Pseudo-marginal MCMC methods for inference in latent variable models

Arnaud Doucet
Department of Statistics, Oxford University
Joint work with George Deligiannidis (Oxford) & Mike Pitt (Kings)

07/07/2016

Organization of the talk

- Latent variable models

Organization of the talk

- Latent variable models
- The pseudo-marginal method

Organization of the talk

- Latent variable models
- The pseudo-marginal method
- Optimal tuning

Organization of the talk

- Latent variable models
- The pseudo-marginal method
- Optimal tuning
- The correlated pseudo-marginal method

Organization of the talk

- Latent variable models
- The pseudo-marginal method
- Optimal tuning
- The correlated pseudo-marginal method
- Illustrations

Latent Variable Models

- Assume

$$X_t \overset{\text{i.i.d.}}{\sim} \mu_\theta\left(\cdot\right), \quad Y_t | \left(X_t = x\right) \sim g_\theta\left(\cdot | x\right) \text{ for } t = 1, ..., T$$

where $\left(X_t\right)_{t\geq 1}$ are latent variables and $\left(Y_t\right)_{t\geq 1}$ correspond to observations.

Latent Variable Models

- Assume

$$X_t \overset{\text{i.i.d.}}{\sim} \mu_\theta(\cdot), \quad Y_t | (X_t = x) \sim g_\theta(\cdot | x) \text{ for } t = 1, \ldots, T$$

where $(X_t)_{t \geq 1}$ are latent variables and $(Y_t)_{t \geq 1}$ correspond to observations.

- The likelihood of $Y_{1:T} = y_{1:T}$ for parameter $\theta \in \mathbb{R}^d$ is

$$p_\theta(y_{1:T}) = \prod_{t=1}^{T} p_\theta(y_t), \text{ where } p_\theta(y_t) = \int \mu_\theta(x_t)\, g_\theta(y_t | x_t)\, \mathrm{d}x_t.$$

Latent Variable Models

- Assume

$$X_t \overset{\text{i.i.d.}}{\sim} \mu_\theta(\cdot), \quad Y_t | (X_t = x) \sim g_\theta(\cdot | x) \text{ for } t = 1, \ldots, T$$

where $(X_t)_{t \geq 1}$ are latent variables and $(Y_t)_{t \geq 1}$ correspond to observations.

- The likelihood of $Y_{1:T} = y_{1:T}$ for parameter $\theta \in \mathbb{R}^d$ is

$$p_\theta(y_{1:T}) = \prod_{t=1}^{T} p_\theta(y_t), \text{ where } p_\theta(y_t) = \int \mu_\theta(x_t)\, g_\theta(y_t | x_t)\, \mathrm{d}x_t.$$

- In many scenarios, $p_\theta(y_{1:T})$ cannot be evaluated exactly.

Example: Multivariate Probit model

- Multivariate latent Gaussian variables

$$X_t = Z_t\beta + \varepsilon_t\,, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, R\right).$$

Example: Multivariate Probit model

- Multivariate latent Gaussian variables

$$X_t = Z_t\beta + \varepsilon_t\,, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, R\right).$$

- Multivariate binary observations

$$Y_{ti} = \mathbb{I}\left(X_{ti} \geq 0\right), \quad i = 1, ..., n$$

Example: Multivariate Probit model

- Multivariate latent Gaussian variables

$$X_t = Z_t \beta + \varepsilon_t\,, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, R\right).$$

- Multivariate binary observations

$$Y_{ti} = \mathbb{I}\left(X_{ti} \geq 0\right), \quad i = 1, ..., n$$

- Likelihood of (β, R) is the product of T integrals of n-dimensional truncated multivariate normals.

State-Space Models

- Assume $\{X_t\}_{t\geq 1}$ is a latent Markov process, i.e. $X_1 \sim \mu_\theta(\cdot)$ and

$$X_{t+1} | (X_t = x) \sim f_\theta(\cdot | x), \quad Y_t | (X_t = x) \sim g_\theta(\cdot | x).$$

State-Space Models

- Assume $\{X_t\}_{t\geq 1}$ is a latent Markov process, i.e. $X_1 \sim \mu_\theta(\cdot)$ and

$$X_{t+1} | (X_t = x) \sim f_\theta(\cdot | x), \quad Y_t | (X_t = x) \sim g_\theta(\cdot | x).$$

- The likelihood of observations $Y_{1:T} = y_{1:T}$ is

$$p_\theta(y_{1:T}) = \int p_\theta(x_{1:T}, y_{1:T}) \mathrm{d}x_{1:T}$$

where

$$p_\theta(x_{1:T}, y_{1:T}) = \mu_\theta(x_1) g_\theta(y_1 | x_1) \prod_{t=2}^{T} f_\theta(x_t | x_{t-1}) g_\theta(y_t | x_t).$$

State-Space Models

- Assume $\{X_t\}_{t\geq 1}$ is a latent Markov process, i.e. $X_1 \sim \mu_\theta(\cdot)$ and

$$X_{t+1} | (X_t = x) \sim f_\theta(\cdot | x), \quad Y_t | (X_t = x) \sim g_\theta(\cdot | x).$$

- The likelihood of observations $Y_{1:T} = y_{1:T}$ is

$$p_\theta(y_{1:T}) = \int p_\theta(x_{1:T}, y_{1:T}) \mathrm{d}x_{1:T}$$

where

$$p_\theta(x_{1:T}, y_{1:T}) = \mu_\theta(x_1) g_\theta(y_1 | x_1) \prod_{t=2}^{T} f_\theta(x_t | x_{t-1}) g_\theta(y_t | x_t).$$

- State-space models are ubiquitous in time series analysis but inference is difficult as $p_\theta(y_{1:T})$ is intractable for non-linear/non-Gaussian models.

- Two species X_s^1 (prey) and X_s^2 (predator)

$$\Pr\left(X_{s+ds}^1 = x_s^1 + 1, X_{s+ds}^2 = x_s^2 \,\middle|\, x_s^1, x_s^2\right) = \alpha\, x_s^1\, ds + o\left(ds\right),$$
$$\Pr\left(X_{s+ds}^1 = x_s^1 - 1, X_{s+ds}^2 = x_s^2 + 1 \,\middle|\, x_s^1, x_s^2\right) = \beta\, x_s^1\, x_s^2\, ds + o\left(ds\right),$$
$$\Pr\left(X_{s+ds}^1 = x_t^1, X_{s+ds}^2 = x_s^2 - 1 \,\middle|\, x_s^1, x_s^2\right) = \gamma\, x_s^2\, ds + o\left(ds\right),$$

observed at discrete times

$$Y_t = X_{\Delta t}^1 + W_t \text{ with } W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right).$$

Stochastic kinetic model - Lotka-Volterra

- Two species X_s^1 (prey) and X_s^2 (predator)

$$\begin{aligned}
\Pr\left(X_{s+ds}^1 = x_s^1+1, X_{s+ds}^2 = x_s^2 \,\middle|\, x_s^1, x_s^2\right) &= \alpha\, x_s^1\, ds + o\left(ds\right), \\
\Pr\left(X_{s+ds}^1 = x_s^1-1, X_{s+ds}^2 = x_s^2+1 \,\middle|\, x_s^1, x_s^2\right) &= \beta\, x_s^1\, x_s^2\, ds + o\left(ds\right), \\
\Pr\left(X_{s+ds}^1 = x_t^1, X_{s+ds}^2 = x_s^2-1 \,\middle|\, x_s^1, x_s^2\right) &= \gamma\, x_s^2\, ds + o\left(ds\right),
\end{aligned}$$

observed at discrete times

$$Y_t = X_{\Delta t}^1 + W_t \text{ with } W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right).$$

- Kinetic rate constants $\theta = (\alpha, \beta, \gamma)$.

Other Applications

- State-space models are ubiquitous: 16,700 hits on Google Scholar since January 2015.

Other Applications

- State-space models are ubiquitous: 16,700 hits on Google Scholar since January 2015.
- **Econometrics**: stochastic volatility models.

Other Applications

- State-space models are ubiquitous: 16,700 hits on Google Scholar since January 2015.
- **Econometrics**: stochastic volatility models.
- **Epidemiology**: disease dynamic models.

Other Applications

- State-space models are ubiquitous: 16,700 hits on Google Scholar since January 2015.
- **Econometrics**: stochastic volatility models.
- **Epidemiology**: disease dynamic models.
- **Ecology**: population dynamics.

Other Applications

- State-space models are ubiquitous: 16,700 hits on Google Scholar since January 2015.
- **Econometrics**: stochastic volatility models.
- **Epidemiology**: disease dynamic models.
- **Ecology**: population dynamics.
- **Environmentrics**: phytoplankton-zooplankton model, paleoclimate reconstruction.

Other Applications

- State-space models are ubiquitous: 16,700 hits on Google Scholar since January 2015.
- **Econometrics**: stochastic volatility models.
- **Epidemiology**: disease dynamic models.
- **Ecology**: population dynamics.
- **Environmentrics**: phytoplankton-zooplankton model, paleoclimate reconstruction.
- **Macroeconomics**: dynamic generalized stochastic equilibrium.

Other Applications

- State-space models are ubiquitous: 16,700 hits on Google Scholar since January 2015.
- **Econometrics**: stochastic volatility models.
- **Epidemiology**: disease dynamic models.
- **Ecology**: population dynamics.
- **Environmentrics**: phytoplankton-zooplankton model, paleoclimate reconstruction.
- **Macroeconomics**: dynamic generalized stochastic equilibrium.
- **Signal Processing**: target tracking.

Other Applications

- State-space models are ubiquitous: 16,700 hits on Google Scholar since January 2015.
- **Econometrics**: stochastic volatility models.
- **Epidemiology**: disease dynamic models.
- **Ecology**: population dynamics.
- **Environmentrics**: phytoplankton-zooplankton model, paleoclimate reconstruction.
- **Macroeconomics**: dynamic generalized stochastic equilibrium.
- **Signal Processing**: target tracking.
- **Systems biology**: stochastic kinetic models.

- Prior distribution of density $p(\theta)$.

Bayesian Inference for Latent Variable Models

- Prior distribution of density $p(\theta)$.
- Likelihood function $p_\theta(y_{1:T})$.

Bayesian Inference for Latent Variable Models

- Prior distribution of density $p(\theta)$.
- Likelihood function $p_\theta(y_{1:T})$.
- Bayesian inference relies on the posterior

$$\pi(\theta) = p(\theta | y_{1:T}) = \frac{p_\theta(y_{1:T})\, p(\theta)}{\int_\Theta p_{\theta'}(y_{1:T})\, p(\theta')\, d\theta'}.$$

Bayesian Inference for Latent Variable Models

- Prior distribution of density $p(\theta)$.
- Likelihood function $p_\theta(y_{1:T})$.
- Bayesian inference relies on the posterior

$$\pi(\theta) = p(\theta|\, y_{1:T}) = \frac{p_\theta(y_{1:T})\, p(\theta)}{\int_\Theta p_{\theta'}(y_{1:T})\, p(\theta')\, \mathrm{d}\theta'}.$$

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

Standard MCMC Approaches

- Standard MCMC schemes target $p\left(\theta, x_{1:T} \mid y_{1:T}\right)$ where

$$p\left(\theta, x_{1:T} \mid y_{1:T}\right) \propto p\left(\theta\right) p_{\theta}\left(x_{1:T}, y_{1:T}\right)$$

  using Gibbs type strategy; i.e. sample alternately $X_{1:T} \sim p_{\theta}\left(\cdot \mid y_{1:T}\right)$ and $\theta \sim p\left(\cdot \mid y_{1:T}, X_{1:T}\right)$.

Standard MCMC Approaches

- Standard MCMC schemes target $p\left(\theta, x_{1:T} \mid y_{1:T}\right)$ where

$$p\left(\theta, x_{1:T} \mid y_{1:T}\right) \propto p\left(\theta\right) p_{\theta}\left(x_{1:T}, y_{1:T}\right)$$

  using Gibbs type strategy; i.e. sample alternately $X_{1:T} \sim p_{\theta}\left(\cdot \mid y_{1:T}\right)$ and $\theta \sim p\left(\cdot \mid y_{1:T}, X_{1:T}\right)$.
- **Problem** 1: it can be difficult to sample $p_{\theta}\left(x_{1:T} \mid y_{1:T}\right)$; e.g. state-space models.

Standard MCMC Approaches

- Standard MCMC schemes target $p\left(\theta, x_{1:T} \mid y_{1:T}\right)$ where

$$p\left(\theta, x_{1:T} \mid y_{1:T}\right) \propto p\left(\theta\right) p_{\theta}\left(x_{1:T}, y_{1:T}\right)$$

  using Gibbs type strategy; i.e. sample alternately $X_{1:T} \sim p_{\theta}\left(\cdot \mid y_{1:T}\right)$ and $\theta \sim p\left(\cdot \mid y_{1:T}, X_{1:T}\right)$.
- **Problem** 1: it can be difficult to sample $p_{\theta}\left(x_{1:T} \mid y_{1:T}\right)$; e.g. state-space models.
- **Problem** 2: Even when it is implementable, Gibbs can converge very slowly.

Standard MCMC Approaches

- Standard MCMC schemes target $p\left(\theta, x_{1:T} \middle| y_{1:T}\right)$ where

$$p\left(\theta, x_{1:T} \middle| y_{1:T}\right) \propto p\left(\theta\right) p_{\theta}\left(x_{1:T}, y_{1:T}\right)$$

  using Gibbs type strategy; i.e. sample alternately $X_{1:T} \sim p_{\theta}\left(\cdot \middle| y_{1:T}\right)$ and $\theta \sim p\left(\cdot \middle| y_{1:T}, X_{1:T}\right)$.
- **Problem** 1: it can be difficult to sample $p_{\theta}\left(x_{1:T} \middle| y_{1:T}\right)$; e.g. state-space models.
- **Problem** 2: Even when it is implementable, Gibbs can converge very slowly.
- Pseudo-marginal methods mimick an algorithm targetting directly $p\left(\theta \middle| y_{1:T}\right)$ instead of $p\left(\theta, x_{1:T} \middle| y_{1:T}\right)$.

Ideal Marginal Metropolis-Hastings algorithm

- Metropolis–Hastings (MH) algorithm simulates an ergodic Markov chain $\{\vartheta_i\}_{i\geq 1}$ of limiting distribution $\pi(\theta)$.

Ideal Marginal Metropolis-Hastings algorithm

- Metropolis–Hastings (MH) algorithm simulates an ergodic Markov chain $\{\vartheta_i\}_{i\geq 1}$ of limiting distribution $\pi\left(\theta\right)$.

Ideal Marginal Metropolis-Hastings algorithm

- Metropolis–Hastings (MH) algorithm simulates an ergodic Markov chain $\{\vartheta_i\}_{i\geq 1}$ of limiting distribution $\pi(\theta)$.

At iteration i

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

Ideal Marginal Metropolis-Hastings algorithm

- Metropolis–Hastings (MH) algorithm simulates an ergodic Markov chain $\{\vartheta_i\}_{i\geq 1}$ of limiting distribution $\pi(\theta)$.

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

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

$$\min\left\{1, \frac{\pi(\vartheta)}{\pi(\vartheta_{i-1})}\frac{q(\vartheta_{i-1}|\,\vartheta)}{q(\vartheta|\,\vartheta_{i-1})}\right\} = \min\left\{1, \frac{p_{\vartheta}(y_{1:T})\,p(\vartheta)}{p_{\vartheta_{i-1}}(y_{1:T})\,p(\vartheta_{i-1})}\frac{q(\vartheta_{i-1}|\,\vartheta)}{q(\vartheta|\,\vartheta_{i-1})}\right\},$$

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

Ideal Marginal Metropolis-Hastings algorithm

- Metropolis–Hastings (MH) algorithm simulates an ergodic Markov chain $\{\vartheta_i\}_{i\geq 1}$ of limiting distribution $\pi(\theta)$.

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

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

$$\min\left\{1, \frac{\pi(\vartheta)}{\pi(\vartheta_{i-1})}\frac{q(\vartheta_{i-1}|\,\vartheta)}{q(\vartheta|\,\vartheta_{i-1})}\right\} = \min\left\{1, \frac{p_{\vartheta}(y_{1:T})\,p(\vartheta)}{p_{\vartheta_{i-1}}(y_{1:T})\,p(\vartheta_{i-1})}\frac{q(\vartheta_{i-1}|\,\vartheta)}{q(\vartheta|\,\vartheta_{i-1})}\right\},$$

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

- **Problem**: MH cannot be implemented if $p_{\vartheta}(y_{1:T})$ cannot be evaluated.

Pseudo-Marginal Metropolis–Hastings algorithm

- **“Idea”**: Replace $p_\vartheta\left(y_{1:T}\right)$ by an estimate $\widehat{p}_\vartheta\left(y_{1:T}\right)$ in MH.

Pseudo-Marginal Metropolis–Hastings algorithm

- **"Idea"**: Replace $p_\vartheta\left(y_{1:T}\right)$ by an estimate $\widehat{p}_\vartheta\left(y_{1:T}\right)$ in MH.

Pseudo-Marginal Metropolis–Hastings algorithm

- **"Idea"**: Replace $p_{\vartheta}\left(y_{1:T}\right)$ by an estimate $\widehat{p}_{\vartheta}\left(y_{1:T}\right)$ in MH.

At iteration i

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

Pseudo-Marginal Metropolis–Hastings algorithm

- **"Idea"**: Replace $p_\vartheta\left(y_{1:T}\right)$ by an estimate $\widehat{p}_\vartheta\left(y_{1:T}\right)$ in MH.

At iteration i

- Sample $\vartheta \sim q\left(\cdot \mid \vartheta_{i-1}\right)$.
- Compute an estimate $\widehat{p}_\vartheta\left(y_{1:T}\right)$ of $p_\vartheta\left(y_{1:T}\right)$.

Pseudo-Marginal Metropolis–Hastings algorithm

- **"Idea"**: Replace $p_\vartheta\left(y_{1:T}\right)$ by an estimate $\widehat{p}_\vartheta\left(y_{1:T}\right)$ in MH.

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

- Sample $\vartheta \sim q\left(\cdot \mid \vartheta_{i-1}\right)$.
- Compute an estimate $\widehat{p}_\vartheta\left(y_{1:T}\right)$ of $p_\vartheta\left(y_{1:T}\right)$.
- With probability

$$
\begin{aligned}
& \min\{1, \underbrace{\frac{p_\vartheta\left(y_{1:T}\right)}{p_{\vartheta_{i-1}}\left(y_{1:T}\right)} \frac{p\left(\vartheta\right)}{p\left(\vartheta_{i-1}\right)} \frac{q\left(\vartheta_{i-1} \mid \vartheta\right)}{q\left(\vartheta \mid \vartheta_{i-1}\right)}}_{\text{exact MH ratio}} \times \underbrace{\frac{\widehat{p}_\vartheta\left(y_{1:T}\right) / p_\vartheta\left(y_{1:T}\right)}{\widehat{p}_{\vartheta_{i-1}}\left(y_{1:T}\right) / p_{\vartheta_{i-1}}\left(y_{1:T}\right)}}_{\text{noise}}\} \\
= \quad & \min\{1, \frac{\widehat{p}_\vartheta\left(y_{1:T}\right) p\left(\vartheta\right)}{\widehat{p}_{\vartheta_{i-1}}\left(y_{1:T}\right) p\left(\vartheta_{i-1}\right)} \frac{q\left(\vartheta_{i-1} \mid \vartheta\right)}{q\left(\vartheta \mid \vartheta_{i-1}\right)}\}
\end{aligned}
$$

set $\vartheta_i = \vartheta$, $\widehat{p}_{\vartheta_i}\left(y_{1:T}\right) = \widehat{p}_\vartheta\left(y_{1:T}\right)$ otherwise set $\vartheta_i = \vartheta_{i-1}$, $\widehat{p}_{\vartheta_i}\left(y_{1:T}\right) = \widehat{p}_{\vartheta_{i-1}}\left(y_{1:T}\right)$.

Key Result

- **Proposition** (Lin, Liu & Sloan, 2000; Andrieu & Roberts, 2009): If $\widehat{p}_{\vartheta}\left(y_{1:T}\right)$ is a non-negative unbiased estimator of $p_{\theta}\left(y_{1:T}\right)$ then the pseudo-marginal MH kernel admits $\pi\left(\theta\right)$ as invariant density.

Key Result

- **Proposition** (Lin, Liu & Sloan, 2000; Andrieu & Roberts, 2009): If $\widehat{p}_{\vartheta}\left(y_{1:T}\right)$ is a non-negative unbiased estimator of $p_{\theta}\left(y_{1:T}\right)$ then the pseudo-marginal MH kernel admits $\pi\left(\theta\right)$ as invariant density.
- Let U be the r.v. such that $\widehat{p}_{\theta}\left(y_{1:T}\right)=\widehat{p}_{\theta}\left(y_{1:T};U\right)$ and $\mathbb{E}\left[\widehat{p}_{\theta}\left(y_{1:T};U\right)\right]=p_{\theta}\left(y_{1:T}\right)$ when $U\sim m\left(\cdot\right)$.

Key Result

- **Proposition** (Lin, Liu & Sloan, 2000; Andrieu & Roberts, 2009): If $\widehat{p}_{\vartheta}\left(y_{1:T}\right)$ is a non-negative unbiased estimator of $p_{\theta}\left(y_{1:T}\right)$ then the pseudo-marginal MH kernel admits $\pi\left(\theta\right)$ as invariant density.
- Let U be the r.v. such that $\widehat{p}_{\theta}\left(y_{1:T}\right)=\widehat{p}_{\theta}\left(y_{1:T};U\right)$ and $\mathbb{E}\left[\widehat{p}_{\theta}\left(y_{1:T};U\right)\right]=p_{\theta}\left(y_{1:T}\right)$ when $U\sim m\left(\cdot\right)$.
- Consider the auxiliary target density on $\Theta\times\mathcal{U}$

$$\overline{\pi}(\theta,u)=\pi\left(\theta\right)\underbrace{\frac{\widehat{p}_{\theta}\left(y_{1:T};u\right)}{p_{\theta}\left(y_{1:T}\right)}m\left(u\right)}_{\int(.)\mathrm{d}u=1}$$

Key Result

- **Proposition** (Lin, Liu & Sloan, 2000; Andrieu & Roberts, 2009): If $\widehat{p}_{\vartheta}\left(y_{1:T}\right)$ is a non-negative unbiased estimator of $p_{\theta}\left(y_{1:T}\right)$ then the pseudo-marginal MH kernel admits $\pi\left(\theta\right)$ as invariant density.
- Let U be the r.v. such that $\widehat{p}_{\theta}\left(y_{1:T}\right)=\widehat{p}_{\theta}\left(y_{1:T};U\right)$ and $\mathbb{E}\left[\widehat{p}_{\theta}\left(y_{1:T};U\right)\right]=p_{\theta}\left(y_{1:T}\right)$ when $U\sim m\left(\cdot\right)$.
- Consider the auxiliary target density on $\Theta\times\mathcal{U}$

$$\overline{\pi}(\theta,u)=\pi\left(\theta\right)\underbrace{\frac{\widehat{p}_{\theta}\left(y_{1:T};u\right)}{p_{\theta}\left(y_{1:T}\right)}m\left(u\right)}_{\int(.)\mathrm{d}u=1}$$

- Pseudo-marginal MH is a standard MH with target $\overline{\pi}(\theta,u)$ and proposal $q\left(\left.\vartheta\right|\theta\right)m(v)$ as

$$\frac{\overline{\pi}(\vartheta,v)}{\overline{\pi}(\theta,u)}\frac{q\left(\left.\theta\right|\vartheta\right)m(u)}{q\left(\left.\vartheta\right|\theta\right)m(v)}=\frac{\widehat{p}_{\vartheta}\left(y_{1:T};v\right)}{\widehat{p}_{\theta}\left(y_{1:T};u\right)}\frac{p\left(\vartheta\right)}{p\left(\theta\right)}\frac{q\left(\left.\theta\right|\vartheta\right)}{q\left(\left.\vartheta\right|\theta\right)}.$$

Importance Sampling Estimator

- For latent variable models, one has

$$p_\theta(y_t) = \int \mu_\theta(x_t)\, g_\theta(y_t | x_t)\, \mathrm{d}x_t.$$

Importance Sampling Estimator

- For latent variable models, one has

$$p_\theta\left(y_t\right) = \int \mu_\theta\left(x_t\right) g_\theta\left(y_t \middle| x_t\right) \mathrm{d}x_t.$$

- An non-negative unbiased estimator is given by

$$\widehat{p}_\theta(y_{1:T}) = \prod_{t=1}^{T} \widehat{p}_\theta(y_t) = \prod_{t=1}^{T} \left\{ \frac{1}{N} \sum_{k=1}^{N} g_\theta\left(y_t \middle| X_t^k\right) \right\}, \quad X_t^k \overset{\text{i.i.d.}}{\sim} \mu_\theta,$$

i.e.

$$m\left(u\right) = \prod_{t=1}^{T} \prod_{k=1}^{N} \mu_\theta\left(x_t^k\right).$$

Importance Sampling Estimator

- For latent variable models, one has

$$p_\theta(y_t) = \int \mu_\theta(x_t)\, g_\theta(y_t | x_t)\, \mathrm{d}x_t.$$

- An non-negative unbiased estimator is given by

$$\widehat{p}_\theta(y_{1:T}) = \prod_{t=1}^{T} \widehat{p}_\theta(y_t) = \prod_{t=1}^{T} \left\{ \frac{1}{N} \sum_{k=1}^{N} g_\theta\left(y_t | X_t^k \right) \right\},\ X_t^k \overset{\text{i.i.d.}}{\sim} \mu_\theta,$$

i.e.

$$m(u) = \prod_{t=1}^{T} \prod_{k=1}^{N} \mu_\theta\left(x_t^k\right).$$

- Computational complexity is $O(NT)$.

Particle Filter Estimator

- For state-space models, previous approach provides an estimator whose relative variance scales typically exponentially with T.

Particle Filter Estimator

- For state-space models, previous approach provides an estimator whose relative variance scales typically exponentially with T.
- An alternative is to use particle filter where

$$\begin{aligned}\widehat{p}_\theta\left(y_{1:T}\right) &= \widehat{p}_\theta(y_1)\prod_{t=2}^{T}\widehat{p}_\theta\left(y_t \middle| y_{1:t-1}\right) \\ &= \prod_{t=1}^{T}\left\{\frac{1}{N}\sum_{k=1}^{N} g_\theta\left(y_t \middle| X_n^k\right)\right\}\end{aligned}$$

where

$$m(u) = \prod_{k=1}^{N}\mu_\theta\left(x_1^k\right)\prod_{t=2}^{T}\{\prod_{k=1}^{N} w_t^{a_{t-1}^k} f\left(x_t^k \middle| x_{t-1}^{a_{t-1}^k}\right)\}$$

with $a_{t-1}^k \in \{1, \ldots, N\}$, $w_t^j \propto g_\theta\left(y_t \middle| X_t^j\right)$, $\sum_j w_t^j = 1$.

Particle Filter Estimator

- For state-space models, previous approach provides an estimator whose relative variance scales typically exponentially with T.
- An alternative is to use particle filter where

$$\begin{aligned}\widehat{p}_\theta\left(y_{1:T}\right) &= \widehat{p}_\theta(y_1)\prod_{t=2}^{T}\widehat{p}_\theta\left(y_t \middle| y_{1:t-1}\right)\\ &= \prod_{t=1}^{T}\left\{\frac{1}{N}\sum_{k=1}^{N} g_\theta\left(y_t \middle| X_n^k\right)\right\}\end{aligned}$$

where

$$m\left(u\right) = \prod_{k=1}^{N}\mu_\theta\left(x_1^k\right)\prod_{t=2}^{T}\{\prod_{k=1}^{N} w_t^{a_{t-1}^k} f\left(x_t^k \middle| x_{t-1}^{a_{t-1}^k}\right)\}$$

with $a_{t-1}^k \in \{1, ..., N\}$, $w_t^j \propto g_\theta\left(y_t \middle| X_t^j\right)$, $\sum_j w_t^j = 1$.

- Computational complexity is $O\left(NT\right)$.

Particle Filter Estimator

- For state-space models, previous approach provides an estimator whose relative variance scales typically exponentially with T.
- An alternative is to use particle filter where

$$\begin{aligned}\widehat{p}_\theta\left(y_{1:T}\right) &= \widehat{p}_\theta(y_1)\prod_{t=2}^{T}\widehat{p}_\theta\left(y_t \mid y_{1:t-1}\right)\\ &= \prod_{t=1}^{T}\left\{\frac{1}{N}\sum_{k=1}^{N} g_\theta\left(y_t \mid X_n^k\right)\right\}\end{aligned}$$

where

$$m\left(u\right)=\prod_{k=1}^{N}\mu_\theta\left(x_1^k\right)\prod_{t=2}^{T}\{\prod_{k=1}^{N} w_t^{a_{t-1}^k} f\left(x_t^k \middle| x_{t-1}^{a_{t-1}^k}\right)\}$$

with $a_{t-1}^k \in \{1,...,N\}$, $w_t^j \propto g_\theta\left(y_t \mid X_t^j\right)$, $\sum_j w_t^j = 1$.

- Computational complexity is $O\left(NT\right)$.
- The estimator $\widehat{p}_\theta(y_{1:T})$ of $p_\theta\left(y_{1:T}\right)$ is unbiased and its relative variance is bounded uniformly over T if $N \propto T$ (Cerou, Del Moral &

Pseudo-Marginal Metropolis–Hastings algorithm

Pseudo-Marginal Metropolis–Hastings algorithm

At iteration i

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

Pseudo-Marginal Metropolis–Hastings algorithm

At iteration i

- Sample $\vartheta \sim q\left(\cdot \middle| \vartheta_{i-1}\right)$.
- Use particle filter to compute an estimate $\widehat{p}_{\vartheta}\left(y_{1:T}\right)$ of $p_{\vartheta}\left(y_{1:T}\right)$.

Pseudo-Marginal Metropolis–Hastings algorithm

At iteration i

- Sample $\vartheta \sim q\left(\cdot\right|\vartheta_{i-1}\left.\right)$.
- Use particle filter to compute an estimate $\widehat{p}_{\vartheta}\left(y_{1:T}\right)$ of $p_{\vartheta}\left(y_{1:T}\right)$.
- With probability

$$\min\{1, \frac{\widehat{p}_{\vartheta}\left(y_{1:T}\right)p\left(\vartheta\right)}{\widehat{p}_{\vartheta_{i-1}}\left(y_{1:T}\right)p\left(\vartheta_{i-1}\right)}\frac{q\left(\vartheta_{i-1}\right|\vartheta\left.\right)}{q\left(\vartheta\right|\vartheta_{i-1}\left.\right)}\}$$

set $\vartheta_i = \vartheta$, $\widehat{p}_{\vartheta_i}\left(y_{1:T}\right) = \widehat{p}_{\vartheta}\left(y_{1:T}\right)$ otherwise set $\vartheta_i = \vartheta_{i-1}$, $\widehat{p}_{\vartheta_i}\left(y_{1:T}\right) = \widehat{p}_{\vartheta_{i-1}}\left(y_{1:T}\right)$.

Empirical performance: Stochastic kinetic model

- Two species X_s^1 (prey) and X_s^2 (predator)

$$
\begin{aligned}
\Pr\left(X_{s+ds}^1 = x_s^1 + 1, X_{s+ds}^2 = x_s^2 \,\middle|\, x_s^1, x_s^2\right) &= \alpha\, x_s^1\, ds + o\left(ds\right), \\
\Pr\left(X_{s+ds}^1 = x_s^1 - 1, X_{s+ds}^2 = x_s^2 + 1 \,\middle|\, x_s^1, x_s^2\right) &= \beta\, x_s^1\, x_s^2\, ds + o\left(ds\right), \\
\Pr\left(X_{s+ds}^1 = x_t^1, X_{s+ds}^2 = x_s^2 - 1 \,\middle|\, x_s^1, x_s^2\right) &= \gamma\, x_s^2\, ds + o\left(ds\right),
\end{aligned}
$$

observed at discrete times

$$
Y_t = X_{\Delta t}^1 + W_t \text{ with } W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right).
$$

Empirical performance: Stochastic kinetic model

- Two species X_s^1 (prey) and X_s^2 (predator)

$$\begin{aligned}
\Pr\left(X^1_{s+ds}{=}x^1_s{+}1, X^2_{s+ds}{=}x^2_s \,\middle|\, x^1_s, x^2_s\right) &= \alpha\, x^1_s\, ds + o\left(ds\right), \\
\Pr\left(X^1_{s+ds}{=}x^1_s{-}1, X^2_{s+ds}{=}x^2_s{+}1 \,\middle|\, x^1_s, x^2_s\right) &= \beta\, x^1_s\, x^2_s\, ds + o\left(ds\right), \\
\Pr\left(X^1_{s+ds}{=}x^1_t, X^2_{s+ds}{=}x^2_s{-}1 \,\middle|\, x^1_s, x^2_s\right) &= \gamma\, x^2_s\, ds + o\left(ds\right),
\end{aligned}$$

observed at discrete times

$$Y_t = X^1_{\Delta t} + W_t \text{ with } W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right).$$

- We are interested in the kinetic rate constants $\theta = (\alpha, \beta, \gamma)$ a priori distributed as (Boys et al., 2008; Kunsch, 2011)

$$\alpha \sim \mathcal{G}(1, 10), \quad \beta \sim \mathcal{G}(1, 0.25), \quad \gamma \sim \mathcal{G}(1, 7.5).$$

Empirical performance: Stochastic kinetic model

- Two species X_s^1 (prey) and X_s^2 (predator)

$$\begin{aligned}
\Pr\left(X_{s+ds}^1 = x_s^1 + 1, X_{s+ds}^2 = x_s^2 \,\middle|\, x_s^1, x_s^2\right) &= \alpha\, x_s^1\, ds + o\left(ds\right), \\
\Pr\left(X_{s+ds}^1 = x_s^1 - 1, X_{s+ds}^2 = x_s^2 + 1 \,\middle|\, x_s^1, x_s^2\right) &= \beta\, x_s^1\, x_s^2\, ds + o\left(ds\right), \\
\Pr\left(X_{s+ds}^1 = x_t^1, X_{s+ds}^2 = x_s^2 - 1 \,\middle|\, x_s^1, x_s^2\right) &= \gamma\, x_s^2\, ds + o\left(ds\right),
\end{aligned}$$

  observed at discrete times

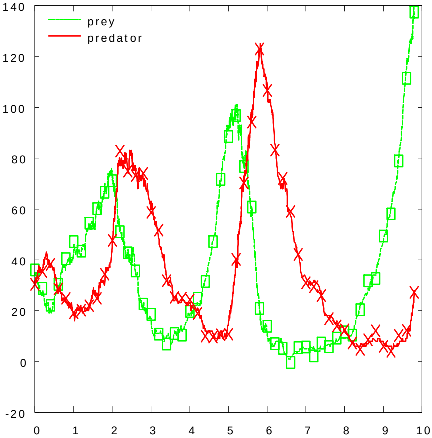
$$Y_t = X_{\Delta t}^1 + W_t \text{ with } W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right).$$

- We are interested in the kinetic rate constants $\theta = (\alpha, \beta, \gamma)$ a priori distributed as (Boys et al., 2008; Kunsch, 2011)

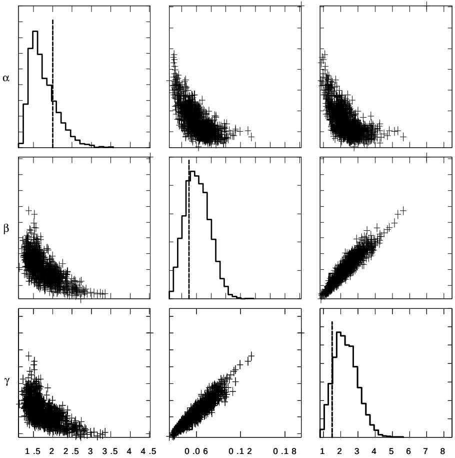
$$\alpha \sim \mathcal{G}(1, 10), \quad \beta \sim \mathcal{G}(1, 0.25), \quad \gamma \sim \mathcal{G}(1, 7.5).$$

- Pseudo-marginal MH with RW proposal, likelihood is approximated using particle filter.

Empirical performance: Stochastic kinetic model

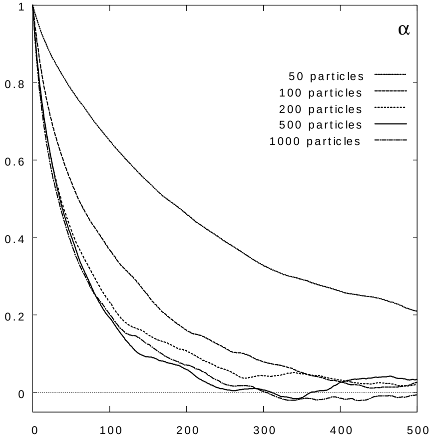


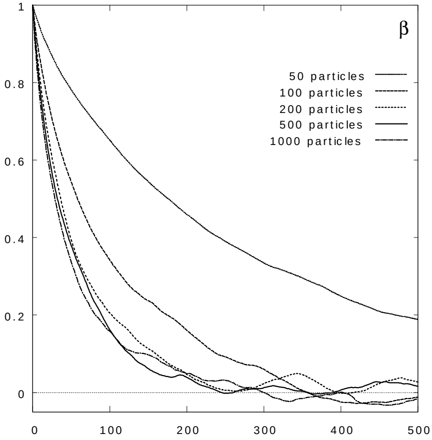
Simulated data



Estimated posteriors

Empirical performance: Stochastic kinetic model





Autocorrelation of α (left) and β (right) for the PM sampler for various N.

Empirical performance: Stochastic volatility model

- Huang & Tauchen, *J. Financial Econometrics* (2005):

$$
\begin{aligned}
\mathrm{d}v_1(s) &= -k_1\{v_1(s)-\mu_1\}\,\mathrm{d}s+\sigma_1\mathrm{d}W_1(s),\\
\mathrm{d}v_2(s) &= -k_2v_2(s)\,\mathrm{d}s+\{1+\beta_{12}v_2(s)\}\,\mathrm{d}W_2(s),\\
\mathrm{d}\log P(s) &= \mu_y\mathrm{d}s+\text{s-exp}\left[\{v_1(s)+\beta_2v_2(s)\}/2\right]\mathrm{d}B(s),
\end{aligned}
$$

with $\phi_1 = \mathrm{corr}\{B(s), W_1(s)\}$ and $\phi_2 = \mathrm{corr}\{B(s), W_2(s)\}$.

Empirical performance: Stochastic volatility model

- Huang & Tauchen, *J. Financial Econometrics* (2005):

$$
\begin{aligned}
\mathrm{d}v_1(s) &= -k_1\{v_1(s)-\mu_1\}\,\mathrm{d}s+\sigma_1\mathrm{d}W_1(s),\\
\mathrm{d}v_2(s) &= -k_2v_2(s)\,\mathrm{d}s+\{1+\beta_{12}v_2(s)\}\,\mathrm{d}W_2(s),\\
\mathrm{d}\log P(s) &= \mu_y\mathrm{d}s+\text{s-exp}\left[\{v_1(s)+\beta_2v_2(s)\}/2\right]\mathrm{d}B(s),
\end{aligned}
$$

  with $\phi_1 = \text{corr}\{B(s), W_1(s)\}$ and $\phi_2 = \text{corr}\{B(s), W_2(s)\}$.
- Euler discretization of the volatilities $v_1(s)$ and $v_2(s)$ provides closed form expression for $Y_t = \log P(\Delta t) - \log P(\Delta(t-1))$.

Empirical performance: Stochastic volatility model

- Huang & Tauchen, *J. Financial Econometrics* (2005):

$$\begin{aligned} \mathrm{d}v_1(s) &= -k_1\{v_1(s)-\mu_1\}\,\mathrm{d}s+\sigma_1\mathrm{d}W_1(s), \\ \mathrm{d}v_2(s) &= -k_2v_2(s)\,\mathrm{d}s+\{1+\beta_{12}v_2(s)\}\,\mathrm{d}W_2(s), \\ \mathrm{d}\log P(s) &= \mu_y\mathrm{d}s+\text{s-exp}\left[\{v_1(s)+\beta_2v_2(s)\}/2\right]\mathrm{d}B(s), \end{aligned}$$

  with $\phi_1 = \mathrm{corr}\{B(s), W_1(s)\}$ and $\phi_2 = \mathrm{corr}\{B(s), W_2(s)\}$.
- Euler discretization of the volatilities $v_1(s)$ and $v_2(s)$ provides closed form expression for $Y_t = \log P(\Delta t) - \log P(\Delta(t-1))$.
- Daily returns $y = (y_1, ..., y_T)$ of the S&P 500 index.

Empirical performance: Stochastic volatility model

- Huang & Tauchen, *J. Financial Econometrics* (2005):

$$
\begin{aligned}
\mathrm{d}v_1(s) &= -k_1\{v_1(s)-\mu_1\}\mathrm{d}s+\sigma_1\mathrm{d}W_1(s),\\
\mathrm{d}v_2(s) &= -k_2v_2(s)\mathrm{d}s+\{1+\beta_{12}v_2(s)\}\mathrm{d}W_2(s),\\
\mathrm{d}\log P(s) &= \mu_y\mathrm{d}s+\text{s-exp}\left[\{v_1(s)+\beta_2v_2(s)\}/2\right]\mathrm{d}B(s),
\end{aligned}
$$

  with $\phi_1 = \mathrm{corr}\{B(s), W_1(s)\}$ and $\phi_2 = \mathrm{corr}\{B(s), W_2(s)\}$.
- Euler discretization of the volatilities $v_1(s)$ and $v_2(s)$ provides closed form expression for $Y_t = \log P(\Delta t) - \log P(\Delta(t-1))$.
- Daily returns $y = (y_1, ..., y_T)$ of the S&P 500 index.
- Bayesian Inference on $\theta = \left(k_1, \mu_1, \sigma_1, k_2, \beta_{12}, \beta_2, \mu_y, \phi_1, \phi_2\right)$.

Empirical performance: Stochastic volatility model

- Huang & Tauchen, *J. Financial Econometrics* (2005):

$$\begin{aligned} \mathrm{d}v_1(s) &= -k_1\{v_1(s)-\mu_1\}\mathrm{d}s+\sigma_1\mathrm{d}W_1(s), \\ \mathrm{d}v_2(s) &= -k_2v_2(s)\mathrm{d}s+\{1+\beta_{12}v_2(s)\}\mathrm{d}W_2(s), \\ \mathrm{d}\log P(s) &= \mu_y\mathrm{d}s+\text{s-exp}\left[\{v_1(s)+\beta_2v_2(s)\}/2\right]\mathrm{d}B(s), \end{aligned}$$

  with $\phi_1 = \text{corr}\{B(s), W_1(s)\}$ and $\phi_2 = \text{corr}\{B(s), W_2(s)\}$.
- Euler discretization of the volatilities $v_1(s)$ and $v_2(s)$ provides closed form expression for $Y_t = \log P(\Delta t) - \log P(\Delta(t-1))$.
- Daily returns $y = (y_1, ..., y_T)$ of the S&P 500 index.
- Bayesian Inference on $\theta = \left(k_1, \mu_1, \sigma_1, k_2, \beta_{12}, \beta_2, \mu_y, \phi_1, \phi_2\right)$.
- Performance of the pseudo-marginal for RW proposal w.r.t σ, standard deviation of $\log \widehat{p}_\theta(y)$ at posterior mean $\overline{\theta}$.

Integrated Autocorrelation Time of Pseudo-Marginal MH

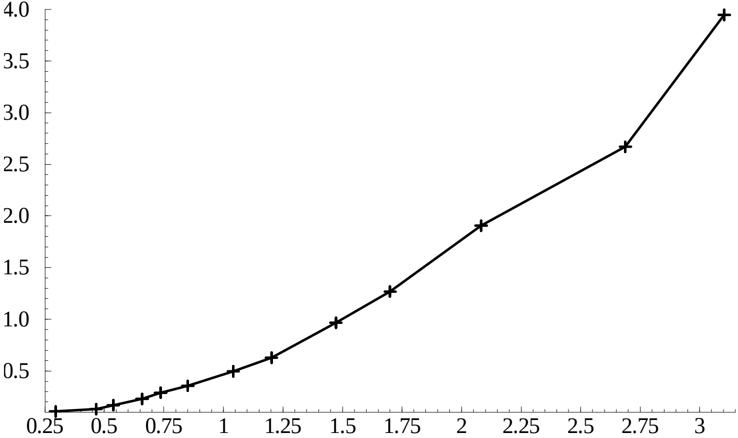


Figure: Average over the 9 parameter components of the log-integrated autocorrelation time of pseudo-marginal chain as a function of σ for $T = 300$.

How precise should the log-likelihood estimator be?

- **Aim**: Minimize the computational time

$$CT_h^Q = IF_h^Q / \sigma^2$$

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

$$IF_h^Q \;=\; \text{Integrated Autocorrelation Time of PM average}$$

How precise should the log-likelihood estimator be?

- **Aim**: Minimize the computational time
$$CT_h^Q = IF_h^Q / \sigma^2$$
as $\sigma^2 \propto 1/N$ and computational efforts proportional to N, where
$$IF_h^Q = \text{Integrated Autocorrelation Time of PM average}$$
- Call the IACT the *inefficiency*
$$IF_h^Q = 1 + 2\sum_{\tau=1}^{\infty} \mathrm{corr}_{\pi,Q}\left\{h\left(\theta_0\right), h\left(\theta_\tau\right)\right\}$$
where Q is the pseudo-marginal kernel given for $(\theta, z) \neq (\vartheta, w)$ by
$$Q\left\{(\theta, z), (\mathrm{d}\vartheta, \mathrm{d}w)\right\} = q(\vartheta|\theta)g_\vartheta(w)\min\left\{1, \frac{\pi\left(\vartheta\right)}{\pi\left(\theta\right)}\exp\left(w - z\right)\right\}\mathrm{d}\vartheta\mathrm{d}w,$$
where
$$z = \log\{\widehat{p}_\theta(y_{1:T})/p_\theta(y_{1:T})\},$$
$$w = \log\{\widehat{p}_\vartheta(y_{1:T})/p_\vartheta(y_{1:T})\}.$$

Computational time for the SV model

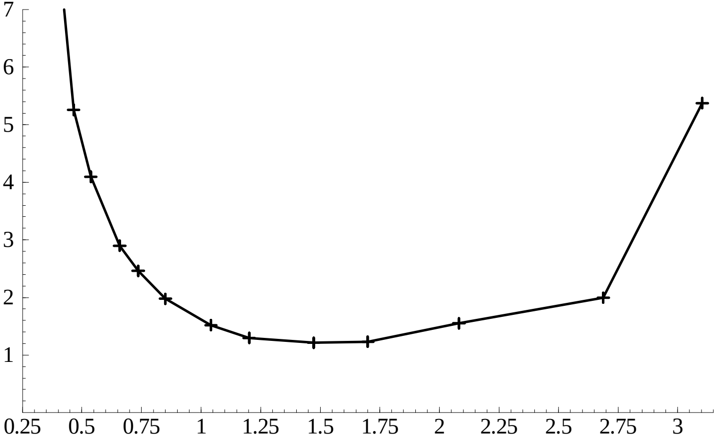


Figure: Computational time as a function of σ

Analysis in the large data regime

- Standard asymptotic study of MCMC relies on $d \longrightarrow \infty$ and independence assumption on the target, interested here in fixed d, large T.

Analysis in the large data regime

- Standard asymptotic study of MCMC relies on $d \rightarrow \infty$ and independence assumption on the target, interested here in fixed d, large T.
- **Assumption 1** - Asymptotic Normality: We have

$$\int \left| p\left(\theta \middle| Y_{1:T}\right) - \phi(\theta; \widehat{\theta}^{T}, \Sigma / T) \right| \mathrm{d}\theta \xrightarrow{P} 0,$$

where $\widehat{\theta}^{T} \xrightarrow{P} \overline{\theta}$ and Σ is a p.d. matrix.

Analysis in the large data regime

- Standard asymptotic study of MCMC relies on $d \to \infty$ and independence assumption on the target, interested here in fixed d, large T.
- **Assumption 1** - Asymptotic Normality: We have

$$\int \left| p\left(\theta \mid Y_{1:T}\right) - \phi(\theta; \widehat{\theta}^{T}, \Sigma / T) \right| \mathrm{d}\theta \xrightarrow{P} 0,$$

where $\widehat{\theta}^{T} \xrightarrow{P} \overline{\theta}$ and Σ is a p.d. matrix.

- **Assumption 2** - CLT: For any θ in a neighbourhood of $\overline{\theta}$,

$$\log \frac{\widehat{p}_{\theta}(Y_{1:T})}{p_{\theta}(Y_{1:T})} \bigg| \mathcal{Y}^{T} \Rightarrow \mathcal{N}\left(-\sigma^{2}\left(\theta\right)/2, \sigma^{2}\left(\theta\right)\right)$$

in probability and $\sigma^{2}\left(\cdot\right)$ continuous at $\overline{\theta}$.

Analysis in the large data regime

- Standard asymptotic study of MCMC relies on $d \longrightarrow \infty$ and independence assumption on the target, interested here in fixed d, large T.
- **Assumption 1** - Asymptotic Normality: We have

$$\int \left| p\left(\theta \middle| Y_{1:T}\right) - \phi(\theta; \widehat{\theta}^{T}, \Sigma/T) \right| \mathrm{d}\theta \xrightarrow{P} 0,$$

  where $\widehat{\theta}^{T} \xrightarrow{P} \overline{\theta}$ and Σ is a p.d. matrix.
- **Assumption 2** - CLT: For any θ in a neighbourhood of $\overline{\theta}$,

$$\log \frac{\widehat{p}_{\theta}(Y_{1:T})}{p_{\theta}(Y_{1:T})} \Bigg| \mathcal{Y}^{T} \Rightarrow \mathcal{N}\left(-\sigma^{2}(\theta)/2, \sigma^{2}(\theta)\right)$$

  in probability and $\sigma^{2}(\cdot)$ continuous at $\overline{\theta}$.
- **Assumption 3** - Proposal: $\vartheta = \theta + \varepsilon/\sqrt{T}$ where $\varepsilon \sim \upsilon(\cdot)$ with $\upsilon(\varepsilon) = \upsilon(-\varepsilon)$.

Analysis in the large data regime

- Assumption 1 holds if for example Bernstein-von Mises holds (in correctly specified/misspecified scenarios).

Analysis in the large data regime

- Assumption 1 holds if for example Bernstein-von Mises holds (in correctly specified/misspecified scenarios).
- Assumption 2 has been shown to hold under regularity assumptions if $N \propto T$ (Berard et al, 2014, Deligiannidis et al, 2015).

Analysis in the large data regime

- Assumption 1 holds if for example Bernstein-von Mises holds (in correctly specified/misspecified scenarios).
- Assumption 2 has been shown to hold under regularity assumptions if $N \propto T$ (Berard et al, 2014, Deligiannidis et al, 2015).
- Assumption 3 can be easily enforced.

Weak convergence

- Let $\left\{\vartheta_i^T, Z_i^T := \log \widehat{p}_{\vartheta_i^T}\left(Y_{1:T}\right)/p_{\vartheta_i^T}\left(Y_{1:T}\right)\right\}_{i\geq 0}$ the stationary PM Markov chain of invariant density $p\left(\theta \middle| Y_{1:T}\right)\exp\left(z\right)g_\theta^T\left(z\right)$.

Weak convergence

- Let $\{\vartheta_i^T, Z_i^T := \log \widehat{p}_{\vartheta_i^T}(Y_{1:T}) / p_{\vartheta_i^T}(Y_{1:T})\}_{i \geq 0}$ the stationary PM Markov chain of invariant density $p(\theta | Y_{1:T}) \exp(z) g_\theta^T(z)$.
- **Proposition** (Schmon et al, 2016): The F.D.D. of the rescaled sequence $\{\widetilde{\vartheta}_i^T = \sqrt{T}(\vartheta_i^T - \widehat{\theta}_T), Z_i^T\}_{i \geq 0}$ converge weakly as $T \to \infty$ to those of a stationary Markov chain of invariant density $\phi\left(\widetilde{\theta}; 0, \Sigma\right) \phi\left(z; -\sigma^2\left(\overline{\theta}\right)/2, \sigma^2\left(\overline{\theta}\right)\right)$ and kernel given by

$$\begin{aligned} \widetilde{Q}\{(\widetilde{\theta}, z), (\mathrm{d}\widetilde{\vartheta}, \mathrm{d}w)\} &= v(\widetilde{\vartheta} - \widetilde{\theta})\phi\left(w; -\sigma^2\left(\overline{\theta}\right)/2, \sigma^2\left(\overline{\theta}\right)\right) \\ &\quad \times \min\left\{1, \frac{\phi(\widetilde{\vartheta}; 0, \Sigma)}{\phi(\widetilde{\theta}; 0, \Sigma)} \exp(w - z)\right\} \mathrm{d}\widetilde{\vartheta}\mathrm{d}w \end{aligned}$$

for $(\widetilde{\theta}, z) \neq (\widetilde{\vartheta}, w)$.

Weak convergence

- These results suggests that a simplified analysis of the PM chain can be performed by looking at

$$\begin{aligned}\widehat{Q}\{(\theta, z), (\mathrm{d}\vartheta, \mathrm{d}w)\} &= q(\vartheta|\theta)\,\phi\left(w; -\sigma^2/2, \sigma^2\right) \\ &\quad \times \min\left\{1, \frac{\pi(\vartheta)}{\pi(\theta)} \exp(w - z)\right\} \mathrm{d}\vartheta \mathrm{d}w,\end{aligned}$$

where $\sigma^2 = \sigma^2\left(\overline{\theta}\right)$.

Weak convergence

- These results suggests that a simplified analysis of the PM chain can be performed by looking at

$$\widehat{Q}\{(\theta, z), (\mathrm{d}\vartheta, \mathrm{d}w)\} = q(\vartheta|\theta)\,\phi\left(w; -\sigma^2/2, \sigma^2\right) \times \min\left\{1, \frac{\pi(\vartheta)}{\pi(\theta)} \exp(w - z)\right\} \mathrm{d}\vartheta \mathrm{d}w,$$

  where $\sigma^2 = \sigma^2\left(\overline{\theta}\right)$.

- It would be more satisfactory to show that

$$\left| IF_h^Q - IF_h^{\widehat{Q}} \right| \to 0$$

  as $T \to \infty$. The analysis relies on (Andrieu & Vihola, 2015) and is much more involved.

Empirical vs Assumed Distributions for SV model

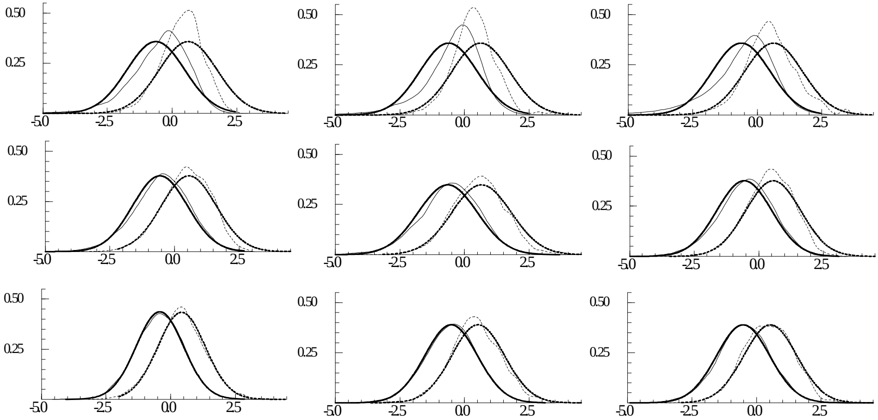


Figure: Empirical distributions (dashed) vs assumed Gaussians (solid) of Z at $\overline{\theta}$ (left) and marginalized over samples from $\pi(\theta)$ (center) and $\int \pi(d\vartheta)\, q(\theta|\,\vartheta)$ (right) for $T = 40$, $T = 300$ and $T = 2700$.

Available Results

- **Aim**: Minimize the computational cost

$$CT_h^{\widehat{Q}}(\sigma) = IF_h^{\widehat{Q}}(\sigma)/\sigma^2.$$

Available Results

- **Aim**: Minimize the computational cost

$$CT_h^{\widehat{Q}}(\sigma) = IF_h^{\widehat{Q}}(\sigma)/\sigma^2.$$

- **Special cases**:

Available Results

- **Aim**: Minimize the computational cost

$$CT_h^{\widehat{Q}}(\sigma) = IF_h^{\widehat{Q}}(\sigma)/\sigma^2.$$

- **Special cases**:

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

Available Results

- **Aim**: Minimize the computational cost

$$CT_h^{\widehat{Q}}(\sigma) = IF_h^{\widehat{Q}}(\sigma)/\sigma^2.$$

- **Special cases**:

1. When $q(\vartheta|\theta) = p(\vartheta|y)$, $\sigma_{\text{opt}} = 0.92$ (Pitt et al., 2012).
2. When $\pi(\theta) = \prod_{i=1}^{d} f(\theta_i)$ and $q(\vartheta|\theta)$ is an isotropic Gaussian random walk then, as $d \to \infty$, diffusion limit suggests $\sigma_{\text{opt}} = 1.81$ (Sherlock et al., 2015).

Sketch of the Analysis

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

Sketch of the Analysis

- For general proposals and targets, direct minimization of $CT_h^{\widehat{Q}}(\sigma) = IF_h^{\widehat{Q}}(\sigma)/\sigma^2$ impossible so minimize an upper bound over it.
- Theoretical study relies on $\overline{\pi}-$invariant kernel Q^* given for $(\theta, z) \neq (\vartheta, w)$ by

$$q(\vartheta|\theta)\phi\left(w; -\sigma^2/2, \sigma^2\right) \min\left\{1, \frac{\pi(\vartheta)}{\pi(\theta)}\right\} \min\left\{1, \exp(w-z)\right\} \mathrm{d}\vartheta \mathrm{d}w,$$

instead of

$$q(\vartheta|\theta)\phi\left(w; -\sigma^2/2, \sigma^2\right) \min\left\{1, \frac{\pi(\vartheta)}{\pi(\theta)} \exp(w-z)\right\} \mathrm{d}\vartheta \mathrm{d}w.$$

Sketch of the Analysis

- For general proposals and targets, direct minimization of $CT_h^{\widehat{Q}}(\sigma) = IF_h^{\widehat{Q}}(\sigma)/\sigma^2$ impossible so minimize an upper bound over it.
- Theoretical study relies on $\overline{\pi}-$invariant kernel Q^* given for $(\theta, z) \neq (\vartheta, w)$ by

$$q(\vartheta|\theta)\phi\left(w; -\sigma^2/2, \sigma^2\right) \min\left\{1, \frac{\pi(\vartheta)}{\pi(\theta)}\right\} \min\left\{1, \exp\left(w - z\right)\right\} \mathrm{d}\vartheta \mathrm{d}w,$$

instead of

$$q(\vartheta|\theta)\phi\left(w; -\sigma^2/2, \sigma^2\right) \min\left\{1, \frac{\pi(\vartheta)}{\pi(\theta)} \exp\left(w - z\right)\right\} \mathrm{d}\vartheta \mathrm{d}w.$$

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

Main Theoretical Result

- **Proposition**: If $IF_h^{Q^*}(\sigma) < \infty$ then $IF_h^{\widehat{Q}}(\sigma) \leq IF_h^{Q^*}(\sigma)$ and

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

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

Main Theoretical Result

- **Proposition**: If $IF_h^{Q^*}(\sigma) < \infty$ then $IF_h^{\widehat{Q}}(\sigma) \leq IF_h^{Q^*}(\sigma)$ and

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

  where $\phi_n(\varphi, P)$ denotes the autocorrelation at lag n under a Markov kernel P.
- $\widetilde{Q}^{\mathrm{EX}}$ and $\widetilde{Q}_{\sigma}^{\mathrm{Z}}$ correspond to the jump kernels associated to Q^{EX} and Q_{σ}^{Z}, $\varrho_{\mathrm{EX}}(\theta)$ and $\varrho_{\mathrm{Z}}^{\sigma}(z)$ are acceptance proba of Q^{EX} and Q_{σ}^{Z}.

Main Theoretical Result

- **Proposition**: If $IF_h^{Q^*}(\sigma) < \infty$ then $IF_h^{\widehat{Q}}(\sigma) \leq IF_h^{Q^*}(\sigma)$ and

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

  where $\phi_n(\varphi, P)$ denotes the autocorrelation at lag n under a Markov kernel P.
- $\widetilde{Q}^{\mathrm{EX}}$ and $\widetilde{Q}_{\sigma}^{\mathrm{Z}}$ correspond to the jump kernels associated to Q^{EX} and Q_{σ}^{Z}, $\varrho_{\mathrm{EX}}(\theta)$ and $\varrho_{\mathrm{Z}}^{\sigma}(z)$ are acceptance proba of Q^{EX} and Q_{σ}^{Z}.
- This identity allows us to "decouple" the influence of the parameter and noise components 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

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

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

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

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

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

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

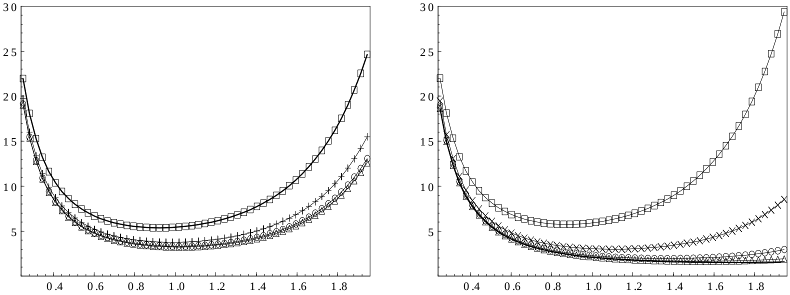
$$\frac{IF_h^{\widehat{Q}}(\sigma)}{IF_h^{\mathrm{EX}}} \leq \frac{IF_h^{Q^*}(\sigma)}{IF_h^{\mathrm{EX}}} \leq \frac{1}{2}(1 + 1/IF_h^{\mathrm{EX}})\pi_Z^{\sigma}(1/\varrho_Z^{\sigma}) - \frac{1}{IF_h^{\mathrm{EX}}}$$

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

$$\frac{IF_h^{Q^*}(\sigma)}{IF_h^{\mathrm{EX}}} \to \frac{1}{\pi_Z^{\sigma}(\varrho_Z^{\sigma})}.$$

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

Bounds on Relative Computational Time



Left: upper bound on $CT_h^{Q^*}(\sigma)\,/\,IF_h^{\mathrm{EX}}$ as a function of σ for $IF_h^{\mathrm{EX}} = 1$ (square), 4 (crosses), 20 (circles), 80 (triangles). Right: upper bounds on $CT_h^{Q^*}(\sigma)\,/\,IF_h^{\mathrm{EX}}$ as a function of σ for $IF^{\mathrm{EX}}_{J,h//\varrho_{\mathrm{EX}}} = 1$ for $IF^{\mathrm{EX}}_{J,h//\varrho_{\mathrm{EX}}} = 1, 4, 20, 80$ and lower bound (solid line).

Practical Guidelines

- For good proposals, select $\sigma \approx 1.0$ whereas for poor proposals, select $\sigma \approx 1.7$.

Practical Guidelines

- For good proposals, select $\sigma \approx 1.0$ whereas for poor proposals, select $\sigma \approx 1.7$.
- When you have no clue about the proposal efficiency,

Practical Guidelines

- For good proposals, select $\sigma \approx 1.0$ whereas for poor proposals, select $\sigma \approx 1.7$.
- When you have no clue about the proposal efficiency,

1. If $\sigma_{\text{opt}} = 1.0$ and you pick $\sigma = 1.7$, computing time increases by $\approx$150%.

Practical Guidelines

- For good proposals, select $\sigma \approx 1.0$ whereas for poor proposals, select $\sigma \approx 1.7$.
- When you have no clue about the proposal efficiency,

1. If $\sigma_{\text{opt}} = 1.0$ and you pick $\sigma = 1.7$, computing time increases by $\approx$150%.
2. If $\sigma_{\text{opt}} = 1.7$ and you pick $\sigma = 1.0$, computing time increases by $\approx$50%.

Practical Guidelines

- For good proposals, select $\sigma \approx 1.0$ whereas for poor proposals, select $\sigma \approx 1.7$.
- When you have no clue about the proposal efficiency,

1. If $\sigma_{\text{opt}} = 1.0$ and you pick $\sigma = 1.7$, computing time increases by $\approx$150%.
2. If $\sigma_{\text{opt}} = 1.7$ and you pick $\sigma = 1.0$, computing time increases by $\approx$50%.
3. If $\sigma_{\text{opt}} = 1.0$ or $\sigma_{\text{opt}} = 1.7$ and you pick $\sigma = 1.2 - 1.3$, 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

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

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

Relative Inefficiency and Computing Time

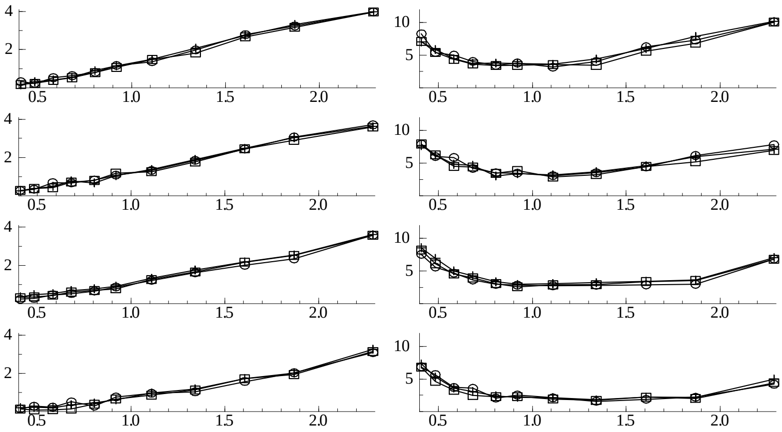


Figure: From left to right: RCT_h^Q vs N, RCT_h^Q vs $\sigma(\overline{\theta})$, RIF_h^Q against N and RIF_h^Q against $\sigma(\overline{\theta})$ for various values of ρ and different parameters.

Discussion

- Simplified quantitative analysis of the pseudo-marginal MH algorithm, useful in large data regime.

Discussion

- Simplified quantitative analysis of the pseudo-marginal MH algorithm, useful in large data regime.
- Optimal σ depends on efficiency of the ideal MH algorithm but $\sigma \approx 1.2$ is a sweet spot.

Discussion

- Simplified quantitative analysis of the pseudo-marginal MH algorithm, useful in large data regime.
- Optimal σ depends on efficiency of the ideal MH algorithm but $\sigma \approx 1.2$ is a sweet spot.
- Pseudo-marginal MH scales in $\mathcal{O}\left(T^2\right)$ as we require $N \propto T$, while simulated likelihood scales in $\mathcal{O}\left(T^{3/2}\right)$, i.e. $N \propto \sqrt{T}$.

Discussion

- Simplified quantitative analysis of the pseudo-marginal MH algorithm, useful in large data regime.
- Optimal σ depends on efficiency of the ideal MH algorithm but $\sigma \approx 1.2$ is a sweet spot.
- Pseudo-marginal MH scales in $\mathcal{O}\left(T^2\right)$ as we require $N \propto T$, while simulated likelihood scales in $\mathcal{O}\left(T^{3/2}\right)$, i.e. $N \propto \sqrt{T}$.
- However, pseudo-marginal MH much more generally applicable than simulated likelihood.

The Correlated Pseudo-Marginal Algorithm

- Reparameterize the likelihood estimator $\widehat{p}_\theta\left(y_{1:T}\right)$ as a function of normal variates $U \sim \mathcal{N}\left(0, I\right)$

$$\widehat{p}_\theta\left(y_{1:T}\right) = \widehat{p}_\theta\left(y_{1:T}; U\right)$$

The Correlated Pseudo-Marginal Algorithm

- Reparameterize the likelihood estimator $\widehat{p}_{\theta}\left(y_{1:T}\right)$ as a function of normal variates $U \sim \mathcal{N}\left(0, I\right)$

$$\widehat{p}_{\theta}\left(y_{1:T}\right) = \widehat{p}_{\theta}\left(y_{1:T}; U\right)$$

- Correlate estimators of $p_{\theta}\left(y_{1:T}\right)$ and $p_{\vartheta}\left(y_{1:T}\right)$ by setting

$$\widehat{p}_{\vartheta}\left(y_{1:T}\right) = \widehat{p}_{\vartheta}\left(y_{1:T}; V\right)$$

where

$$V = \rho U + \sqrt{1-\rho^{2}}\varepsilon,\ \varepsilon \sim \mathcal{N}\left(0, I\right)$$

for $\rho \in \left(-1, 1\right)$.

The Correlated Pseudo-Marginal Algorithm

- Reparameterize the likelihood estimator $\widehat{p}_{\theta}\left(y_{1:T}\right)$ as a function of normal variates $U \sim \mathcal{N}\left(0, I\right)$

$$\widehat{p}_{\theta}\left(y_{1:T}\right) = \widehat{p}_{\theta}\left(y_{1:T}; U\right)$$

- Correlate estimators of $p_{\theta}\left(y_{1:T}\right)$ and $p_{\vartheta}\left(y_{1:T}\right)$ by setting

$$\widehat{p}_{\vartheta}\left(y_{1:T}\right) = \widehat{p}_{\vartheta}\left(y_{1:T}; V\right)$$

  where

$$V = \rho U + \sqrt{1-\rho^{2}}\varepsilon,\ \varepsilon \sim \mathcal{N}\left(0, I\right)$$

  for $\rho \in \left(-1, 1\right)$.

- In practice, ρ will be select close to 1.

Correlated Pseudo-Marginal Metropolis–Hastings algorithm

Correlated Pseudo-Marginal Metropolis–Hastings algorithm

At iteration i

- Sample $\vartheta \sim q\left(\cdot \middle| \vartheta_{i-1}\right)$ and $V = \rho U_{i-1} + \sqrt{1-\rho^2}\varepsilon$, $\varepsilon \sim \mathcal{N}\left(0, I\right)$.

Correlated Pseudo-Marginal Metropolis–Hastings algorithm

At iteration i

- Sample $\vartheta \sim q\left(\cdot \middle| \vartheta_{i-1}\right)$ and $V = \rho U_{i-1} + \sqrt{1-\rho^2}\varepsilon$, $\varepsilon \sim \mathcal{N}\left(0, I\right)$.
- Compute the estimate $\widehat{p}_{\vartheta}\left(y_{1:T}; V\right)$ of $p_{\vartheta}\left(y_{1:T}\right)$.

Correlated Pseudo-Marginal Metropolis–Hastings algorithm

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

- Sample $\vartheta \sim q\left(\cdot \mid \vartheta_{i-1}\right)$ and $V = \rho U_{i-1} + \sqrt{1-\rho^2}\varepsilon$, $\varepsilon \sim \mathcal{N}\left(0, I\right)$.
- Compute the estimate $\widehat{p}_{\vartheta}\left(y_{1:T}; V\right)$ of $p_{\vartheta}\left(y_{1:T}\right)$.
- With probability

$$\min\{1, \frac{\widehat{p}_{\vartheta}\left(y_{1:T}; V\right)}{\widehat{p}_{\vartheta_{i-1}}\left(y_{1:T}; U_{i-1}\right)} \frac{p\left(\vartheta\right)}{p\left(\vartheta_{i-1}\right)} \frac{q\left(\vartheta_{i-1} \mid \vartheta\right)}{q\left(\vartheta \mid \vartheta_{i-1}\right)}\}$$

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

Analysis in the large data regime - i.i.d. case

Proposition. Let $N = N(T) \rightarrow \infty$ as $T \rightarrow \infty$ with $N = o(T)$. When $U \sim \overline{\pi}(\cdot|\theta)$ and $V = \rho_T U + \sqrt{1-\rho_T^2}\varepsilon$ with $\rho_T = \exp\left(-\psi \frac{N}{T}\right)$ then as $T \rightarrow \infty$

$$\log\left\{ \frac{\widehat{p}_{\theta+\xi/\sqrt{T}}(y_{1:T};V)}{\widehat{p}_{\theta}(y_{1:T};U)} \Big/ \frac{p_{\theta+\xi/\sqrt{T}}(y_{1:T})}{p_{\theta}(y_{1:T})} \right\} \Bigg| \mathcal{Y}^T, \mathcal{U}^T \Rightarrow \mathcal{N}\left(-\frac{\kappa^2(\theta)}{2}, \kappa^2(\theta)\right).$$

- This CLT is conditional on the observation sequence and the current auxiliary variables.

Analysis in the large data regime - i.i.d. case

Proposition. Let $N = N(T) \to \infty$ as $T \to \infty$ with $N = o(T)$. When $U \sim \overline{\pi}(\cdot|\theta)$ and $V = \rho_T U + \sqrt{1-\rho_T^2}\varepsilon$ with $\rho_T = \exp\left(-\psi \frac{N}{T}\right)$ then as $T \to \infty$

$$\log\left\{\frac{\widehat{p}_{\theta+\xi/\sqrt{T}}(y_{1:T};V)}{\widehat{p}_\theta(y_{1:T};U)} \Big/ \frac{p_{\theta+\xi/\sqrt{T}}(y_{1:T})}{p_\theta(y_{1:T})}\right\}\Bigg| \mathcal{Y}^T, \mathcal{U}^T \Rightarrow \mathcal{N}\left(-\frac{\kappa^2(\theta)}{2}, \kappa^2(\theta)\right).$$

- This CLT is conditional on the observation sequence and the current auxiliary variables.
- Asymptotically the distribution of the log-ratio decouples from the current location of the Markov chain.

Analysis in the large data regime - i.i.d. case

Proposition. Let $N = N(T) \to \infty$ as $T \to \infty$ with $N = o(T)$. When $U \sim \overline{\pi}(\cdot|\theta)$ and $V = \rho_T U + \sqrt{1-\rho_T^2}\varepsilon$ with $\rho_T = \exp\left(-\psi \frac{N}{T}\right)$ then as $T \to \infty$

$$\log\left\{\frac{\widehat{p}_{\theta+\xi/\sqrt{T}}(y_{1:T};V)}{\widehat{p}_{\theta}(y_{1:T};U)} \Big/ \frac{p_{\theta+\xi/\sqrt{T}}(y_{1:T})}{p_{\theta}(y_{1:T})}\right\}\Bigg|\, \mathcal{Y}^T, \mathcal{U}^T \Rightarrow \mathcal{N}\left(-\frac{\kappa^2(\theta)}{2}, \kappa^2(\theta)\right).$$

- This CLT is conditional on the observation sequence and the current auxiliary variables.
- Asymptotically the distribution of the log-ratio decouples from the current location of the Markov chain.
- The asymptotic variance is $O(1)$ even for $N \sim \log(T)$.

Analysis in the large data regime

- **Assumption 1** - Asymptotic Normality: We have

$$\int \left| p\left(\theta \middle| Y_{1:T}\right) - \phi(\theta; \widehat{\theta}^{T}, \Sigma / T) \right| \mathrm{d}\theta \xrightarrow{P} 0,$$

where $\widehat{\theta}^{T} \xrightarrow{P} \overline{\theta}$ and Σ is a p.d. matrix.

Analysis in the large data regime

- **Assumption 1** - Asymptotic Normality: We have

$$\int \left| p\left(\theta \right| Y_{1:T}\right) - \phi(\theta; \widehat{\theta}^{T}, \Sigma / T) \Big| \, \mathrm{d}\theta \xrightarrow{P} 0,$$

  where $\widehat{\theta}^{T} \xrightarrow{P} \overline{\theta}$ and Σ is a p.d. matrix.
- **Assumption 2** - Proposal: $\vartheta = \theta + \xi / \sqrt{T}$ where $\varepsilon \sim \upsilon\left(\cdot\right)$ with $\upsilon\left(\xi\right) = \upsilon\left(-\xi\right)$.

- **Assumption 1** - Asymptotic Normality: We have

$$\int \left| p\left(\theta \middle| Y_{1:T}\right) - \phi(\theta; \widehat{\theta}^{T}, \Sigma/T) \right| \mathrm{d}\theta \xrightarrow{P} 0,$$

  where $\widehat{\theta}^{T} \xrightarrow{P} \overline{\theta}$ and Σ is a p.d. matrix.
- **Assumption 2** - Proposal: $\vartheta = \theta + \xi/\sqrt{T}$ where $\varepsilon \sim \upsilon\left(\cdot\right)$ with $\upsilon\left(\xi\right) = \upsilon\left(-\xi\right)$.
- **Assumption 3** - For any θ in a neighbourhood of $\overline{\theta}$, the conditional CLT holds and $\kappa^2\left(\cdot\right)$ is continuous at $\overline{\theta}$.

Weak convergence

- Let $\{\vartheta_i^T\}_{i\geq 0}$ the stationary non-Markovian sequence of the correlated PM of invariant density $p\left(\theta \mid Y_{1:T}\right)$.

Weak convergence

- Let $\{\vartheta_i^T\}_{i\geq 0}$ the stationary non-Markovian sequence of the correlated PM of invariant density $p\left(\theta \middle| Y_{1:T}\right)$.
- **Proposition** (Deligiannidis et al., 2016): The F.D.D. of the rescaled sequence $\{\widetilde{\vartheta}_i^T = \sqrt{T}(\vartheta_i^T - \widehat{\theta}_T)\}_{i\geq 0}$ converge weakly as $T \to \infty$ to those of a stationary Markov chain of invariant density $\phi\left(\widetilde{\theta}; 0, \Sigma\right)$ and kernel given for $\widetilde{\theta} \neq \widetilde{\vartheta}$ by

$$\widetilde{Q}(\widetilde{\theta}, \mathrm{d}\widetilde{\vartheta}) = \upsilon(\widetilde{\vartheta} - \widetilde{\theta})\mathbb{E}_R\left[\min\left\{1, \frac{\phi(\widetilde{\vartheta}; 0, \Sigma)}{\phi(\widetilde{\theta}; 0, \Sigma)} R\right\}\right] \mathrm{d}\widetilde{\vartheta}$$

where $R \sim \mathcal{N}(-\kappa^2\left(\overline{\theta}\right)/2, \kappa^2\left(\overline{\theta}\right))$.

Weak convergence

- Let $\{\vartheta_i^T\}_{i\geq 0}$ the stationary non-Markovian sequence of the correlated PM of invariant density $p\left(\theta \middle| Y_{1:T}\right)$.
- **Proposition** (Deligiannidis et al., 2016): The F.D.D. of the rescaled sequence $\{\widetilde{\vartheta}_i^T = \sqrt{T}(\vartheta_i^T - \widehat{\theta}_T)\}_{i\geq 0}$ converge weakly as $T \to \infty$ to those of a stationary Markov chain of invariant density $\phi\left(\widetilde{\theta}; 0, \Sigma\right)$ and kernel given for $\widetilde{\theta} \neq \widetilde{\vartheta}$ by

$$\widetilde{Q}(\widetilde{\theta}, \mathrm{d}\widetilde{\vartheta}) = \upsilon(\widetilde{\vartheta} - \widetilde{\theta})\mathbb{E}_R\left[\min\left\{1, \frac{\phi(\widetilde{\vartheta}; 0, \Sigma)}{\phi(\widetilde{\theta}; 0, \Sigma)} R\right\}\right] \mathrm{d}\widetilde{\vartheta}$$

where $R \sim \mathcal{N}(-\kappa^2\left(\overline{\theta}\right)/2, \kappa^2\left(\overline{\theta}\right))$.
- These results suggests that a simplified analysis of the CPM chain can be performed by looking at

$$\widehat{Q}(\theta, \mathrm{d}\vartheta) = q\left(\vartheta \middle| \theta\right) \mathbb{E}_R\left[\min\left\{1, \frac{\pi\left(\vartheta\right)}{\pi\left(\theta\right)} R\right\}\right] \mathrm{d}\vartheta$$

where $R \sim \mathcal{N}(-\kappa^2/2, \kappa^2)$.

Breakdown

- An analysis based on this limiting kernel shows that one should select $\kappa^2 \approx 4.5$ to optimize the performance of the algorithm at fixed computational complexity.

Breakdown

- An analysis based on this limiting kernel shows that one should select $\kappa^2 \approx 4.5$ to optimize the performance of the algorithm at fixed computational complexity.
- Too good to be true? Can I really pick N arbitrarily?

Breakdown

- An analysis based on this limiting kernel shows that one should select $\kappa^2 \approx 4.5$ to optimize the performance of the algorithm at fixed computational complexity.
- Too good to be true? Can I really pick N arbitrarily?
- Weak convergence does NOT show that $\left| IF_h^Q - IF_h^{\widehat{Q}} \right| \to 0$.

Breakdown

- An analysis based on this limiting kernel shows that one should select $\kappa^2 \approx 4.5$ to optimize the performance of the algorithm at fixed computational complexity.
- Too good to be true? Can I really pick N arbitrarily?
- Weak convergence does NOT show that $\left| IF_h^Q - IF_h^{\widehat{Q}} \right| \to 0.$
- Informally, we have for $h(\theta) = \theta$

$$\mathbb{C}\text{ov}\left(\theta_0, \theta_\tau\right) \approx \underbrace{\mathbb{E}\left(\mathbb{C}\left(\theta_0, \theta_\tau \middle| U_0, U_\tau\right)\right)}_{\text{fast}} + \underbrace{\mathbb{C}\left(\mathbb{E}\left(\theta_0 \middle| U_0\right), \mathbb{E}\left(\theta_\tau \middle| U_\tau\right)\right)}_{\text{slow}}$$

where $\mathbb{E}\left(\theta_0 \middle| U_0\right) \approx \widehat{\theta}^T + \Sigma / T\ \nabla_\theta \log \widehat{p}_\theta\left(y_{1:T}; U\right) / p_\theta\left(y_{1:T}\right)\Big|_{\widehat{\theta}^T}$ and $IF_h^Q \to \infty$ if $N/\sqrt{T} \to 0$.

Breakdown

- An analysis based on this limiting kernel shows that one should select $\kappa^2 \approx 4.5$ to optimize the performance of the algorithm at fixed computational complexity.
- Too good to be true? Can I really pick N arbitrarily?
- Weak convergence does NOT show that $\left| IF_h^Q - IF_h^{\widehat{Q}} \right| \to 0.$
- Informally, we have for $h(\theta) = \theta$

$$\mathbb{C}\text{ov}(\theta_0, \theta_\tau) \approx \underbrace{\mathbb{E}\left(\mathbb{C}\left(\theta_0, \theta_\tau \middle| U_0, U_\tau\right)\right)}_{\text{fast}} + \underbrace{\mathbb{C}\left(\mathbb{E}\left(\theta_0 \middle| U_0\right), \mathbb{E}\left(\theta_\tau \middle| U_\tau\right)\right)}_{\text{slow}}$$

where $\mathbb{E}\left(\theta_0 \middle| U_0\right) \approx \widehat{\theta}^T + \Sigma / T \, \nabla_\theta \log \widehat{p}_\theta\left(y_{1:T}; U\right) / p_\theta\left(y_{1:T}\right)\Big|_{\widehat{\theta}^T}$. and $IF_h^Q \to \infty$ if $N/\sqrt{T} \to 0$.

- To ensure IF_h^Q, we need at least $N \propto \sqrt{T}$ and we conjecture it is sufficient.

Example: Gaussian Latent Variable Model

- Consider the toy model

$$X_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1), \quad Y_t \,|\, X_t \sim \mathcal{N}(X_t, \sigma^2).$$

Example: Gaussian Latent Variable Model

- Consider the toy model

$$X_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1), \quad Y_t | X_t \sim \mathcal{N}(X_t, \sigma^2).$$

- The likelihood can be computed exactly, allowing to implement the "exact" MH algorithm.

Example: Gaussian Latent Variable Model

- Consider the toy model

$$X_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1), \quad Y_t \,|\, X_t \sim \mathcal{N}(X_t, \sigma^2).$$

- The likelihood can be computed exactly, allowing to implement the "exact" MH algorithm.
- The likelihood estimator is based on importance sampling.

Example: Gaussian Latent Variable Model

- Consider the toy model

$$X_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1), \;\; Y_t | X_t \sim \mathcal{N}(X_t, \sigma^2).$$

- The likelihood can be computed exactly, allowing to implement the "exact" MH algorithm.
- The likelihood estimator is based on importance sampling.
- Integrated Autocorrelation Time is referred to as the Inefficiency IF.

Example: Gaussian Latent Variable Model

MH ($T = 8192$)		IF(θ)	
		15.6	
PM ($\rho = 0.0$)			
N		RIF(θ)	RCT(θ)
5000		2.2	11210
CPM ($\rho = 0.9963$)			
N	κ	RIF(θ)	RCT(θ)
9	3.1	14.0	126.2
12	2.7	8.3	99.7
20	2.2	4.7	93.3
25	2.0	2.8	69.3
35	1.7	1.7	61.1
56	1.3	1.6	87.0
80	1.1	1.1	89.0
120	0.9	0.9	113.5

Here $\mathrm{RIF} = \mathrm{IF}/\mathrm{IF}_{MH}$ and $\mathrm{RCT} = N \times \mathrm{RIF}$.

Discussion

- In i.i.d. case, very substantial improvement over the PM algorithm can be achieved by introducing a correlation scheme.

Discussion

- In i.i.d. case, very substantial improvement over the PM algorithm can be achieved by introducing a correlation scheme.
- Analysis suggests that complexity is $O\left(T\sqrt{T}\right)$ vs $O\left(T^2\right)$.

Discussion

- In i.i.d. case, very substantial improvement over the PM algorithm can be achieved by introducing a correlation scheme.
- Analysis suggests that complexity is $O\left(T\sqrt{T}\right)$ vs $O\left(T^2\right)$.
- In state-space models, implementation relies on non-standard particle filter scheme (Hilbert sorting): our analysis does not hold experimentally for state dimension > 1 and theoretically and but still substantial gains.

Discussion

- In i.i.d. case, very substantial improvement over the PM algorithm can be achieved by introducing a correlation scheme.
- Analysis suggests that complexity is $O\left(T\sqrt{T}\right)$ vs $O\left(T^2\right)$.
- In state-space models, implementation relies on non-standard particle filter scheme (Hilbert sorting): our analysis does not hold experimentally for state dimension > 1 and theoretically and but still substantial gains.
- Novel pseudo-marginal scheme using Conditional Sequential Monte Carlo (Andrieu, A.D., Yildirim, 2016) appears to suggest $O\left(T\right)$ is feasible.

Experimental results using conditional SMC

	Novel c-SMC PM		Standard PM	
	σ_v^2	σ_w^2	σ_v^2	σ_w^2
$T = 1000$	17.7	23.5	71.2	59.2
$T = 2000$	17.5	23.7	759.0	757.9
$T = 5000$	17.6	23.7	5808.6	5663.5
$T = 10000$	17.6	23.6	7368.1	7176.9

Estimated IACT on a nonlinear state-space model for $N = 200$ for novel c-SMC PM algorithm and $N = 2000$ for standard PM algorithm

Some References

- C. Andrieu, A.D. & R. Holenstein, “Particle Markov chain Monte Carlo Methods”, *JRSS* B, 2010.
- C. Andrieu & G.O. Roberts, “The Pseudo-Marginal Algorithm for Bayesian Computation”, *Ann. Stat.*, 2009.
- J. Berard, P. Del Moral & A.D., “A Lognormal CLT for Particle Approximations of Normalizing Constants”, *Electronic J. Proba.*, 2014.
- A.D., M.K. Pitt, G. Deligiannidis and R. Kohn, “Efficient Implementation of Markov Chain Monte Carlo when Using an Unbiased Likelihood Estimator”, *Biometrika*, 2015.
- L. Lin, K. Lin & J. Sloan, “A Noisy Monte Carlo Algorithm”, *Phys. Rev.* D, 2000.
- M.K. Pitt, R. Silva, P. Giordani & R. Kohn, “On Some Properties of MCMC Simulation Methods Based of the Particle Filter”, *J. Econometrics*, 2012.
- C. Sherlock, A. Thiery, G.O. Roberts & J.S. Rosenthal, “On the Efficiency of the RW Pseudo-Marginal MH”, *Ann. Stat.*, 2015.