Reduced-Variance Estimation with Intractable Likelihoods

Antonietta Mira

IDIDS, Universita' Svizzera Italiana, Switzerland
DISAT, Insubria University, Como, Italia

Joint with N. Friel (UCD, Dublin) and C. Oates (UTS, Sidney)

Warwick - CRiSM, April 22, 2016

Overview of talk

- PART I:
  Zero Variance MCMC for intractable problems
- PART II:
  Reduced Variance MCMC for doubly intractable problems

Monte Carlo Integration

Let μ_g be the expected value of a function $g : \mathbb{R}^d \to \mathbb{R}$, under π:

$$\mu_g := \mathbb{E}_\pi [g(X)] = \frac{\int_{\mathbb{R}^d} g(X)\pi(x)dx}{\int_{\mathbb{R}^d} \pi(x)dx}$$

and let $X_1, ..., X_N$ be a sequence of iid draws from π.
An unbiased estimator of μ_g is:

$$\hat{\mu}_g := \frac{1}{N}\sum_{i=1}^{N} g(X_i)$$

with variance

$$\mathbb{V}(\hat{\mu}_g) = \frac{1}{N}\sigma_g^2$$

where σ_g^2 is the variance of g under π.
Therefore if g has finite variance, $\hat{\mu}_g$ is a consistent estimator of μ_g

If drawing iid from π is difficult, build an ergodic Markov chain $\{X_n\}$, stationary wrt π

MCMC Integration

Theorem (Tierney 1996). Suppose the ergodic Markov chain $\{X_n\}$, with stationary distribution π and a real valued function g, satisfy one of the following conditions:

- The chain is uniformly ergodic and $\mathbb{E}_\pi\left[g(X)^2\right] < \infty$
- The chain is geometrically ergodic and $\mathbb{E}_\pi\left[|g(X)|^{2+\epsilon}\right] < \infty$ for some $\epsilon > 0$

Then

$$
\begin{aligned}
lim_{N\to\infty} N\mathbb{V}\left(\hat{\mu}_g\right) &= \mathbb{E}_\pi\left[\left(g(X_0)-\mu_g\right)^2\right] + 2\sum_{k=1}^{+\infty}\mathbb{E}_\pi\left[\left(g(X_k)-\mu_g\right)\left(g(X_0)-\mu_g\right)\right] \\
&= \sigma_g^2[1+2\sum_{k=1}^{+\infty}\rho_g(k)] = \sigma_g^2\,\tau_g = V(g,P)
\end{aligned}
$$

is well defined, non-negative and finite, and

$$
\sqrt{N}\left(\hat{\mu}_g - \mu_g\right) \xrightarrow{L} \mathcal{N}(0, \sigma_g^2\,\tau_g)
$$

- Delayed rejection strategy $\longrightarrow$ reduce τ_g $\longrightarrow$ by modifying P
- Zero variance strategy $\longrightarrow$ reduce σ_g^2 $\longrightarrow$ by modifying g

Control Variate Method in MC

In MC simulation, control variate are used to reduce the variance of MC estimators. Assume Z is a random variable with known mean, and correlated with $g(X)$:

$$\begin{aligned} \mathbb{E}(Z) &= 0 \\ \mathbb{C}\text{ov}(g(X), Z) &= \sigma_{g,Z} \neq 0 \end{aligned}$$

By exploiting the correlation of $g(X)$ and Z, we can build new **unbiased** estimators of μ_g, with **lower variances**. Let's define:

$$\tilde{g}(X) := g(X) + aZ$$

where $a \in \mathbb{R}$. Obviously

$$\begin{aligned} \mu_{\tilde{g}} := E\left[\tilde{g}(X)\right] &= \mu_g \\ \sigma^2_{\tilde{g}} &= \sigma^2_g + a^2\sigma^2_Z + 2a\sigma_{g,Z} \end{aligned}$$

Control Variate Method in MC

By minimizing $\sigma^2_{\tilde{g}}$ w.r.t. a, it can be shown that the optimal choice of a is

$$a = -\frac{\sigma_{g,Z}}{\sigma^2_z}$$

that reduces the variance of $\sigma^2_{\tilde{g}}$ to $\left(1 - \rho^2_{g,Z}\right)\sigma^2_g$. Therefore

$$\hat{\mu}_{\tilde{g}} := \frac{1}{N}\sum_{i=1}^{N} \tilde{g}(X_i)$$

is a new unbiased estimator of μ_g, with variance

$$\mathbb{V}\left(\hat{\mu}_{\tilde{g}}\right) = \frac{1}{N}\sigma^2_{\tilde{g}} = \frac{1}{N}\left(1 - \rho^2_{g,Z}\right)\sigma^2_g \quad \leq \quad \frac{1}{N}\sigma^2_g = \mathbb{V}\left(\hat{\mu}_g\right)$$

First order ZV for MCMC

Under regularity conditions, the score has zero mean

$$z(x) := -\frac{1}{2}\nabla \ln \pi(x)$$

Use it as a control variate!

First order ZV for MCMC

$\tilde{g}(X) = g(X) + \Delta_x[P(x)] + \nabla_x[P(x)] \cdot z(x)$
where $P(x)$ is polynomial in x
If $P(x)$ is a first degree polynomial:

$$\tilde{g}(X) = g(X) + a^T z(x)$$

The optimal value of a is:

$$a = -\Sigma_{\mathbf{zz}}^{-1}\sigma_{\mathbf{z}g}, \qquad \text{where} \qquad \Sigma_{\mathbf{zz}} = \mathbb{E}(zz^T), \quad \sigma_{\mathbf{z}g} = \mathbb{E}(zg)$$

The optimal a is estimated using the existing MCMC simulation

Second order ZV for MCMC

If $P(x)$ is a second degree polynomial:

$\tilde{g}(X) = g(X) - \frac{1}{2}\text{tr}(B) + (a + Bx)^T z = g(X) + g^T y$
where g and y are column vectors with $\frac{1}{2}d(d+3)$ elements:

- $g := [a^T \ b^T \ c^T]^T$: where $b := \text{diag}(B)$, and c is a column vector with $\frac{1}{2}d(d-1)$ elements; The element ij of matrix B (for $i \in \{2, ..., d\}$, and $j < i$), is the element $\frac{1}{2}(2d-j)(j-1) + (i-j)$ of vector c.
- $y := [z^T \ u^T \ v^T]^T$: where
  $u := x * z - \frac{1}{2}\mathbf{i}$ (where "$*$" = Hadamard product, and $\mathbf{i}$ = vector of ones), and
  v is a column vector with $\frac{1}{2}d(d-1)$ elements;
  $x_i z_j + x_j z_i$ (for $i \in \{2, ..., d\}$, and $j < i$), is the element $\frac{1}{2}(2d-j)(j-1) + (i-j)$ of vector v

With a polynomial of order 2 we have $\frac{d(d+3)}{2}$ control variates
With a polynomial of order p, we get $\binom{d+p}{p} - 1$ control variates

Probit Model

$$y_i|\mathbf{x}_i \sim \mathcal{B}(1, p_i), \quad p_i = \Phi(\mathbf{x}_i^T \beta)$$

where $\beta \in \mathbb{R}^d$ is the vector of parameters of the model.

The likelihood function is:

$$l(\beta|\mathbf{y}, \mathbf{X}) \propto \prod_{i=1}^{n} \left[\Phi(\mathbf{x}_i^T \beta)\right]^{y_i} \left[1 - \Phi(\mathbf{x}_i^T \beta)\right]^{1-y_i}$$

With flat priors the Bayesian estimator of each parameter, β_d, is $\mathbb{E}_\pi[\beta_k|\mathbf{y}, \mathbf{X}]$, $k = 1, 2, \cdots, d$

We have:

$$\widetilde{g}_k(\beta) \quad = \quad g_k(\beta) + \sum_{j=1}^{d} a_{j,k} z_j \qquad \text{where} \qquad z_j = -\frac{1}{2} \sum_{i=1}^{n} x_{ij} \frac{\phi(\mathbf{x}_i^T \beta)}{\Phi(\mathbf{x}_i^T \beta)}$$

Existence of MLE implies unbiasedness of the ZVMCMC estimator. The bank dataset from Flury and Riedwyl (1988), contains the measurements of 4 variables on 200 Swiss banknotes (100 genuine and 100 counterfeit).

The four measured variables x_i $(i = 1, 2, 3, 4)$, are the length of the bill, the width of the left and the right edge, and the bottom margin width.

These variables are used in a probit model as the regressors, and the type of the banknote y_i, as the response variable (0 for genuine and 1 for counterfeit)

Zero-Variance MCMC

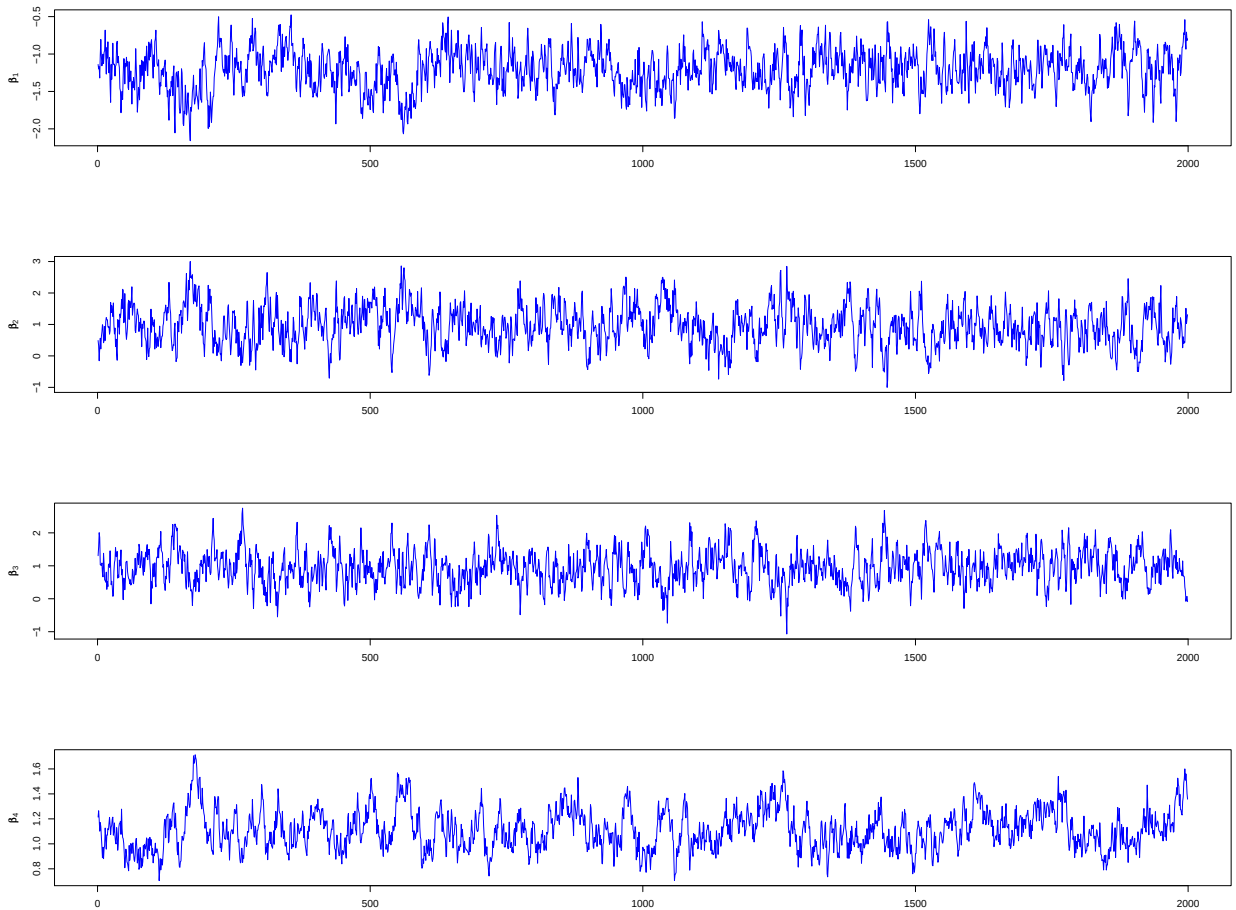
In the first stage (2000 steps), we run an MCMC simulation
to estimate the optimal coefficients of the polynomial

In the second stage (2000 steps), we run another MCMC simulation, independent of the first one, to estimate $\hat{\mu}_{\tilde{g}}$ using the estimated coefficients obtained in the first stage

We use the Albert and Chib Gibbs sampler
We try both first and second order ZV

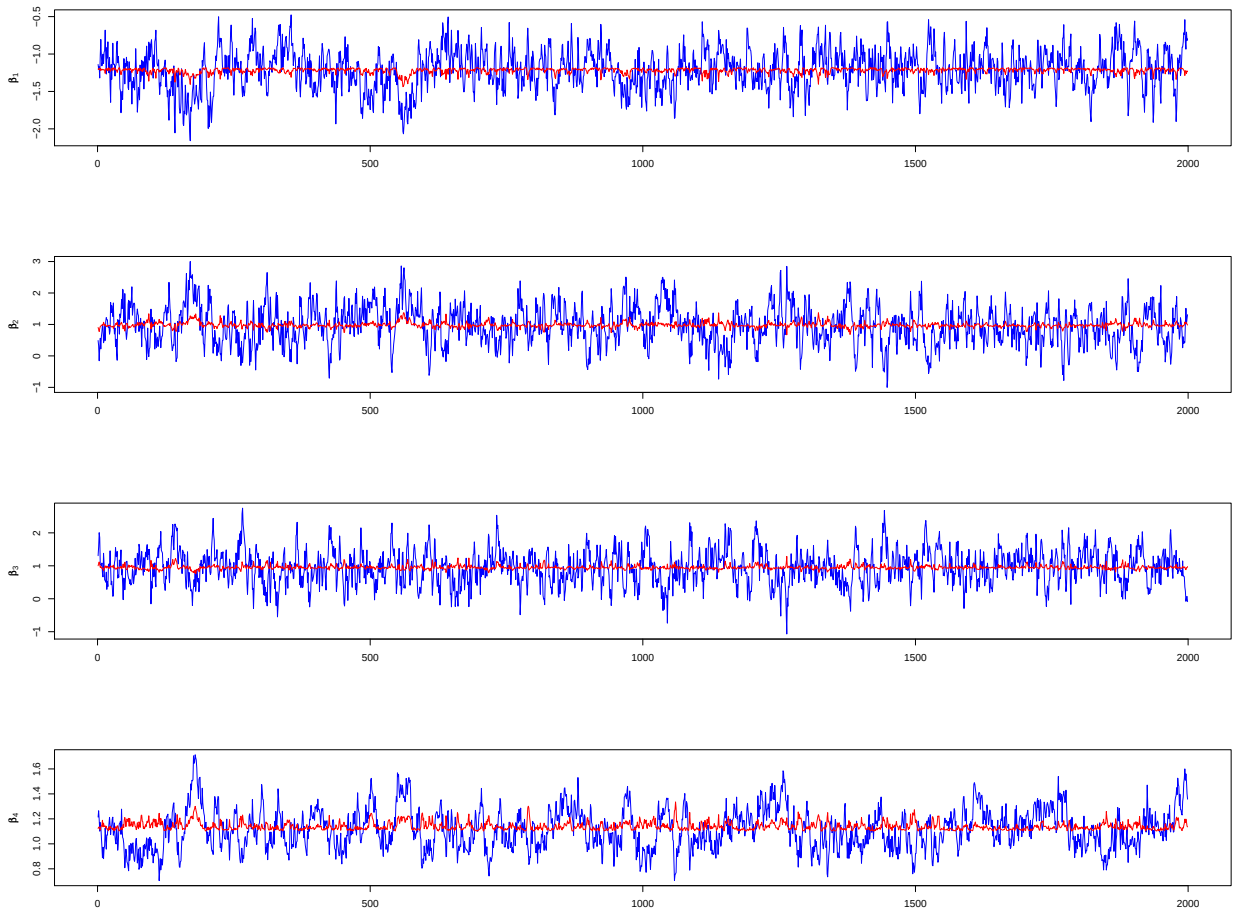
To compare performance: compute the ratio of Sokal's estimates of variances of the ordinary MCMC and ZV-MCMC.

Ordinary MCMC



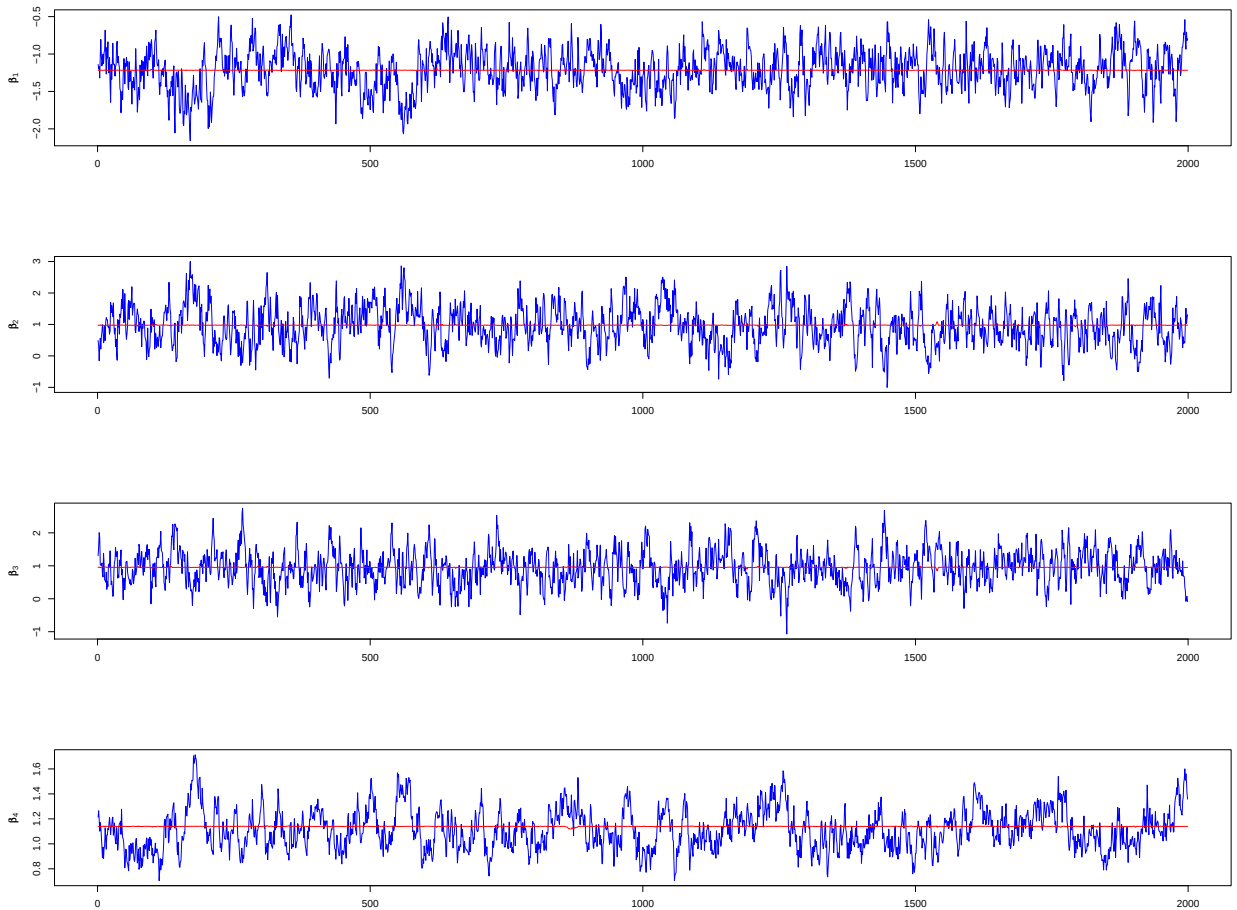
Ordinary and ZV-MCMC: 1st degree P(x)

Variance Reduction Ratios: 25-100



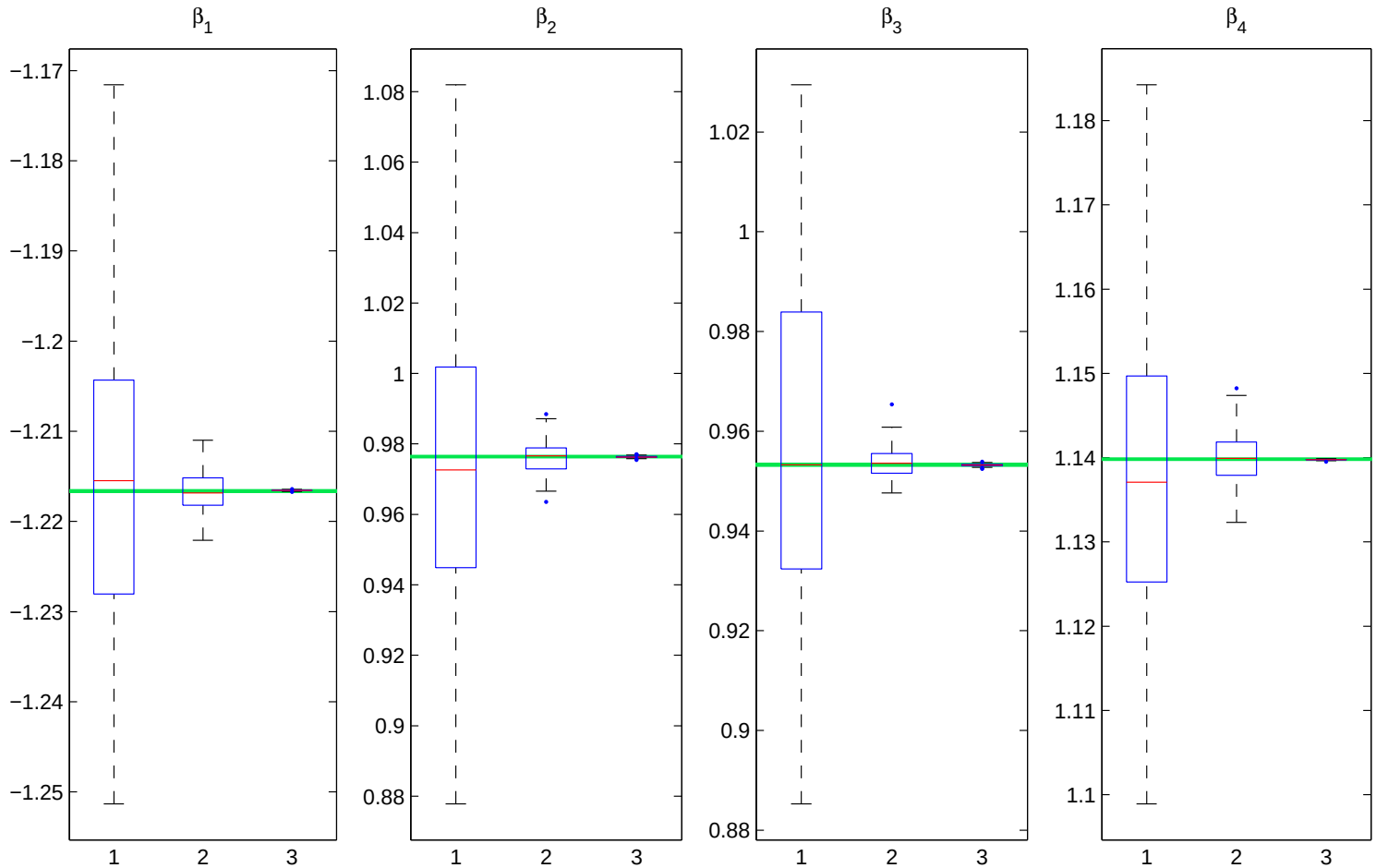
Ordinary and ZV-MCMC: 2nd degree P(x)

Variance Reduction Ratios: 25000-90000



Ordinary and ZV-MCMC: A Monte Carlo Study

Ordinary MCMC estimates (1)
First order ZV-MCMC estimates (2)
Second order ZV-MCMC estimates (3)
+ 95% confidence region obtained by an ordinary MCMC of length 10^8 (green)



Logit Model

$$y_i|\mathbf{x}_i \sim \mathcal{B}(1, p_i), \quad p_i = \frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)}$$

where $\beta \in \mathbb{R}^d$ is the vector of parameters of the model.

With flat priors the Bayesian estimator of each parameter, β_d, is $\mathbb{E}_\pi[\beta_k|\mathbf{y}, \mathbf{X}]$ $k = 1, 2, \cdots, d$

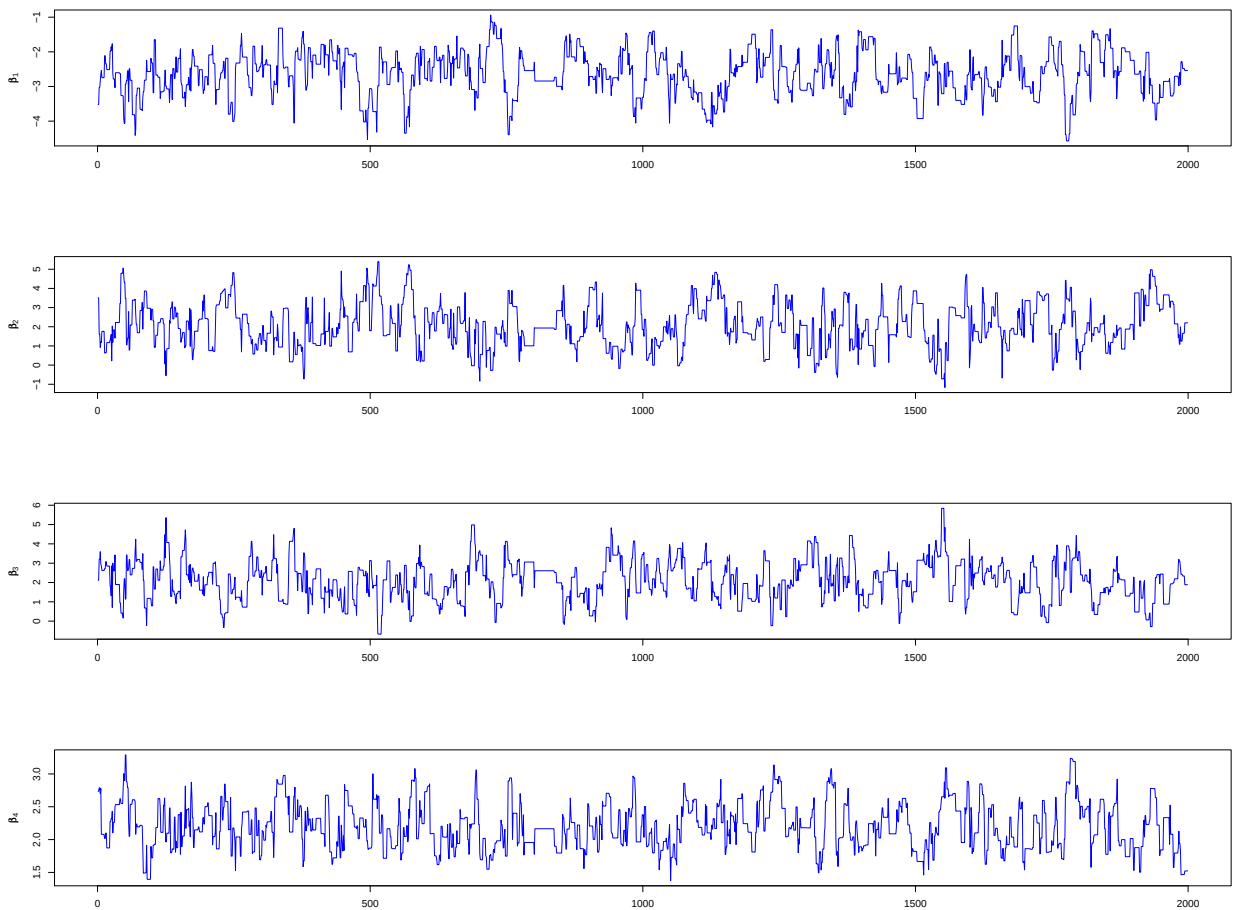
We have:

$$\widetilde{g}_k(\beta) \quad = \quad g_k(\beta) + \sum_{j=1}^{d} a_{j,k} z_j \qquad \text{where} \qquad z_j = \frac{1}{2} \sum_{i=1}^{n} x_{ij} \frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)}$$

Existence of MLE implies finiteness of $2 + \epsilon$-th moment of the control variates and thus a CLT

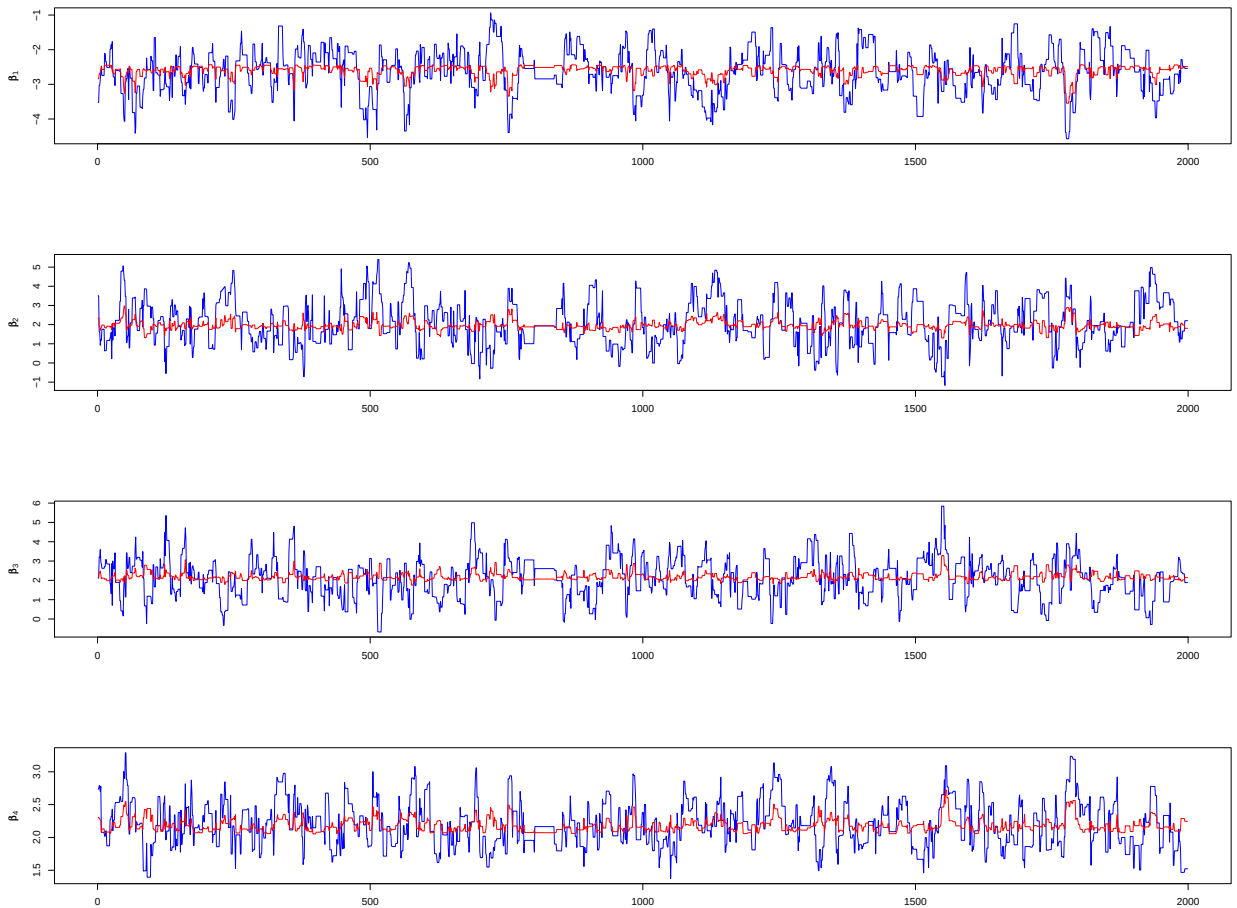
Example: we use the same Swiss banknotes dataset

Ordinary MCMC



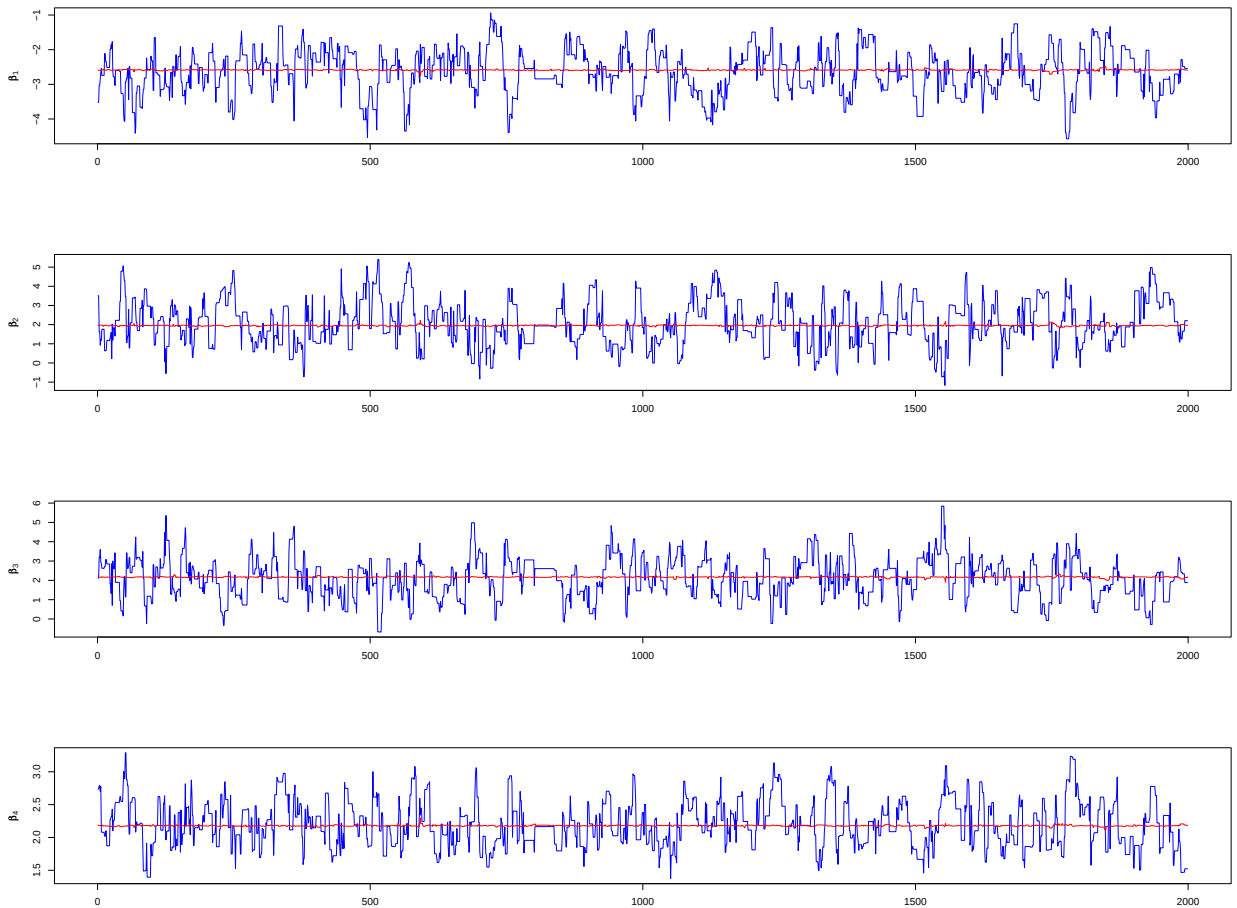
Ordinary and ZV-MCMC: 1st degree P(x)

Variance Reduction Ratios: 10-40



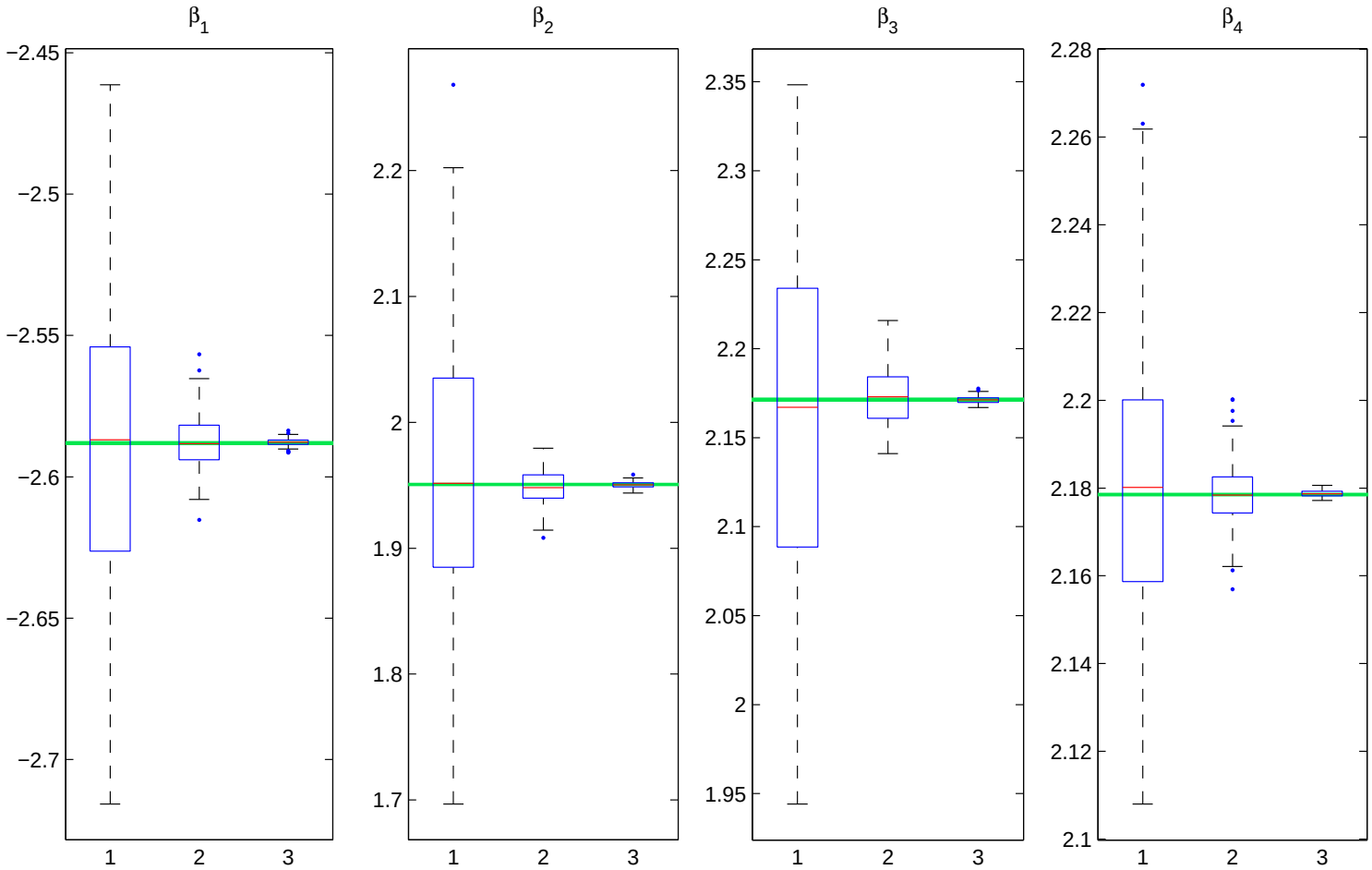
Ordinary and ZV-MCMC: 2nd degree P(x)

Variance Reduction Ratios: 2000-6000



Ordinary and ZV-MCMC: A Monte Carlo Study

Ordinary MCMC estimates (1)
First order-MCMC estimates (2)
Second order ZV-MCMC estimates (3)
+ 95% confidence region obtained by an ordinary MCMC of length 10^8 (green)



GARCH Model

Let S_t be the price of an asset at time t and let the daily returns r_t be:

$$r_t = \frac{S_t - S_{t-1}}{S_t}.$$

The Normal-GARCH(1, 1) can be formulated as:

$$\begin{aligned} r_{t+1}|\mathcal{F}_t &\sim \mathcal{N}(0, h_{t+1}^2) \\ h_{t+1}^2 &= \omega + \alpha r_t^2 + \beta h_t^2 \end{aligned}$$

where $x := (\omega, \alpha, \beta)$ are the parameters of the model, and $\omega > 0$, $\alpha \geq 0$, and $\beta \geq 0$. Using non informative independent priors for parameters, the posterior is:

$$\pi(\omega, \alpha, \beta|\mathbf{r}) \propto \exp\left[-\frac{1}{2}\left(\frac{\omega^2}{\sigma_\omega^2} + \frac{\alpha^2}{\sigma_\alpha^2} + \frac{\beta^2}{\sigma_\beta^2}\right)\right]\left(\prod_{t=1}^{T} h_t\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\frac{r_t^2}{h_t}\right)$$

The Bayesian estimators of the GARCH parameters are $\mathbb{E}_\pi[\omega|\mathbf{r}]$, $\mathbb{E}_\pi[\alpha|\mathbf{r}]$ and $\mathbb{E}_\pi[\beta|\mathbf{r}]$. The control variates are:

$$z_j = -\frac{1}{2}\frac{\partial \ln \pi}{\partial x_j} = \frac{x_j}{2\sigma^2_{x_j}} + \frac{1}{4}\sum_{t=1}^{T}\frac{1}{h_t}\frac{\partial h_t}{\partial x_j} - \frac{1}{4}\sum_{t=1}^{T}\frac{r_t^2}{h_t^2}\frac{\partial h_t}{\partial x_j} \qquad \text{for} \quad j = 1, 2, 3$$

(where $x_1 = \omega$, $x_2 = \alpha$, and $x_3 = \beta$.), and:

$$\frac{\partial h_t}{\partial x_1} = \frac{\partial h_t}{\partial \omega} = \frac{1-\beta^{t-1}}{1-\beta}$$

$$\frac{\partial h_t}{\partial x_2} = \frac{\partial h_t}{\partial \alpha} = \begin{cases} 0 & t = 1 \\ r_{t-1}^2 + \beta\frac{\partial h_{t-1}}{\partial \alpha} & t > 1 \end{cases}$$

$$\frac{\partial h_t}{\partial x_3} = \frac{\partial h_t}{\partial \beta} = \begin{cases} 0 & t = 1 \\ h_{t-1} + \beta\frac{\partial h_{t-1}}{\partial \beta} & t > 1 \end{cases}$$

Example: We fit a Normal-GARCH(1, 1) to the daily returns of the Deutsche Mark vs British Pound (DEM/GBP) exchange rates from Jan. 1985, to Dec. 1987 (750 obs)
We have used the MH algorithm for estimating GARCH models proposed in Ardia D., Financial Risk Management with Bayesian Estimation of GARCH Models r to estimate the optimal parameters of the polynomial.
In the second stage we run an independent MCMC simulation and compute $\tilde{g}_j(x)$.

Estimates of Parameters:

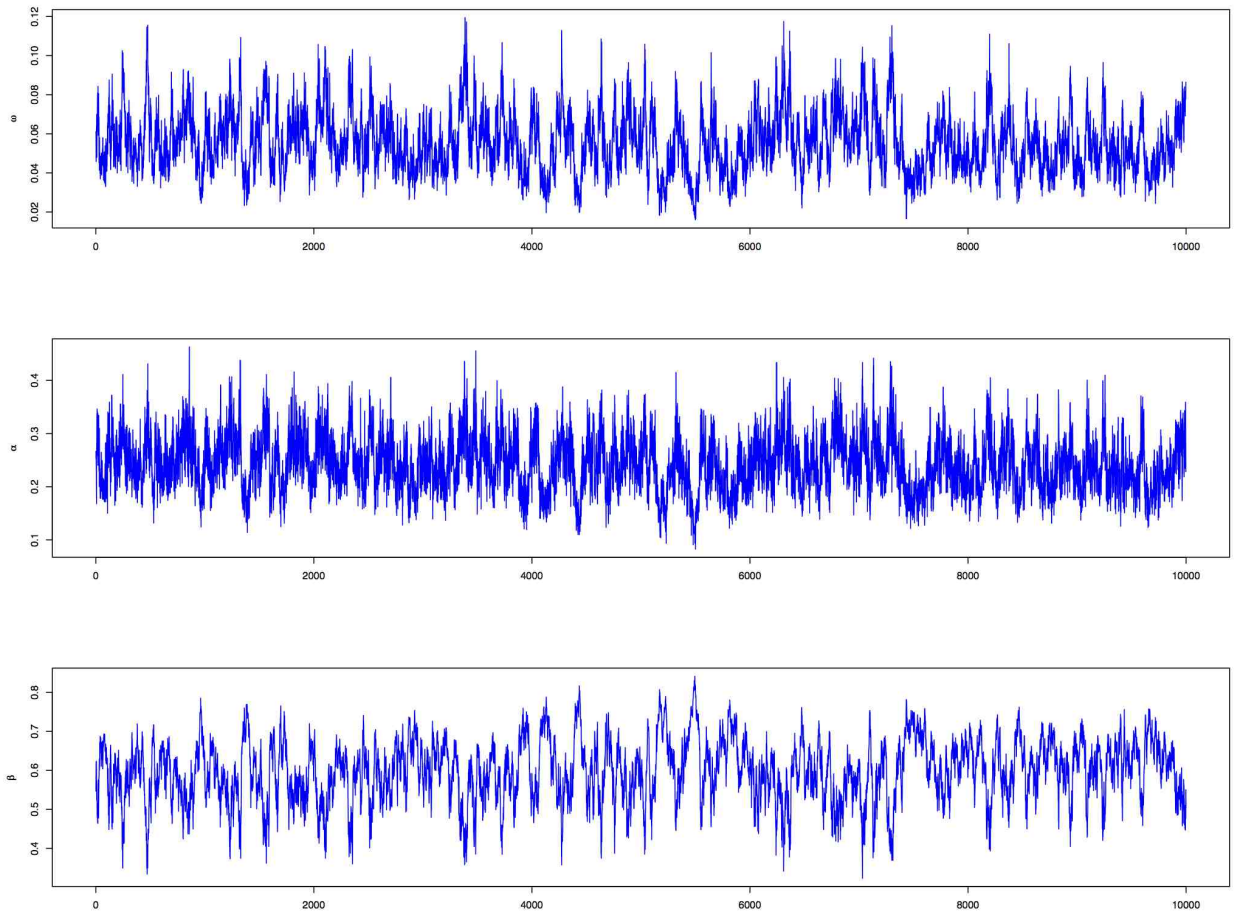
Method	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\beta}$
MLE	0.0445	0.2104	0.6541
MCMC	0.0568	0.2494	0.5873

Variance Reduction:

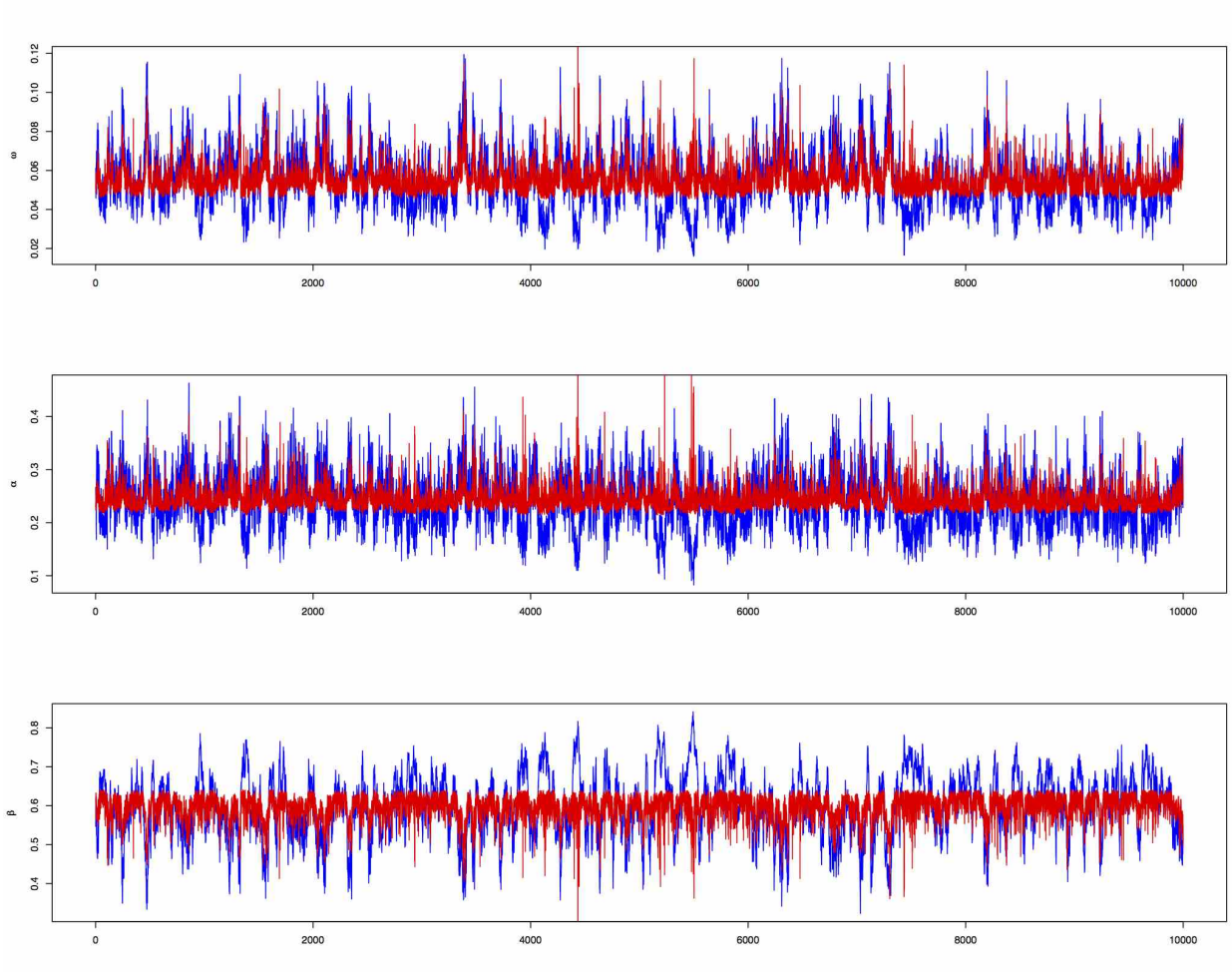
(Sokal estimate of std. error of MC estimator)^2 / (Sokal estimate of std. error of ZV-MC estimator)^2

	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\beta}$
1st Degree $P(x)$	9	20	12
2nd Degree $P(x)$	2,070	12,785	11,097
3rd Degree $P(x)$	28,442	70,325	30,281

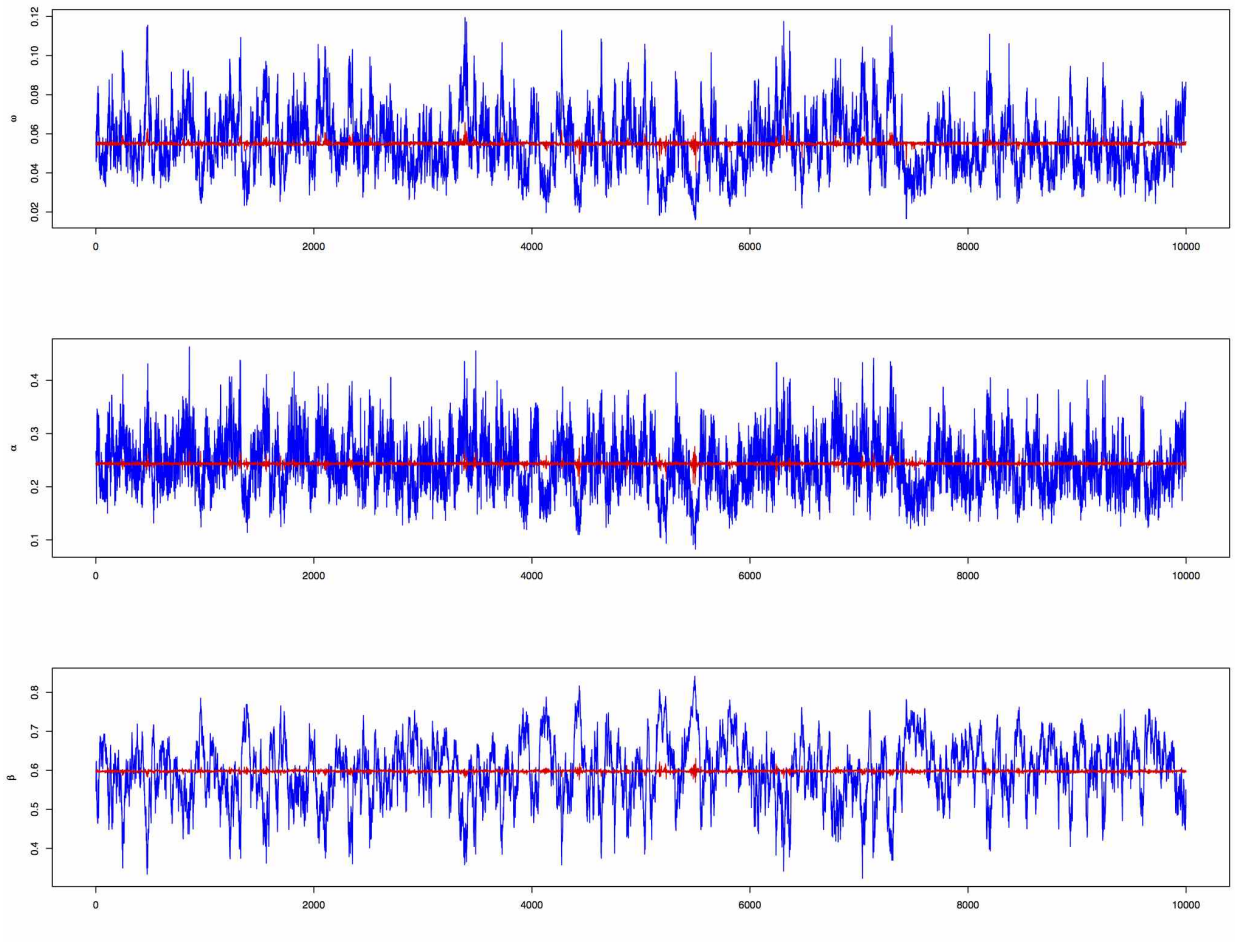
Ordinary MCMC



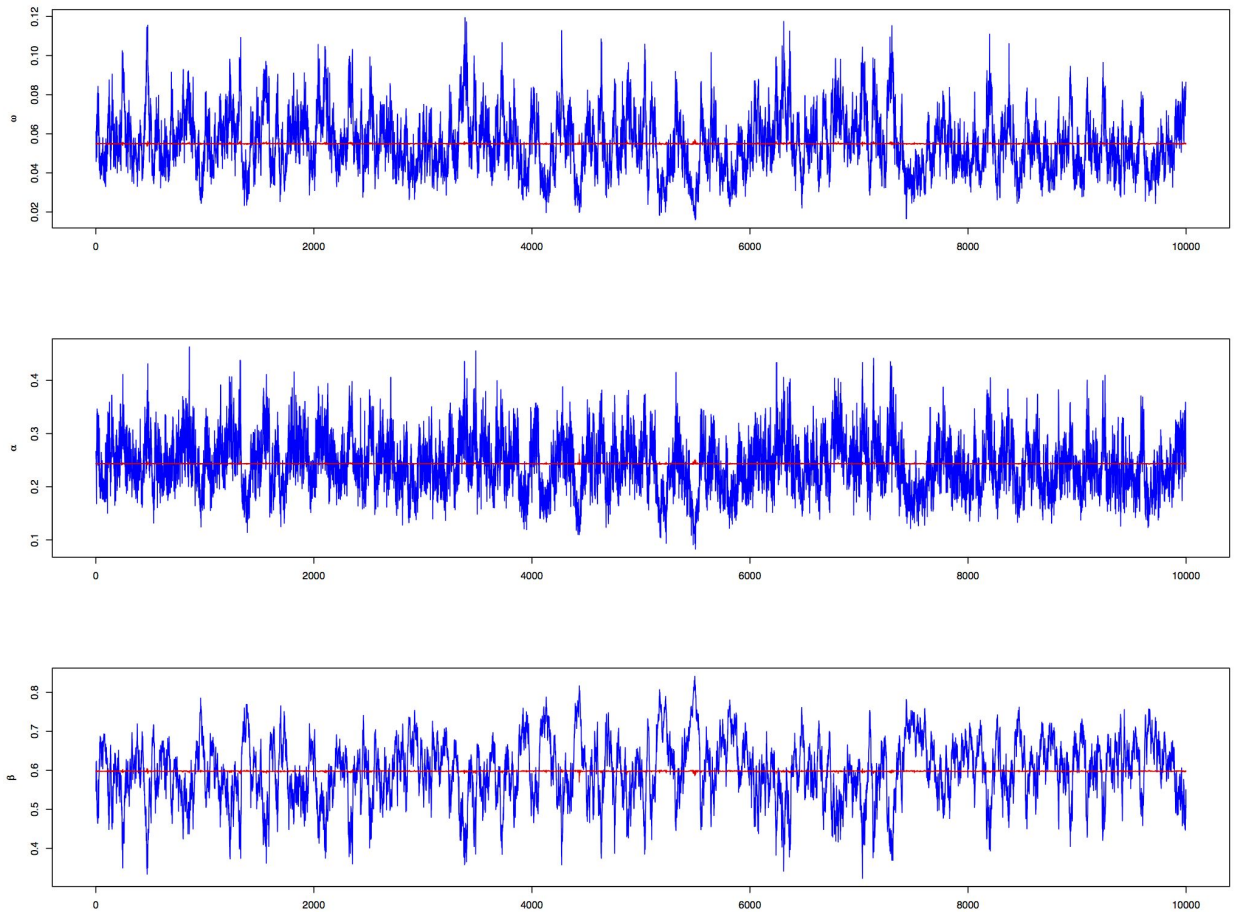
Ordinary and ZV-MCMC: 1st degree P(x)



Ordinary and ZV-MCMC: 2nd degree P(x)

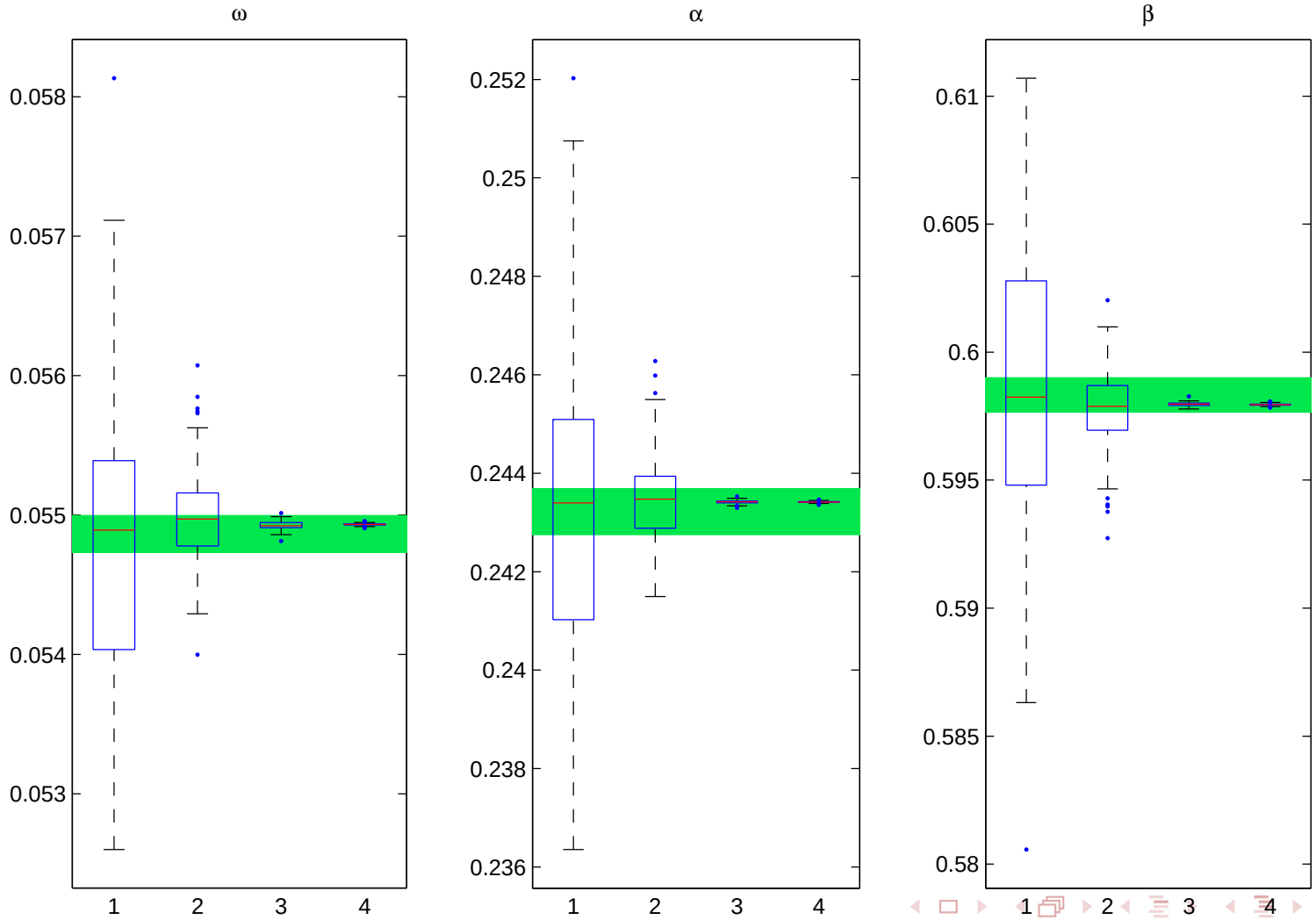


Ordinary and ZV-MCMC: 3rd degree P(x)



Ordinary and ZV-MCMC: A Monte Carlo Study

Ordinary MCMC estimates (1)
First, second, third order ZV-MCMC estimates (2, 3, 4)
+ 95% confidence region obtained by an ordinary MCMC of length 10^8 (green)



ZV - HMCMC

The ZV strategy is efficiently combined with Hamiltonian MC, MALA and variations by Girolami et al. without exceeding the computational requirements since its main ingredient (the score function) is exploited twice:

- to guide the MC towards relevant portion of the state space via a clever proposal, that exploits the geometry of the target and achieves convergence in fewer iterations
- to post-process the MC to reduce the variance of the resulting estimators

Zero Variance Differential Geometric MCMC, BA, 2014, Papamarkou, Mira, and Girolami

Conclusions - PART I

- Conditions for unbiasedness and CLT for ZV estimators
- Significant variance reduction
- Negligible additional computational costs
- Can control the variance only of the observables of real interest
- ZV is efficiently combined with Differential Geometric MCMC

Overview PART II

- Statistical models with intractable likelihood functions abound.
- It is possible to significantly reduce the Monte Carlo variance of resulting Bayesian estimators.
- Our methodology is compatible with many existing algorithms to carry out Bayesian inference for intractable LHD models

Bayesian inference with Intractable likelihoods

- Suppose we have data y, and a likelihood function f with parameters θ:

$$\pi(\theta|y) \propto f(y|\theta)p(\theta)$$

Bayesian inference with Intractable likelihoods

- Suppose we have data y, and a likelihood function f with parameters θ:

$$\pi(\theta|y) \propto f(y|\theta)p(\theta)$$

- However it turns out that there are many statistical models for which the likelihood function cannot be evaluated.

$$\pi(\theta|y) \propto f(y|\theta)p(\theta).$$

Bayesian inference with Intractable likelihoods

- Suppose we have data y, and a likelihood function f with parameters θ:

$$\pi(\theta|y) \propto f(y|\theta)p(\theta)$$

- However it turns out that there are many statistical models for which the likelihood function cannot be evaluated.

$$\pi(\theta|y) \propto f(y|\theta)p(\theta).$$

- This extra level of intractability is sometimes due to the complicated dependency in the data, or even due to the sheer volume of the data.

The predominant sources of intractability can be classified as follows:

Type I: The need to compute a normalisation constant $z(\theta) = \int q_\theta(y')dy'$ that depends on parameters θ, where $f(y|\theta) = q_\theta(y)/z(\theta)$

Type II: The need to marginalise over latent variables x, such that $f(y|\theta) = \int p(y|x,\theta)p(x|\theta)dx$.

Bayesian estimation in either of these settings is extremely challenging as many established techniques are incompatible with intractable likelihoods.

Type I intractability

Here we focus on Gibbs random fields where data y arise from the model,

$$f(\mathbf{y}|\boldsymbol{\theta}) = \frac{q_\theta(y)}{z(\boldsymbol{\theta})} = \frac{\exp\{\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y})\}}{z(\boldsymbol{\theta})}$$

such that the partition function

$$z(\boldsymbol{\theta}) = \int \exp\{\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y})\} d\mathbf{y}$$

is intractable.

Example: Spatial statistics – Ising model

- Defined on a lattice $y = \{y_1, \ldots, y_n\}$.
- Lattice points y_i take values $\{-1, 1\}$.
-

$$f(y|\theta) \propto q_\theta(y) = \exp\left\{\frac{1}{2}\theta_1 \sum_{i\sim j} y_i y_j\right\}.$$

Here $\sim$ means "is a neighbour of".

- The normalising constant

$$z(\theta) = \sum_{y_1} \cdots \sum_{y_n} q_\theta(y).$$

is intractable for moderately small n.

Metropolis-Hastings algorithm

Doubly-intractable posterior:

$$\pi(\theta|y) \propto f(y|\theta)p(\theta)$$

---

1 **for** $i = 1, \ldots, I$ **do**

2 Draw $\theta' \sim h(\theta'|\theta^{(i)})$;

3 With probability

$$\min\left(1, \frac{q_{\theta'}(y)}{q_{\theta^{(i)}}(y)} \frac{p(\theta')}{p(\theta^{(i)})} \times \frac{z(\theta^{(i)})}{z(\theta')}\right)$$

set $\theta^{(i+1)} = \theta'$, otherwise set $\theta^{(i+1)} = \theta^{(i)}$;

4 **end**

Exchange algorithm

Augmented posterior distribution:

$$\pi(\theta', y', \theta|y) \propto f(y|\theta)p(\theta)h(\theta'|\theta)f(y'|\theta')$$

---

for $i = 1, \ldots, I$ **do**

Draw $\theta' \sim h(\theta'|\theta^{(i)})$;

Draw $y' \sim f(\cdot|\theta')$;

With probability

$$\min\left(1, \frac{q_{\theta'}(y)}{q_{\theta^{(i)}}(y)} \frac{p(\theta')}{p(\theta^{(i)})} \frac{q_{\theta^{(i)}}(y')}{q_{\theta'}(y')} \times \frac{z(\theta^{(i)})}{z(\theta')} \frac{z(\theta')}{z(\theta^{(i)})}\right)$$

set $\theta^{(i+1)} = \theta'$, otherwise set $\theta^{(i+1)} = \theta^{(i)}$;

end

Type II intractability

Type II intractability arises from the need to marginalise over latent variables $\mathbf{x}$ such that the marginal likelihood

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}$$

is unavailable in closed form.

Type II intractability
Example: Hidden Markov model

In a hidden Markov model the parameters $\boldsymbol{\theta}$ that specify a Markov chain may be of interest, whilst the latent sample path $\{\mathbf{x}_n\}_{n=0}^N$ of the Markov chain that gives rise to observations $\{\mathbf{y}\}_{n=0}^N$ may not be of interest and must be marginalised.

Even in discrete cases where $\mathbf{x}_n \in \mathcal{X}$ for a finite state space $\mathcal{X}$, the number of possible samples paths $\{\mathbf{x}_n\}_{n=0}^N$ grows exponentially quickly in N and this renders the marginalisation

$$p(\{\mathbf{y}\}_{n=0}^N|\boldsymbol{\theta}) = \sum_{\mathbf{x}_0,\ldots,\mathbf{x}_N \in \mathcal{X}} p(\{\mathbf{y}\}_{n=0}^N|\{\mathbf{x}_n\}_{n=0}^N, \boldsymbol{\theta}) p(\{\mathbf{x}_n\}_{n=0}^N|\boldsymbol{\theta})$$

computationally intractable.

Type II intractability

- A popular approach to inference under Type II intractability is the pseudo-marginal MCMC of Andrieu and Roberts (2009)
- This replaces the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ in the Metropolis acceptance ratio with an unbiased estimate
- The unbiased estimator of $p(\mathbf{y}|\boldsymbol{\theta})$ can either be obtained by forward-simulation from $p(\mathbf{x}|\boldsymbol{\theta})$, or using importance sampling techniques.

Reduced variance estimation with intractable likelihoods

Problem of interest:

Estimate the posterior expectation $\mu = \mathbb{E}_{\theta|y}[g(\theta)]$ for some known function $g : \Theta \to \mathbb{R}$ where data y arises from an intractable likelihood of either Type I or Type II.

We focus on reducing in the Monte Carlo variance of the estimate of μ through the use of control variates.

Control variates

The approach involves constructing a function

$$\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + h(\boldsymbol{\theta})$$

that satisfies

$$\mathbb{E}_{\theta|y}[\tilde{g}(\boldsymbol{\theta})] = \mathbb{E}_{\theta|y}[g(\boldsymbol{\theta})]$$

and so

$$\mathbb{E}_{\theta|y}[h(\theta)] = 0.$$

Control variates

In many cases it is possible to choose $h(\theta)$ such that the variance $\mathbb{V}_{\theta|y}[\tilde{g}(\theta)] < \mathbb{V}_{\theta|y}[g(\theta)]$, leading to a reduced variance MC estimator

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} \tilde{g}(\theta^{(i)}) \tag{1}$$

where $\theta^{(1)}, \ldots, \theta^{(n)}$ are samples from $\pi(\theta|y)$.

Control variates

In many cases it is possible to choose $h(\theta)$ such that the variance $\mathbb{V}_{\theta|y}[\tilde{g}(\theta)] < \mathbb{V}_{\theta|y}[g(\theta)]$, leading to a reduced variance MC estimator

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} \tilde{g}(\theta^{(i)}) \tag{1}$$

where $\theta^{(1)}, \ldots, \theta^{(n)}$ are samples from $\pi(\theta|y)$.

In classical literature $h(\theta)$ is formed as a sum $\phi_1 h_1(\theta) + \ldots \phi_m h_m(\theta)$ where the $h_i(\theta)$ have zero mean under $\pi(\theta|y)$ and are known as "control variates", whilst ϕ_i are coefficients that must be specified.

Control variates

We focus on control variates expressed as functions of the score of the posterior

$$u(\theta) = \nabla_\theta \log \pi(\theta|y).$$

Control variates

We focus on control variates expressed as functions of the score of the posterior

$$u(\theta) = \nabla_\theta \log \pi(\theta|y).$$

Further, we propose,

$$h(\theta) = \Delta_\theta[P(\theta)] + \nabla_\theta[P(\theta)] \cdot u(\theta)$$

where $P(\theta)$ belongs to the family of polynomials in θ.

Type I intractability:

$$\log \pi(\theta|y) = \theta^T s(y) - \log z(\theta) + \log p(\theta) + C$$

where C is a constant in θ, yielding

$$u(\theta|y) = s(y) - \nabla_\theta \log z(\theta) + \nabla_\theta \log p(\theta).$$

Type I intractability:

$$\log \pi(\theta|y) = \theta^T s(y) - \log z(\theta) + \log p(\theta) + C$$

where C is a constant in θ, yielding

$$u(\theta|y) = s(y) - \nabla_\theta \log z(\theta) + \nabla_\theta \log p(\theta).$$

Although, $u(\theta|y)$ is unavailable, since $z(\theta)$ is intractable, we can estimate it via Monte Carlo simulation.

Estimating the score

Type I intractability:

An unbiased estimate for $u(\theta|y)$ can be computed as follows:

$$\begin{aligned}
\nabla_\theta \log z(\theta) &= \frac{1}{z(\theta)} \nabla_\theta z(\theta) \\
&= \frac{1}{z(\theta)} \nabla_\theta \int \exp(\theta^T s(y)) dy \\
&= \frac{1}{z(\theta)} \int s(y) \exp(\theta^T s(y)) dy \\
&= \mathbb{E}_{Y|\theta}[s(Y)].
\end{aligned}$$

Estimating the score

Type I intractability:

An unbiased estimate for $u(\theta|y)$ can be computed as follows:

$$\begin{aligned}
\nabla_\theta \log z(\theta) &= \frac{1}{z(\theta)} \nabla_\theta z(\theta) \\
&= \frac{1}{z(\theta)} \nabla_\theta \int \exp(\theta^T s(y)) dy \\
&= \frac{1}{z(\theta)} \int s(y) \exp(\theta^T s(y)) dy \\
&= \mathbb{E}_{Y|\theta}[s(Y)].
\end{aligned}$$

We estimate the score function by exploiting multiple forward-simulations

$$\hat{u}(\theta|y) := s(y) - \frac{1}{K} \sum_{k=1}^{K} s(Y_k) + \nabla_\theta \log p(\theta)$$

Type II intractability:

Similarly the score function

$$u(\theta|y) = \nabla_\theta \log \int p(y, x|\theta)p(x|\theta)dx + \nabla_\theta \log p(\theta).$$

is unavailable, but again it can be estimated unbiased.

Estimating the score

Type II:

$$\begin{aligned}
\mathbf{u}(\boldsymbol{\theta}|\mathbf{y}) = \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y})}{p(\boldsymbol{\theta}|\mathbf{y})} \\
&= \frac{1}{p(\boldsymbol{\theta}|\mathbf{y})} \nabla_{\boldsymbol{\theta}} \int p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}) d\mathbf{x} \\
&= \int \frac{[\nabla_{\boldsymbol{\theta}} p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})]}{p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})} \frac{p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})}{p(\boldsymbol{\theta}|\mathbf{y})} d\mathbf{x} \\
&= \int [\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})] p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) d\mathbf{x} \\
&= \mathbb{E}_{\mathbf{X}|\boldsymbol{\theta},\mathbf{y}}[\mathbf{u}(\boldsymbol{\theta}, \mathbf{X})]
\end{aligned}$$

Estimating the score

Type II:

$$\begin{aligned}
\mathbf{u}(\boldsymbol{\theta}|\mathbf{y}) = \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y})}{p(\boldsymbol{\theta}|\mathbf{y})} \\
&= \frac{1}{p(\boldsymbol{\theta}|\mathbf{y})} \nabla_{\boldsymbol{\theta}} \int p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}) d\mathbf{x} \\
&= \int \frac{[\nabla_{\boldsymbol{\theta}} p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})]}{p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})} \frac{p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})}{p(\boldsymbol{\theta}|\mathbf{y})} d\mathbf{x} \\
&= \int [\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})] p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) d\mathbf{x} \\
&= \mathbb{E}_{\mathbf{X}|\boldsymbol{\theta},\mathbf{y}}[\mathbf{u}(\boldsymbol{\theta}, \mathbf{X})]
\end{aligned}$$

yielding a simulation-based estimator

$$\hat{u}(\theta|y) := \frac{1}{K} \sum_{k=1}^{K} u(\theta, X_k)$$

where X_i are independent from the posterior conditional $p(x|\theta, y)$

Control variates for intractable likelihoods

Reduced-variance control variates are then constructed using

$$\hat{h}(\theta|y) := \Delta_\theta[P(\theta)] + \nabla_\theta[P(\theta)] \cdot \hat{u}(\theta|y),$$

where again $P \in \mathcal{P}$ is a polynomial.

We require that $\mathbb{E}_{\theta, Y_1, \ldots, Y_K|y}[\hat{h}(\theta|y)] = 0$.

This is guaranteed under mild assumptions

Optimising the tuning parameters

Our proposed estimator has two tuning parameters:
(i) the polynomial coefficients ϕ
(ii) the number K of forward-simulations from $Y|\theta$

Optimality is defined as maximising the variance reduction factor

$$R := \frac{\mathbb{V}_{\theta, Y_1, \ldots, Y_K|y}[g(\theta)]}{\mathbb{V}_{\theta, Y_1, \ldots, Y_K|y}[g(\theta) + \hat{h}(\theta|y)]}.$$

Optimal polynomial coefficients ϕ

For general degree polynomials $P(\theta|\phi)$ with coefficients ϕ we can write $\hat{h}(\theta|y) = \phi^T m(\theta, \hat{u})$, where in the case of, eg, degree-one polynomials $m(\theta, \hat{u}) = \hat{u}$

Lemma

The variance reduction factor R is maximised over all possible coefficients ϕ by the choice

$$\phi^*(y) := -\mathbb{V}^{-1}_{\theta, Y_1, \ldots, Y_K|y}[m(\boldsymbol{\theta}, \hat{u})]\mathbb{E}_{\theta, Y_1, \ldots, Y_K|y}[g(\theta)m(\boldsymbol{\theta}, \hat{u})]$$

and at the optimal value $\phi = \phi^*$ we have

$$R^{-1} = 1 - \rho(K)^2$$

where $\rho(K) = Corr_{\theta, Y_1, \ldots, Y_K|y}[g(\theta), \hat{h}(\theta|y)]$.

Choosing the number of forward-simulations K

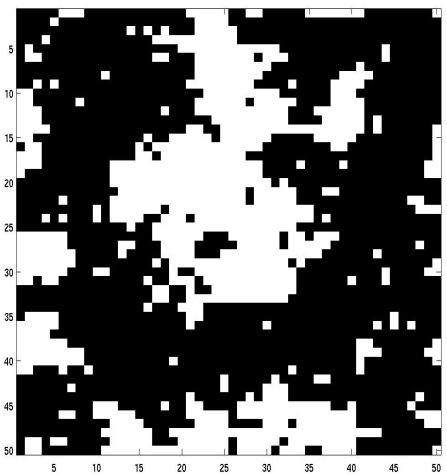
Our main results here may be concisely summarised as follows: For a fixed computational cost,

1. For serial computation, choose $K = 1$. (This typically requires no additional computation since one forward-simulation $\mathbf{Y}$ is generated as part of the exchange algorithm or pseudo-marginal algorithm.)

2. For parallel computation, choose $K = K_0$ equal to the number of available cores.

Application: Ising model (Type I intractability)

- Defined on a lattice $y = \{y_1, \ldots, y_n\}$.
- Lattice points y_i take values $\{-1, 1\}$.
-



$$f(y|\theta) \propto q_\theta(y) = \exp\left\{\frac{1}{2}\theta_1 \sum_{i\sim j} y_i y_j\right\}.$$

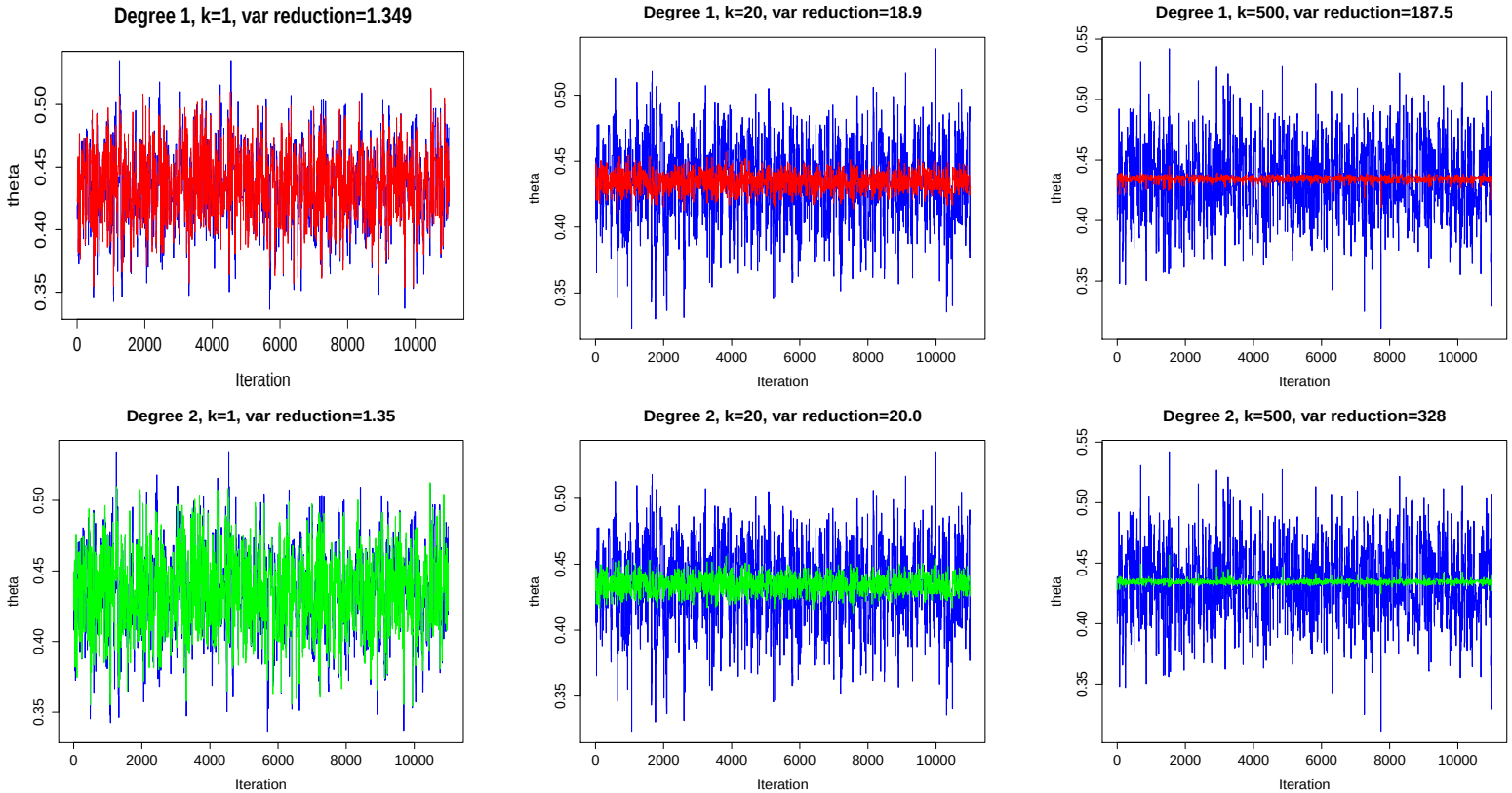
Here $\sim$ means "is a neighbour of".

- With intractable normalising constant

$$z(\theta) = \sum_{y_1} \cdots \sum_{y_n} q_\theta(y).$$

Experiment: Data were simulated from an Ising model defined on a 16×16 lattice which is sufficiently small to allow a very accurate estimate of $E_\pi\theta$.

Application: Ising model



As the number of forward-simulations, K, increases, the precision of the controlled estimate of $E_{\theta|y}(\theta)$ improves

The degree-2 polynomial yields improved precision compared to the degree-1 polynomial, particularly for larger K

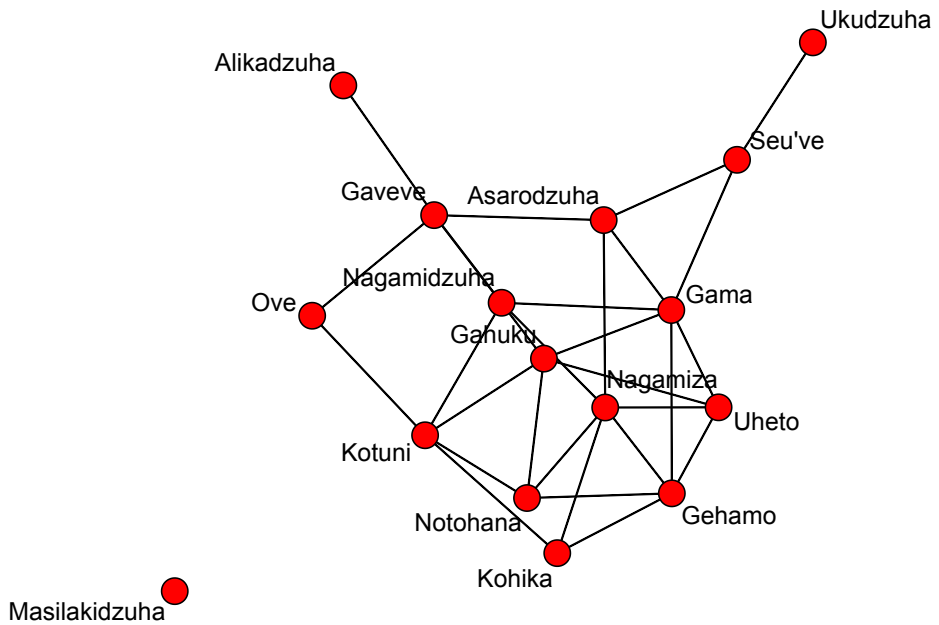
Application: Exponential random graph models (Type I)

- Consider a graph $\{y_{ij}\}$ defined on n nodes.
- $y_{ij} = 1$, if nodes i and j are connected by an edge. Otherwise, $y_{ij} = 0$.
- $$f(y|\theta_1, \theta_2) \propto \exp\{\theta_1 s_1(\mathbf{y}) + \theta_2 s_2(\mathbf{y})\}$$

  where
  $$s_1(\mathbf{y}) = \sum_{i<j} y_{ij} \quad \text{and} \quad s_2(\mathbf{y}) = \sum_{i<j<k} y_{ik} y_{jk}$$
- The normalising constant
  $$z(\theta) = \sum_{\textit{all possible graphs}} \exp\{\theta^t s(y)\}$$

  - $2^{\binom{n}{2}}$ possible undirected graphs of n nodes
  - Calculation of $z(\theta)$ is infeasible for non-trivially small graphs

Exponential random graph models (Type I intractability)



Gamaneg graph: The vertices represent 16 sub-tribes of the Eastern central highlands of New Guinea and edges represent an antagonistic relationship between two sub-tribes

- Goal: estimate the posterior mean $\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[\theta_i]$, for $i = 1, 2$
- The exchange algorithm was run for $I = 11,000$ iterations, where at each iteration $K = 500$ forward-simulations were used to estimate the score.

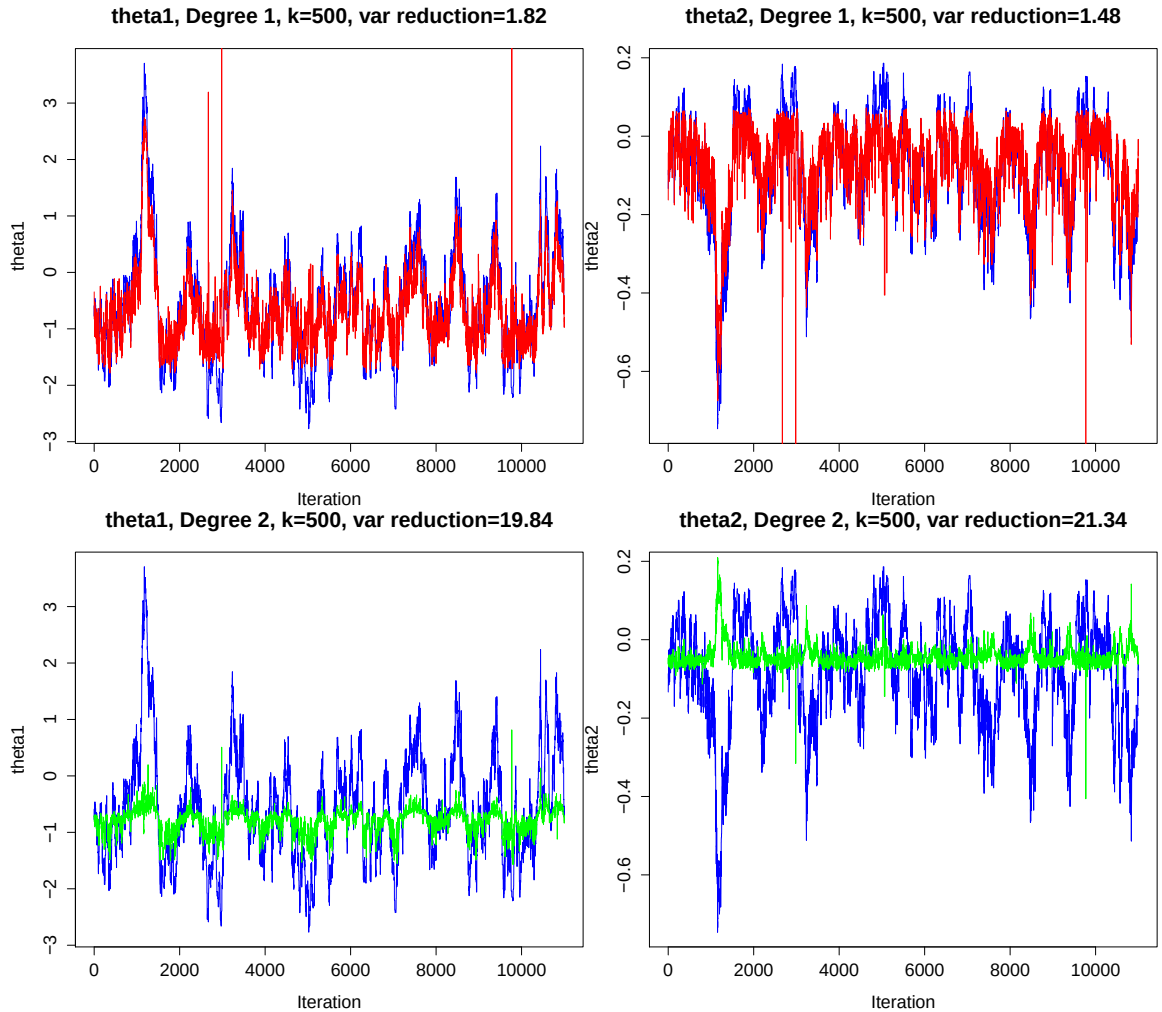


Figure: ERGM: A variance reduction of about 20 times is possible using a degree-two polynomial for each of the two parameters

Application: Nonlinear SDE (Type II)

Goal: Bayesian inference for a system of nonlinear SDEs:

$$d\mathbf{X}(t) = \boldsymbol{\alpha}(\mathbf{X}(t);\boldsymbol{\theta})dt + \boldsymbol{\beta}^{1/2}(\mathbf{X}(t);\boldsymbol{\theta})d\mathbf{W}(t), \qquad \mathbf{X}(0) = \mathbf{X}_0$$

- $\mathbf{X}(t)$ is a stochastic process taking values in $\mathbb{R}^d$,
- $\boldsymbol{\alpha} : \mathcal{X} \times \Theta \to \mathcal{X}$ is a drift function,
- $\boldsymbol{\beta} : \mathcal{X} \times \Theta \to \mathcal{X} \times \mathcal{X}$ is a diffusion function,
- $\mathbf{W}(t)$ is a d-dimensional Weiner process,
- $\boldsymbol{\theta} \in \Theta$ are unknown model parameters
- $\mathbf{X}_0 \in \mathcal{X}$ is a known initial state

Application: Nonlinear SDE

Introduce a fine discretisation $t_1, \ldots, t_T$ of time with mesh size δt. Write $\mathbf{X}_i = \mathbf{X}(t_i)$. Use Euler approximation to the SDE likelihood:

$$p(\mathbf{X}|\boldsymbol{\theta}) \propto \prod_{i=2}^{T} \mathcal{N}(\mathbf{X}_i|\mathbf{X}_{i-1} + \boldsymbol{\alpha}_i \delta t, \boldsymbol{\beta}_i \delta t)$$

where $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}(\mathbf{X}_{i-1}; \boldsymbol{\theta})$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}(\mathbf{X}_{i-1}; \boldsymbol{\theta})$.

Partition $\mathbf{X} = [\mathbf{X}^o \; \mathbf{X}^u]$ such that $\mathbf{y} = \mathbf{X}^o$ are observed (for simplicity without noise) and $\mathbf{x} = \mathbf{X}^u$ are unobserved. Estimate the score using
$\nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}|\mathbf{y}) \approx \frac{1}{K} \sum_{k=1}^{K} \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \mathbf{x}^{(k)}|\mathbf{y})$
where $\mathbf{x}^{(k)}$ are samples from $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ obtained using a Metropolis-Hastings sampler with "diffusion bridge" proposals

Application: Nonlinear stochastic differential equations

Consider the specific example of the susceptible-infected-recovered (SIR) model from epidemiology, which has a stochastic representation given by

$$\boldsymbol{\alpha}(\mathbf{X};\boldsymbol{\theta}) = \begin{bmatrix} -\theta_1 X_1 X_2 \\ \theta_1 X_1 X_2 - \theta_2 X_2 \end{bmatrix},$$

$$\boldsymbol{\beta}(\mathbf{X};\boldsymbol{\theta}) = \frac{1}{N}\begin{bmatrix} \theta_1 X_1 X_2 & -\theta_1 X_1 X_2 \\ -\theta_1 X_1 X_2 & \theta_1 X_1 X_2 + \theta_2 X_2 \end{bmatrix}$$

where $N = 1,000$ is a fixed population size and the rates $\boldsymbol{\theta}$ are unknown

Application: Nonlinear stochastic differential equations

- We estimate the posterior mean of $\boldsymbol{\theta}$, taking $g(\boldsymbol{\theta}) = \theta_j$ for $j = 1, 2, 3$.
- Here each θ_j was assigned an independent Gamma prior with shape and scale parameters equal to 2.
- Data were generated using the initial condition $\mathbf{X}_0 = [0.99, 0.01]$ and parameters $\boldsymbol{\theta} = [0.5, 0.25]$.
- Observations were made at 20 evenly spaced intervals in the period from $t = 0$ to $t = 35$.
- Five latent data points were introduced between each observed data point, so that the latent process has dimension $2 \times (20 - 1) \times 5 = 190$.
- At each Monte Carlo iteration we sampled $K = 100$ realisations of the latent data process $\mathbf{X}^u$ using MCMC.

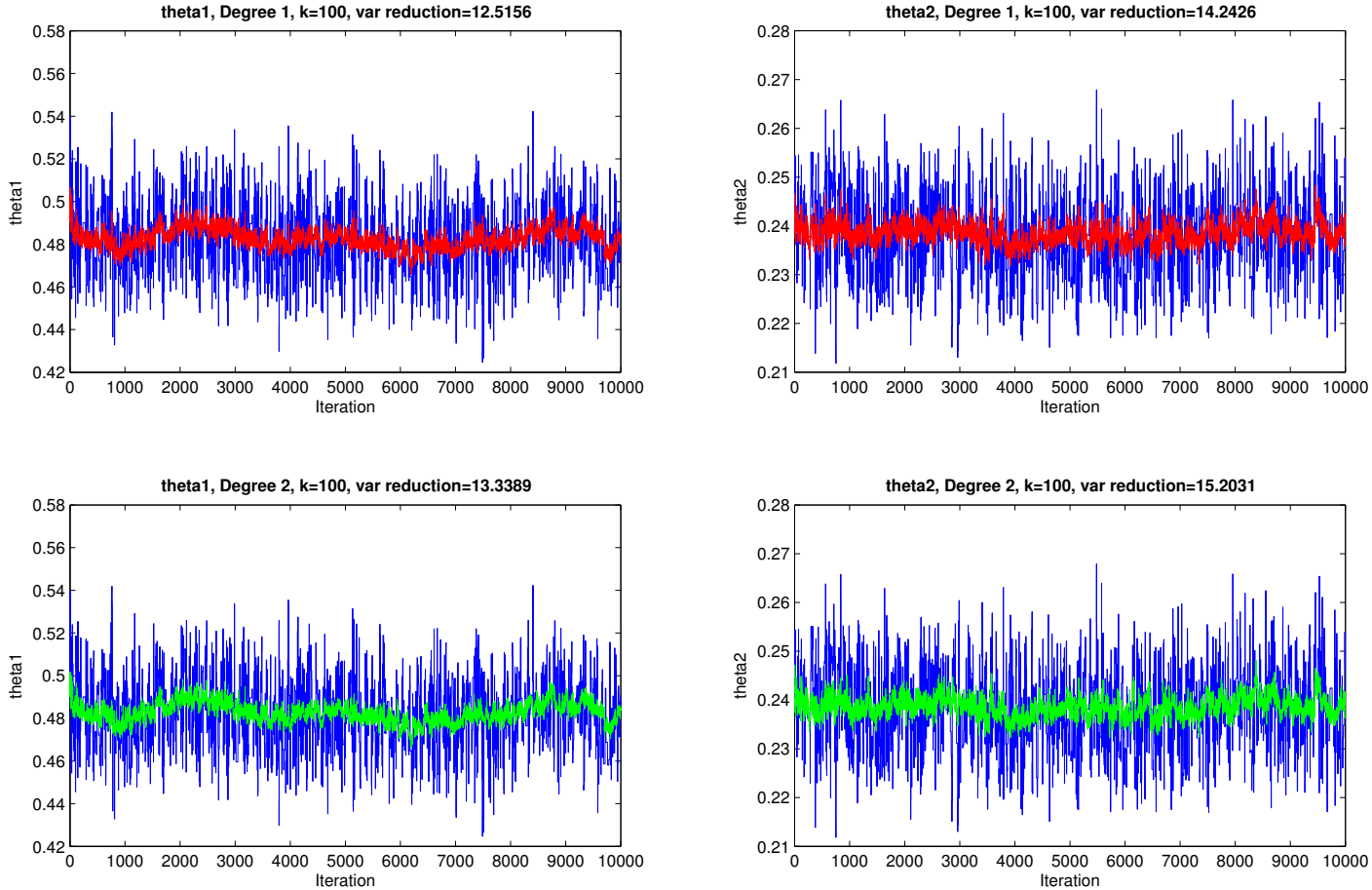


Figure: The top row: trace plot for θ_1 and θ_2 in uncontrolled (blue) and controlled (red/green) versions for a degree-one polynomial. Bottom row is similar but for a degree-two polynomial

Conclusions - PART II

- Exploit simulation from the LHD to achieve reduced-variance estimation in models that have intractable LHD
- $K = 1$ forward-simulation provides the optimal variance reduction per unit (serial) computation
- When multi-core processing architectures are available, additional variance reduction can be achieved
- Reduced-variance estimator can leverage the simulation stage of the exchange algorithm, or the sampling stage of the pseudo-marginal algorithm, to achieve approx ZV with essentially no additional computational effort

References

- Andrieu and Roberts (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*.
- Friel, Mira, Oates (2015) Exploiting Multi-Core Architectures for Reduced-Variance Estimation with Intractable Likelihoods. Bayesian Analysis, 2015
- Murray, Ghahramani, and MacKay. (2006) *MCMC for doubly-intractable distributions*. In Proceedings of the 22nd annual conference on uncertainty in artificial intelligence