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)

MCQMC, 19/08/2016

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 $(Y_t)_{t\geq 1}$ are i.i.d. random variables such that

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

where $(X_t)_{t\geq 1}$ are latent variables.

Latent Variable Models

- Assume $(Y_t)_{t\geq 1}$ are i.i.d. random variables such that

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

  where $(X_t)_{t\geq 1}$ are latent variables.
- The likelihood of $\theta \in \Theta \subseteq \mathbb{R}^d$ associated to observations $Y_{1:T} = y_{1:T}$ 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 $(Y_t)_{t\geq 1}$ are i.i.d. random variables such that

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

  where $(X_t)_{t\geq 1}$ are latent variables.
- The likelihood of $\theta \in \Theta \subseteq \mathbb{R}^d$ associated to observations $Y_{1:T} = y_{1:T}$ 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; e.g. multivariate probit model.

Latent Variable Models

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

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

Latent Variable 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 $\theta \in \mathbb{R}^d$ associated to $Y_{1:T} = y_{1:T}$ is

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

where

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

Latent Variable 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 $\theta \in \mathbb{R}^d$ associated to $Y_{1:T} = y_{1:T}$ 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 a very popular class of time series models but inference is difficult as $p_\theta(y_{1:T})$ is intractable for non-linear/non-Gaussian models.

Bayesian Inference

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

Bayesian Inference

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

Bayesian Inference

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

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

Bayesian Inference

- 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

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

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

Standard MCMC Approaches

- Standard MCMC schemes target

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

  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.

Standard MCMC Approaches

- Standard MCMC schemes target

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

  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.

Standard MCMC Approaches

- Standard MCMC schemes target

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

  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 mimic 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

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

Ideal Marginal Metropolis-Hastings

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

Ideal Marginal Metropolis-Hastings

- 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

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

At iteration i

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

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

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

Ideal Marginal Metropolis-Hastings

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

At iteration i

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

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

  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

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

Pseudo-Marginal Metropolis–Hastings

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

Pseudo-Marginal Metropolis–Hastings

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

At iteration i

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

Pseudo-Marginal Metropolis–Hastings

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

At iteration i

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

Pseudo-Marginal Metropolis–Hastings

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

At iteration i

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

$$
\begin{aligned}
& 1 \wedge \frac{\widehat{p}_\vartheta(y_{1:T})\, p(\vartheta)}{\widehat{p}_{\vartheta_{i-1}}(y_{1:T})\, p(\vartheta_{i-1})} \frac{q(\vartheta_{i-1} \mid \vartheta)}{q(\vartheta \mid \vartheta_{i-1})} \\
= \quad & 1 \wedge \frac{p_\vartheta(y_{1:T})\, p(\vartheta)}{p_{\vartheta_{i-1}}(y_{1:T})\, p(\vartheta_{i-1})} \frac{q(\vartheta_{i-1} \mid \vartheta)}{q(\vartheta \mid \vartheta_{i-1})} \frac{\widehat{p}_\vartheta(y_{1:T}) / p_\vartheta(y_{1:T})}{\widehat{p}_{\vartheta_{i-1}}(y_{1:T}) / p_{\vartheta_{i-1}}(y_{1:T})},
\end{aligned}
$$

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

Pseudo-Marginal Metropolis–Hastings

- **Proposition** (Lin, Liu & Sloan, *Phys. Rev.* D, 2000): If $\widehat{p}_\theta\left(y_{1:T}\right)$ 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.

Pseudo-Marginal Metropolis–Hastings

- **Proposition** (Lin, Liu & Sloan, *Phys. Rev.* D, 2000): If $\widehat{p}_\theta\left(y_{1:T}\right)$ 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 $\mathcal{U}$-valued 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 $U \sim m_\theta\left(\cdot\right)$.

Pseudo-Marginal Metropolis–Hastings

- **Proposition** (Lin, Liu & Sloan, *Phys. Rev.* D, 2000): If $\widehat{p}_\theta\left(y_{1:T}\right)$ 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 $\mathcal{U}$-valued 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 $U \sim m_\theta\left(\cdot\right)$.
- Pseudo-marginal MH is a standard MH on $\Theta \times \mathcal{U}$ with target $\overline{\pi}(\theta, u)$ and proposal $q\left(\vartheta \middle| \theta\right) m_\vartheta(v)$ where

$$\overline{\pi}(\theta, u) = \pi\left(\theta\right) \frac{\widehat{p}_\theta\left(y_{1:T}; u\right)}{p_\theta\left(y_{1:T}\right)} m_\theta\left(u\right)$$

satisfies

$$\int \overline{\pi}(\theta, u) \mathrm{d}u = \pi\left(\theta\right).$$

Likelihood Estimators

- For latent variable models, one has

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

Likelihood Estimators

- For latent variable models, one has

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

- A non-negative unbiased estimator is given by

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

i.e. $U = \left(X_1^1, ..., X_1^N, \ldots, X_T^1, ..., X_T^N\right)$ with

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

Likelihood Estimators

- For latent variable models, one has

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

- A non-negative unbiased estimator is given by

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

  i.e. $U = \left(X_1^1, ..., X_1^N, \ldots, X_T^1, ..., X_T^N\right)$ with

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

- Computational complexity is $O(NT)$.

Likelihood Estimators

- For state-space models, an alternative is to use particle filters where

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

where

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

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

Likelihood Estimators

- For state-space models, an alternative is to use particle filters where

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

  where

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

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

- The estimator is unbiased (Del Moral, 1998), relative variance is bounded uniformly over T if $N \propto T$ (Cerou, Del Moral & Guyader, 2011), SQMC could also be used (Gerber & Chopin, *JRSS* B 2015).

Likelihood Estimators

- For state-space models, an alternative is to use particle filters where

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

  where

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

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

- The estimator is unbiased (Del Moral, 1998), relative variance is bounded uniformly over T if $N \propto T$ (Cerou, Del Moral & Guyader, 2011), SQMC could also be used (Gerber & Chopin, *JRSS* B 2015).
- Computational complexity is $O\left(NT\right)$.

Examples of recent applications

- Measuring the impact of Ebola control measures in Sierra Leone, *PNAS*, 2016.

Examples of recent applications

- Measuring the impact of Ebola control measures in Sierra Leone, *PNAS*, 2016.
- Dynamic prediction pools: an investigation of financial frictions and forecasting performance, *J. Econometrics*, 2016.

Examples of recent applications

- Measuring the impact of Ebola control measures in Sierra Leone, *PNAS*, 2016.
- Dynamic prediction pools: an investigation of financial frictions and forecasting performance, *J. Econometrics*, 2016.
- Capturing the dynamics of pathogens with many strains, *J. Mathematical Biology*, 2016.

Examples of recent applications

- Measuring the impact of Ebola control measures in Sierra Leone, *PNAS*, 2016.
- Dynamic prediction pools: an investigation of financial frictions and forecasting performance, *J. Econometrics*, 2016.
- Capturing the dynamics of pathogens with many strains, *J. Mathematical Biology*, 2016.
- Monte Carlo estimation of stage structured development from cohort data, *Ecology*, 2016.

Examples of recent applications

- Measuring the impact of Ebola control measures in Sierra Leone, *PNAS*, 2016.
- Dynamic prediction pools: an investigation of financial frictions and forecasting performance, *J. Econometrics*, 2016.
- Capturing the dynamics of pathogens with many strains, *J. Mathematical Biology*, 2016.
- Monte Carlo estimation of stage structured development from cohort data, *Ecology*, 2016.
- Bayesian analysis of entry games: A simulated likelihood approach without simulation errors, *Global Economic Review*, 2016.

Examples of recent applications

- Measuring the impact of Ebola control measures in Sierra Leone, *PNAS*, 2016.
- Dynamic prediction pools: an investigation of financial frictions and forecasting performance, *J. Econometrics*, 2016.
- Capturing the dynamics of pathogens with many strains, *J. Mathematical Biology*, 2016.
- Monte Carlo estimation of stage structured development from cohort data, *Ecology*, 2016.
- Bayesian analysis of entry games: A simulated likelihood approach without simulation errors, *Global Economic Review*, 2016.
- Controlling procedural modeling programs. *SIGGRAPH*, 2015.

Empirical performance: Stochastic volatility model

- SV model with leverage (Huang & Tauchen, 2005):

$$\begin{aligned} \mathrm{d}v_1(s) &= -k_1\{v_1(s)-\mu_1\}\mathrm{d}s+\sigma_1\mathrm{d}W_1(s), \\ \mathrm{d}v_2(s) &= -k_2 v_2(s)\mathrm{d}s+\{1+\beta_{12}v_2(s)\}\mathrm{d}W_2(s), \\ \mathrm{d}\log P(s) &= \mu_y\mathrm{d}s+\text{s-exp}\left[\{v_1(s)+\beta_2 v_2(s)\}/2\right]\mathrm{d}B(s), \end{aligned}$$

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

Empirical performance: Stochastic volatility model

- SV model with leverage (Huang & Tauchen, 2005):

$$\begin{aligned} \mathrm{d}v_1(s) &= -k_1\{v_1(s)-\mu_1\}\mathrm{d}s+\sigma_1\mathrm{d}W_1(s), \\ \mathrm{d}v_2(s) &= -k_2 v_2(s)\mathrm{d}s+\{1+\beta_{12}v_2(s)\}\mathrm{d}W_2(s), \\ \mathrm{d}\log P(s) &= \mu_y\mathrm{d}s+\text{s-exp}\left[\{v_1(s)+\beta_2 v_2(s)\}/2\right]\mathrm{d}B(s), \end{aligned}$$

  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

- SV model with leverage (Huang & Tauchen, 2005):

$$\begin{aligned} \mathrm{d}v_1(s) &= -k_1\{v_1(s)-\mu_1\}\mathrm{d}s+\sigma_1\mathrm{d}W_1(s), \\ \mathrm{d}v_2(s) &= -k_2 v_2(s)\mathrm{d}s+\{1+\beta_{12}v_2(s)\}\mathrm{d}W_2(s), \\ \mathrm{d}\log P(s) &= \mu_y\mathrm{d}s+\text{s-exp}\left[\{v_1(s)+\beta_2 v_2(s)\}/2\right]\mathrm{d}B(s), \end{aligned}$$

  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.

Empirical performance: Stochastic volatility model

- SV model with leverage (Huang & Tauchen, 2005):

$$\begin{aligned} \mathrm{d}v_1(s) &= -k_1\{v_1(s)-\mu_1\}\mathrm{d}s+\sigma_1\mathrm{d}W_1(s), \\ \mathrm{d}v_2(s) &= -k_2 v_2(s)\mathrm{d}s+\{1+\beta_{12}v_2(s)\}\mathrm{d}W_2(s), \\ \mathrm{d}\log P(s) &= \mu_y\mathrm{d}s+\text{s-exp}\left[\{v_1(s)+\beta_2 v_2(s)\}/2\right]\mathrm{d}B(s), \end{aligned}$$

  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 $\theta = \left(k_1, \mu_1, \sigma_1, k_2, \beta_{12}, \beta_2, \mu_y, \phi_1, \phi_2\right)$.

Empirical performance: Stochastic volatility model

- SV model with leverage (Huang & Tauchen, 2005):

$$\begin{aligned} \mathrm{d}v_1(s) &= -k_1\{v_1(s)-\mu_1\}\mathrm{d}s+\sigma_1\mathrm{d}W_1(s), \\ \mathrm{d}v_2(s) &= -k_2 v_2(s)\mathrm{d}s+\{1+\beta_{12}v_2(s)\}\mathrm{d}W_2(s), \\ \mathrm{d}\log P(s) &= \mu_y\mathrm{d}s+\text{s-exp}\left[\{v_1(s)+\beta_2 v_2(s)\}/2\right]\mathrm{d}B(s), \end{aligned}$$

  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 MH 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

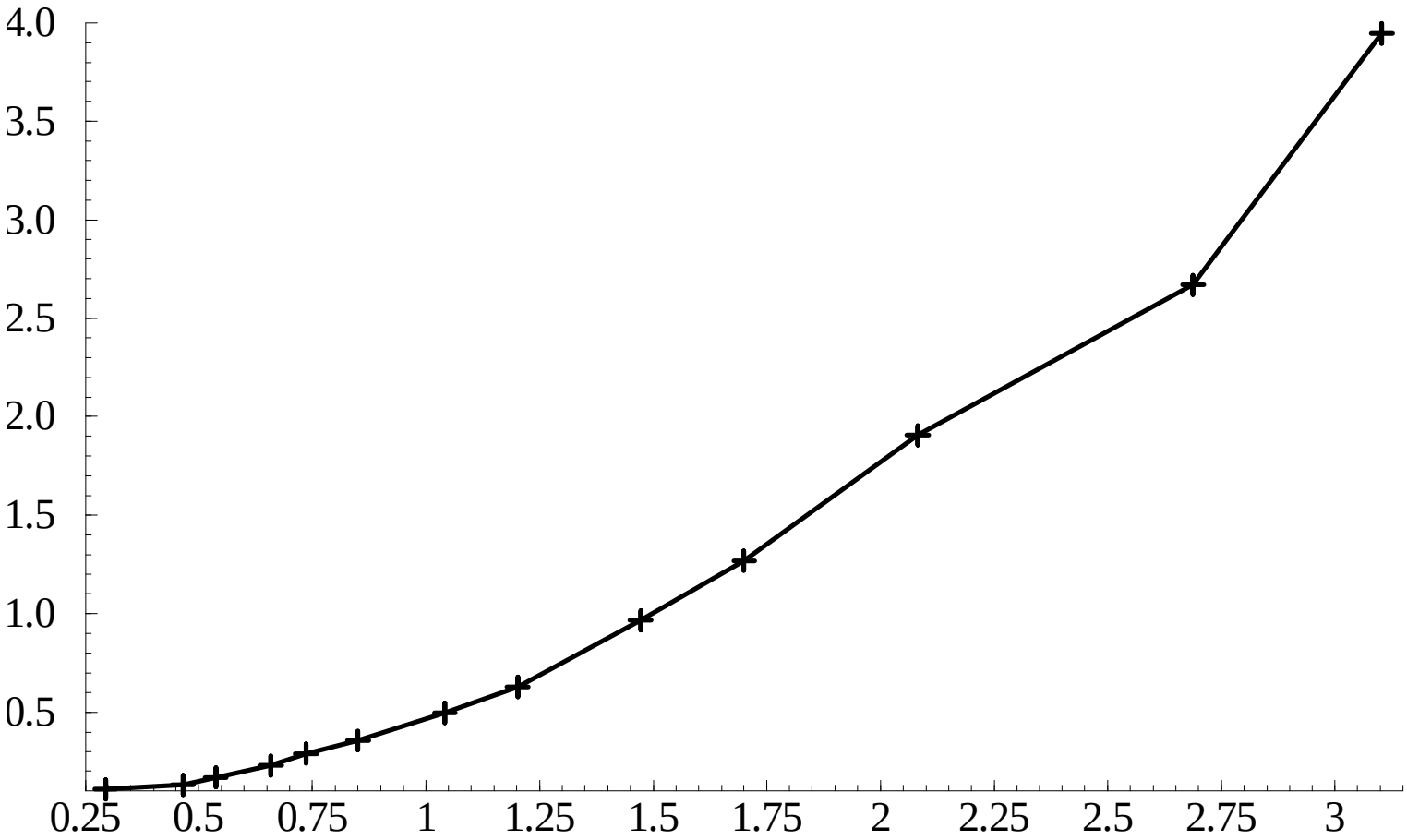

Figure: Average over the 9 parameter components of the log-integrated autocorrelation time of pseudo-marginal chain as a function of σ for $T = 300$.

How precise should the log-likelihood estimator be?

- **Aim**: Minimize the "computational time"

$$\mathrm{CT}\left(Q, h\right) = \mathrm{IACT}\left(Q, h\right) / \sigma^2$$

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

$$\mathrm{IACT}\left(Q, h\right) \;=\; \text{Integrated Autocorrelation Time of } \left\{h\left(\vartheta_i\right)\right\}_{i \geq 1}$$

How precise should the log-likelihood estimator be?

- **Aim**: Minimize the "computational time"

$$\mathrm{CT}\left(Q,h\right)=\mathrm{IACT}\left(Q,h\right)/\sigma^2$$

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

$$\mathrm{IACT}\left(Q,h\right) \;=\; \text{Integrated Autocorrelation Time of } \left\{h\left(\vartheta_i\right)\right\}_{i\geq 1}$$

- The IACT is

$$\mathrm{IACT}\left(Q,h\right)=1+2\sum_{\tau=1}^{\infty}\mathrm{corr}_{\overline{\pi},Q}\left\{h\left(\theta_0\right),h\left(\theta_\tau\right)\right\}$$

where, for $(\theta,z)\neq(\vartheta,w)$, the PM kernel Q is

$$Q\left\{\left(\theta,z\right),\left(\mathrm{d}\vartheta,\mathrm{d}w\right)\right\}=q(\vartheta|\theta)g_\vartheta(w)\left\{1\wedge\frac{\pi\left(\vartheta\right)}{\pi\left(\theta\right)}\exp\left(w-z\right)\right\}\mathrm{d}\vartheta\mathrm{d}w$$

with

$$z=\log\{\widehat{p}_\theta(y_{1:T})/p_\theta(y_{1:T})\},\qquad w=\log\{\widehat{p}_\vartheta(y_{1:T})/p_\vartheta(y_{1:T})\}.$$

Computational time for the SV model

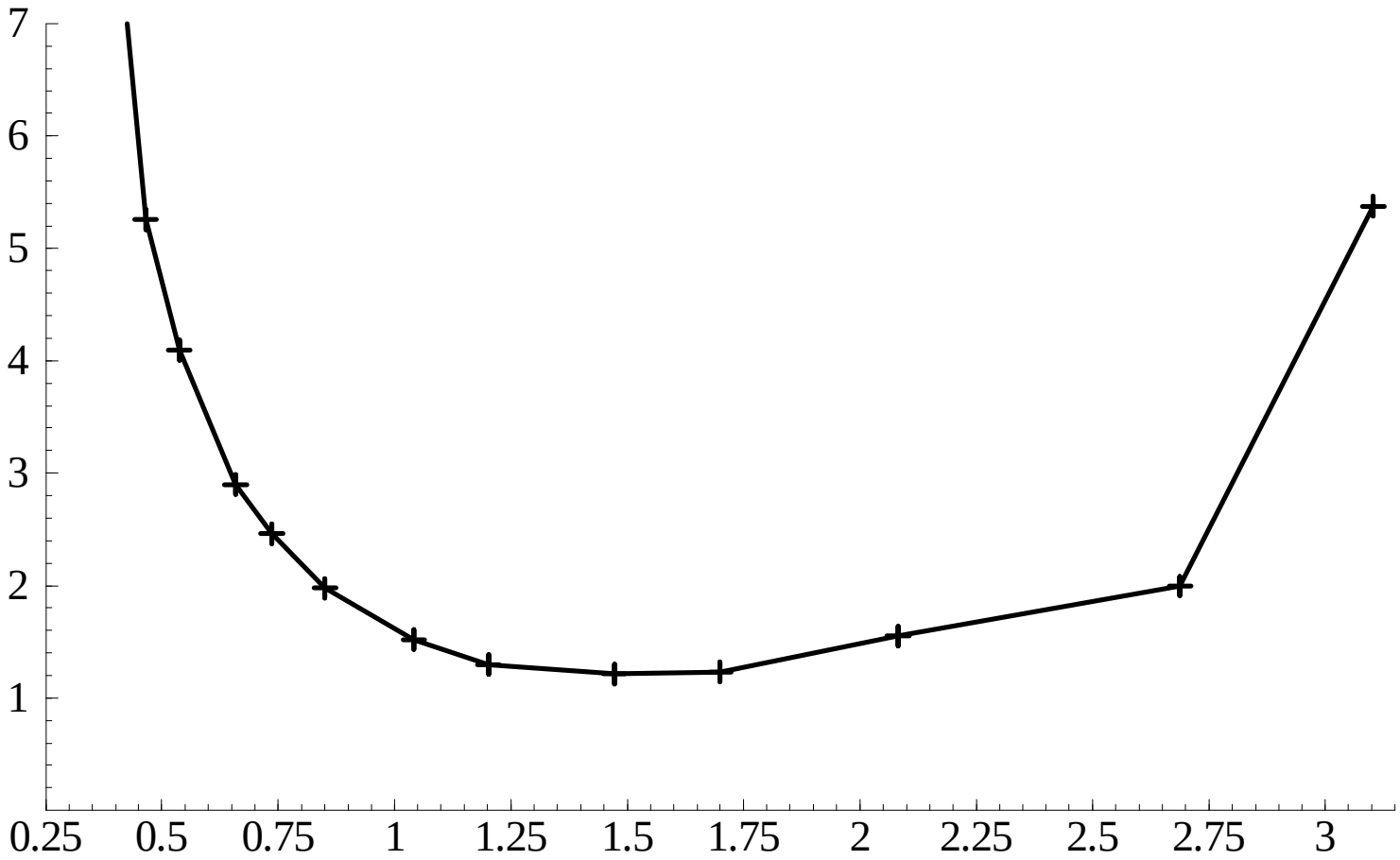

Figure: Computational time as a function of σ

Large sample analysis of the pseudo-marginal kernel

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

Large sample analysis of the pseudo-marginal kernel

- 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.

Large sample analysis of the pseudo-marginal kernel

- 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** - *Proposal*: $\vartheta = \theta + \varepsilon / \sqrt{T}$ where $\varepsilon \sim v\left(\cdot\right)$ with $v\left(\varepsilon\right) = v\left(-\varepsilon\right)$.

Large sample analysis of the pseudo-marginal kernel

- 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

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

  where $\widehat{\theta}^T \xrightarrow{P} \overline{\theta}$ and Σ is a p.d. matrix.
- **Assumption 2** - *Proposal*: $\vartheta = \theta + \varepsilon / \sqrt{T}$ where $\varepsilon \sim v\left(\cdot\right)$ with $v\left(\varepsilon\right) = v\left(-\varepsilon\right)$.
- **Proposition** (Berard, Del Moral & D., *EJP*, 2014): Under regularity conditions, we have if $N \propto T$

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

Large sample analysis of the pseudo-marginal kernel

- 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 targetting $\pi^T\left(\theta, z\right) = p\left(\theta \middle| Y_{1:T}\right) \exp\left(z\right) g_\theta^T\left(z\right)$.

Large sample analysis of the pseudo-marginal kernel

- 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 targetting $\pi^T(\theta, z) = p(\theta|\, Y_{1:T}) \exp(z)\, g_\theta^T(z)$.
- **Proposition**: The F.D.D. of $\{\widetilde{\vartheta}_i^T = \sqrt{T}(\vartheta_i^T - \widehat{\theta}_T), Z_i^T\}_{i\geq 0}$ 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

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

Large sample analysis of the pseudo-marginal kernel

- 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 targetting $\pi^T(\theta, z) = p(\theta | Y_{1:T}) \exp(z) g_\theta^T(z)$.
- **Proposition**: The F.D.D. of $\{\widetilde{\vartheta}_i^T = \sqrt{T}(\vartheta_i^T - \widehat{\theta}_T), Z_i^T\}_{i \geq 0}$ 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

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

- Let $\sigma^2 = \sigma^2\left(\overline{\theta}\right)$, it suggests a simplified analysis based on

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

Available Results

- **Aim**: Minimize the computational cost

$$\mathrm{CT}(\widehat{Q}_\sigma, h) = \frac{\mathrm{IACT}(\widehat{Q}_\sigma, h)}{\sigma^2}.$$

Available Results

- **Aim**: Minimize the computational cost

$$\mathrm{CT}(\widehat{Q}_\sigma, h) = \frac{\mathrm{IACT}(\widehat{Q}_\sigma, h)}{\sigma^2}.$$

- **Special cases**:

Available Results

- **Aim**: Minimize the computational cost

$$\mathrm{CT}(\widehat{Q}_\sigma, h) = \frac{\mathrm{IACT}(\widehat{Q}_\sigma, h)}{\sigma^2}.$$

- **Special cases**:

1. When $q(\vartheta|\theta) = p(\vartheta|y)$, $\sigma_{\mathrm{opt}} = 0.9$ (Pitt et al., 2012).

Available Results

- **Aim**: Minimize the computational cost

$$\mathrm{CT}(\widehat{Q}_\sigma, h) = \frac{\mathrm{IACT}(\widehat{Q}_\sigma, h)}{\sigma^2}.$$

- **Special cases**:

1. When $q(\vartheta|\theta) = p(\vartheta|y)$, $\sigma_{\text{opt}} = 0.9$ (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 $d \longrightarrow \infty$, diffusion limit suggests $\sigma_{\text{opt}} = 1.8$ (Sherlock et al., 2015).

Sketch of the Analysis

- For general proposals and targets, direct minimization of $\mathrm{CT}(\widehat{Q}_\sigma, h)$ impossible so minimize an upper bound over it.

Sketch of the Analysis

- For general proposals and targets, direct minimization of $\mathrm{CT}(\widehat{Q}_\sigma, h)$ impossible so minimize an upper bound over it.
- Theoretical study relies on $\overline{\pi}-$invariant kernel Q_σ^* 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 $\mathrm{CT}(\widehat{Q}_\sigma, h)$ impossible so minimize an upper bound over it.
- Theoretical study relies on $\overline{\pi}-$invariant kernel Q_σ^* 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

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

- We have $\mathrm{IACT}(\widehat{Q}_\sigma, h) \leq \mathrm{IACT}(Q_\sigma^*, h)$ by Peskun (1973) so $\mathrm{CT}(\widehat{Q}_\sigma, h) \leq \mathrm{CT}(Q_\sigma^*, h)$ and exact expression of $\mathrm{IACT}(Q_\sigma^*, h)$ is available.

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%.

Discussion

- Many sharp qualitative results for the pseudo-marginal MH algorithm have been obtained (Andrieu & Roberts, AoS 2009; Andrieu & Vihola, AAP 2015).

Discussion

- Many sharp qualitative results for the pseudo-marginal MH algorithm have been obtained (Andrieu & Roberts, AoS 2009; Andrieu & Vihola, AAP 2015).
- Simplified quantitative analysis of the pseudo-marginal MH algorithm, useful in large data regime.

Discussion

- Many sharp qualitative results for the pseudo-marginal MH algorithm have been obtained (Andrieu & Roberts, AoS 2009; Andrieu & Vihola, AAP 2015).
- 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 $\sigma \approx 1.2 - 1.3$ is a sweet spot.

Discussion

- Many sharp qualitative results for the pseudo-marginal MH algorithm have been obtained (Andrieu & Roberts, AoS 2009; Andrieu & Vihola, AAP 2015).
- 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 $\sigma \approx 1.2 - 1.3$ is a sweet spot.
- Pseudo-marginal MH scales in $\mathcal{O}\left(T^2\right)$ at each iteration as we require $N \propto T$.

Discussion

- Many sharp qualitative results for the pseudo-marginal MH algorithm have been obtained (Andrieu & Roberts, AoS 2009; Andrieu & Vihola, AAP 2015).
- 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 $\sigma \approx 1.2 - 1.3$ is a sweet spot.
- Pseudo-marginal MH scales in $\mathcal{O}\left(T^2\right)$ at each iteration as we require $N \propto T$.
- **Problem**: the ratio $p_\vartheta\left(y_{1:T}\right) / p_\theta\left(y_{1:T}\right)$ is estimated by estimating independently the numerator and denominator.

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 $\rho \in (-1, 1)$.

The Correlated Pseudo-Marginal Algorithm

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

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

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

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

where

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

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

- In practice, ρ will be selected 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

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 $p_{\vartheta}\left(y_{1:T}\right)$.
- With probability

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

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

Large sample analysis of the correlated PM - i.i.d. case

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

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

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

Large sample analysis of the correlated PM - i.i.d. case

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

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

- 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.

Large sample analysis of the correlated PM - i.i.d. case

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

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

- 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 $N \sim \log(T)$.

Large sample analysis of the correlated PM - i.i.d. case

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

Large sample analysis of the correlated PM - i.i.d. case

- 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**. The F.D.D. of $\{\widetilde{\vartheta}_i^T = \sqrt{T}(\vartheta_i^T - \widehat{\theta}_T)\}_{i\geq 0}$ converge weakly as $T \longrightarrow \infty$ 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

$$\widetilde{Q}(\widetilde{\theta}, \mathrm{d}\widetilde{\vartheta}) = v(\widetilde{\vartheta} - \widetilde{\theta})\mathbb{E}_{R\sim\mathcal{N}\left(-\kappa^2\left(\overline{\theta}\right)/2, \kappa^2\left(\overline{\theta}\right)\right)}\left[\min\left\{1, \frac{\phi(\widetilde{\vartheta}; 0, \Sigma)}{\phi(\widetilde{\theta}; 0, \Sigma)} R\right\}\right] \mathrm{d}\widetilde{\vartheta}.$$

Large sample analysis of the correlated PM - i.i.d. case

- Let $\{\vartheta_i^T\}_{i\geq 0}$ the stationary *non-Markovian* sequence of the correlated PM of invariant density $p\left(\theta|\, Y_{1:T}\right)$.
- **Proposition**. The F.D.D. of $\{\widetilde{\vartheta}_i^T = \sqrt{T}(\vartheta_i^T - \widehat{\theta}_T)\}_{i\geq 0}$ 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

$$\widetilde{Q}(\widetilde{\theta}, \mathrm{d}\widetilde{\vartheta}) = v(\widetilde{\vartheta} - \widetilde{\theta})\mathbb{E}_{R\sim\mathcal{N}\left(-\kappa^2\left(\bar{\theta}\right)/2, \kappa^2\left(\bar{\theta}\right)\right)}\left[\min\left\{1, \frac{\phi(\widetilde{\vartheta}; 0, \Sigma)}{\phi(\widetilde{\theta}; 0, \Sigma)} R\right\}\right] \mathrm{d}\widetilde{\vartheta}.$$

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

$$\widehat{Q}(\theta, \mathrm{d}\vartheta) = q\left(\vartheta|\,\theta\right) \mathbb{E}_{R\sim\mathcal{N}\left(-\kappa^2\left(\bar{\theta}\right)/2, \kappa^2\left(\bar{\theta}\right)\right)}\left[\min\left\{1, \frac{\pi\left(\vartheta\right)}{\pi\left(\theta\right)} R\right\}\right] \mathrm{d}\vartheta$$

and $\kappa \approx 2.1$ minimizes computational time.

Limitations of Weak Convergence

- Previous result does not imply convergence of IACT.

Limitations of Weak Convergence

- Previous result does not imply convergence of IACT.
- In the CPM context, the auxiliary variables U^T evolves according to an AR scheme with persistency $\approx 1 - N_T / T$.

Limitations of Weak Convergence

- Previous result does not imply convergence of IACT.
- In the CPM context, the auxiliary variables U^T evolves according to an AR scheme with persistency $\approx 1 - N_T / T$.
- When N_T grows to slowly with T, correlations decay too slowy leading to increasing IACT.

Limitations of Weak Convergence

- Previous result does not imply convergence of IACT.
- In the CPM context, the auxiliary variables U^T evolves according to an AR scheme with persistency $\approx 1 - N_T / T$.
- When N_T grows to slowly with T, correlations decay too slowy leading to increasing IACT.
- For the i.i.d. latent variable case, $\mathrm{IACT}(Q, h)$ is driven by the IACT of $\left\{\Psi_T\left(U_i^T\right) := \nabla_\theta \log \widehat{p}\left(Y_{1:T}; U_i^T\right)\right\}_{i \geq 1}$.

Limitations of Weak Convergence

- Previous result does not imply convergence of IACT.
- In the CPM context, the auxiliary variables U^T evolves according to an AR scheme with persistency $\approx 1 - N_T / T$.
- When N_T grows to slowly with T, correlations decay too slowy leading to increasing IACT.
- For the i.i.d. latent variable case, $\mathrm{IACT}(Q, h)$ is driven by the IACT of $\left\{\Psi_T\left(U_i^T\right) := \nabla_\theta \log \widehat{p}\left(Y_{1:T}; U_i^T\right)\right\}_{i \geq 1}$.
- **Proposition**. Let $N_T \propto T^\alpha$ for $0 < \alpha < 1$ then $\mathrm{IACT}(Q, h) \gtrsim T^{1-2\alpha}$.

Limitations of Weak Convergence

- Previous result does not imply convergence of IACT.
- In the CPM context, the auxiliary variables U^T evolves according to an AR scheme with persistency $\approx 1 - N_T / T$.
- When N_T grows to slowly with T, correlations decay too slowy leading to increasing IACT.
- For the i.i.d. latent variable case, $\mathrm{IACT}(Q, h)$ is driven by the IACT of $\left\{\Psi_T\left(U_i^T\right) := \nabla_\theta \log \widehat{p}\left(Y_{1:T}; U_i^T\right)\right\}_{i \geq 1}$.
- **Proposition**. Let $N_T \propto T^\alpha$ for $0 < \alpha < 1$ then $\mathrm{IACT}(Q, h) \gtrsim T^{1-2\alpha}$.
- This result suggests we need at least $\sqrt{T} / N_T = O(1)$.

Example: Gaussian Latent Variable Model

MH ($T = 8192$)		IACT(θ)	
		15.6	
PM ($\rho = 0.0$)			
N		RIACT(θ)	RCT(θ)
5000		2.2	11210
CPM ($\rho = 0.9963$)			
N	κ	RIACT(θ)	RCT(θ)
10	3.1	14.0	126.2
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

Here $\mathrm{RIACT} = \mathrm{IACT}/\mathrm{IACT}_{MH}$ and $\mathrm{RCT} = N \times \mathrm{RIACT}$. Improvement by 180 fold.

Example: Noisy Autoregressive Model

MH ($T = 16,000$)		IACT(θ)	
		5.8	
PM ($\rho = 0.0$)			
N		RIACT(θ)	RCT(θ)
2500		3.1	8427.0
CPM ($\rho = 0.9965$)			
N	κ	RIACT(θ)	RCT(θ)
6	6.7	43.8	262.8
10	3.3	8.7	86.7
16	1.9	6.0	85.8
22	1.3	3.9	85.6
35	0.8	2.4	85.0
40	0.7	2.4	94.8

Improvement by 100 fold.

Discussion

- Large sample analysis of pseudo-marginal algorithm provides useful guidelines, overall complexity is $O\left(T^2\right)$.

Discussion

- Large sample analysis of pseudo-marginal algorithm provides useful guidelines, overall complexity is $O\left(T^2\right)$.
- Correlated pseudo-marginal can achieve very substantial improvement.

Discussion

- Large sample analysis of pseudo-marginal algorithm provides useful guidelines, overall complexity is $O\left(T^2\right)$.
- Correlated pseudo-marginal can achieve very substantial improvement.
- Analysis shows $\sqrt{T}/N_T = O(1)$ is necessary and we conjecture it is sufficient leading to complexity $O(T^{3/2})$ vs $O\left(T^2\right)$.

Discussion

- Large sample analysis of pseudo-marginal algorithm provides useful guidelines, overall complexity is $O\left(T^2\right)$.
- Correlated pseudo-marginal can achieve very substantial improvement.
- Analysis shows $\sqrt{T}/N_T = O(1)$ is necessary and we conjecture it is sufficient leading to complexity $O(T^{3/2})$ vs $O\left(T^2\right)$.
- Implementation for state-space models in state dimension $n > 1$ relies on non-standard particle scheme (Gerber & Chopin, 2015): our analysis does not capture these cases, experimental results suggest $O(T^{1+\frac{n}{n+1}})$.

Some References

- C. Andrieu & M Vihola, "Convergence properties of pseudo-margninal MCMC", *Ann. Applied Proba.*, 2015.
- J. Berard, P. Del Moral & A.D., "A Lognormal CLT for Particle Approximations of Normalizing Constants", *Elec. J. Proba.*, 2014.
- G. Deligiannidis, A.D. and M.K. Pitt, "The Correlated Pseudo-marginal Method", arXiv:1511.04992, 2015.
- 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.
- C. Sherlock, A. Thiery, G.O. Roberts & J.S. Rosenthal, "On the Efficiency of the RW Pseudo-Marginal MH", *Ann. Stat.*, 2015.