un réel quelconque et π une loi de probabilité donnée, on appelle alors $\hat{f}$ un mélange de Gibbs de la collection d'estimateurs $\left\{\hat{f}_m, m \in \mathcal{M}\right\}$ si il est de la forme :

$$\widehat{f} = \sum_{m\in\mathcal{M}} w_m \widehat{f}_m, \quad \text{with } w_m = \frac{\pi_m e^{-\beta\hat{r}_m/\sigma^2}}{\mathcal{Z}}, \quad \text{where } \mathcal{Z} = \sum_{m\in\mathcal{M}} \pi_m e^{-\beta\hat{r}_m/\sigma^2} \tag{4}$$

Nous verrons par la suite comment obtenir une expression explicite de cet estimateur dans des cadres spécifiques.

3.2 Méthode Markov Chain Monte Carlo (MCMC)

Losque $|m|$ grandit, nous rencontrons le même problème abordé auparavant : nous pouvons nous intéresser à une autre méthode appelée Méthode MCMC qui nous permettra d'approximer w.

3.2.1 Metropolis-Hastings

Soit $F : \mathcal{M} \to \mathbb{R}$ et w une loi de probabilité tel que $w_m > 0$ pour tout m.

L'algorithme de Metropolis-Hastings va alors permettre d'approcher la quantité :

$$\mathbb{E}_w[F] := \sum_{m\in\mathcal{M}} w_m F(m)$$

L'idée de cet algorithme est de générer une chaîne de Markov ergodique $(M_t)_{t\in\mathbb{N}}$ avec une distribution stationnaire w, et ensuite approcher $\mathbb{E}_w[F]$ par $\frac{1}{T}\sum_{t=1}^{T} F(M_t)$, pour T assez grand.
En effet, le théorème ergodique nous assure que pour toute chaîne de Markov ergoqique $(M_t)_{t\in\mathbb{N}}$ à distribution stationnaire w, nous avons :

$$\frac{1}{T}\sum_{t=1}^{T} F(M_t) \overset{p.s.}{\to} \mathbb{E}_w[F] \tag{5}$$

Nous ne rentrerons pas dans les détails techniques de cet algorithme, et nous nous contenterons juste de le présenter :

Algorithm 1 Metropolis-Hastings Algorithm

Initialization. Pick an arbitrary $M_1 \in \mathcal{M}$ and choose a burn-in time T_0.
For $t = 0, \ldots, T$ do :
Perform Updates :
(1) From the current state M_t, generate M'_{t+1} according to the distribution.
(2) Set $p_{t+1} = 1 \wedge \frac{w_{M'_{t+1}} \Gamma\left(M'_{t+1}, M_t\right)}{w_{M_t} \Gamma\left(M_t, M'_{t+1}\right)}$

(3) $\begin{cases} \text{With probability } p_{t+1} : \text{set } M_{t+1} = M'_{t+1} \\ \text{else : set } \mathrm{M}_{t+1} = M_t \end{cases}$

Output. $\frac{1}{T-T_0} \sum_{t=T_0+1}^{T} F\left(M_t\right)$.

Dans notre cas, on cherche à estimer la distribution de Gibbs, on aura alors $F(m) = \widehat{f}_m$ et on pourra ainsi calculer une approximation de w.

Malgré les difficultés apparentes à calculer ces estimateurs, une hypothèse sur X va nous faciliter la tâche !

4 Étude de cas : X Orthonormale

4.1 Sélection de modèle

Nous sommes ici en train de considérer le problème 1, et allons voir comment cette hypothèse va nous aider en pratique pour cet estimateur :

Pour cette partie, nous allons prendre comme loi de probabilité :

$$\pi_m = (1 + 1/p)^{-p} p^{-|m|}$$

et allons supposer $pen(m) = \lambda |m|$, avec $\lambda = K(1 + \sqrt{2 \log(p)})^2$.

On pose alors $\hat{m}_\lambda$ le minimiseur de notre problème, et nous obtenons le théorème suivant :

Theorem 1. *Sous les conditions ci-dessus, nous avons :*

$$\widehat{m}_\lambda = \left\{ j \in m : \left(\mathbf{X}_j^T Y\right)^2 > \lambda \sigma^2 \right\}$$

Démonstration. Nous allons faire la preuve pour $\sigma = 1$, car c'est le cas qui nous intéresse. Le terme à minimiser dans (1) est alors égal à :

$$\|Y - \Pi_{S_m} Y\|^2 + \mathrm{pen}(m) \tag{6}$$

Calculons le premier terme :

$$\|Y - \Pi_{S_m} Y\|^2 = \|Y\|^2 - 2 < Y, \Pi_{S_m} Y > + \|\Pi_{S_m} Y\|^2$$

Puisque Π_{S_m} est une projection, nous avons alors :

$$\Pi_{S_m} = \Pi_{S_m}^2 = \Pi_{S_m}^T \Pi_{S_m} \Rightarrow \quad < Y, \Pi_{S_m} Y >= Y^T \Pi_{S_m} Y = Y^T \Pi_{S_m}^T \Pi_{S_m} Y = \|\Pi_{S_m} Y\|^2$$

.

Finalement, nous obtenons :

$$\|Y - \Pi_{S_m} Y\|^2 = \|Y\|^2 - \|\Pi_{S_m} Y\|^2 \tag{7}$$

De plus, par orthonormalité de X, nous savons que les colonnes de X forment une famille orthonormale, et donc :

$$\Pi_{S_m} Y = \sum_{j \in m} < Y, X_j > X_j \iff \|\Pi_{S_m}\|^2 = \sum_{j \in m} \left(X_j^T Y\right)^2$$

Le second terme, quand à lui, donne :

$$\text{pen}(m) = \lambda |m| = \sum_{j \in m} \lambda$$

Alors, $(6) = \|Y\|^2 + \sum_{j \in m} \left(\lambda - \left(X_j^T Y\right)^2 \right)$

Minimiser cette formule en m revient alors à minimiser la somme. On peut faire cela en prenant uniquement les termes strictement négatifs de la somme.

On a alors :

$$\widehat{m}_\lambda = \left\{ j \in m : \lambda - \left(\mathbf{X}_j^T Y\right)^2 < 0 \right\} \iff \widehat{m}_\lambda = \left\{ j \in m : \left(\mathbf{X}_j^T Y\right)^2 > \lambda \right\}$$

□

Il suffit ensuite de faire le maximum de vraisemblance sur la matrice réduite suivies des étapes classiques...
On remarquera de plus que la complexité est maintenant linéaire, comparée à une complexité exponentielle auparavant.

4.2 Agrégation de modèles

Nous allons voir comment obtenir une expression un peu plus sympathique à l'aide de l'hypothèse sur X :

Theorem 2. *En posant $Z_j =< Y, X_j >$, et $\pi_m = (1+1/p)^{-p} p^{-|m|}$, l'estimateur de Gibbs est donné par :*

$$\hat{f} = \sum_{j=1}^{p} \frac{\exp\left(\beta Z_j^2/\sigma^2\right)}{\exp(2\beta + \log(p)) + \exp\left(\beta Z_j^2/\sigma^2\right)} Z_j X_j$$

Démonstration. Tout d'abord, en utilisant l'équation (7), nous trouvons que :

$$\widehat{r}_m = \left\| Y - \widehat{f}_m \right\|^2 + 2d_m\sigma^2 - n\sigma^2 = \|Y\|^2 - \left\| \widehat{f}_m \right\|^2 + (2|m| - n)\sigma^2$$

.

D'autre part, on injecte w_m par sa formule dans l'expression (4) :

$$\mathcal{Z}\widehat{f} = \sum_{m\in\mathcal{M}} w_m \widehat{f}_m = \sum_{m\in\mathcal{M}} \pi_m \exp\left(-\frac{\beta\hat{r}_m}{\sigma^2}\right) \hat{f}_m$$

Puis, en développant $\hat{r}_m$, on obtient :

$$\mathcal{Z}\widehat{f} = \left(1+\frac{1}{p}\right)^{-p} \exp\left(n\beta - \frac{\beta\|Y\|^2}{\sigma^2}\right) \underbrace{\sum_{m\in\mathcal{M}} p^{-|m|} \exp\left(\frac{\beta}{\sigma^2}\left\|\hat{f}_m\right\|^2 - 2\beta|m|\right)}_{T_1} \hat{f}_m$$

D'autre part,

$$\mathcal{Z} = \sum_{m\in\mathcal{M}} \pi_m \exp\left(-\frac{\beta\hat{r}_m}{\sigma^2}\right)$$

$$\mathcal{Z} = \left(1+\frac{1}{p}\right)^{-p} \exp\left(n\beta - \frac{\beta\|Y\|^2}{\sigma^2}\right) \underbrace{\sum_{m\in\mathcal{M}} p^{-|m|} \exp\left(\frac{\beta}{\sigma^2}\left\|\hat{f}_m\right\|^2 - 2\beta|m|\right)}_{T_1}$$

Alors, par identification, on obtient :

$$\hat{f} = \sum_{m\in\mathcal{M}} \frac{p^{-|m|}}{T_1} \exp\left(\frac{\beta}{\sigma^2}\left\|\hat{f}_m\right\|^2 - 2\beta|m|\right) \hat{f}_m$$

Par définition de $\widehat{f}_m$, nous pouvons injecter son expression explicite :

$$\begin{aligned}\hat{f} &= \sum_{m\in\mathcal{M}} \frac{p^{-|m|}}{T_1} \exp\left(\frac{\beta}{\sigma^2}\left\|\hat{f}_m\right\|^2 - 2\beta|m|\right) \sum_{j=1}^{p} Z_j X_j 1_{j\in m} \\ &= \sum_{j=1}^{p}\left(\sum_{m\in\mathcal{M}} \frac{p^{-|m|}}{T_1} \exp\left(\frac{\beta}{\sigma^2}\left\|\hat{f}_m\right\|^2 - 2\beta|m|\right) 1_{j\in m}\right) Z_j X_j\end{aligned}$$

Étudions maintenant le terme exponentiel dans la formule, que nous allons expliciter en utilisant la définition de $\widehat{f}_m$:

$$\begin{aligned}\exp\left(\frac{\beta}{\sigma^2}\left\|\hat{f}_m\right\|^2 - 2\beta|m|\right) 1_{j\in m} &= \exp\left(\frac{\beta}{\sigma^2}\sum_{k\in m} Z_k^2 - 2\beta|m|\right) 1_{j\in m} \\ &= \exp\left(\frac{\beta}{\sigma^2} Z_j^2 + \frac{\beta}{\sigma^2}\sum_{\substack{k\in m\\ k\neq j}} Z_k^2 - 2\beta|m|\right) 1_{j\in m} \\ &= \exp\left(\frac{\beta}{\sigma^2} Z_j^2\right) \exp\left(\frac{\beta}{\sigma^2}\sum_{\substack{k\in m\\ k\neq j}} Z_k^2 - 2\beta|m|\right) 1_{j\in m}\end{aligned}$$

En injectant cette dernière égalité dans l'expression de $\widehat{f}$, on obtient :

$$\hat{f} = \sum_{j=1}^{p}\left(\sum_{\substack{m\in M\\ j\in m}} \frac{p^{-|m|}}{T_1} \exp\left(\frac{\beta}{\sigma^2}\sum_{\substack{k\in m\\ k\neq j}} Z_k^2 - 2\beta|m|\right)\right) \exp\left(\frac{\beta}{\sigma^2} Z_j^2\right) Z_j X_j$$

Pour finir cette preuve, il faut montrer que :

$$T_1 = \sum_{\substack{m\in M\\ j\in m}} p^{-|m|} \exp\left(\frac{\beta}{\sigma^2}\sum_{\substack{k\in m\\ k\neq j}} Z_k^2 - 2\beta|m|\right) \times \left(p\exp(2\beta) + \exp\left(\beta Z_j^2/\sigma^2\right)\right)$$

On notera par la suite $\mathcal{Q}_|$ le terme de droite dans l'égalité précédente. Alors, en fixant j,

nous obtenons :

$$\begin{aligned}\mathcal{Q}_j &= \sum_{\substack{m\in\mathcal{M}\\ j\in m}} p^{-(|m|-1)} \exp\left(\frac{\beta}{\sigma^2}\sum_{\substack{k\in m\\ k\neq j}} Z_k^2 - 2\beta(|m|-1)\right) + \sum_{\substack{m\in\mathcal{M}\\ j\in m}} p^{-|m|}\exp\left(\frac{\beta}{\sigma^2}\sum_{k\in m} Z_k^2 - 2\beta|m|\right)\\ &= \sum_{m\in\mathcal{M}_{jc}} p^{-|m|}\exp\left(\frac{\beta}{\sigma^2}\sum_{k\in m} Z_k^2 - 2\beta\,|m|\right) + \sum_{\substack{m\in\mathcal{M}\\ j\in m}} p^{-|m|}\exp\left(\frac{\beta}{\sigma^2}\sum_{k\in m} Z_k^2 - 2\beta|m|\right)\end{aligned}$$

, où l'on a juste distribué le terme de droite de T_1 et défini l'ensemble $\mathcal{M}_{j^c} = \mathcal{M}\backslash\{j\}$.

Par définition, ces deux ensembles forment ainsi une partition et donc on a :

$$\begin{aligned}\mathcal{Q}_j &= \sum_{m\in\mathcal{M}} p^{-|m|}\exp\left(\frac{\beta}{\sigma^2}\sum_{k\in m} Z_k^2 - 2\beta|m|\right)\\ &= \sum_{m\in\mathcal{M}} p^{-|m|}\exp\left(\frac{\beta}{\sigma^2}\left\|\hat{f}_m\right\|^2 - 2\beta|m|\right)\end{aligned}$$

Nous retrouvons bien T_1 et concluons ainsi cette preuve.

□

Regardons maintenant ce que ces deux estimateurs donnent en pratique !

5 Application

Les mêmes données ont été simulées pour les deux estimateurs, testons-donc leurs efficacités :

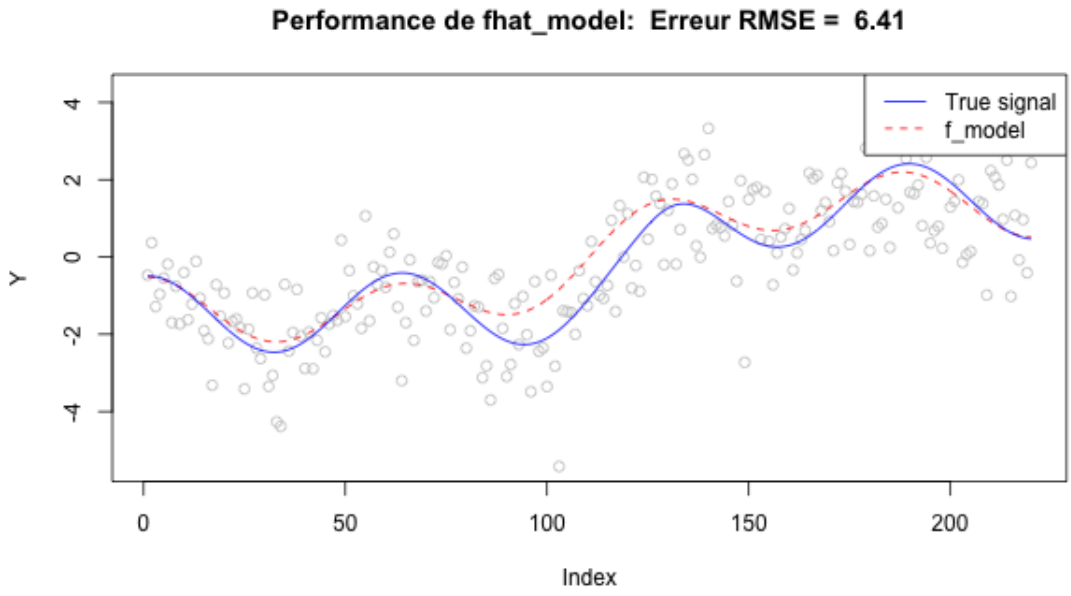
FIGURE 1 – Initialisation du signal

```
x <- seq(-4*pi,3*pi, 1/10)
n <- length(x)
f = cos(x) + atan(x)
Y = f + rnorm(n)
X=ComputeCosMat(n)
Z= t(X)%*%Y
beta = 0.25
```

5.1 Sélection de modèle

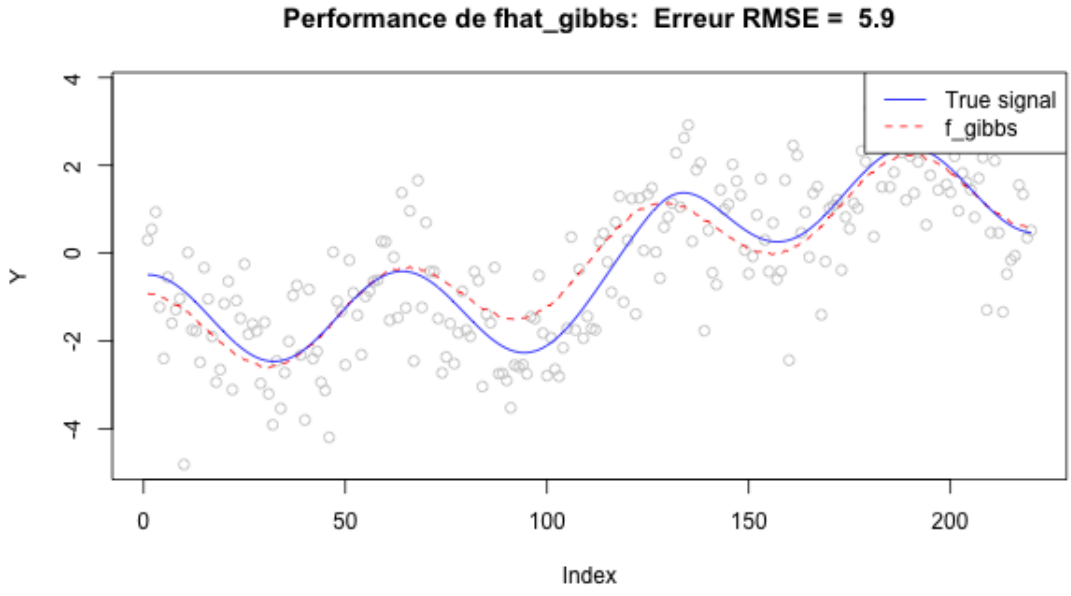
Cet estimateur semble plutôt bien fonctionner et produit des résultats satisfaisants :

FIGURE 2 – Performance de f_{model}



5.2 Agrégation de modèles

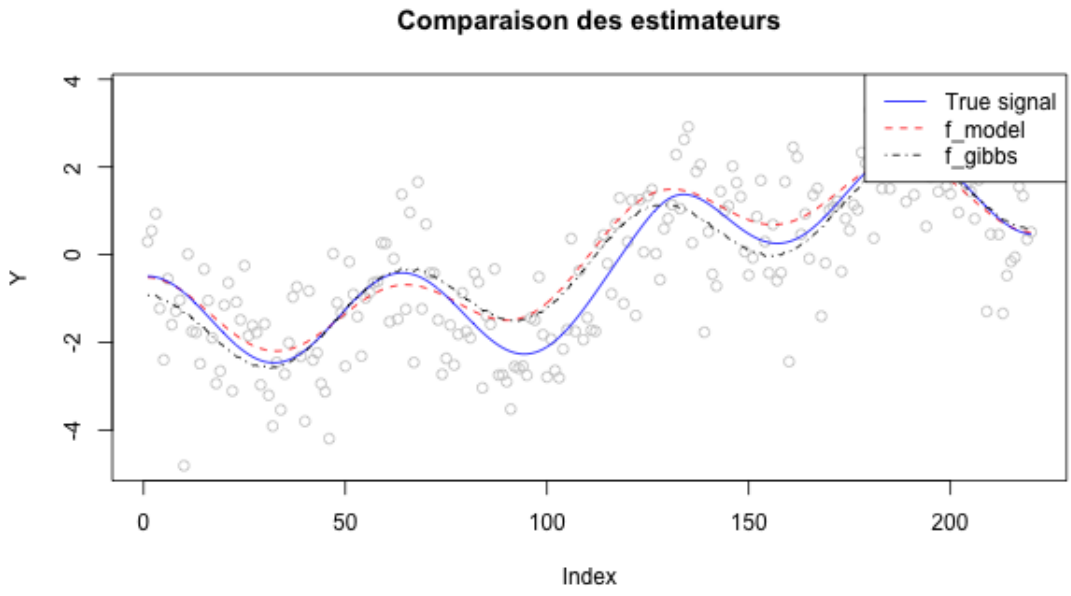
FIGURE 3 – Performance de f_{Gibbs}



On remarque que ce dernier est légèrement mieux sur cette instance, mais les deux sont généralement équivalents si l'on relance la simulation.

6 Comparaison

FIGURE 4 – Comparaison des deux estimateurs



On peut voir que l'estimateur de Gibbs est légèrement plus précis.

7 Conclusion

Au cours de ce projet, nous avons ainsi pu étudier de près les vertus de ces deux estimateurs ainsi que leurs performances dans un cadre très spécifique. Par soucis de temps, nous trouvons ça dommage que nous n'avons pas pu implémenter la Méthode Metropolis-Hastings car les résultats auraient sûrement été très intéressants. De plus, nous aurions aimés étudier de plus près le cas où les deux estimateurs ont des performances équivalentes et comprendre un peu plus les raisons derrière ce résultat (Voir Remarque $p.62$ de l'article).