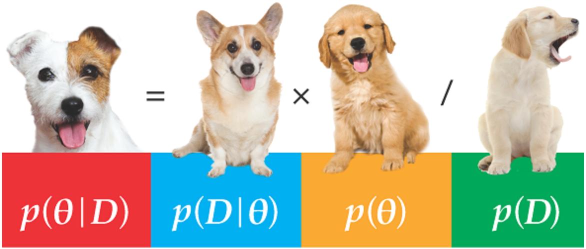
Importance Sampling appliqué à la régression Bridge Bayésienne

Gautier Appert

Stage ENSAE ParisTech 2A

16 décembre 2015



Objectif

Référence

- *The Bayesian Bridge (2014)*, Nicholas G. Polson, James G. Scott, Jesse Windle [a].
- Utilisation d'un Gibbs Sampler s'appuyant sur la structure hiérarchique du modèle Bridge bayésien.
- Package "BayesBridge" sur R associé à l'article.

---

a. Lien arXiv : `http://arxiv.org/pdf/1109.2279v2.pdf`

Objectif

- Utilisation d'une méthode beaucoup plus simple appelée Importance Sampling.
- Reproduire les résultats (applications) de l'article avec cette méthode et comparaison.
- Limites de l'importance sampling lorsque la dimension du modèle augmente $\Rightarrow$ SMC tempering.

Régression Bridge

Modèle

- Modèle linéaire Gaussien : $y = X\beta + \epsilon$
- $\epsilon \sim \mathcal{N}(0_{\mathbb{R}^n}, \sigma^2 I_n)$
- $X = [x^1|\dots|x^p] \in \mathcal{M}_{n,p}(\mathbb{R})$, $\beta \in \mathbb{R}^p$ et $y \in \mathbb{R}^n$

Régression Bridge

$$\mathcal{P}_\delta : \min_{\beta} \quad \|y - X\beta\|_2^2$$

$$\text{s.c.} \quad \sum_{j=1}^{p} |\beta_j|^\alpha \leq \delta$$

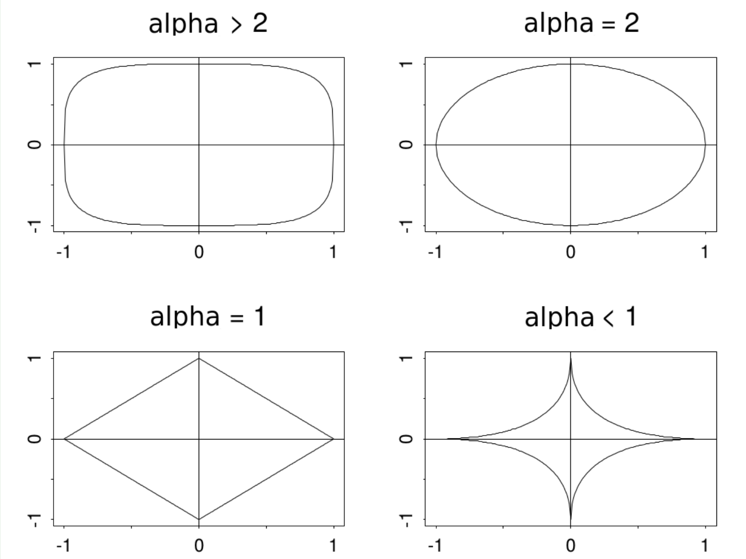
où le paramètre $\alpha \in [0, 1]$

Remarques

- Pour $\alpha < 1$ le problème n'est pas convexe contrairement à la régression Ridge ou Lasso
- $\alpha = 1 \Rightarrow$ fournit la pénalisation **Lasso** (le pb devient convexe).
- $\alpha = 0 \Rightarrow$ la régression sur les meilleurs sous-ensembles des variables, $\{\mathcal{M}_k \in \mathcal{P}(\{1, \dots, p\}) | k \in [\![1; p]\!]\}$.

Régression Bridge

Représentation graphique de la pénalisation



Régression Bridge

Fonction objectif (Lagrangien), $\alpha \in [0,1]$ et $\nu \in \mathbb{R}_+$

$$\hat{\beta}_{bridge}(\nu) \in \arg\min_{\beta} \; \mathcal{L}_y(\beta,\nu) \triangleq \frac{1}{2\sigma^2}\|y - X\beta\|_2^2 + \nu \sum_{j=1}^{p} |\beta_j|^{\alpha}$$

Remarques

- Les Problèmes $\mathcal{P}_\delta$ et $\min_{\beta} \mathcal{L}_y(\beta,\nu)$ sont équivalents.
- Le problème n'est pas convexe $\Rightarrow$ problème d'optimisation plus difficile d'un point de vue numérique.
- Le passage dans une formulation bayésienne du modèle peut donc être un moyen de palier ce problème.

Régression Bridge Bayésienne

Loi a priori dans le modèle Bridge bayésien

- On impose la loi a priori suivante sur $\beta = (\beta_1, \ldots, \beta_p)'$:

$$\pi(\beta|\alpha,\nu) = \prod_{j=1}^{p} \pi(\beta_j|\alpha,\nu) \propto \prod_{j=1}^{p} \exp\{-\nu|\beta_j|^{\alpha}\} \propto \exp\left\{-\nu \sum_{j=1}^{p} |\beta_j|^{\alpha}\right\}$$

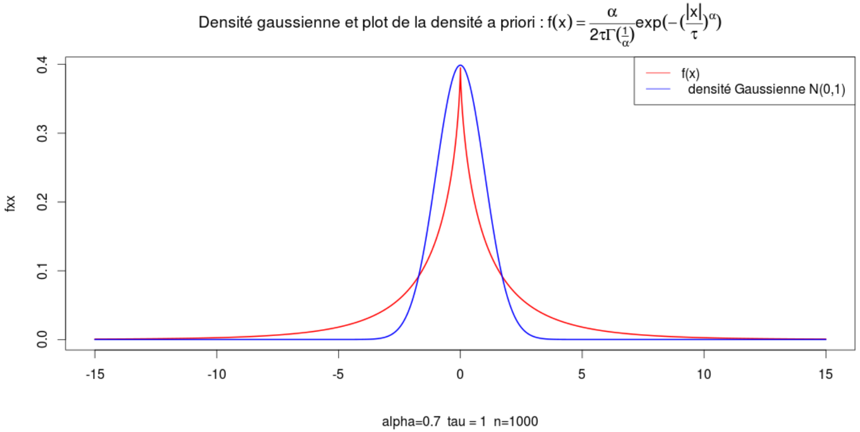
Remarques

- Chaque β_j suit une loi normale généralisée : $f(x) = K \exp\{-\frac{|x-\mu|^{\alpha}}{\sigma}\}\ 1_{\mathbb{R}}(x)$
- La constante de normalisation est calculable : $\mathcal{K}(\alpha,\nu) = \frac{\alpha \cdot \nu^{1/\alpha}}{2 \cdot \Gamma(1/\alpha)}$.
- Cet a priori peut être vu comme un mélange (continue) de Gaussiennes [b] :

$$\exp\{-|z|^{q}\} \propto \int_0^{\infty} \frac{1}{\sqrt{2\pi s}} \exp\{-\frac{z^2}{2s}\} g(s) ds$$

b. Théorème de Bernstein sur les fonctions totalement monotones (Article).

Régression Bridge Bayésienne



Régression Bridge Bayésienne

Vraisemblance du modèle linéaire Gaussien

$$L(y|\beta,\sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\|y - X\beta\|_2^2\right\}.$$

La distribution a posteriori de β (formule de **Bayes**)

$$\pi(\beta|y,\sigma^2,\alpha,\nu) \propto L(y|\beta,\sigma^2)\cdot\pi(\beta|\alpha,\nu) \propto \exp\{-\mathcal{L}_y(\beta,\nu)\}.$$

Remarques

- Maximiser la densité a posteriori $\pi(\beta|y,\sigma^2,\alpha,\nu)$ est équivalent à maximiser le Lagrangien $\mathcal{L}_y(\beta,\nu)$ du problème $\mathcal{P}_\delta$.
- la densité a posteriori n'est pas une loi connue $\Rightarrow$ Gibbs Sampler, Importance Sampling , SMC tempering, ...

Estimation Bayésienne

- $\hat{\beta}_{map} \overset{def}{\in} \underset{\beta}{\arg\max}\ \pi(\beta|y,\sigma^2,\alpha,\nu) = K(y,\sigma^2,\alpha,\nu)\cdot\exp\{-\mathcal{L}_y(\beta,\nu)\}$
- $\hat{\beta}_{Bayes} = \mathbb{E}^{\pi}_{\beta|y,\sigma^2,\alpha,\nu}[\beta|y,\sigma^2,\alpha,\nu] = \int_{\mathbb{R}^p}\beta\cdot\pi(\beta|y,\sigma^2,\alpha,\nu)\ \mathrm{d}\beta.$

Régression Bridge Bayésienne

Cadre plus général avec les paramètres σ^2, α et ν inconnus.

Le Bridge vu comme un Modèle Hiérarchique Bayésien

- $y|\beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n) \equiv$ vraisemblance
- $\beta|\alpha, \nu \sim \pi(.|\alpha, \nu) \equiv$ a priori
- $\alpha, \nu \sim \pi(., .) \equiv$ hyper a priori

Formule de Bayes dans le cas d'un Modèle Hiérarchique Bayésien

$$\pi(\beta, \alpha, \nu, \sigma^2|y) \propto L(y|\beta, \sigma^2)\pi(\beta|\alpha, \nu)\pi(\alpha)\pi(\nu)\pi(\sigma^2)$$

Espérance a posteriori

$$\mathbb{E}^{\pi}_{\beta,\alpha,\nu,\sigma^2|y}[h(\beta, \alpha, \nu, \sigma^2)|y]$$

$$\propto \int_{\Theta} h(\beta, \alpha, \nu, \sigma^2)L(y|\beta, \sigma^2)\pi(\beta|\alpha, \nu)\pi(\alpha)\pi(\nu)\pi(\sigma^2)\ \mathrm{d}\alpha\mathrm{d}\nu\mathrm{d}\sigma^2\mathrm{d}\beta$$

On s'intéresse à l'estimation de quantités sous la forme d'espérance a posteriori.

Estimation bayésienne

Soit $h : \mathbb{R}^p \to \mathbb{R}^q$:

$$\mathcal{I}(\alpha,\nu,\sigma^2) = \mathbb{E}^{\pi}_{\beta|y,\sigma^2,\alpha,\nu}[h(\beta)|y,\sigma^2,\alpha,\nu] = \int_{\mathbb{R}^p} h(\beta)\cdot\pi(\beta|y,\sigma^2,\alpha,\nu)\ \mathrm{d}\beta$$

$$= \frac{\int_{\mathbb{R}^p} h(\beta)\cdot L(y|\beta,\sigma^2)\cdot\pi(\beta|\alpha,\nu)\ \mathrm{d}\beta}{\int_{\mathbb{R}^p} L(y|\beta,\sigma^2)\pi(\beta|\alpha,\nu)\ \mathrm{d}\beta}$$

Importance Sampling

Soit $q : \mathbb{R}^p \to \mathbb{R}^+$ une densité instrumentale (proposal) facile à simuler :

$$\mathcal{I}(\alpha,\nu,\sigma^2) = \frac{\int_{\mathbb{R}^p} h(\beta)\cdot\frac{L(y|\beta,\sigma^2)\cdot\pi(\beta|\alpha,\nu)}{q(\beta)}\cdot q(\beta)\ \mathrm{d}\beta}{\int_{\mathbb{R}^p} \frac{L(y|\beta,\sigma^2)\cdot\pi(\beta|\alpha,\nu)}{q(\beta)}\cdot q(\beta)\ \mathrm{d}\beta} = \frac{\mathbb{E}_q\left(h(\beta)\cdot\frac{L(y|\beta,\sigma^2)\cdot\pi(\beta|\alpha,\nu)}{q(\beta)}\right)}{\mathbb{E}_q\left(\frac{L(y|\beta,\sigma^2)\cdot\pi(\beta|\alpha,\nu)}{q(\beta)}\right)}$$

Poids d'importance

On pose :

$$\omega(\beta) = \frac{L(y|\beta, \sigma^2) \cdot \pi(\beta|\alpha, \nu)}{q(\beta)}$$

Importance Sampling (suite)

$$\mathcal{I}(\alpha, \nu, \sigma^2) = \frac{\mathbb{E}_q\left(h(\beta) \cdot \omega(\beta)\right)}{\mathbb{E}_q\left(\omega(\beta)\right)}$$

Estimateur auto-normalisé :

$$\hat{\mathcal{I}}_{\text{AIS}} = \frac{\displaystyle\sum_{k=1}^{N} h(\beta^{(k)}) w(\beta^{(k)})}{\displaystyle\sum_{k=1}^{N} w(\beta^{(k)})} \xrightarrow{\text{p.s}} \mathcal{I}(\alpha, \nu, \sigma^2)$$

Avec $\{\beta^{(1)}, \dots, \beta^{(N)}\} \overset{\text{iid}}{\sim} q$.
Justification basée sur la **loi forte des grands nombres** et la **Delta-Méthode**.

Importance Sampling (suite)

- On peut obtenir une estimation non paramétrique des densités a posteriori marginales à l'aide de l'estimateur **Parzen Rosenblatt** sur l'échantillon pondéré $\{\beta^{(k)}, \omega(\beta^{(k)})\}_{k=1}^{N}$.
- Resampling des $\{\beta^{(k)}\}_{k=1}^{N}$ suivant la distribution empirique :

$$\hat{\pi}(\tilde{\beta}) = \sum_{k=1}^{N} \omega(\beta^{(k)}) \cdot \delta_{\beta^{(k)}}(\tilde{\beta}) \approx \pi(\beta|y)$$

Comment choisir la fonction d'importance $q(\beta)$?

- Fonction d'importance minimisant la variance asymptotique : $q_{opt}(\beta) \propto \pi(\beta|y) \cdot |h(\beta) - \mathcal{I}| \Rightarrow$ pas d'intérêt pratique.
- Condition de support : $\text{supp}\{q(\beta)\} \supset \text{supp}\{h(\beta)L(y|\beta,\sigma^2)\pi(\beta|\alpha,\nu)\}$.
- Faire en sorte que les poids $w(\beta) = \frac{L(y|\beta)\pi(\beta|\alpha,\nu)}{q(\beta)}$ soit bornés $\Rightarrow$ estimateur à variance asymptotique finie.
- q facilement simulable à l'aide d'un générateur rapide et fiable.
- Choisir une fonction d'importance q qui "s'approche" le plus possible de la densité a posteriori.

Astuce typique dans les modèles linéaires bayésiens

- Si l'on impose une loi a priori impropre $\pi(\beta|\alpha,\nu) = 1$.
- On obtient $\pi(\beta|y,\sigma) \propto L(y|\beta,\sigma^2) \propto \exp\{-\frac{1}{2\sigma^2}\|y - X\beta\|_2^2\}$.
- Or il est possible de vérifier que la vraisemblance $L(y|\beta,\sigma^2)$ est équivalente en β à la densité de la loi normale multivariée $\mathcal{N}\{\hat{\beta}_{\mathrm{ols}}, \mathbb{V}(\hat{\beta}_{\mathrm{ols}})\}$.

Fonction instrumentale

- On peut donc proposer comme fonction instrumentale la densité de la loi normale multivariée, $q(\beta) = \mathcal{N}\{\beta|\hat{\beta}_{\mathrm{ols}}, \mathbb{V}(\hat{\beta}_{\mathrm{ols}})\}$, facilement simulable sur **R** à l'aide du package **mvtnorm**.

Poid d'importance

On pose :

$$w(\beta) = \frac{L(y|\beta,\sigma^2)\pi(\beta|\alpha,\nu)}{\mathcal{N}\{\beta|\hat{\beta}_{\text{ols}},\mathbb{V}(\hat{\beta}_{\text{ols}})\}} \propto \pi(\beta|\alpha,\nu).$$

les poids sont bornés par : $\max_{\beta} \pi(\beta|\alpha,\nu) = \pi(0|\alpha,\nu) = \frac{\alpha^p.\nu^{p/\alpha}}{2^p.[\Gamma(1/\alpha)]^p}$.

Estimateur AIS

On peut alors employer l'estimateur autonormalisé pour l'estimation de $\mathcal{I}(\alpha,\nu,\sigma^2)$:

$$\hat{\mathcal{I}}_{AIS} = \frac{\sum_{k=1}^{N} h(\beta^{(k)})\pi(\beta^{(k)}|\alpha,\nu)}{\sum_{k=1}^{N} \pi(\beta^{(k)}|\alpha,\nu)}$$

Où $\{\beta^{(1)},\ldots,\beta^{(N)}\} \overset{iid}{\sim} \mathcal{N}\{\hat{\beta}_{\text{ols}},\mathbb{V}(\hat{\beta}_{\text{ols}})\}$.

Critère $\mathbb{ESS}$ (*taille effective d'échantillon*)

$$\mathbb{ESS} \overset{\text{def}}{=} \frac{\left\{ \displaystyle\sum_{k=1}^{N} \omega(\beta^{(k)}) \right\}^2}{\displaystyle\sum_{k=1}^{N} \omega(\beta^{(k)})^2} \in [\![1; N]\!]$$

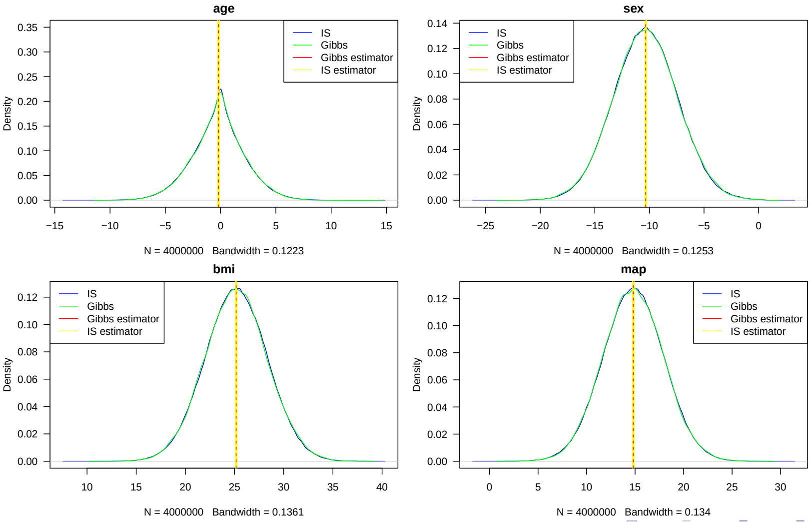
Critère $\mathbb{EF}$ (facteur d'efficacité)

$$\mathbb{EF} = \mathbb{ESS}/N \equiv \mathbb{ESS}\%$$

TABLE : $\mathbb{ESS}$ en % dans le cas où α, ν et σ^2 sont fixés.

	Diabete	Ozone	Boston
$\mathbb{EF}$	13%	90%	93%

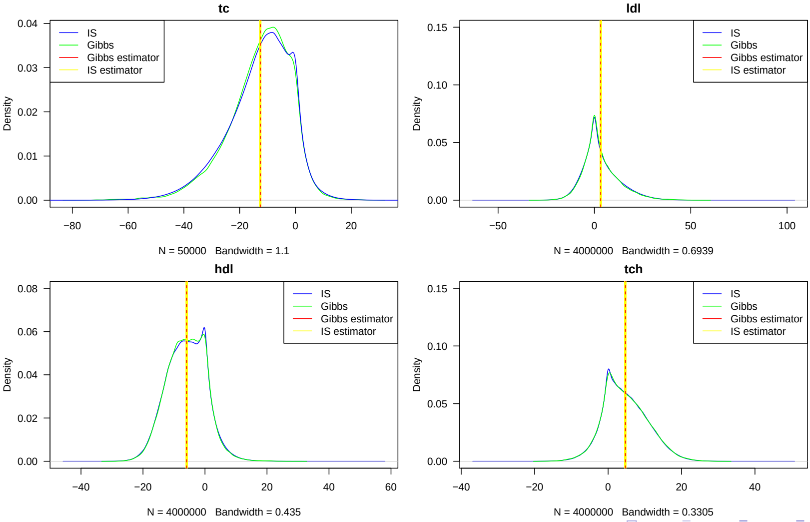
On fixe $\hat{\sigma}^2 = \frac{1}{n}\sum \hat{\epsilon}_i^2$ et $\alpha = 0.5$, $\nu = 4.5$.



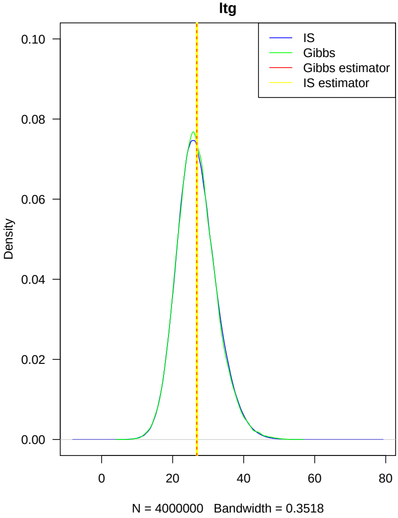
Application sur les données du diabète

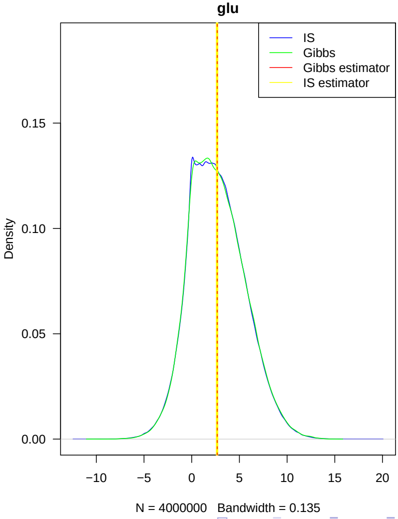
Paramètre σ^2 et hyperparamètres α et ν fixés

On fixe $\hat{\sigma}^2 = \frac{1}{n}\sum \hat{\epsilon}_i^2$ et $\alpha = 0.5$, $\nu = 4.5$.

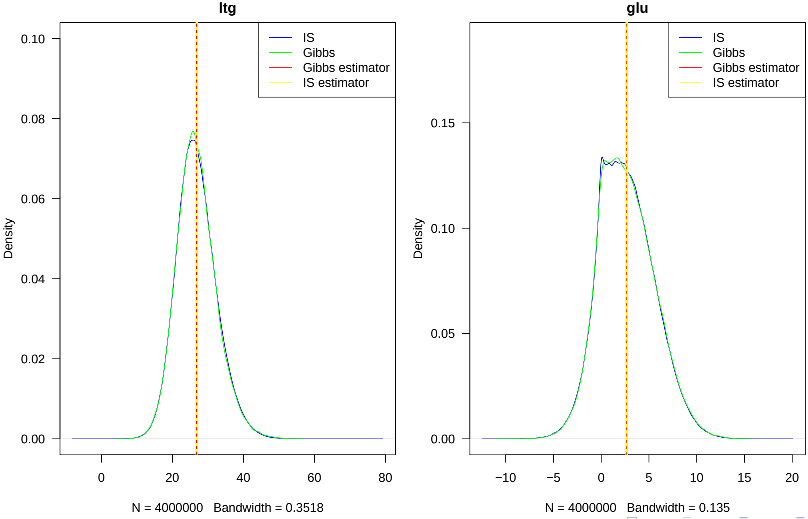


On fixe $\hat{\sigma}^2 = \frac{1}{n}\sum \hat{\epsilon}_i^2$ et $\alpha = 0.5$, $\nu = 4.5$.





On fixe $\hat{\sigma}^2 = \frac{1}{n}\sum \hat{\epsilon}_i^2$ et $\alpha = 0.5$, $\nu = 4.5$.



Comment se comporte la méthode importance sampling lorsque la dimension du modèle augmente ?

Traitement préalable des DATA *Boston* et *Ozone*

- On ajoute toutes les variables croisées $(x^i \cdot x^j)_{i \neq j}$ et carrées $(x^i)^2$.
- On obtient **103** variables explicatives pour la base *Boston*.
- On obtient **54** variables explicatives pour la base de données *Ozone*.

TABLE : $\mathbb{EF}$ lorsque la dimension du modèle augmente.

	Boston	Ozone
$\mathbb{EF}$	0.001%	0.001%

Variance des poids

$$\mathbb{V}_g(\omega(\beta)) = \mathbb{E}_g[\omega(\beta)^2] - \mathbb{E}_g^2[\omega(\beta)].$$

$$\mathbb{E}_g[\omega(\beta)^2] \geq 1 \iff \prod_{j=1}^{p} \mathbb{E}_g \left(\frac{f_1(\beta_j)}{g_1(\beta_j)} \right) \geq 1 \iff C^p \geq 1.$$

Monte Carlo Séquentiel

Rappel de la problématique

- On souhaite simuler les paramètres $\beta^{(k)}$, appelés **"particules"**, suivant la distribution cible $\pi(\beta|y,\alpha,\nu,\sigma^2) \propto L(y|\beta,\sigma^2)\cdot\pi(\beta|\alpha,\nu)$.

Remarques

- Les données ne possèdent pas de temporalité.
- Comment appliquer l'Algorithme SMC dans notre problème statique ?

Création (artificielle) d'une séquence de distribution : SMC tempering

- On définit une séquence de distribution pour tout $t \in [\![1;T]\!]$,

$$\boldsymbol{\pi_t(\beta) \propto q(\beta)^{1-\gamma_t}\{L(y|\beta,\sigma^2)\cdot\pi(\beta|\alpha,\nu)\}^{\gamma_t}}$$

- $q(\beta) \equiv$ proposal.
- On définit $(\gamma_t)_{t\in[\![1;T]\!]}$ tel que $\gamma_0 = 0 < \gamma_1 < \cdots < \gamma_T = 1$, suite croissante appelée "la température".

Propriété directe

- $\pi_0(\beta) \propto q(\beta)$ (proposal)
- $\pi_T(\beta) \propto \pi(\beta|y,\alpha,\nu,\sigma^2)$ (distribution a posteriori)

Choix du proposal $q(\beta)$?

- $q(\beta) = \mathcal{N}\{\beta|\hat{\beta}_{\text{ols}}, \mathbb{V}(\hat{\beta}_{\text{ols}})\}$.
- Dans le cas où σ^2 est fixé $\Rightarrow$ simplifications avec la vraisemblance $L(y|\beta, \sigma^2)$.
- On obtient la séquence de distribution suivante :

$$\boldsymbol{\pi_t(\beta) \propto \mathcal{N}\{\beta|\hat{\beta}_{\text{ols}}, \mathbb{V}(\hat{\beta}_{\text{ols}})\} \cdot \pi(\beta|\alpha, \nu)^{\gamma_t}} \ .$$

Propriété

- La séquence de distribution $\pi_t(\beta)$ permet d'atteindre la distribution cible $\pi(\beta|y, \alpha, \nu, \sigma^2)$ au fur et à mesure.
- Evolution de la distribution auxiliaire, $\pi_0(\beta) \longrightarrow \ldots \longrightarrow \pi_t(\beta) \longrightarrow \pi_{t+1}(\beta) \longrightarrow \ldots \longrightarrow \pi_T(\beta)$.

Algorithme SMC tempering avec α, σ^2, ν fixés

Initialisation (t=0) :

- $\beta^{(1)}, \dots, \beta^{(N)} \sim \mathcal{N}(\beta | \hat{\beta}_{ols}, \sigma^2 (X^T X)^{-1})$ (proposal)

For t=1 to T :

- **Poids** : $\omega_t(\beta^{(k)}) = \frac{\pi_t(\beta^{(k)})}{\pi_{t-1}(\beta^{(k)})} \propto \exp\{-(\gamma_t - \gamma_{t-1}) \cdot \nu \sum_{j=1}^{p} |\beta_j|^{\alpha}\}$ (poids des particules)
- **Resampling** : $\mathbb{P}(\hat{\beta}^{(k)} = \beta^{(k)}) = \frac{\omega_t(\beta^{(k)})}{\sum_k \omega_t(\beta^{(k)})}$ (weighted boostrap)
- **Mutation/Move** : MCMC Hastings par marche aléatoire : $\xi^{(k)} = \hat{\beta}^{(k)} + \epsilon$ où $\epsilon \sim \mathcal{N}(0, \mathbb{V}(\hat{\beta}) \cdot \kappa)$ avec $\kappa = 0.05$

$$\tilde{\beta}^{(k)} = \begin{cases} \xi^{(k)} & \text{avec probabilité } \rho(\xi^{(k)}, \hat{\beta}^{(k)}) = \min\left\{1, \frac{\pi_t(\xi^{(k)})}{\pi_t(\hat{\beta}^{(k)})}\right\} \\ \hat{\beta}^{(k)} & \text{avec probabilité } 1 - \rho(\xi^{(k)}, \hat{\beta}^{(k)}) \end{cases}$$

End For
Return : $\tilde{\beta}$

Monte Carlo Séquentiel

Propriété

- A chaque itération on simule suivant la distribution empirique :

$$\hat{\pi}_t(\tilde{\beta}) = \sum_{k=1}^{N} \omega_t(\beta^{(k)}) \cdot \delta_{\beta^{(k)}}(\tilde{\beta}) \approx \pi_t(\tilde{\beta})$$

- A l'itération $t = T$ on a :

$$\hat{\pi}_T(\tilde{\beta}) = \sum_{k=1}^{N} \omega_T(\beta^{(k)}) \cdot \delta_{\beta^{(k)}}(\tilde{\beta}) \approx \pi(\beta|y, \alpha, \nu, \sigma^2)$$

Amélioration : comment choisir la séquence $(\gamma_t)_{t \in [\![1;T]\!]}$?

Algorithme SMC tempering Adaptatif avec α, σ^2, ν fixés

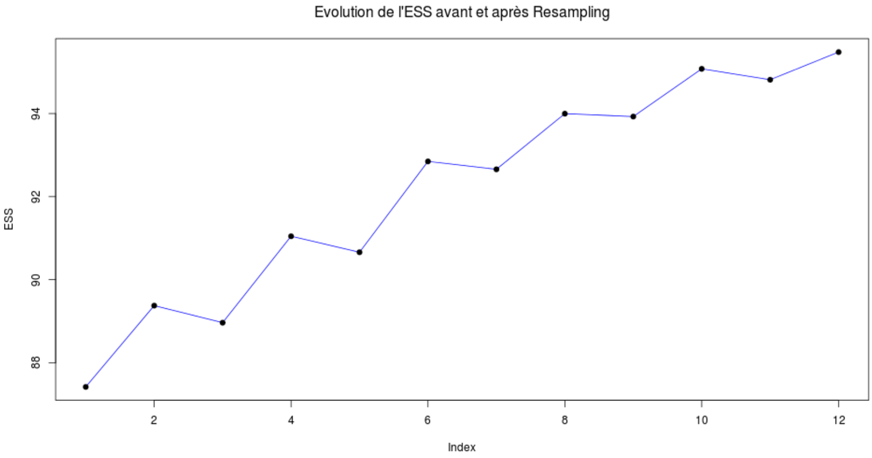
Au lieu de choisir une suite $(\gamma_t)_t$ arbitaire il est possible de rendre l'algorithme SMC adaptatif en résolvant à chaque itération t l'équation suivante en γ :

$$\mathbb{EF}(\gamma) = 98\%$$

Il est possible de prendre une autre valeur supérieure à 98%.

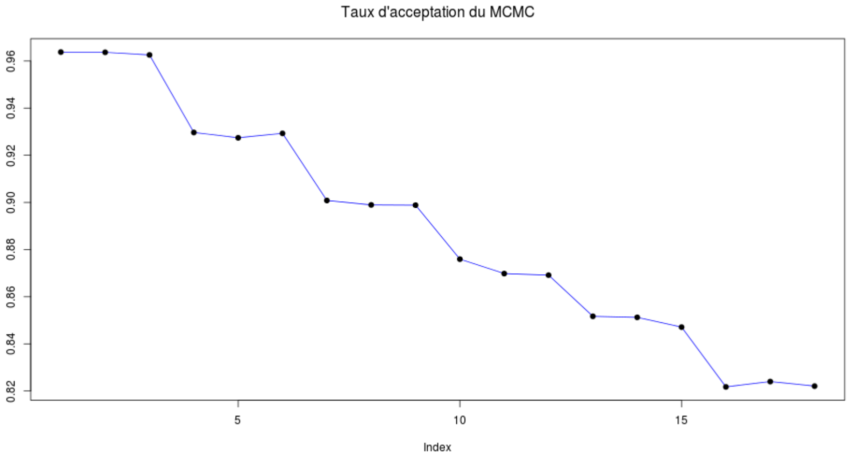
Diabete Data : résultats dans le cas particulier où σ^2, α et ν sont fixés

$(\gamma_t)_{t \in [\![1;T]\!]} = [0; 0.2; 0.4; 0.6; 0.8; 1]$ et $\alpha = 0.5$, $\nu = 4.5$ et $\hat{\sigma}^2 = \hat{\sigma}^2_{\mathrm{ML}}$



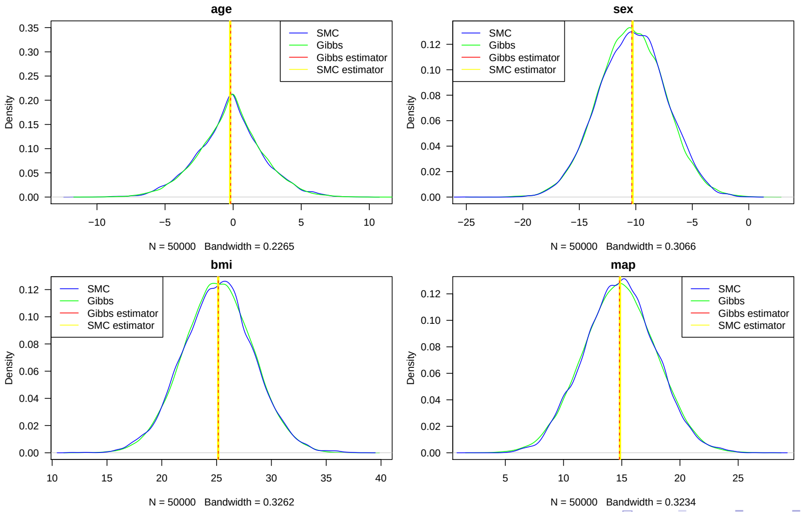
Diabete Data : résultats dans le cas particulier où σ^2, α et ν sont fixés

$(\gamma_t)_{t\in[\![1;T]\!]} = [0; 0.2; 0.4; 0.6; 0.8; 1]$ et $\alpha = 0.5$, $\nu = 4.5$ et $\hat{\sigma}^2 = \hat{\sigma}^2_{\mathrm{ML}}$



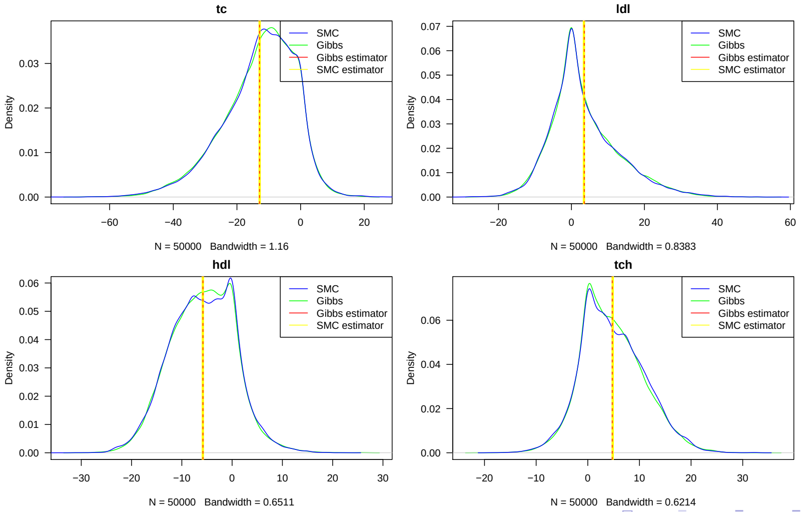
Diabete Data : résultats dans le cas particulier où σ^2, α et ν sont fixés

$(\gamma_t)_{t\in[\![1;T]\!]} = [0; 0.2; 0.4; 0.6; 0.8; 1]$ et $\alpha = 0.5$, $\nu = 4.5$ et $\hat{\sigma}^2 = \hat{\sigma}^2_{\mathrm{ML}}$



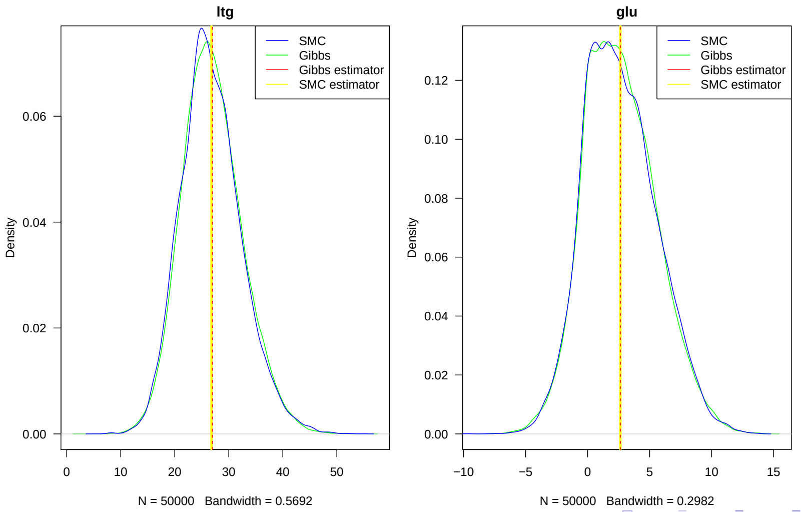
Diabete Data : résultats dans le cas particulier où σ^2, α et ν sont fixés

$(\gamma_t)_{t\in[\![1;T]\!]} = [0; 0.2; 0.4; 0.6; 0.8; 1]$ et $\alpha = 0.5$, $\nu = 4.5$ et $\hat{\sigma}^2 = \hat{\sigma}^2_{\mathrm{ML}}$



Diabete Data : résultats dans le cas particulier où σ^2, α et ν sont fixés

$(\gamma_t)_{t\in[\![1;T]\!]} = [0; 0.2; 0.4; 0.6; 0.8; 1]$ et $\alpha = 0.5$, $\nu = 4.5$ et $\hat{\sigma}^2 = \hat{\sigma}^2_{\mathrm{ML}}$



Ouverture

Considérer le cas général avec σ^2, α et ν inconnus

- Importance sampling avec hyperprior sur α et ν et un prior sur σ^2.
- Même idée pour le SMC tempering.
- Appliquer le SMC tempering lorsque la dimension du modèle est grande.
- Rendre le SMC tempering adaptatif.

Ouverture


Optima
KIA
07-07
INDIANA
10
1
CELEBRATE THE ARTS
53
I'd rather be doing
Bayesian statistics