Efficient Implementation of MCMC When Using An Unbiased Likelihood Estimator

Arnaud Doucet
University of Oxford
Joint work with M. Pitt, G. Deligiannidis & R. Kohn

Tokyo, 24/07/14

Bayesian Inference

- Likelihood function $p_\theta(y)$ where $\theta \in \Theta \subseteq \mathbb{R}^d$.

Bayesian Inference

- Likelihood function $p_\theta(y)$ where $\theta \in \Theta \subseteq \mathbb{R}^d$.
- Prior distribution of density $p(\theta)$.

Bayesian Inference

- Likelihood function $p_\theta(y)$ where $\theta \in \Theta \subseteq \mathbb{R}^d$.
- Prior distribution of density $p(\theta)$.
- Bayesian inference relies on the posterior

$$\pi(\theta) = p(\theta|\, y) = \frac{p_\theta(y)\, p(\theta)}{\int_\Theta p_{\theta'}(y)\, p(\theta')\, d\theta'}.$$

Bayesian Inference

- Likelihood function $p_\theta(y)$ where $\theta \in \Theta \subseteq \mathbb{R}^d$.
- Prior distribution of density $p(\theta)$.
- Bayesian inference relies on the posterior

$$\pi(\theta) = p(\theta|\, y) = \frac{p_\theta(y)\, p(\theta)}{\int_\Theta p_{\theta'}(y)\, p(\theta')\, d\theta'}.$$

- For non-trivial models, inference relies typically on MCMC.

Metropolis-Hastings algorithm

Metropolis-Hastings algorithm

At iteration i

- Sample $\vartheta \sim q\left(\cdot \middle| \vartheta_{i-1}\right)$.

Metropolis-Hastings algorithm

At iteration i

- Sample $\vartheta \sim q\left(\cdot \mid \vartheta_{i-1}\right)$.
- With probability

$$1 \wedge \frac{\pi\left(\vartheta\right)}{\pi\left(\vartheta_{i-1}\right)} \frac{q\left(\vartheta_{i-1} \mid \vartheta\right)}{q\left(\vartheta \mid \vartheta_{i-1}\right)} = 1 \wedge \frac{p_{\vartheta}\left(y\right) p\left(\vartheta\right)}{p_{\vartheta_{i-1}}\left(y\right) p\left(\vartheta_{i-1}\right)} \frac{q\left(\vartheta_{i-1} \mid \vartheta\right)}{q\left(\vartheta \mid \vartheta_{i-1}\right)},$$

set $\vartheta_i = \vartheta$, otherwise set $\vartheta_i = \vartheta_{i-1}$.

Metropolis-Hastings algorithm

<u>*At iteration i*</u>

- Sample $\vartheta \sim q\left(\cdot|\,\vartheta_{i-1}\right)$.
- With probability

$$1 \wedge \frac{\pi\left(\vartheta\right)}{\pi\left(\vartheta_{i-1}\right)} \frac{q\left(\vartheta_{i-1}|\,\vartheta\right)}{q\left(\vartheta|\,\vartheta_{i-1}\right)} = 1 \wedge \frac{p_{\vartheta}\left(y\right) p\left(\vartheta\right)}{p_{\vartheta_{i-1}}\left(y\right) p\left(\vartheta_{i-1}\right)} \frac{q\left(\vartheta_{i-1}|\,\vartheta\right)}{q\left(\vartheta|\,\vartheta_{i-1}\right)},$$

  set $\vartheta_i = \vartheta$, otherwise set $\vartheta_i = \vartheta_{i-1}$.
- **Problem**: Metropolis-Hastings (MH) cannot be implemented if $p_{\vartheta}\left(y\right)$ cannot be evaluated.

Intractable Likelihood Function

- For latent variable models, one has

$$p_\theta(y) = \int p_\theta(x, y)\, dx$$

where the integral cannot often be evaluated.

Intractable Likelihood Function

- For latent variable models, one has

$$p_\theta(y) = \int p_\theta(x, y)\, dx$$

  where the integral cannot often be evaluated.
- A standard "solution" consists of using MCMC to sample from

$$p(\theta, x|\, y) \propto p_\theta(x, y)\, p(\theta)$$

  by updating iterately x and θ.

Intractable Likelihood Function

- For latent variable models, one has

$$p_\theta(y) = \int p_\theta(x, y)\, dx$$

where the integral cannot often be evaluated.

- A standard "solution" consists of using MCMC to sample from

$$p(\theta, x \mid y) \propto p_\theta(x, y)\, p(\theta)$$

by updating iterately x and θ.

- Gibbs sampling strategies can be slow mixing and difficult to put in practice.

Intractable Likelihood Function

- For latent variable models, one has

$$p_\theta(y) = \int p_\theta(x, y)\, dx$$

  where the integral cannot often be evaluated.
- A standard "solution" consists of using MCMC to sample from

$$p(\theta, x | y) \propto p_\theta(x, y)\, p(\theta)$$

  by updating iterately x and θ.
- Gibbs sampling strategies can be slow mixing and difficult to put in practice.
- Could we use approximations of $p_\theta(y)$ within MH instead?

Pseudo-Marginal MH algorithm

- **Key Idea**: Replace $p_\vartheta(y)$ by an estimate $\widehat{p}_\vartheta(y)$ in MH.

Pseudo-Marginal MH algorithm

- **Key Idea**: Replace $p_\vartheta(y)$ by an estimate $\widehat{p}_\vartheta(y)$ in MH.

Pseudo-Marginal MH algorithm

- **Key Idea**: Replace $p_\vartheta(y)$ by an estimate $\widehat{p}_\vartheta(y)$ in MH.

At iteration i

- Sample $\vartheta \sim q(\cdot|\vartheta_{i-1})$.

Pseudo-Marginal MH algorithm

- **Key Idea**: Replace $p_\vartheta(y)$ by an estimate $\widehat{p}_\vartheta(y)$ in MH.

At iteration i

- Sample $\vartheta \sim q(\cdot|\,\vartheta_{i-1})$.
- Compute an estimate $\widehat{p}_\vartheta(y)$ of $p_\vartheta(y)$.

Pseudo-Marginal MH algorithm

- **Key Idea**: Replace $p_\vartheta(y)$ by an estimate $\widehat{p}_\vartheta(y)$ in MH.

<u>*At iteration i*</u>

- Sample $\vartheta \sim q(\cdot|\,\vartheta_{i-1})$.
- Compute an estimate $\widehat{p}_\vartheta(y)$ of $p_\vartheta(y)$.
- With probability

$$1 \wedge \frac{\widehat{p}_\vartheta(y)\,p(\vartheta)}{\widehat{p}_{\vartheta_{i-1}}(y)\,p(\vartheta_{i-1})}\frac{q(\vartheta_{i-1}|\,\vartheta)}{q(\vartheta|\,\vartheta_{i-1})},$$

set $\vartheta_i = \vartheta$, $\widehat{p}_{\vartheta_i}(y) = \widehat{p}_\vartheta(y)$ otherwise set $\vartheta_i = \vartheta_{i-1}$, $\widehat{p}_{\vartheta_i}(y) = \widehat{p}_{\vartheta_{i-1}}(y)$.

Pseudo-Marginal MH algorithm

- **Key Idea**: Replace $p_\vartheta(y)$ by an estimate $\widehat{p}_\vartheta(y)$ in MH.

<u>*At iteration i*</u>

- Sample $\vartheta \sim q(\cdot|\,\vartheta_{i-1})$.
- Compute an estimate $\widehat{p}_\vartheta(y)$ of $p_\vartheta(y)$.
- With probability

$$1 \wedge \underbrace{\frac{p(y;\vartheta)}{p(y;\vartheta_{i-1})}\frac{p(\vartheta)}{p(\vartheta_{i-1})}\frac{q(\vartheta_{i-1}|\,\vartheta)}{q(\vartheta|\,\vartheta_{i-1})}}_{\text{exact MH ratio}} \times \underbrace{\frac{\widehat{p}(y;\vartheta)/p(y;\vartheta)}{\widehat{p}(y;\vartheta_{i-1})/p(y;\vartheta_{i-1})}}_{\text{noise}}$$

set $\vartheta_i = \vartheta$, $\widehat{p}_{\vartheta_i}(y) = \widehat{p}_\vartheta(y)$ otherwise set $\vartheta_i = \vartheta_{i-1}$, $\widehat{p}_{\vartheta_i}(y) = \widehat{p}_{\vartheta_{i-1}}(y)$.

Importance Sampling Estimator

- For latent variable models, one has

$$p_\theta(y) = \int p_\theta(x, y)\, dx = \int \frac{p_\theta(x, y)}{q_\theta(x)} q_\theta(x) dx$$

where $q_\theta(x)$ is an importance sampling density.

Importance Sampling Estimator

- For latent variable models, one has

$$p_\theta(y) = \int p_\theta(x, y)\, dx = \int \frac{p_\theta(x, y)}{q_\theta(x)} q_\theta(x) dx$$

where $q_\theta(x)$ is an importance sampling density.

- An unbiased estimator is given by

$$\widehat{p}_\theta(y) = \frac{1}{N} \sum_{k=1}^{N} \frac{p_\theta\left(X^k, y\right)}{q_\theta(X^k)}, \qquad X^k \overset{\text{i.i.d.}}{\sim} q_\theta(\cdot)$$

Sequential Monte Carlo Estimator

- $\{X_t\}_{t\geq 1}$ is a $\mathbb{X}$-valued latent Markov process with $X_1 \sim \mu(\cdot;\theta)$ and $X_{t+1}|X_t \sim f\left(\cdot|\, X_t;\theta\right)$.

Sequential Monte Carlo Estimator

- $\{X_t\}_{t\geq 1}$ is a $\mathbb{X}$-valued latent Markov process with $X_1 \sim \mu(\cdot;\theta)$ and $X_{t+1}|X_t \sim f\left(\cdot|\, X_t;\theta\right)$.
- Observations $\{Y_t\}_{t\geq 1}$ are conditionally independent given $\{X_t\}_{t\geq 0}$ with $Y_t|\,\{X_k\}_{k\geq 0} \sim g(\cdot|X_t,\theta)$.

Sequential Monte Carlo Estimator

- $\{X_t\}_{t\geq 1}$ is a $\mathbb{X}$-valued latent Markov process with $X_1 \sim \mu(\cdot;\theta)$ and $X_{t+1}|X_t \sim f\left(\cdot|\, X_t;\theta\right)$.
- Observations $\{Y_t\}_{t\geq 1}$ are conditionally independent given $\{X_t\}_{t\geq 0}$ with $Y_t|\,\{X_k\}_{k\geq 0} \sim g(\cdot|X_t,\theta)$.
- Likelihood of $y_{1:T} = (y_1, ..., y_T)$ is

$$p(y_{1:T};\theta) = \int_{\mathbb{X}^T} p(x_{1:T}, y_{1:T};\theta)dx_{1:T}.$$

Sequential Monte Carlo Estimator

- $\{X_t\}_{t\geq 1}$ is a $\mathbb{X}$-valued latent Markov process with $X_1 \sim \mu(\cdot;\theta)$ and $X_{t+1}|X_t \sim f(\cdot|X_t;\theta)$.
- Observations $\{Y_t\}_{t\geq 1}$ are conditionally independent given $\{X_t\}_{t\geq 0}$ with $Y_t|\{X_k\}_{k\geq 0} \sim g(\cdot|X_t,\theta)$.
- Likelihood of $y_{1:T} = (y_1, ..., y_T)$ is

$$p(y_{1:T};\theta) = \int_{\mathbb{X}^T} p(x_{1:T}, y_{1:T};\theta)dx_{1:T}.$$

- SMC provides an unbiased estimator of relative variance $\mathcal{O}(T/N)$ where N is the number of particles.

Main Result

- **Proposition**: Let $\widehat{p}_{\vartheta}(y)$ be a non-negative unbiased estimator then the pseudo-marginal MH kernel admits an invariant distribution admitting $\pi(\theta)$ as a marginal.

Main Result

- **Proposition**: Let $\widehat{p}_{\vartheta}(y)$ be a non-negative unbiased estimator then the pseudo-marginal MH kernel admits an invariant distribution admitting $\pi(\theta)$ as a marginal.
- "Proof". Define $Z = \log \widehat{p}(y;\theta) / p(y;\theta)$ and an auxiliary target density on $\Theta \times \mathbb{R}$

$$\overline{\pi}(\theta, z) = \pi(\theta) \underbrace{\exp(z) g_{\theta}(z)}_{\text{unbiasedness} \Leftrightarrow \int (\cdot) dz = 1}$$

where $Z \sim g_{\theta}$; e.g. importance sampling or particle filter.

Main Result

- **Proposition**: Let $\widehat{p}_{\vartheta}(y)$ be a non-negative unbiased estimator then the pseudo-marginal MH kernel admits an invariant distribution admitting $\pi(\theta)$ as a marginal.
- "Proof". Define $Z = \log \widehat{p}(y; \theta) / p(y; \theta)$ and an auxiliary target density on $\Theta \times \mathbb{R}$

$$\overline{\pi}(\theta, z) = \pi(\theta) \underbrace{\exp(z) g_{\theta}(z)}_{\text{unbiasedness} \Leftrightarrow \int (\cdot) dz = 1}$$

where $Z \sim g_{\theta}$; e.g. importance sampling or particle filter.
- Pseudo marginal MH is MH of target $\overline{\pi}(\theta, z)$ and proposal $q(\theta, \vartheta) g_{\vartheta}(z)$ as

$$\frac{\overline{\pi}(\vartheta, Z)}{\overline{\pi}(\vartheta_{i-1}, Z_{i-1})} \frac{q(\vartheta_{i-1} | \vartheta) g_{\vartheta_{i-1}}(Z_{i-1})}{q(\vartheta | \vartheta_{i-1}) g_{\vartheta}(Z)} = \frac{\widehat{p}(y; \vartheta)}{\widehat{p}(y; \vartheta_{i-1})} \frac{p(\vartheta)}{p(\vartheta_{i-1})} \frac{q(\vartheta_{i-1} | \vartheta)}{q(\vartheta | \vartheta_{i-1})}.$$

A Nonlinear State-Space Model

- Standard non-linear model

$$X_t = \frac{1}{2}X_{t-1} + 25\frac{X_{t-1}}{1+X_{t-1}^2} + 8\cos(1.2t) + V_t, \quad V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_V^2\right),$$

$$Y_t = \frac{1}{20}X_t^2 + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_W^2\right).$$

A Nonlinear State-Space Model

- Standard non-linear model

$$X_t = \tfrac{1}{2}X_{t-1} + 25\tfrac{X_{t-1}}{1+X_{t-1}^2} + 8\cos(1.2t) + V_t, \quad V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_V^2\right),$$

$$Y_t = \tfrac{1}{20}X_t^2 + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_W^2\right).$$

- $T = 200$ data points with $\theta = \left(\sigma_V^2, \sigma_W^2\right) = (10, 10)$.

A Nonlinear State-Space Model

- Standard non-linear model

$$X_t = \tfrac{1}{2}X_{t-1} + 25\tfrac{X_{t-1}}{1+X_{t-1}^2} + 8\cos(1.2t) + V_t, \quad V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_V^2\right),$$

$$Y_t = \tfrac{1}{20}X_t^2 + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_W^2\right).$$

- $T = 200$ data points with $\theta = \left(\sigma_V^2, \sigma_W^2\right) = (10, 10)$.
- Difficult to perform standard MCMC as $p\left(x_{1:T} \middle| y_{1:T}, \theta\right)$ is highly multimodal.

A Nonlinear State-Space Model

- Standard non-linear model

$$X_t = \frac{1}{2}X_{t-1} + 25\frac{X_{t-1}}{1+X_{t-1}^2} + 8\cos(1.2t) + V_t, \quad V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_V^2\right),$$

$$Y_t = \frac{1}{20}X_t^2 + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_W^2\right).$$

- $T = 200$ data points with $\theta = \left(\sigma_V^2, \sigma_W^2\right) = (10, 10)$.
- Difficult to perform standard MCMC as $p\left(x_{1:T} \middle| y_{1:T}, \theta\right)$ is highly multimodal.
- We sample from $p\left(\theta \middle| y_{1:T}\right)$ using a random walk pseudo-marginal MH where $p_\theta\left(y_{1:T}\right)$ is estimated using SMC with N particles.

A Nonlinear State-Space Model

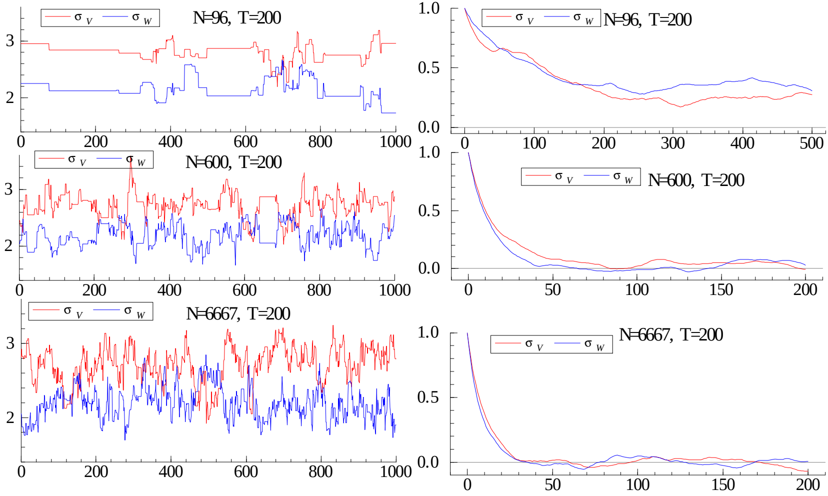


Figure: Autocorrelation of $\left\{\sigma_V^{(i)}\right\}$ and $\left\{\sigma_W^{(i)}\right\}$ of the MH sampler for various N.

Optimal Tuning of the Pseudo-marginal MH

- A key issue from a practical point of view is the selection of N.

Optimal Tuning of the Pseudo-marginal MH

- A key issue from a practical point of view is the selection of N.
- If N is too small, then the algorithm mixes poorly and will require many MCMC iterations.

Optimal Tuning of the Pseudo-marginal MH

- A key issue from a practical point of view is the selection of N.
- If N is too small, then the algorithm mixes poorly and will require many MCMC iterations.
- If N is too large, then each pseudo-marginal MH iteration is expensive.

Inefficiency of the Pseudo-marginal MH

- Consider the particle MH chain $\{\theta_i, Z_i\}$ of $\overline{\pi}-$invariant transition kernel Q

$$
\begin{aligned}
Q\{(\theta, z), (d\vartheta, dw)\} &= q(\vartheta|\theta)g_\vartheta(w)\min\{1, r(\theta,\vartheta)\exp(w-z)\}\,d\vartheta dw \\
&\quad + \left\{1 - \varrho_Q(\theta, z)\right\}\delta_{(\theta,z)}(d\vartheta, dw)
\end{aligned}
$$

where $r(\theta, \vartheta) = \pi(\vartheta)q(\theta|\vartheta)/\{\pi(\theta)q(\vartheta|\theta)\}$.

Inefficiency of the Pseudo-marginal MH

- Consider the particle MH chain $\{\theta_i, Z_i\}$ of $\overline{\pi}-$invariant transition kernel Q

$$Q\left\{(\theta, z), (d\vartheta, dw)\right\} = q(\vartheta|\theta)g_\vartheta(w)\min\left\{1, r(\theta, \vartheta)\exp(w - z)\right\} d\vartheta dw + \left\{1 - \varrho_Q(\theta, z)\right\}\delta_{(\theta,z)}(d\vartheta, dw)$$

where $r(\theta, \vartheta) = \pi(\vartheta)q(\theta|\vartheta)/\left\{\pi(\theta)q(\vartheta|\theta)\right\}$.

- **Proposition** (KV 1986). Let $h : \Theta \to \mathbb{R}$, $\pi(h) = \mathbb{E}_\pi[h(\theta)]$ and $\widehat{\pi}_n(h) = n^{-1}\sum_{i=1}^n h(\theta_i)$. If $\{\theta_i, Z_i\}$ is stationary and ergodic, $\mathbb{V}_\pi[h(\theta)] < \infty$ and $IF_h^Q(\sigma) = 1 + 2\sum_{\tau=1}^\infty corr_{\overline{\pi},Q}\left\{h(\theta_0), h(\theta_\tau)\right\} < \infty$ then

$$\sqrt{n}\left\{\widehat{\pi}_n(h) - \pi(h)\right\} \to \mathcal{N}\left(0, \mathbb{V}_\pi[h(\theta)]\ IF_h^Q(\sigma)\right).$$

Inefficiency of the Pseudo-marginal MH

- Consider the particle MH chain $\{\theta_i, Z_i\}$ of $\overline{\pi}-$invariant transition kernel Q

$$Q\left\{(\theta,z),(d\vartheta,dw)\right\} = q(\vartheta|\theta)g_\vartheta(w)\min\left\{1,r(\theta,\vartheta)\exp(w-z)\right\}d\vartheta dw + \left\{1-\varrho_Q(\theta,z)\right\}\delta_{(\theta,z)}(d\vartheta,dw)$$

  where $r(\theta,\vartheta) = \pi(\vartheta)q(\theta|\vartheta)/\left\{\pi(\theta)q(\vartheta|\theta)\right\}$.

- **Proposition** (KV 1986). Let $h:\Theta\rightarrow\mathbb{R}$, $\pi(h)=\mathbb{E}_\pi[h(\theta)]$ and $\widehat{\pi}_n(h)=n^{-1}\sum_{i=1}^n h(\theta_i)$. If $\{\theta_i, Z_i\}$ is stationary and ergodic, $\mathbb{V}_\pi[h(\theta)]<\infty$ and $IF_h^Q(\sigma)=1+2\sum_{\tau=1}^\infty corr_{\overline{\pi},Q}\left\{h(\theta_0),h(\theta_\tau)\right\}<\infty$ then

$$\sqrt{n}\left\{\widehat{\pi}_n(h)-\pi(h)\right\}\rightarrow\mathcal{N}\left(0,\mathbb{V}_\pi[h(\theta)]\ IF_h^Q(\sigma)\right).$$

- The **Integrated Autocorrelation Time** IF_h^Q is a measure of **inefficiency** of Q which **we want to minimize for a fixed computational budget**.

Aim of the Analysis

- **Simplifying Assumption**: The noise Z is independent of θ and Gaussian; i.e. $Z \sim \mathcal{N}\left(-\sigma^2/2; \sigma^2\right)$:

$$\overline{\pi}\left(\theta, z\right) = \pi\left(\theta\right) \underbrace{\exp\left(z\right) g_\sigma\left(z\right)}_{\pi_{Z,\sigma}(z)} = \pi\left(\theta\right) \mathcal{N}\left(z; \sigma^2/2; \sigma^2\right).$$

Aim of the Analysis

- **Simplifying Assumption**: The noise Z is independent of θ and Gaussian; i.e. $Z \sim \mathcal{N}\left(-\sigma^2/2; \sigma^2\right)$:

$$\overline{\pi}\left(\theta, z\right) = \pi\left(\theta\right) \underbrace{\exp\left(z\right) g_{\sigma}\left(z\right)}_{\pi_{Z,\sigma}(z)} = \pi\left(\theta\right) \mathcal{N}\left(z; \sigma^2/2; \sigma^2\right).$$

- **Aim**: Minimize the computational cost

$$CT_h^Q\left(\sigma\right) = IF_h^Q\left(\sigma\right)/\sigma^2$$

as $\sigma^2 \propto 1/N$ and computational efforts proportional to N.

Aim of the Analysis

- **Simplifying Assumption**: The noise Z is independent of θ and Gaussian; i.e. $Z \sim \mathcal{N}\left(-\sigma^2/2; \sigma^2\right)$:

$$\overline{\pi}\left(\theta, z\right) = \pi\left(\theta\right) \underbrace{\exp\left(z\right) g_\sigma\left(z\right)}_{\pi_{Z,\sigma}(z)} = \pi\left(\theta\right) \mathcal{N}\left(z; \sigma^2/2; \sigma^2\right).$$

- **Aim**: Minimize the computational cost

$$CT_h^Q\left(\sigma\right) = IF_h^Q\left(\sigma\right)/\sigma^2$$

as $\sigma^2 \propto 1/N$ and computational efforts proportional to N.

- **Special cases**:

Aim of the Analysis

- **Simplifying Assumption**: The noise Z is independent of θ and Gaussian; i.e. $Z \sim \mathcal{N}\left(-\sigma^2/2; \sigma^2\right)$:

$$\overline{\pi}\left(\theta, z\right) = \pi\left(\theta\right) \underbrace{\exp\left(z\right) g_{\sigma}\left(z\right)}_{\pi_{Z,\sigma}(z)} = \pi\left(\theta\right) \mathcal{N}\left(z; \sigma^2/2; \sigma^2\right).$$

- **Aim**: Minimize the computational cost

$$CT_h^Q\left(\sigma\right) = IF_h^Q\left(\sigma\right)/\sigma^2$$

as $\sigma^2 \propto 1/N$ and computational efforts proportional to N.

- **Special cases**:

1. When $q(\vartheta|\theta) = p\left(\vartheta|\, y\right)$, $\sigma_{\text{opt}} = 0.92$ (Pitt et al., 2012).

Aim of the Analysis

- **Simplifying Assumption**: The noise Z is independent of θ and Gaussian; i.e. $Z \sim \mathcal{N}\left(-\sigma^2/2; \sigma^2\right)$:

$$\overline{\pi}\left(\theta, z\right) = \pi\left(\theta\right) \underbrace{\exp\left(z\right) g_\sigma\left(z\right)}_{\pi_{Z,\sigma}\left(z\right)} = \pi\left(\theta\right) \mathcal{N}\left(z; \sigma^2/2; \sigma^2\right).$$

- **Aim**: Minimize the computational cost

$$CT_h^Q\left(\sigma\right) = IF_h^Q\left(\sigma\right)/\sigma^2$$

as $\sigma^2 \propto 1/N$ and computational efforts proportional to N.

- **Special cases**:

1. When $q(\vartheta|\theta) = p\left(\vartheta|\, y\right)$, $\sigma_{\text{opt}} = 0.92$ (Pitt et al., 2012).
2. When $\pi\left(\theta\right) = \prod_{i=1}^{d} f\left(\theta_i\right)$ and $q(\vartheta|\theta)$ is an isotropic Gaussian random walk then, as $d \to \infty$, $\sigma_{\text{opt}} = 1.81$ (Sherlock, Thiery, Roberts & Rosenthal, 2014).

Sketch of the Analysis

- For general proposals and targets, direct minimization of $CT_h^Q(\sigma) = IF_h^Q(\sigma)/\sigma^2$ impossible so minimize an upper bound over it.

Sketch of the Analysis

- For general proposals and targets, direct minimization of $CT_h^Q(\sigma) = IF_h^Q(\sigma)/\sigma^2$ impossible so minimize an upper bound over it.
- We introduce an auxiliary $\overline{\pi}-$invariant kernel

$$\begin{aligned} Q^*\{(\theta, z), (d\vartheta, dw)\} &= q(\vartheta|\theta) g_\sigma(w) \alpha_{\mathrm{EX}}(\theta, \vartheta)\, \alpha_{\mathrm{Z}}(z, w)\, d\vartheta dw \\ &\quad + \left\{1 - \varrho_{\mathrm{EX}}(\theta)\, \varrho_{\mathrm{Z},\sigma}(z)\right\} \delta_{(\theta,z)}(d\vartheta, dw) \end{aligned}$$

where

$$\alpha_{\mathrm{EX}}(\theta, \vartheta) = \min\{1, r(\theta, \vartheta)\}, \quad \alpha_{\mathrm{Z}}(z, w) = \min\{1, \exp(w - z)\}$$

Sketch of the Analysis

- For general proposals and targets, direct minimization of $CT_h^Q(\sigma) = IF_h^Q(\sigma)/\sigma^2$ impossible so minimize an upper bound over it.
- We introduce an auxiliary $\overline{\pi}-$invariant kernel

$$\begin{aligned} Q^*\left\{(\theta, z), (d\vartheta, dw)\right\} &= q(\vartheta|\theta) g_\sigma(w) \alpha_{\mathrm{EX}}(\theta, \vartheta) \alpha_{\mathrm{Z}}(z, w)\, d\vartheta dw \\ &\quad + \left\{1 - \varrho_{\mathrm{EX}}(\theta) \varrho_{\mathrm{Z},\sigma}(z)\right\} \delta_{(\theta, z)}(d\vartheta, dw) \end{aligned}$$

  where

$$\alpha_{\mathrm{EX}}(\theta, \vartheta) = \min\left\{1, r(\theta, \vartheta)\right\}, \quad \alpha_{\mathrm{Z}}(z, w) = \min\left\{1, \exp(w - z)\right\}$$

- Peskun's theorem (1973) guarantees that $IF_h^Q(\sigma) \le IF_h^{Q^*}(\sigma)$ so that $CT_h^Q(\sigma) \le CT_h^{Q^*}(\sigma)$.

Sketch of the Analysis

- Let $(\theta_i, Z_i)_{i \geq 1}$ be generated by Q^*.

Sketch of the Analysis

- Let $(\theta_i, Z_i)_{i\geq 1}$ be generated by Q^*.
- Denote $\left(\widetilde{\theta}_i, \widetilde{Z}_i\right)_{i\geq 1}$ the accepted proposals and $(\tau_i)_{i\geq 1}$ the associated sojourn times; i.e. $\left(\widetilde{\theta}_1, \widetilde{Z}_1\right) = (\theta_1, Z_1) = \cdots = (\theta_{\tau_1}, Z_{\tau_1})$, $\left(\widetilde{\theta}_2, \widetilde{Z}_2\right) = (\theta_{\tau_{1+1}}, Z_{\tau_1+1}) = \cdots = (\theta_{\tau_2}, Z_{\tau_2})$ and so on where $\left(\widetilde{\theta}_{i+1}, \widetilde{Z}_{i+1}\right) \neq \left(\widetilde{\theta}_i, \widetilde{Z}_i\right)$ a.s.

Sketch of the Analysis

- Let $(\theta_i, Z_i)_{i\geq 1}$ be generated by Q^*.
- Denote $\left(\widetilde{\theta}_i, \widetilde{Z}_i\right)_{i\geq 1}$ the accepted proposals and $(\tau_i)_{i\geq 1}$ the associated sojourn times; i.e. $\left(\widetilde{\theta}_1, \widetilde{Z}_1\right) = (\theta_1, Z_1) = \cdots = (\theta_{\tau_1}, Z_{\tau_1})$, $\left(\widetilde{\theta}_2, \widetilde{Z}_2\right) = (\theta_{\tau_{1+1}}, Z_{\tau_1+1}) = \cdots = (\theta_{\tau_2}, Z_{\tau_2})$ and so on where $\left(\widetilde{\theta}_{i+1}, \widetilde{Z}_{i+1}\right) \neq \left(\widetilde{\theta}_i, \widetilde{Z}_i\right)$ a.s.
- $IF_h^{Q^*}(\sigma)$ can be re-expressed in terms of $IF_{h/(\varrho_{\mathrm{EX}}\varrho_{\mathrm{Z}})}^{\widetilde{Q}^*}(\sigma)$ where

$$
\begin{aligned}
\widetilde{Q}^*\left\{(\theta, z), (d\vartheta, dw)\right\} &= \widetilde{Q}^{\mathrm{EX}}(\theta, d\vartheta)\, \widetilde{Q}_\sigma^{\mathrm{Z}}(z, dw) \\
&= \frac{q(d\vartheta|\theta)\alpha_{\mathrm{EX}}(\theta, \vartheta)}{\varrho_{\mathrm{EX}}(\theta)} \frac{g_\sigma(dw)\alpha_{\mathrm{Z}}(z, w)}{\varrho_{\mathrm{Z},\sigma}(z)}
\end{aligned}
$$

Main Result

- **Proposition**: Under weak assumptions, we have $IF_h^{Q}(\sigma) \leq IF_h^{Q^*}(\sigma)$ where

$$
\begin{aligned}
IF_h^{Q^*}(\sigma) \;=\;& 2\frac{\left\{1 + IF_h^{\mathrm{EX}}\right\}}{1 + IF_{h/\varrho_{\mathrm{EX}}}^{\widetilde{Q}^{\mathrm{EX}}}}\left\{\pi_{\mathrm{Z},\sigma}\left(1/\varrho_{\mathrm{Z},\sigma}\right) - 1/\pi_{\mathrm{Z},\sigma}\left(\varrho_{\mathrm{Z},\sigma}\right)\right\} \\
& \times \sum_{n=0}^{\infty} \phi_n\left(h/\varrho_{\mathrm{EX}}, \widetilde{Q}^{\mathrm{EX}}\right)\phi_n\left(1/\varrho_{\mathrm{Z}}, \widetilde{Q}_{\sigma}^{\mathrm{Z}}\right) \\
& + \frac{1 + IF_h^{\mathrm{EX}}}{\pi_{\mathrm{Z},\sigma}\left(\varrho_{\mathrm{Z},\sigma}\right)} - 1,
\end{aligned}
$$

where $\phi_n(\varphi, P)$ denotes the autocorrelation at lag n under a Markov kernel P.

Main Result

- **Proposition**: Under weak assumptions, we have $IF_h^{Q}(\sigma) \leq IF_h^{Q^*}(\sigma)$ where

$$
\begin{aligned}
IF_h^{Q^*}(\sigma) &= 2\frac{\left\{1+IF_h^{\mathrm{EX}}\right\}}{1+IF_{h/\varrho_{\mathrm{EX}}}^{\widetilde{Q}^{\mathrm{EX}}}}\left\{\pi_{\mathrm{Z},\sigma}\left(1/\varrho_{\mathrm{Z},\sigma}\right)-1/\pi_{\mathrm{Z},\sigma}\left(\varrho_{\mathrm{Z},\sigma}\right)\right\} \\
&\times\sum_{n=0}^{\infty}\phi_n\left(h/\varrho_{\mathrm{EX}},\widetilde{Q}^{\mathrm{EX}}\right)\phi_n\left(1/\varrho_{\mathrm{Z}},\widetilde{Q}_{\sigma}^{\mathrm{Z}}\right) \\
&+\frac{1+IF_h^{\mathrm{EX}}}{\pi_{\mathrm{Z},\sigma}\left(\varrho_{\mathrm{Z},\sigma}\right)}-1,
\end{aligned}
$$

  where $\phi_n(\varphi, P)$ denotes the autocorrelation at lag n under a Markov kernel P.

- This identity allows us to "decouple" the influence of the parameter and of the noise on $IF_h^{Q^*}(\sigma)$.

Simpler Bounds on the Relative Inefficiency

- If $IF_{h/\varrho_{\mathrm{EX}}}^{\widetilde{Q}^{\mathrm{EX}}} \geq 1$, e.g. $\widetilde{Q}^{\mathrm{EX}}$ is a positive kernel, then

$$\frac{IF_h^{Q}(\sigma)}{IF_h^{\mathrm{EX}}} \leq \frac{IF_h^{Q^*}(\sigma)}{IF_h^{\mathrm{EX}}} \leq \frac{1}{2}\left(1 + \frac{1}{IF_h^{\mathrm{EX}}}\right)\pi_{\mathsf{Z},\sigma}\left(1/\varrho_{\mathsf{Z},\sigma}\right) - \frac{1}{IF_h^{\mathrm{EX}}}$$

and the bound is tight as $IF_h^{\mathrm{EX}} \to 1$ or $\sigma \to 0$.

Simpler Bounds on the Relative Inefficiency

- If $IF_{h/\varrho_{\mathrm{EX}}}^{\widetilde{Q}^{\mathrm{EX}}} \geq 1$, e.g. $\widetilde{Q}^{\mathrm{EX}}$ is a positive kernel, then

$$\frac{IF_h^Q(\sigma)}{IF_h^{\mathrm{EX}}} \leq \frac{IF_h^{Q^*}(\sigma)}{IF_h^{\mathrm{EX}}} \leq \frac{1}{2}\left(1 + \frac{1}{IF_h^{\mathrm{EX}}}\right)\pi_{Z,\sigma}\left(1/\varrho_{Z,\sigma}\right) - \frac{1}{IF_h^{\mathrm{EX}}}$$

  and the bound is tight as $IF_h^{\mathrm{EX}} \to 1$ or $\sigma \to 0$.
- As $IF_{J,h/\varrho_{\mathrm{EX}}}^{\mathrm{EX}} \to \infty$,

$$\frac{IF_h^{Q^*}(\sigma)}{IF_h^{\mathrm{EX}}} \to \frac{1}{\pi_{Z,\sigma}\left(\varrho_{Z,\sigma}\right)}.$$

Simpler Bounds on the Relative Inefficiency

- If $IF_{h/\varrho_{\text{EX}}}^{\widetilde{Q}^{\text{EX}}} \geq 1$, e.g. $\widetilde{Q}^{\text{EX}}$ is a positive kernel, then

$$\frac{IF_h^Q(\sigma)}{IF_h^{\text{EX}}} \leq \frac{IF_h^{Q^*}(\sigma)}{IF_h^{\text{EX}}} \leq \frac{1}{2}\left(1 + \frac{1}{IF_h^{\text{EX}}}\right)\pi_{Z,\sigma}\left(1/\varrho_{Z,\sigma}\right) - \frac{1}{IF_h^{\text{EX}}}$$

  and the bound is tight as $IF_h^{\text{EX}} \to 1$ or $\sigma \to 0$.

- As $IF_{J,h/\varrho_{\text{EX}}}^{\text{EX}} \to \infty$,

$$\frac{IF_h^{Q^*}(\sigma)}{IF_h^{\text{EX}}} \to \frac{1}{\pi_{Z,\sigma}\left(\varrho_{Z,\sigma}\right)}.$$

- Results used to minimize w.r.t σ upper bounds on $CT_h^Q(\sigma) = IF_h^Q(\sigma)/\sigma^2$.

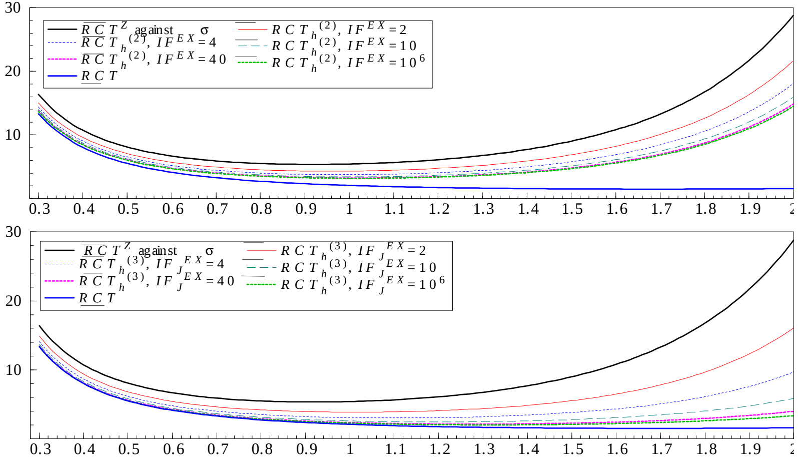


Figure: Bounds on $IF_h^Q(\sigma) \,/\, \left(\sigma^2 IF_h^{\mathrm{EX}}\right)$

Practical Guidelines

- For good proposals, select $\sigma \approx 1$ whereas for poor proposals, select $\sigma \approx 1.7$.

Practical Guidelines

- For good proposals, select $\sigma \approx 1$ whereas for poor proposals, select $\sigma \approx 1.7$.
- When you have no clue about the proposal efficiency,

Practical Guidelines

- For good proposals, select $\sigma \approx 1$ whereas for poor proposals, select $\sigma \approx 1.7$.
- When you have no clue about the proposal efficiency,

1. If $\sigma_{\text{opt}} = 1$ and you pick $\sigma = 1.7$, computing time increases by $\approx$150%.

Practical Guidelines

- For good proposals, select $\sigma \approx 1$ whereas for poor proposals, select $\sigma \approx 1.7$.
- When you have no clue about the proposal efficiency,

1. If $\sigma_{\text{opt}} = 1$ and you pick $\sigma = 1.7$, computing time increases by $\approx$150%.
2. If $\sigma_{\text{opt}} = 1.7$ and you pick $\sigma = 1$, computing time increases by $\approx$50%.

Practical Guidelines

- For good proposals, select $\sigma \approx 1$ whereas for poor proposals, select $\sigma \approx 1.7$.
- When you have no clue about the proposal efficiency,

1. If $\sigma_{\text{opt}} = 1$ and you pick $\sigma = 1.7$, computing time increases by $\approx$150%.
2. If $\sigma_{\text{opt}} = 1.7$ and you pick $\sigma = 1$, computing time increases by $\approx$50%.
3. If $\sigma_{\text{opt}} = 1$ or $\sigma_{\text{opt}} = 1.7$ and you pick $\sigma = 1.2$, computing time increases by $\approx$15%.

Example: Noisy Autoregressive Example

- Consider

$$
\begin{aligned}
X_t = &\quad \mu(1-\phi) + \phi X_t + V_t, \quad V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_\eta^2\right), \\
Y_t = &\quad X_t + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_\varepsilon^2\right),
\end{aligned}
$$

where $\theta = \left(\phi, \mu, \sigma_\eta^2\right)$.

Example: Noisy Autoregressive Example

- Consider

$$X_t = \mu(1-\phi) + \phi X_t + V_t, \quad V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_\eta^2\right),$$

$$Y_t = X_t + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_\varepsilon^2\right),$$

where $\theta = \left(\phi, \mu, \sigma_\eta^2\right)$.

- Likelihood can be computed exactly using Kalman.

Example: Noisy Autoregressive Example

- Consider

$$X_t = \mu(1-\phi) + \phi X_t + V_t, \quad V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_\eta^2\right),$$

$$Y_t = X_t + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_\varepsilon^2\right),$$

  where $\theta = \left(\phi, \mu, \sigma_\eta^2\right)$.
- Likelihood can be computed exactly using Kalman.
- Autoregressive Metropolis proposal of coefficient ρ for ϑ based on multivariate t-distribution.

Example: Noisy Autoregressive Example

- Consider

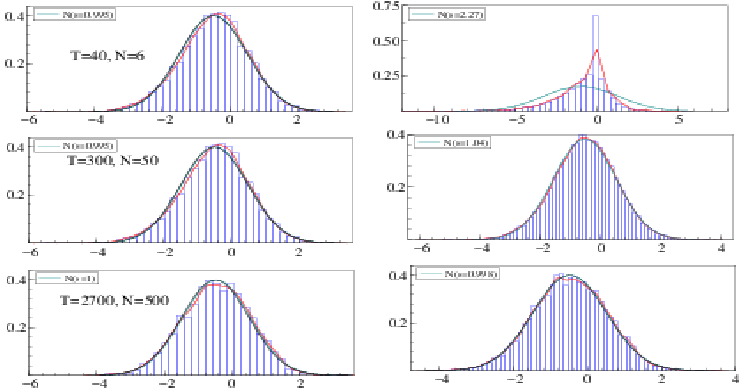
$$X_t = \mu(1-\phi) + \phi X_t + V_t, \quad V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_\eta^2\right),$$

$$Y_t = X_t + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_\varepsilon^2\right),$$

  where $\theta = \left(\phi, \mu, \sigma_\eta^2\right)$.

- Likelihood can be computed exactly using Kalman.
- Autoregressive Metropolis proposal of coefficient ρ for ϑ based on multivariate t-distribution.
- N is selected so as to obtain $\sigma\left(\overline{\theta}\right) \approx$constant where $\overline{\theta}$ posterior mean.

Empirical vs Asymptotic Distribution of Log-Likelihood Estimator



Empirical distribution of Z at posterior mean (left) and marginalized over samples from $\pi q(\vartheta) = \int \pi(\theta)\, q(\theta, \vartheta)\, d\theta$.

Relative Inefficiency and Computing Time

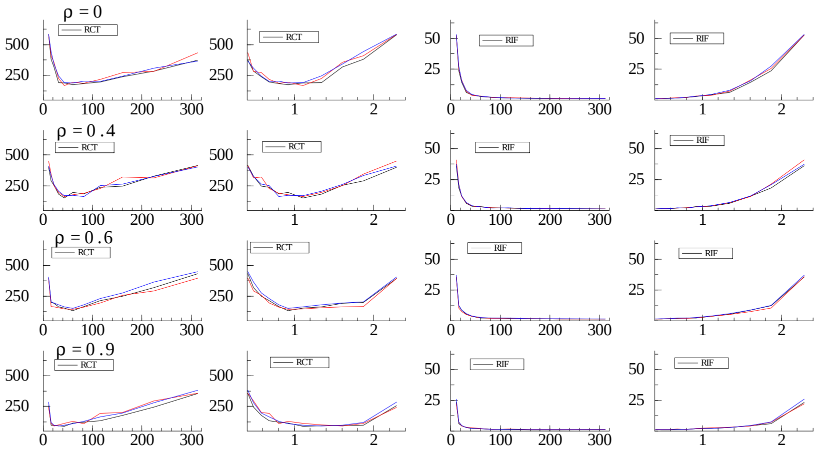


Figure: From left to right: RCT_h^Q vs N, RCT_h^Q vs $\sigma(\overline{\theta})$, RIF_h^Q against N and RIF_h^Q against $\sigma(\overline{\theta})$ for various values of ρ and different parameters.

Discussion

- Simplified quantitative analysis of the particle MH algorithm.

Discussion

- Simplified quantitative analysis of the particle MH algorithm.
- Particle MH scales roughly in $O\left(T^2\right)$.

Discussion

- Simplified quantitative analysis of the particle MH algorithm.
- Particle MH scales roughly in $O\left(T^2\right)$.
- Particle Gibbs sampling displays better theoretical properties: scaling?