Variance reduction strategies for MCMC simulation

Antonietta Mira
University of Lugano, Switzerland

joint with L. Tierney and C. Geyer
R. Solgi, D. Imparato, M. Girolami, T. Papamarkou

ICMS - Edinburgh - 5.6.4
Monte Carlo Integration

Let $\mu_f$ be the expected value of a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, under the (possibly not normalized) measure $\pi$:

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

and let $X_1, X_2, \ldots, X_N$ be a sequence of iid draws from $\pi$. Then an unbiased estimator of $\mu_f$ is:

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

with variance

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

where $\sigma_f^2$ is the variance of $f$ under $\pi$.

Therefore if $f$ has finite variance, $\hat{\mu}_f$ is a consistent estimator of $\mu_f$. 
If drawing iid from $\pi$ is difficult, we can build an **ergodic Markov chain** $X_1, X_2, ..., X_N$, with stationary distribution $\pi$.

In this case $\hat{\mu}_f$ is (asymptotically) an unbiased estimator of $\mu_f$, and:

$$ \text{V}(\hat{\mu}_f) \approx \frac{1}{N} \sigma^2_f \tau_f $$

$$ \tau_f = 1 + 2 \sum_{k=1}^{+\infty} \rho_f(k) $$

$$ = \text{number of correlated samples with same variance-reduction power as 1 iid sample} \tag{1} $$

where $\rho_f(k)$ is the autocorrelation of $f$ at lag $k$ along the chain

Therefore solely finiteness of $\sigma^2_f$ cannot guarantee the consistency of $\hat{\mu}_f$

We need “joint” conditions on both $f(X)$ and the Markov chain
Theorem (Tierney 1996)

Suppose an ergodic Markov chain \( \{X_n\} \), with stationary distribution \( \pi \) and a real valued function \( f \), satisfy one of the following conditions:

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

Then

\[
\lim_{N \to \infty} N \mathbb{V}(\hat{\mu}_f) = \mathbb{E}_\pi [(f(X_0) - \mu_f)^2] + 2 \sum_{k=1}^{+\infty} \mathbb{E}_\pi [(f(X_k) - \mu_f)(f(X_0) - \mu_f)]
\]

\[
= \sigma_f^2 [1 + 2 \sum_{k=1}^{+\infty} \rho_f(k)] = \sigma_f^2 \tau_f
\]

is well defined, non-negative and finite, and

\[
\sqrt{N} (\hat{\mu}_f - \mu_f) \overset{L}{\to} \mathcal{N}(0, \sigma_f^2 \tau_f)
\]
Asymptotic variance in CLT of MCMC estimators

\[ \sigma_f^2 \tau_f = \sigma_f^2 [1 + 2 \sum_{k=1}^{+\infty} \rho_f(k)] = V(f, P) \]

- Delayed rejection strategy \( \rightarrow \) reduce \( \tau_f \) \( \rightarrow \) by modifying \( P \)
- Zero variance strategy \( \rightarrow \) reduce \( \sigma_f^2 \) \( \rightarrow \) by modifying \( f \)
**Regular MH**

Current position $X_t = x$

1. propose a candidate $y \sim q(x, \cdot)$
2. with probability $\alpha(x, y)$ accept $y$: $X_{t+1} = y$
3. otherwise stay where you are: $X_{t+1} = x$

the acceptance probability $\alpha(x, y)$ preserves **REVERSIBILITY** wrt $\pi$:

\[
\int_{(x,y)\in A \times B} \pi(dx) q(x, dy) \alpha(x, y) = \\
\int_{(x,y)\in A \times B} \pi(dy) q(y, dx) \alpha(y, x)
\]
Delayed Rejection MH

Current position $X_t = x$
(1) propose a candidate move $y \sim q_1(x, \cdot)$
(2) with probability $\alpha(x, y)$ let $X_{t+1} = y$

(3) if $y$ is rejected propose a new candidate move $z \sim q_2(x, y, \cdot)$
(4) with probability $\alpha(x, y, z)$ let $X_{t+1} = z$

(5) keep proposing candidates until acceptance
(5’) interrupt the delaying process and set $X_{t+1} = x$

The acceptance probabilities are computed so that reversibility w.r.t. $\pi$ is preserved separately at each stage
Acceptance probabilities

- First stage acc. probability: \[ \alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)} \frac{q_1(y, x)}{q_1(x, y)} \]

- Second stage acc. probability: \[ \alpha(x, y, z) = 1 \wedge \frac{\pi(z)}{\pi(x)} \frac{q_1(z, y)}{q_1(x, y)} \frac{[1 - \alpha(z, y)]}{[1 - \alpha(x, y)]} \frac{q_2(z, y, x)}{q_2(x, y, z)} \]
Delayed Rejection MH

Recursive formulas for accept. probabilities in higher order stages

The DR-MH dominates the MH algorithm in the Peskun (Peskun 1973, Biometrika) and the Covariance ordering (Mira 2001, Stat. Science)
PESKUN-TIERNEY ORDERING

DEFINITIONS

$P$ dominates $Q$ off-diagonally, $P \succeq_P Q$, iff

- finite state spaces
  $P(x, y) \geq Q(x, y)$ \quad $x \neq y$

- general state spaces
  $P(x, B) \geq Q(x, B)$ \quad $x \notin B$

THEOREM
If $P$ and $Q$ are reversible w.r.t. $\pi$ then

$P \succeq_P Q$

$\downarrow$

$V(f, P) \leq V(f, Q) \quad \forall f \in L^2(\pi)$

INTUITION: when $X_{t+1} = X_t$ we fail to explore the state space and increase the covariance along the sample path of the chain
COVARIANCE ORDERING

DEFINITION

\( P \) dominates \( Q \) in the covariance sense, \( \boxed{P \succeq_c Q} \)
if the lag 1 auto-covariances are ordered:

\[
\text{cov}_\pi [f(X_0), f(X_1)] \leq \text{Cov}_\pi [f(Y_0), f(Y_1)], \quad \forall f \in L^2(\pi)
\]

where \( X \sim P \) and \( Y \sim Q \)
or, equivalently, if \( \sigma(Q - P) \subseteq [0, \infty) \)

THEOREM

If \( P \) and \( Q \) are reversible w.r.t. \( \pi \) then
\[
\boxed{P \succeq_c Q} \quad \leftrightarrow \quad V(f, P) \leq V(f, Q) \quad \forall f \in L^2(\pi)
\]
I can order all the MC of Peskun + more

I can order the eigenvalues: $\lambda_{1,P} \geq \lambda_{2,P} \cdots$

$P \succeq_{C} Q \Rightarrow \lambda_{i,P} \leq \lambda_{i,Q}, \quad \forall i$

I can say something about non-reversible MC

Extended to continuous time MC
Peskun says that whenever $X_{t+1} = X_t$
- fail to explore the state space
- increase autocorrelation along sample path
- increase the variance of MCMC estimators

In the MH algorithm this happens every time a candidate move is rejected

Thus we can beat MH in the Peskun sense by diminishing the rejection frequency

Whether delaying rejection is useful in practice depends on whether the reduction in variance compensates the additional computational cost
ADJUSTING THE PROPOSAL DIST.

One possible reason for rejection in MH algorithms is that the proposal is locally badly calibrated to the target. With the delaying strategy you have freedom to use intuition in designing the way proposals at later stages "learn" from previous mistakes. Validity is ensured by using the correct acceptance probability.

- easy first and more computationally intensive proposals only if really needed
- bold first and more timid moves later
- independence + rnd walk proposals: the rnd walk gives protection against the potentially poor behavior of an independence chain with bad proposal distribution
- trust region based proposals: start with a local quadratic approximation of log (π) and gradually reduce the region supporting the proposal
- griddy proposals: select a point from the previously rejected ones with probability ∝ π(y_j) and add to the point a random increment
Simulation results

Target: posterior in a hierarchical logistic regression model for credit scoring

2-stages random walk DR-MH compared with MH with $\sigma = \frac{\sigma_1 + \sigma_2}{2}$

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\beta}_2$</th>
<th>$\hat{\beta}_3$</th>
<th>$\hat{\beta}_4$</th>
<th>$\mu_\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MH</td>
<td>6.5</td>
<td>22.9</td>
<td>21.1</td>
<td>5.7</td>
<td>17.2</td>
</tr>
<tr>
<td>DR</td>
<td>3.1</td>
<td>6.9</td>
<td>9.7</td>
<td>1.9</td>
<td>8.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\alpha}_1$</th>
<th>$\hat{\alpha}_2$</th>
<th>$\hat{\alpha}_3$</th>
<th>$\hat{\alpha}_4$</th>
<th>$\hat{\alpha}_5$</th>
<th>$\hat{\alpha}_6$</th>
<th>$\hat{\alpha}_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MH</td>
<td>9.8</td>
<td>15.3</td>
<td>20.0</td>
<td>18.7</td>
<td>20.7</td>
<td>23.7</td>
<td>25.6</td>
</tr>
<tr>
<td>DR</td>
<td>4.7</td>
<td>8.8</td>
<td>7.4</td>
<td>11.9</td>
<td>10.5</td>
<td>14.1</td>
<td>11.7</td>
</tr>
</tbody>
</table>

Values obtained averaging over 50 simulations of length 1500 after a 150 burn-in steps
Extensions

- DR + AM = DRAM
  combine local and global calibration
- DR + RJ = DRRJ
  extension when the dimensions of the target are not fixed
- DR + variational Monte Carlo
- DR for doubly intractable problems
- Others have used DR: 300 (DRAM) + 150 (DRRJ) + 100 (DR)
- 2 stage and second antithetic move is best: M. Bdard, R. Douc, E. Moulines
  Scaling Analysis of Delayed Rejection MCMC
Conclusions

- The **covariance ordering** gives a nec. and suff. condition for **absolute efficiency**
- The **delaying rejection** strategy improves Metropolis-Hastings-Green algorithms in terms of absolute efficiency modulo extra computational effort

Easy to **modify an existing code** for a MH algorithm to allow delaying rejection
For MC simulation, control (and antithetic) variate methods have been used in order to reduce the variance of MC estimators.\footnote{Ripley B. D., Stochastic Simulation, Wiley, 1983.}

Assume $Z$ is a random variable with zero (or known) mean, and correlated with $f(X)$:

\[
\begin{align*}
E(Z) &= 0 \\
\text{Cov}(f(X), Z) &= \sigma_{f,Z} \neq 0
\end{align*}
\]

By exploiting the correlation of $f(X)$ and $Z$, we can build new \textbf{unbiased} estimators of $\mu_f$, with \textbf{lower variances}. Let’s define:

\[
\tilde{f}(X) := f(X) + aZ
\]

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

\[
\begin{align*}
\mu_{\tilde{f}} := E[\tilde{f}(X)] &= \mu_f \\
\sigma^2_{\tilde{f}} &= \sigma^2_f + a^2 \sigma^2_Z + 2a \sigma_{f,Z}
\end{align*}
\]
Control Variate Method in MC

By minimizing $\sigma^2_f$ w.r.t. $a$, it can be shown that the optimal choice of $a$ is

$$a = -\frac{\sigma_{f,Z}}{\sigma^2_Z}$$

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

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

is a new unbiased estimator of $\mu_f$, with variance

$$\mathbb{V} (\hat{\mu}_f) = \frac{1}{N} \sigma^2_{\tilde{f}} = \frac{1}{N} \left(1 - \rho^2_{f,Z}\right)\sigma^2_f \leq \frac{1}{N} \sigma^2_f = \mathbb{V} (\hat{\mu}_f)$$
Main Idea

Under regularity conditions, the score has zero mean

Use it as a control variate!
Zero Variance Principle for MC

For MC simulation, Assaraf and Caffarel (1999)\(^2\) propose the ZV principle based on control variates. Let \(H\) be a Hermitian operator, that satisfies the following condition:

\[
\int H(x, y) \sqrt{\pi(y)} dy = 0
\]

and define:

\[
\tilde{f}(x) := f(x) + \varepsilon(x)
\]

\[
\varepsilon(x) := \frac{\int H(x, y) \psi(y) dy}{\sqrt{\pi(x)}}
\]

where \(\psi(x)\) is a trial function.

Under mild conditions, \(\mu_{\tilde{f}} = \mu_f\) (2)

and depending on the choice of \(H\) and \(\psi\)

\[
\sigma^2_{\tilde{f}} << \sigma^2_f
\]

Zero Variance Principle for MC

Zero-variance can be reached iff $H$ and $\psi$ satisfy the following equation:

$$\int H(x, y) \sqrt{\psi(y)} dy = - [f(x) - \mu_f] \sqrt{\pi(x)}$$

However the unknown $\mu_f$ appears in this equation! Therefore we cannot solve it for optimal $H$ and $\psi$, unless we know $\mu_f$

In practice:

- select an operator $H$
- specify a family of functions for $\psi$ indexed by a vector of parameters, $a$
- find expressions for optimal parameters, $a$ the minimize $\sigma(\tilde{f})$
- estimate the optimal $a$ using the existing MCMC simulation
- estimate $\mu_f$ with $\mu_{\tilde{f}}$
Choice of \((H, \psi)\)

Assaraf and Caffarel (1999) proposed different operators \(H\):
- the Schrödinger-type Hamiltonian

\[
H \equiv -\frac{1}{2} \Delta + \frac{1}{2} \frac{1}{\sqrt{\pi(x)}} \Delta \sqrt{\pi(x)}
\]  

(3)

- Alternative operator used in Dellaportas et al. (JRSSB, 12):

\[
H(x, y) = \sqrt{\frac{\pi(x)}{\pi(y)}} [P(x, y) - \delta(x - y)],
\]  

(4)

where \(P(x, y)\) is the kernel of the chain.
We chose the Schrödinger-type Hamiltonian operator and parametrize \(\psi\) as:

\[
\psi(x) = P(x) \sqrt{\pi(x)}
\]

where \(P(x)\) is a polynomial function.
Using the Schrödinger-type Hamiltonian $H$ and $\psi(x) = P(x)\sqrt{\pi(x)}$, the following holds:

**Corollary 1**

Let $\pi$ be a unidimensional target distribution, and $l$ and $u$ be the lower and upper bound of its support. The expected value of $\varepsilon(X)$ is equal to zero, if

$$\pi(u)P'(u) = \pi(l)P'(l)$$

**Corollary 2**

Let $\pi$ be a $d$-dimensional target distribution, and $\partial\Omega$ be the boundary of its support. The expected value of $\varepsilon(X)$ is equal to zero, if

$$\forall x \in \partial\Omega, \forall j \in \{1, ..., d\}; \quad \pi(x)\frac{\partial P(x)}{\partial x_j} = 0$$
Using the same $H$ and $\psi(x) = P(x) \sqrt{\pi(x)}$, we have:

$$\tilde{f}(x) = f(x) - \frac{1}{2} \Delta P(x) + \nabla P(x) \cdot z,$$

where

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

Therefore for linear $P(x)$,

$$P(x) = a^T x$$

we have:

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

Therefore under this setting we have $d$ control variates.

The optimal value of $a$ is:

$$a = -\Sigma_{zz}^{-1} \sigma_{zf}, \quad \text{where} \quad \Sigma_{zz} = \mathbb{E}(zz^T), \quad \sigma_{zf} = \mathbb{E}(zf)$$

The optimal $a$ is estimated using the existing MCMC simulation.
Quadratic $P(x)$

For quadratic $P(x)$,

$$P(x) = a^T x + \frac{1}{2} x^T B x$$

we have:

$$\tilde{f}(x) = f(x) - \frac{1}{2} \text{tr}(B) + (a + B x)^T z = f(x) + g^T y$$

where $g$ and $y$ are column vectors with $\frac{1}{2} d(d+3)$ elements defined as:

- $g := [a^T b^T c^T]^T$: 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 $i \in \{2, \ldots, d\}$, 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 $i \in \{2, \ldots, d\}$, 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, and in general, with a polynomial of order $p$, we get $\binom{d+p}{p} - 1$ control variates.
Corollary 3

Let \( \{X_n\} \) be an ergodic Markov chain, stationary wrt \( \pi \) and \( f \) be a real valued function. The ZV-MCMC estimator \( \hat{\mu}_f \) is an unbiased consistent estimator of \( \mu_f \) if in addition to the condition of Corollary 1 or 2, one of the following conditions holds:

For linear \( P(x) \):
- The chain is uniformly ergodic and \( \mathbb{E}_\pi \left[ z_j^2 \right] < \infty \) for all \( j \in \{1, \ldots, d\} \)
- The chain is geometrically ergodic and \( \mathbb{E}_\pi \left[ |z_j|^{2+\epsilon} \right] < \infty \) for all \( j \in \{1 \ldots d\} \) and some \( \epsilon > 0 \)

For quadratic \( P(x) \):
- The chain is uniformly ergodic and \( \mathbb{E}_\pi \left[ x_i^2 z_j^2 \right] < \infty \) for all \( i, j \in \{1, \ldots, d\} \)
- The chain is geometrically ergodic and \( \mathbb{E}_\pi \left[ |x_i z_j|^{2+\epsilon} \right] < \infty \) for all \( i, j \in \{1 \ldots d\} \) and some \( \epsilon > 0 \)
To guarantee a CLT for $\hat{\mu}_{\tilde{f}}$ we need to prove the existence of 2nd moment of $f(X)$ (for uniformly ergodic chains) or $2 + \epsilon$-th moment of $|f(X)|$ (for geometrically ergodic chains).

For linear $P(x)$ we have:

$$\tilde{f}(x) = f(x) + a^T z, \quad z := -\frac{1}{2} \nabla \ln \pi(x)$$

Therefore a sufficient condition for $E_{\pi} \left[ |\tilde{f}(X)|^{2+\epsilon} \right] < \infty$ is:

$$E_{\pi} \left[ |z_j|^{2+\epsilon} \right] = \int_{\mathbb{R}^d} \left| \frac{\partial \ln \pi(x)}{\partial x_j} \right|^{2+\epsilon} \pi(x) dx < \infty \quad \forall j \in \{1, \ldots, d\}$$

Setting $\epsilon = 0$, this functional is the *Fisher Information with respect to location* $= I[\pi]$. 
Finiteness of $I[\pi]$

To prove CLT for ZV-MCMC must investigate the finiteness of $I[\pi]$ w.r.t. the target $\pi$

$I[\pi]$ is finite for common densities, such as Gaussian, exponential, Boltzmann, Laplace, ... 

For exponential families we have:

**Theorem**

Let $\pi \propto \exp(\theta \cdot T - K_p(\theta))$ belong to a $d$–dimensional exponential family, where $p$ is such that $\frac{\partial \log p}{\partial x_j} \in L^2(\pi)$.

Then, $I[\pi] < \infty$ if and only if

$$\forall i, \forall k, \frac{\partial T_k}{\partial x_i} \in L^2(\pi)$$
Zero Variance MCMC in a Bayesian setting

Target \( \pi = \text{posterior} \):

\[
\pi(\theta|x) \propto l(\theta|x)\pi_0(\theta).
\]

The CVs depend on the data \( x \): \( z_j = z_j(\theta;x) \)
Therefore, CLT conditions for ZVMCMC should be verified for any \( x \in \mathcal{X} \), i.e., the finiteness of a \( d \)-dimensional integral in the variable \( \theta \) has to be proved for different \( x \).
To this aim the following approach can be used:

- Suppose that \( x \) is s.t. the MLE \( l(\theta|x) \) exists.
- Express the integral of interest in hyper-spherical coordinates as

\[
\int_{[0,2\pi]^{d-1}} \int_{0}^{\infty} K(\rho, \theta;x) d\rho d\theta := \int_{[0,2\pi]^{d-1}} A(\theta;x) d\theta,
\]

where \( \rho \in (0, +\infty) \), \( \theta \in [0, 2\pi]^{d-1} \)
- prove that \( A(\theta;x) < \infty \) for any \( x \) and for any \( \theta \) in the compact set \([0, 2\pi]^{d-1}\).

We proved that a CLT for ZVMCMC holds for the Bayesian estimation of many models like probit, logit and GARCH.
Probit Model

\[ y_i | x_i \sim B(1, p_i), \quad p_i = \Phi(x_i^T \beta) \]

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

The likelihood function is:

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

We use flat priors and the Bayesian estimator of each parameter, \( \beta_d \), is

\[
\mathbb{E}_\pi[\beta_k | y, X], \quad k = 1, 2, \cdots, d
\]
Probit Model

Using the Schrödinger-type Hamiltonian $H$, $\psi_k(\beta) = P_k(\beta)\sqrt{\pi(\beta)}$, where $P_k(\beta) = \sum_{j=1}^{d} a_{j,k}\beta_j$ we have:

$$\tilde{f}_k(\beta) = f_k(\beta) + \sum_{j=1}^{d} a_{j,k}z_j \quad \text{where} \quad z_j = -\frac{1}{2} \sum_{i=1}^{n} x_{ij} \frac{\phi(x_i^T\beta)}{\Phi(x_i^T\beta)}$$

It can be easily shown that the existence of MLE implies the unbiasedness of the ZVMCMC estimator.
**Example:**³ 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).

Probit Model

Zero-Variance MCMC

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

In the second stage (2000 steps), we run another MCMC simulation, independent of the first one, to estimate $\hat{\mu}_f$ using the trial functions obtained in the first stage.

We use the Albert and Chib\textsuperscript{4} Gibbs sampler. We try both a linear and a quadratic trial function $P(\beta)$.

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)
Linear ZV-MCMC estimates (2)
Quadratic ZV-MCMC estimates (3)
+ 95% confidence region obtained by an ordinary MCMC of length $10^8$ (green)
Logit Model

\[ y_i|x_i \sim B(1, p_i), \quad p_i = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \]

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

We use flat priors and the Bayesian estimator of each parameter, \( \beta_d \), is
\[ \mathbb{E}_\pi [\beta_k|y, X] \quad k = 1, 2, \ldots, d \]

Using the operator \( H \) and trial function \( \psi_k(\beta) = P_k(\beta) \sqrt{\pi(\beta)} \), we have:

\[ \widetilde{f}_k(\beta) = f_k(\beta) + \sum_{j=1}^d a_{j,k} z_j \quad \text{where} \quad z_j = \frac{1}{2} \sum_{i=1}^n x_{ij} \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \]
**Logit Model**

Existence of MLE in the logit model implies the finiteness of $2 + \epsilon$-th moment of the control variates. Therefore, there is a CLT for $\tilde{f}_k$.

**Example:** We fit a logit model to the same dataset of Swiss banknotes of Flury and Riedwyl (1988), introduced in the previous example.
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)
Linear ZV-MCMC estimates (2)
Quadratic 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:

$$r_{t+1}|\mathcal{F}_t \sim \mathcal{N}(0, h_{t+1}^2)$$

$$h_{t+1}^2 = \omega + \alpha r_t^2 + \beta h_t^2$$

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:

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

The Bayesian estimators of the GARCH parameters are \( \mathbb{E}_\pi [\omega | r] \), \( \mathbb{E}_\pi [\alpha | r] \) and \( \mathbb{E}_\pi [\beta | r] \). Using the usual operator \( H \) and trial function \( \psi \), the control variates are:

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

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

\[
\begin{align*}
\frac{\partial h_t}{\partial x_1} &= \frac{\partial h_t}{\partial \omega} = \frac{1 - \beta^{t-1}}{1 - \beta} \\
\frac{\partial h_t}{\partial x_2} &= \frac{\partial h_t}{\partial \alpha} = \begin{cases} 0 & t = 1 \\
r_t^2 + \beta \frac{\partial h_{t-1}}{\partial \alpha} & t > 1 \end{cases} \\
\frac{\partial h_t}{\partial x_3} &= \frac{\partial h_t}{\partial \beta} = \begin{cases} 0 & t = 1 \\
h_{t-1} + \beta \frac{\partial h_{t-1}}{\partial \beta} & t > 1 \end{cases}
\end{align*}
\]
GARCH Model

**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 to estimate the optimal parameters of the trial function. In the second stage we run an independent MCMC simulation and compute $\tilde{f}_j(x)$.

**Estimates of Parameters:**

<table>
<thead>
<tr>
<th>Method</th>
<th>$\hat{\omega}$</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>0.0445</td>
<td>0.2104</td>
<td>0.6541</td>
</tr>
<tr>
<td>MCMC</td>
<td>0.0568</td>
<td>0.2494</td>
<td>0.5873</td>
</tr>
</tbody>
</table>

**Variance Reduction:**

$\frac{(\text{Sokal estimate of std. error of MC estimator})^2}{(\text{Sokal estimate of std. error of ZV-MC estimator})^2}$

<table>
<thead>
<tr>
<th>Degree</th>
<th>$\hat{\omega}$</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>9</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td>2nd</td>
<td>2,070</td>
<td>12,785</td>
<td>11,097</td>
</tr>
<tr>
<td>3rd</td>
<td>28,442</td>
<td>70,325</td>
<td>30,281</td>
</tr>
</tbody>
</table>
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)
Linear, Quadratic, Cubic ZV-MCMC estimates (2, 3, 4)
+ 95% confidence region obtained by an ordinary MCMC of length $10^8$ (green)
Example from Dellaportas and Kontoyiannis (2010): model

\[ Y \sim \mathcal{N}(m, V) \]

with priors:

\[ m \sim \text{Cauchy}(0, 1) \]
\[ V \sim \text{IG}(1, 1) \]

Therefore the posterior is:

\[ \pi(m, V) = \prod_{i=1}^{n} \phi(y_i; m, V) \ f_{\text{Cauchy}}(m; 0, 1) \ f_{\text{IG}}(V; 1, 1) \]

and the control variates are:

\[ z_1 = -\frac{1}{2} \frac{\partial \pi}{\partial m} = \frac{m}{m^2 + 1} - \frac{\sum y_i - nm}{2V} \]
\[ z_2 = -\frac{1}{2} \frac{\partial \pi}{\partial V} = \frac{n + 4}{4V} - \frac{2 + \sum y_i^2 + nm^2 - 2m \sum y_i}{4V^2} \]
Metropolis-within-Gibbs Sampling from a Heavy Tailed Distribution

Variance Reduction:

\[
\frac{\text{Sokal estimate of std. error of } \text{MCMC estimator}}{\text{Sokal estimate of std. error of } \text{ZV-MCMC estimator}}^2
\]

<table>
<thead>
<tr>
<th>Degree</th>
<th>Function</th>
<th>(V)</th>
<th>(m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>(P(x))</td>
<td>1.1</td>
<td>101</td>
</tr>
<tr>
<td>2nd</td>
<td>(P(x))</td>
<td>39</td>
<td>2800</td>
</tr>
<tr>
<td>3rd</td>
<td>(P(x))</td>
<td>107000</td>
<td>113000</td>
</tr>
</tbody>
</table>
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

Conclusions

- 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