Pseudo-marginal MH using averages of unbiased estimators

Joint work with Alex Thiery (NUS)

Chris Sherlock

Department of Mathematics and Statistics
Lancaster University

April 2016

Example set up

Imagine:

y	data;
x	parameters of a model (interest);
v	auxiliary (latent) variables (nuisance)
$p(y\|x,v) = p(y\|v,x)p_V(\mathrm{d}v\|x)$	model
$\pi_0(x)$	prior

Ideally we'd use the Metropolis-Hastings (MH) algorithm to target

$$\pi(x) \propto \pi_0(x)p(y|x) = \pi_0(x)\int p(y|x,v)p_V(\mathrm{d}v|x),$$

but the integral is intractable.

We can, however create a non-negative, unbiased estimator of $p(y|x)$, for example

$$\hat{p}(y|x,V) := p(y|x,V) \quad \text{where} \quad V \sim p_V(\mathrm{d}v|x).$$

The PMMH algorithm

Now, let $\hat{p}(y|x, V) \geq 0$ be any unbiased estimator of $p(y|x)$, where $V \sim p_V(dv|x)$ are auxiliary variables (e.g. from importance sampling; particle filter; Rhee/Glynn). Then

$$\hat{\pi}(x; V) = \pi_0(x)\hat{p}(y|x, V)$$

is an unbiased estimator of $\pi(x)$ up to some fixed constant.

Given a current value, x and a realisation $\hat{\pi} = \hat{\pi}(x; v)$, one iteration of the PMMH algorithm is:

PMMH Algorithm

1. Sample x' from some density $q(x, x')$.
2. Sample $\hat{\pi}'$ from unbiased estimator, $\hat{\pi}(x'; V')$ of $\pi(x')$.
3. Let
$$\alpha = 1 \wedge \frac{\hat{\pi}' q(x', x)}{\hat{\pi} q(x, x')}.$$
4. W.p. α set $x \leftarrow x'$ and $\hat{\pi} \leftarrow \hat{\pi}'$ else keep x and $\hat{\pi}$ unchanged.

Averages of estimators

Instead of a single realisation, $\hat{\pi}(x; v)$, of an unbiased estimator, we could create m such realisations, $\hat{\pi}(x; v_1), \ldots, \hat{\pi}(x; v_m)$. Their average

$$\hat{\pi}_m = \frac{1}{m} \sum_{j=1}^{m} \hat{\pi}(x; v_j)$$

is a realisation from a new unbiased estimator, which could be used in a PMMH algorithm.

Is this worth doing?

Outline

The normalised weight, W

The PMMH algorithm creates a Markov chain on (x, v); the stationary distribution is: $p_V(x, \mathrm{d}v)\hat{\pi}(x; v)\mathrm{d}x$.

Let $W := \hat{\pi}(x; V)/\pi(x) \in \mathsf{W}$, so (WLOG) $\mathbb{E}[W] = 1$. The PMMH creates a Markov chain on (x, w); the stationary distribution is:

$$\tilde{\pi}(\mathrm{d}x, \mathrm{d}w) := \pi(x)\mathrm{d}x\mathrm{q}_1(x, \mathrm{d}w)w.$$

Given a current value, x and a realisation $\hat{\pi} = \pi(x)w$, one iteration of the PMMH algorithm is:

PMMH Algorithm

1. Sample x' from some density $q(x, x')$.
2. Sample w' from $\mathrm{q}(x', \mathrm{d}w')$.
3. Let
$$\alpha = 1 \wedge \frac{\pi(x')w'q(x', x)}{\pi(x)wq(x, x')} = 1 \wedge r(x, x')\frac{w'}{w}.$$
4. W.p. α set $x \leftarrow x'$ and $w \leftarrow w'$ else keep x and w unchanged.

Vector of normalised weights, $\underline{W}$

Alternatively we could sample a vector of m estimates, $\underline{W}$ from

$$\mathrm{q}(x, \mathrm{d}\underline{w}) := \prod_{j=1}^{m} \mathrm{q}_1(x, \mathrm{d}w_j).$$

$\frac{1}{m}\sum_{j=1}^{m} w_j$ represents a realisation from a new unbiased estimator. The stationary distribution is

$$\tilde{\pi}(\mathrm{d}x, \mathrm{d}\underline{w}) := \pi(x)\mathrm{d}x\mathrm{q}(x, \mathrm{d}\underline{w})\frac{1}{m}\sum_{j=1}^{m} w_j.$$

Denote the kernels by $\mathrm{P}_1(x, w; \mathrm{d}x', \mathrm{d}w')$ and $\mathrm{P}_m(x, \underline{w}; \mathrm{d}x', \mathrm{d}\underline{w}')$.

Measures of interest

Conditional acceptance probability:

$$\alpha(x, x'|\mathrm{P}) := \int \mathrm{q}(x, \mathrm{d}w)w\mathrm{q}(x', \mathrm{d}w')\left[1 \wedge r(x, x')\frac{w'}{w}\right]$$

Dirichlet form:

$$\mathcal{E}_{\mathrm{P}}(f) := \frac{1}{2}\int \pi(x)\mathrm{d}xq(x, x')\mathrm{d}x' \int \mathrm{q}(x, \mathrm{d}w)w\mathrm{q}(x', \mathrm{d}w') \left[1 \wedge r(x, x')\frac{w'}{w}\right]\left[f(x, w) - f(x', w')\right]^2.$$

Spectral gap:

$$\inf_{f \in L_0^2(\tilde{\pi}), \langle f, f\rangle = 1} \mathcal{E}_{\mathrm{P}}(f).$$

Asymptotic variance:

$$\mathrm{Var}(f, \mathrm{P}) := \lim_{n\to\infty} \mathrm{Var}\left(n^{-1/2}\sum_{i=1}^{n} f(X_i)\right).$$

Andrieu and Vihola, 2015.

AV2015: Theorem 10 + Corollary 31

1. For any $x, x' \in \mathsf{X}$ the conditional acceptance rates satisfy $\alpha^*(x, x'|\mathsf{P}_m) \geq \alpha^*(x, x'|\mathsf{P}_1)$.
2. For any $f : \mathsf{X} \to \mathbb{R}$, the Dirichlet forms satisfy $\mathcal{E}_{\mathsf{P}_m}(f) \geq \mathcal{E}_{\mathsf{P}_1}(f)$.
3. $\mathrm{Gap}(\mathsf{P}_m) \geq \mathrm{Gap}(\mathsf{P}_1)$.
4. For any $f : \mathsf{X} \to \mathbb{R}$ with $\mathrm{Var}_\pi(f) < \infty$, the asymptotic variances satisfy $\mathrm{Var}(f, \mathsf{P}_m) \leq \mathrm{Var}(f, \mathsf{P}_1)$.

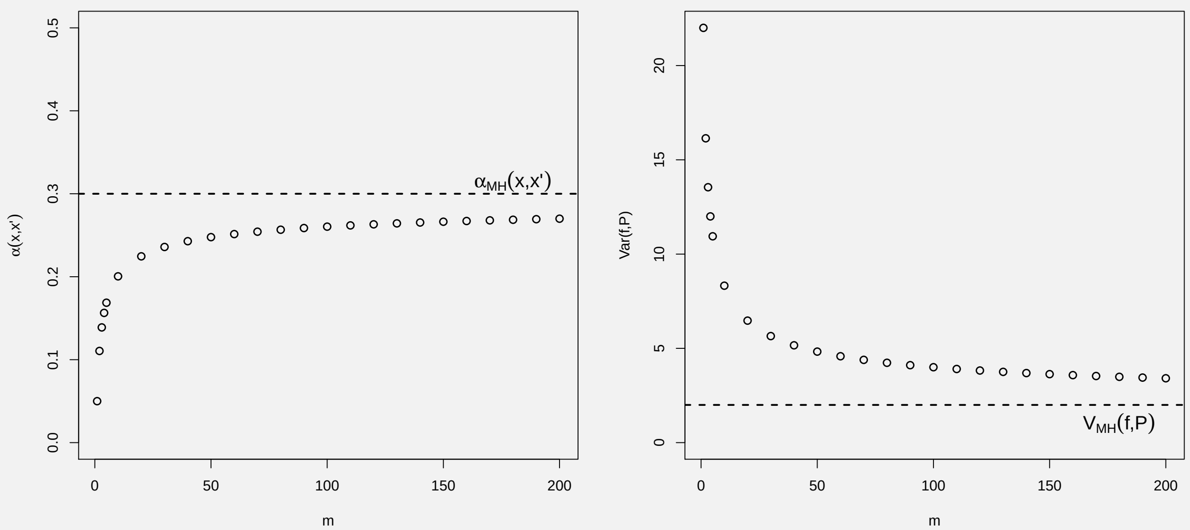
Does not require independence; W must arise from an exchangeable distribution.

How much better is P_m than P_1? Does it justify the extra computational effort?

PMMH and averages | Existing theory | First result | A tighter result? | Simulation study | Summary

Heuristics

Andrieu and Vihola (2016): PMMH is never as good as ideal MH.



Suppose sampling $W_1, \ldots W_m$ takes m times the computational effort of sampling W_1. For a given computational budget, # iterations reduced by a factor of m, so we need $m\mathrm{Var}(f, P_m) < \mathrm{Var}(f, P_1)$ for averaging to be worthwhile.

Previous work

Sherlock, Thiery, Roberts and Rosenthal (2013) [ArXiv vn 1 of 2015 paper] examines the PMRWM as $d \to \infty$.

Empirically: if $W_j \sim \mathrm{Gam}(a, a)$ iid, $m\mathrm{Var}(f, \mathrm{P}_m) \geq \mathrm{Var}(f, \mathrm{P}_1)$. Same for $W_j = (a, b)$ w.p. $(1-p, p)$ iid (with $a(1-p) + bp = 1$).

Bornn, Pillai, Smith and Woodard (2014): ABC-MCMC with a uniform error window and assumption that P_m is non-negative definite then $(2m-1)\mathrm{Var}(f, \mathrm{P}_m) \geq \mathrm{Var}(f, \mathrm{P}_1)$.

Our result

Theorem 1

1. For any $x, x' \in \mathsf{X}$ the conditional acceptance rates satisfy $\alpha^*(x, x'|\mathrm{P}_m) \leq m\alpha^*(x, x'|\mathrm{P}_1)$.
2. For any $f : \mathsf{X} \to \mathbb{R}$, the Dirichlet forms satisfy $\mathcal{E}_{\mathrm{P}_m}(f) \leq m\mathcal{E}_{\mathrm{P}_1}(f)$.
3. For any $f : \mathsf{X} \to \mathbb{R}$ with $\mathrm{Var}_\pi(f) < \infty$, $m\mathrm{Var}(f, \mathrm{P}_m) \geq \mathrm{Var}(f, \mathrm{P}_1) - (m-1)\mathrm{Var}_\pi(f)$.

Does not require independence; W must arise from an exchangeable distribution (two proofs).

If P_m is non-negative definite, then $(2m-1)\mathrm{Var}(f, \mathrm{P}_m) \geq \mathrm{Var}(f, \mathrm{P}_1)$.

Direct proof: key tools (1)

Consider an extended statespace $(\mathsf{X} \times \mathsf{W}^m \times \mathsf{K})$, where $\mathsf{K} = \{1, 2, \ldots, m\}$.
Let $r = r(x, x') = \pi(x')q(x', x)/(\pi(x)q(x, x'))$. Define $\mathrm{Q}_1(x, \underline{w}, k; \mathrm{d}x', \mathrm{d}\underline{w}', k')$ as

$$q(x, x')\mathrm{q}(x', \mathrm{d}\underline{w}')\mathrm{q}_1(\underline{w}', k')\alpha_1(x, \underline{w}, k; x', \underline{w}', k') \\ + (1 - \overline{\alpha}_1(x, \underline{w}, k))\delta((x', \underline{w}', k') - (x, \underline{w}, k)),$$

where $\overline{\alpha}_1(x, \underline{w}, k)$ is acc. prob from $(x, \underline{w}, k)$ and

$$\mathrm{q}_1(\underline{w}; k) = \begin{cases} \frac{1}{m} & k \in \mathsf{K} \\ 0 & \text{otherwise}, \end{cases} , \ \alpha_1(x, \underline{w}, k; x', \underline{w}', k') = 1 \wedge \left[r \frac{w'_{k'}}{w_k} \right]$$

Lemma: $\{(X_t, W_{t,K_t})\}_{t=1}^{\infty}$ under Q_1 is $\stackrel{\mathcal{D}}{=}$ $\{(X_t, W_t)\}_{t=1}^{\infty}$ under P_1.

Direct proof: key tools (2)

Define $\mathrm{Q}_m(x, \underline{w}, k; \mathrm{d}x', \mathrm{d}\underline{w}', k')$ as

$$q(x, x')\mathrm{q}(x', \mathrm{d}\underline{w}')\mathrm{q}_m(\underline{w}', k')\alpha_m(x, \underline{w}, k; x', \underline{w}', k') \\ + (1 - \overline{\alpha}_m(x, \underline{w}, k))\delta((x', \underline{w}', k') - (x, \underline{w}, k)),$$

where $\overline{\alpha}_m(x, \underline{w}, k)$ is acc. prob from $(x, \underline{w}, k)$ and

$$\mathrm{q}_m(\underline{w}; k) = \begin{cases} \frac{w_k}{\sum_{j=1}^m w_j} & k \in \mathsf{K} \\ 0 & \text{otw.} \end{cases} , \alpha_m(x, \underline{w}, k; x', \underline{w}', k') = 1 \wedge \left[r \frac{\sum_{j=1}^m w'_j}{\sum_{j=1}^m w_j} \right]$$

Lemma: the joint distribution of $\{(X_t, \sum_{j=1}^m W_{t,j})\}_{t=1}^{\infty}$ is the same under Q_m and P_m.

Key Steps

Proposition

Q_1 and Q_m both have an invariant distribution of

$$\tilde{\pi}_m(x, \underline{w}, k) := \pi(x)\mathrm{q}(x; \underline{w})\mathrm{q}_1(\underline{w}; k)w_k.$$

Proposition

$$\mathrm{q}_1(\underline{w}', k')\alpha_1(x, \underline{w}, k; x', \underline{w}', k') \geq \frac{1}{m}\mathrm{q}_m(\underline{w}', k')\alpha_m(x, \underline{w}, k; x', \underline{w}', k').$$

This leads directly to our results on $\alpha^*(x, x')$ and $\mathcal{E}$. Our result for Var follows from a simple (but neat!) Lemma in Andrieu, Lee and Vihola (2015).

A tighter result?

We have: $m\mathrm{Var}(f, \mathrm{P}_m) \geq \mathrm{Var}(f, \mathrm{P}_1) - (m-1)\mathrm{Var}_\pi(f)$.

Qn: $m\mathrm{Var}(f, \mathrm{P}_m) \geq \mathrm{Var}(f, \mathrm{P}_1)$ would be better! Is it true?

Counter example

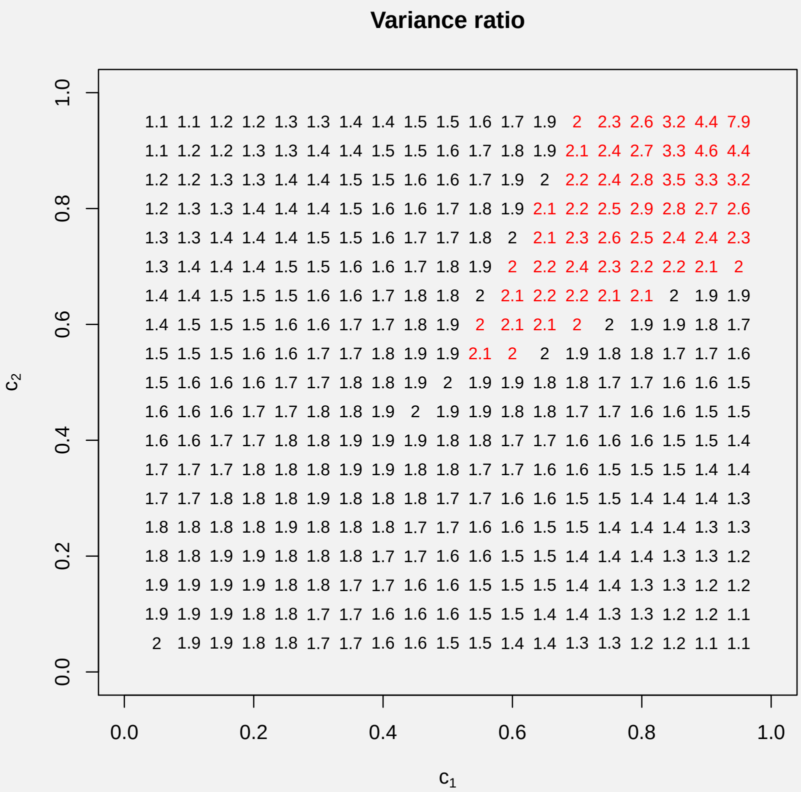
$$\mathsf{X} = \{1, 2\},\ q(1, 2) = c_1,\ q(2, 1) = c_2,\ \pi = (0.5, 0.5).$$
$$m = 2,\ \mathsf{W} = \{0, 2\}$$

$$\mathrm{q}(x, (0, 2)) = \mathrm{q}(x, (2, 0)) = 0.5,\ \mathrm{q}(x, (0, 0)) = \mathrm{q}(x, (2, 2)) = 0.$$

$$f(x) = 2x - 1.$$

Counter example: plot

The ratio $\mathrm{Var}(f, \mathrm{P}_1)/\mathrm{Var}(f, \mathrm{P}_2)$ as a function of (c_1, c_2).



Tighter result?

Qn: $m\mathrm{Var}(f, \mathrm{P}_m) \geq \mathrm{Var}(f, \mathrm{P}_1)$ would be better! Is it true?
A1: Not for general exchangeable weights.

Qn What if the weights are independent?

Consider the kernels on the extended statespace:

$$m\mathrm{Var}(f, \mathrm{Q}_m) - \mathrm{Var}(f, \mathrm{Q}_1) = \langle f, Af \rangle$$

where

$$A := 2m(I - \mathrm{Q}_m)^{-1} - 2(I - \mathrm{Q}_1)^{-1} - (m-1)I.$$

Qn: Does A have any negative eigenvalues?
A: Yes, for some (c_1, c_2), and some independent $\underline{W}$ distributions.

So $\exists$ functions $f(x, \underline{w}, k)$ for which $m\mathrm{Var}(f, \mathrm{Q}_m) < \mathrm{Var}(f, \mathrm{Q}_1)$.

Tighter result?

Qn: $m\text{Var}(f, \text{P}_m) \geq \text{Var}(f, \text{P}_1)$ would be better! Is it true?
A1: Not for general exchangeable weights.
A2: Not with independent weights for $f : \text{X} \times \text{W}^m \times \text{K} \to \mathbb{R}$.

Qn: What about functions $f(x)$ and with independent weights?

A: ??? - we have not been able to find a counter example.

Simulation study

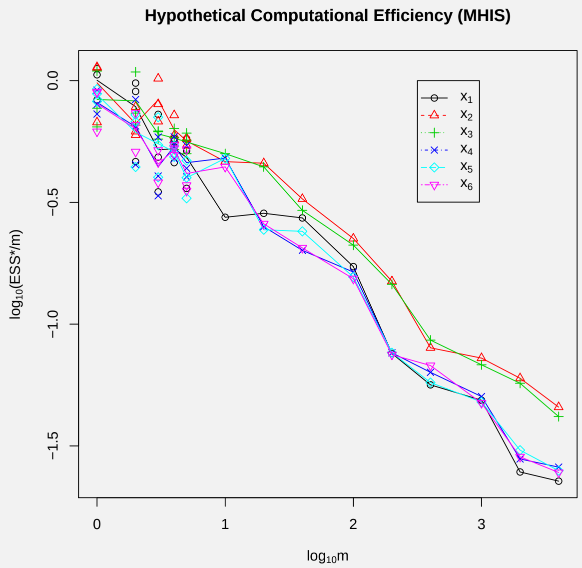
Gaussian-process logistic regression.

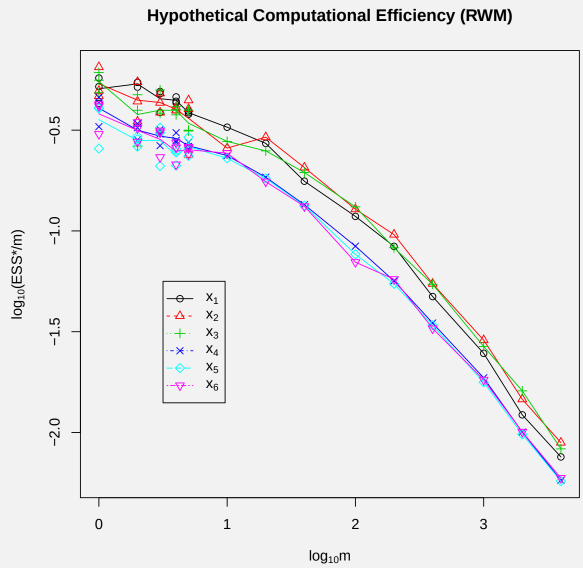
1. Independence sampler.
2. RWM with scaling optimal for the marginal algorithm.

Graphs showing

$$\frac{1}{m}\text{ESS}.$$

Simulation study: ESS/m



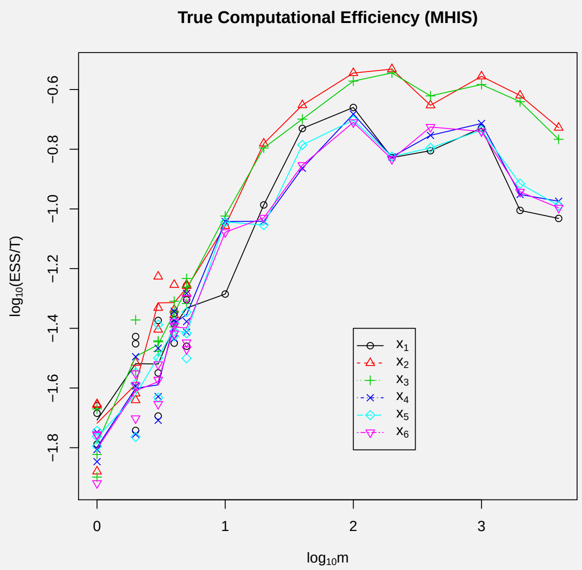


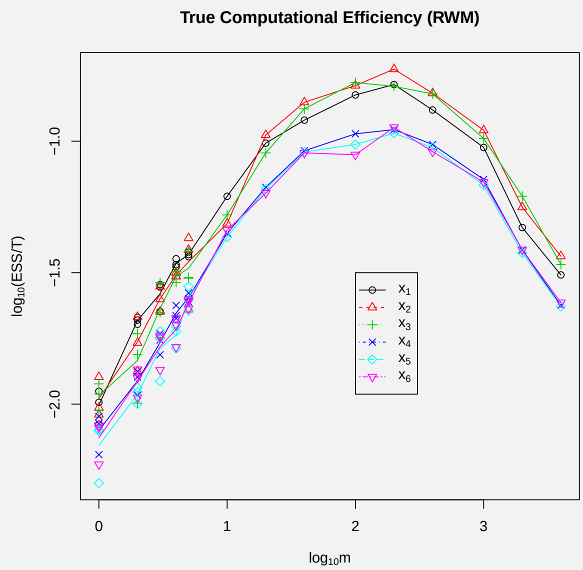
Qn: Never worth taking an average?

PMMH and averages | Existing theory | First result | A tighter result? | Simulation study | Summary

Simulation study: ESS/T

Graphs show ESS/T_{cpu}.





Qn: Worth taking an average?
A: Yes, when there is a set-up cost.

Summary

We provide upper bounds on the efficiency of the PMMH when using the average of m exchangeable unbiased estimators compared to using just 1 of the estimators.

If there is no start-up cost then there is little gain in using $m > 1$.

This is entirely different from the choice of the number of particles in particle-marginal MH: choose m such that $\mathrm{Var_q}(\log W) = \mathcal{O}(1)$.

Thank you for your attention!