Use of Importance sampling with MCMC algorithms

J. Rousseau

CEREMADE, Université Paris-Dauphine

Warwick

Joint work with C. Guihennec, R. McVinish, K. Mengersen, D. Nur

Outline

Outline

Motivation

- **Repeated MCMC under different samples :**

• Bayesian p-values. $X^o =$ observed sample
Let $H_\pi(X) = E^\pi[h(\theta)|X]$ be a test statistic and
$P[H(X) > H(X^o)] = p(X^o)$: a p value to evaluate $p(X^o)$
compute

- For $j = 1, ...J$ $X^{(j)} \sim P$ and compute $H(X^{(j)})$

BUT... $H(X^{(t)})$ evaluated using MCMC $\forall t \longrightarrow$ Time consuming
• Bayesian cross validation : Need of computing $H(X^{(-l)})$, where $X^{(-l)} \subset X$ for many $(-l)$.
• Evaluation of procedures by simulations

- **prior sensitivity analysis :** Need to compute $H_{\pi_j}(X)$ for different π_j.

Motivation

▶ **Repeated MCMC under different samples :**

• Bayesian p-values. X^o = observed sample

Let $H_\pi(X) = E^\pi[h(\theta)|X]$ be a test statistic and

$P[H(X) > H(X^o)] = p(X^o)$: a p value to evaluate $p(X^o)$

compute

- For $j = 1, ...J$ $X^{(j)} \sim P$ and compute $H(X^{(j)})$
- $$\hat{p}(X^o) = \frac{1}{J}\sum_{j=1}^{J} \mathbb{I}_{H(X^{(j)})>H(X^o)}$$

BUT... $H(X^{(t)})$ evaluated using MCMC $\forall t \longrightarrow$ Time consuming

• Bayesian cross validation : Need of computing $H(X^{(-l)})$, where $X^{(-l)} \subset X$ for many $(-l)$.

• Evaluation of procedures by simulations

▶ **prior sensitivity analysis :** Need to compute $H_{\pi_j}(X)$ for different π_j.

Outline

Framework

- **Bayesian model**

$$X \sim f_\theta, \quad \theta \in \Theta, \quad \theta \sim \pi,$$

- **Object of interest** $H_\pi(X) = E^\pi[h(\theta)|X]$
- **Evaluation with MCMC** $(\theta^t)_{t=1}^T = \text{MC}\,(\pi(.|X))$

$$\hat{H}_\pi(X) = \frac{1}{T}\sum_t h(\theta^t)$$

- **New sample** : $Y \overset{d}{=} X$

$$H_\pi(Y) = \frac{\int_\Theta h(\theta)[f(Y|\theta)/f(X|\theta)]d\pi(\theta|X)}{\int_\Theta [f(Y|\theta)/f(X|\theta)]d\pi(\theta|X)}$$

$$H_\pi(Y) = \frac{\int_\Theta h(\theta)[f(Y|\theta)/f(X|\theta)]d\pi(\theta|X)}{\int_\Theta [f(Y|\theta)/f(X|\theta)]d\pi(\theta|X)}$$

So No need to run a new MCMC : use IS on $(\theta^t)_t$:

$$\hat{H} = \frac{\sum_{t=1}^T h(\theta^t)w(\theta^t, y, x)}{\sum_{t=1}^T w(\theta^t, y, x)} = \frac{\bar{hw}}{\bar{w}},$$

where

$$w(\theta, y, x) = \frac{f_{\theta^t}(y)}{f_{\theta^t}(x)}, \quad \text{or} \quad w(\theta, \pi', \pi) = \frac{\pi'(\theta)}{\pi(\theta)},$$

- **Much quicker**
- **How good/bad is it ?**

Outline

Convergence

- **Back to the theory on MCMC convergence :**

• Consistency (in T) : OK if MC ergodic

• rate : In the original MC : estimation of the function

$$\tilde{h}(\theta) = h(\theta)\frac{f(Y|\theta)}{f(X|\theta)}$$

Usual tools • Asymptotic Variance

$$\gamma^2 = \frac{m_\pi(x)^2}{m_{\pi'}(y)^2}\left[H_\pi^2 \text{var}(\bar{w}) + \text{var}(\bar{hw}) - 2\text{cov}(\bar{w}, \bar{hw})H_\pi\right].$$

• Variance estimation : Same asy var as

$$Z(\theta_t) = \frac{\mathbb{E}_\pi(h(\theta)w(\theta))}{\mathbb{E}_\pi(w(\theta))}\left(\frac{w(\theta_t)}{\mathbb{E}_\pi(w(\theta))} - \frac{h(\theta_t)w(\theta_t)}{\mathbb{E}_\pi(h(\theta)w(\theta))}\right)$$

Outline

Behaviour of the weights

$$w(\theta, y, x) = \frac{f_{\theta^t}(y)}{f_{\theta^t}(x)}, \quad \tilde{w}(\theta) = \pi(\theta|x)/\pi(\theta|y)$$

$$x = (x_1, ..., x_n) \stackrel{d}{=} y = (y_1, ..., y_n) \text{ + regul. condits} \Rightarrow$$

$$\tilde{w}(\theta) = e^{n(\hat{\theta}^x - \hat{\theta}^y)' I(\theta_0)(\theta - \hat{\theta}^x)}(1 + O(n^{-1/2})),$$

$$\text{var}_a s\,(\tilde{w}(\theta) \mid x, y) \quad = \quad \exp\left\{n(\hat{\theta}^x - \hat{\theta}^y)' I(\theta_0)(\hat{\theta}^x - \hat{\theta}^y)\right\} - 1.$$

- **Stability** $\hat{\theta}^x \neq \hat{\theta}^y \longrightarrow$ Instability.

Outline

Stabilization by recentering

- **Simple stabilization**

$$\theta'_t = \theta_t + \hat{\theta}^y - \hat{\theta}^x, \quad w'(\theta_t) = \frac{\pi(\theta'_t) f(y^n|\theta'_t)}{\pi(\theta_t) f(x^n|\theta_t)} = 1 + 0_P(n^{-1/2})$$

- Very effective if posterior not too strongly multimodal
- Often $\hat{\theta}^x$ complicated to calculate : Two-step procedure

- **Compute centering**

$$\tilde{\theta}^y = \frac{\sum_t \theta_t w_t}{\sum_t w_t}$$

- **Apply centering**

$$\theta'_t = \theta_t + \tilde{\theta}^y - \tilde{\theta}^x$$

- **Conditions** x and y have marginally the same distribution but are not nece. indpdt.

see also MacEachern+Perrugia

Outline

Regression example (explicit calculations)

- **Problem :** Test for

$$\mathrm{H}_0 : Y_i \sim \mathcal{N}(\beta_0, \sigma^2) \quad \mathrm{H}_1 : Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

- **test procedure :** $|\beta_1| < \epsilon$ versus$|\beta_1| \geq \epsilon \Rightarrow$

$$\mathrm{H}_0 \quad \text{iff } H(Y) = E^{\pi}\left[\beta_1^2 | Y\right] < \epsilon$$

- **pb : Choice of ϵ ?** $\Rightarrow$ use of p-value

Under H_0 $\theta = (\beta_0, \sigma)$ unknown $\Rightarrow$ Use of conditional predictive p-value (Bayarri+Berger, Robbins et al., Robert + Rousseau, Fraser + Rousseau)

$$p(Y^o) = \int_{\Theta} P_\theta[H(Y) > H(Y^0)|\hat{\theta}] d\pi_0(\theta|\hat{\theta}) = P_\theta[H(Y) > H(Y^0)|\hat{\theta}]$$

Special case here : $\pi(\beta_0, \beta_1, \sigma) \propto 1/\sigma$ and $n = 250$.

$$H(y) = \mathbb{E}_\pi\left(\beta_1^2 \mid y, x\right) = \hat{\beta}_1^2 + \frac{\sum(y_i - \hat{y}_i)^2}{(n-4)\sum(x_i - \bar{x})^2},$$

- **Algorithm for each p-value**

• $[Y^1|\hat{\beta}_0, \hat{\sigma}] \sim f(Y|\hat{\theta}) = \mathcal{U}_E$ + MCMC $\pi(\psi|Y^1)$ $\psi = (\beta_0, \beta_1, \sigma)$,

$$\rightarrow \psi^t, \quad t = 1, ..., T$$

• For $j = 2, ..., J$ Simulate $[Y^j|\hat{\beta}_0, \hat{\sigma}] \overset{iid}{\sim} f(Y|\hat{\theta})$
• $\forall j = 2, ..., J$ compute $w_t(\psi^t, Y^j, Y^1)$ and $\hat{\psi}^j$ and

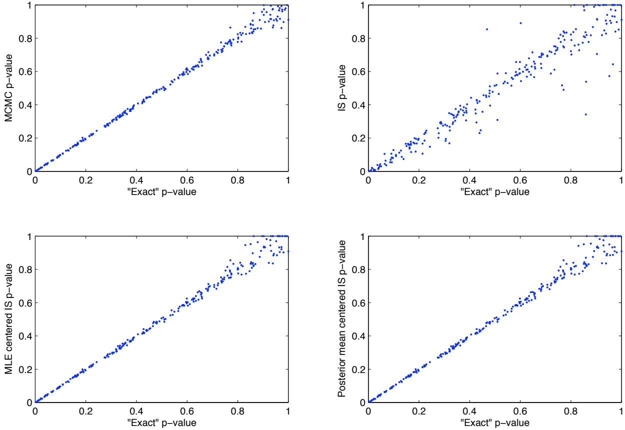
$$H_S(Y^j) = \frac{\sum_t w_t(\beta_1^t)^2}{\sum_t w_t}, \quad \text{and also } \mathbb{E}_\pi\left(\beta_1^2 \mid Y^j\right)$$

$$H_{CS1} = \frac{\sum_t w_t'((\beta_1^t)')^2}{\sum_t w_t'}, \quad (\psi^t)' = \psi^t + \hat{\psi}^j - \hat{\psi}^1$$

$$H_{CS2} = \frac{\sum_t w_t''((\beta_1^t)'')^2}{\sum_t w_t''}, \quad (\psi^t)'' = \psi^t + E^\pi[\psi|Y^j] - E^\pi[\psi|Y^1]$$

Results

250 p-values and for each : $J = 1000$ (M.C samples) and $T = 10^5$ (MCMC samples) with burn-in = 1000



Weights

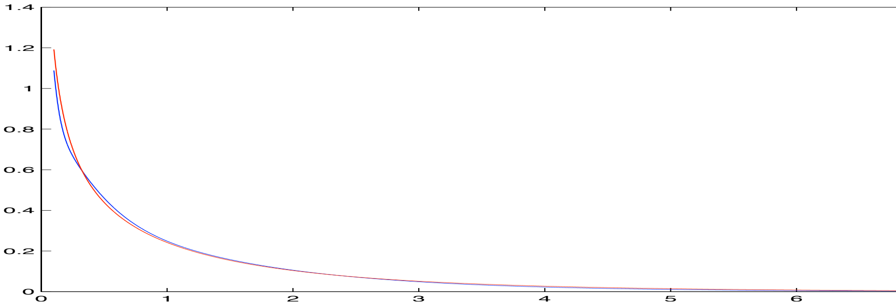


FIG.: Blue line is a kernel density estimate of the calculated variance of the normalised IS weights. Red line is the density of a χ^2_1 distribution.

Remarks

- Weights : close to asymptotic

Remarks

- Weights : close to asymptotic
- IS simple : OK but not marvelous

Remarks

- Weights : close to asymptotic
- IS simple : OK but not marvelous
- IS recentered (MLE or posterior) : much better

Remarks

- Weights : close to asymptotic
- IS simple : OK but not marvelous
- IS recentered (MLE or posterior) : much better
- posterior distribution : close to Gaussian $\Rightarrow$ perfect for recentering.

GOF : Nonparametric example

- **Problem**

$$H_0 : f_* \in \mathcal{F} \quad \text{against} \quad H_1 : f_* \notin \mathcal{F}, \quad \mathcal{F} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$$

- **Nonparametric model for $\mathcal{F}^c$** (VW+RR+R)

$$F(y \mid \theta) \sim G_\psi, \quad \psi \in \mathcal{S} \quad \text{on } [0,1] \quad \exists \psi_0; g_{\psi_0} \equiv 1$$

If $Y \sim f_\theta$ then $F(Y \mid \theta) \sim \mathcal{U}(0,1)$. Model :

$$f_*(y \mid \theta, \psi) = f(y \mid \theta) g(F(y \mid \theta) \mid \psi), \theta \in \Theta, \quad \psi \in \mathcal{S}$$

- **prior on H_1**

$$d\pi_1(\theta, \psi) = d\pi_0(\theta) d\pi(\psi)$$

Test

- **Test statistic** (Bayesian) $H(x) = \mathbb{E}_{\pi_1}[d(1, g(\cdot \mid \psi)) \mid x]$
- *p*-**value**

$$p(x^0) = \int_\Theta P_\theta[H(y) > H(x^0)|\hat{\theta}]\pi_0(\theta)d\theta$$

- **Set up here** $f_\theta \equiv \exp(\theta)$ $\pi_0 = \Gamma(\gamma_1, \gamma_2)$ $\pi_1 =$ mixture of triangular distributions (fixed partition, random weights) $\psi = (k, \omega)$, $k \in \mathbb{N}$, $\omega \in \mathcal{S}_k = \{z \in [0,1]^k; \sum_{i=0}^k z_i = 1\}$,

$$\begin{aligned} g(y \mid \omega, k) &= \sum_{i=0}^{k} \omega_i h_i(y; k), \quad \pi(k) = C(\rho)\rho^k, \quad k \geq 1, \\ \pi(\omega \mid k) &= \mathcal{D}(\alpha_{0,k}, \ldots, \alpha_{k,k}), \end{aligned}$$

Algorithm

- $y^1|\hat{\theta}^x \sim f(y|\hat{\theta}^y = \hat{\theta}^x)$

Algorithm

- $y^1|\hat{\theta}^x \sim f(y|\hat{\theta}^y = \hat{\theta}^x)$
- MCMC = RJMCMC $\eta^t = (\theta^t, k^t, \omega^t)$ for $\pi(\eta|y^1)$

$$\text{Compute} \quad H(y^1)$$

Algorithm

- $y^1|\hat{\theta}^x \sim f(y|\hat{\theta}^y = \hat{\theta}^x)$
- MCMC = RJMCMC $\eta^t = (\theta^t, k^t, \omega^t)$ for $\pi(\eta|y^1)$

$$\text{Compute} \quad H(y^1)$$

- $\forall j = 2, ..., J \; y^j \stackrel{d}{=} y^1$ (iid)

$$\text{Compute} \quad w_t(\eta^t, y^j, y^1) \Rightarrow H(y^j) \Rightarrow p(x^0) = \frac{\sum_j \mathbb{I}_{H(y^j) > H(x^0)}}{J}$$

Algorithm

- $y^1|\hat{\theta}^x \sim f(y|\hat{\theta}^y = \hat{\theta}^x)$
- MCMC = RJMCMC $\eta^t = (\theta^t, k^t, \omega^t)$ for $\pi(\eta|y^1)$

$$\text{Compute} \quad H(y^1)$$

- $\forall j = 2, ..., J$ $y^j \stackrel{d}{=} y^1$ (iid)

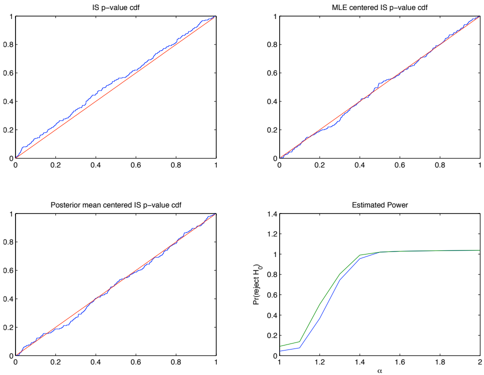
$$\text{Compute} \quad w_t(\eta^t, y^j, y^1) \Rightarrow H(y^j) \Rightarrow p(x^0) = \frac{\sum_j \mathbb{I}_{H(y^j) > H(x^0)}}{J}$$

- Recentering : only on $k = 2$: MLE or posterior mean of (θ, w_1) Because

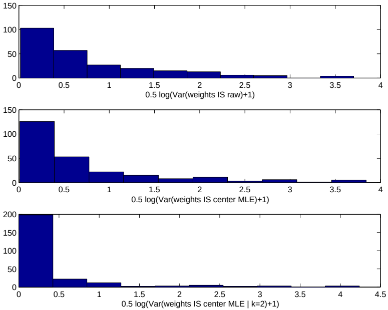
$$P^\pi[k = 2|y_1, ..., y_n] = 1 + o_P(1), \quad \text{if} \quad y = (y_1, ..., y_n) \in H_0$$

results

$n = 250$, $x^0 \sim \exp(1/4)$, $J = 1000$, $T = 100000$ with burn-in = 5000 and 250 p-values



weights


0.5 log(Var(weights IS raw)+1)
0.5 log(Var(weights IS center MLE)+1)
0.5 log(Var(weights IS center MLE | k=2)+1)

conclusion

- IS in MCMC for repeated sampling : promising

conclusion

- IS in MCMC for repeated sampling : promising
- curse of dimensionality not so severe

conclusion

- IS in MCMC for repeated sampling : promising
- curse of dimensionality not so severe
- Much quicker

conclusion

- IS in MCMC for repeated sampling : promising
- curse of dimensionality not so severe
- Much quicker
- possible simple improvements : recentering (+ rescaling)

conclusion

- IS in MCMC for repeated sampling : promising
- curse of dimensionality not so severe
- Much quicker
- possible simple improvements : recentering (+ rescaling)
- Other improvements : If data set y^j too different from y^1 IS not too good : consider $(y^1, ..., y^{J_0})$ Guihenneuc et al.
  - for each : 1 MCMC
  - New y^j : choose *best* $y^l, l = 1, ..., J_0$ compute IS with it.

conclusion

- IS in MCMC for repeated sampling : promising
- curse of dimensionality not so severe
- Much quicker
- possible simple improvements : recentering (+ rescaling)
- Other improvements : If data set y^j too different from y^1 IS not too good : consider $(y^1, ..., y^{J_0})$ Guihenneuc et al.
  - for each : 1 MCMC
  - New y^j : choose *best* $y^l, l = 1, ..., J_0$ compute IS with it.
- Excellent for prior sensitivity analysis

conclusion

- IS in MCMC for repeated sampling : promising
- curse of dimensionality not so severe
- Much quicker
- possible simple improvements : recentering (+ rescaling)
- Other improvements : If data set y^j too different from y^1 IS not too good : consider $(y^1, ..., y^{J_0})$ Guihenneuc et al.
  - for each : 1 MCMC
  - New y^j : choose *best* $y^l, l = 1, ..., J_0$ compute IS with it.
- Excellent for prior sensitivity analysis
- Non stationarity $\Rightarrow$ bad.