A. Mira – R. Solgi – D. Imparato

Zero Variance Markov Chain Monte Carlo for Bayesian Estimators

2011/9

UNIVERSITÀ DELL'INSUBRIA
FACOLTÀ DI ECONOMIA

http://eco.uninsubria.it

In questi quaderni vengono pubblicati i lavori dei docenti della Facoltà di Economia dell'Università dell'Insubria. La pubblicazione di contributi di altri studiosi, che abbiano un rapporto didattico o scientifico stabile con la Facoltà, può essere proposta da un professore della Facoltà, dopo che il contributo sia stato discusso pubblicamente. Il nome del proponente è riportato in nota all'articolo. I punti di vista espressi nei quaderni della Facoltà di Economia riflettono unicamente le opinioni degli autori, e non rispecchiano necessariamente quelli della Facoltà di Economia dell'Università dell'Insubria.

These Working papers collect the work of the Faculty of Economics of the University of Insubria. The publication of work by other Authors can be proposed by a member of the Faculty, provided that the paper has been presented in public. The name of the proposer is reported in a footnote. The views expressed in the Working papers reflect the opinions of the Authors only, and not necessarily the ones of the Economics Faculty of the University of Insubria.

Printed in Italy in March 2011
Università degli Studi dell'Insubria
Via Monte Generoso, 71, 21100 Varese, Italy

Zero Variance Markov Chain Monte Carlo for Bayesian Estimators

Antonietta Mira[*], Reza Solgi[†], Daniele Imparato[‡]

March 10, 2011

Abstract

A general purpose variance reduction technique for Markov chain Monte Carlo (MCMC) estimators, based on the zero-variance principle introduced in the physics literature, is proposed to evaluate the expected value, μ_f, of a function f with respect to a, possibly unnormalized, probability distribution π. In this context, a control variate approach, generally used for Monte Carlo simulation, is exploited by replacing f with a different function, $\tilde{f}$. The function $\tilde{f}$ is constructed so that its expectation, under π, equals μ_f, but its variance with respect to π is much smaller. Theoretically, an optimal re-normalization μ_f exists which may lead to zero variance; in practice, a suitable approximation for it must be investigated.

In this paper, an efficient class of re-normalized $\tilde{f}$ is investigated, based on a polynomial parametrization. We find that a low-degree polynomial (1st, 2nd or 3rd degree) can lead to dramatically huge variance reduction of the resulting zero-variance MCMC estimator. General formulas for the construction of the control variates in this context are given. These allow for an easy implementation of the method in very general settings regardless of the form of the target/posterior distribution (only differentiability is required) and of the MCMC algorithm implemented (in particular, no reversibility is needed).

[*]Università dell'Insubria, Varese. E-mail: antonietta.mira@uninsubria.it
[†]Istituto di Finanza, Università di Lugano. E-mail: reza.solgi@usi.ch
[‡]Università dell'Insubria, Varese. E-mail: daniele.imparato@uninsubria.it

Conditions for asymptotic unbiasedness of the zero-variance estimator are derived in the general setting of zero-variance principle. A central limit theorem is also proved under regularity conditions.

The potential of the new idea is illustrated with real applications to Bayesian inference for probit, logit and GARCH models. For all these models, CLT and unbiasedness for the zero-variance estimator are proved.

Keywords: Control variates; GARCH models; Logistic regression; Metropolis-Hastings algorithm; Variance reduction.

1 General idea

The expected value of a function f with respect to a, possibly unnormalized, probability distribution π

$$\mu_f = \frac{\int f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}}{\int \pi(\mathbf{x})d\mathbf{x}} \tag{1}$$

is to be evaluated. Markov chain Monte Carlo (MCMC) methods estimate integrals using a large but finite set of points, $\mathbf{x}^i, i = 1, \cdots, N$, collected along the sample path of an ergodic Markov chain having π (normalized) as its unique stationary and limiting distribution $\hat{\mu}_f = \sum_{i=1}^{N} f(\mathbf{x}^i)/N$.

In this paper a general method is suggested to reduce the MCMC error by replacing f with a different function, $\tilde{f}$, obtained by properly re-normalizing f. The function $\tilde{f}$ is constructed so that its expectation, under π, equals μ_f, but its variance with respect to π is much smaller. To this aim, a standard variance reduction technique introduced for Monte Carlo (MC) simulation, known as control variates (Ripley 1987), is exploited.

In the univariate setting, a control variate is a random variable, z, with zero (or known) mean under π, and correlated with $f(x)$: $\sigma(f, z) \neq 0$. By exploiting this correlation, a new unbiased estimator of μ_f, with lower variance, can be built. Let $a \in \mathbb{R}$ and define $\tilde{f}(x) = f(x) + az$. By construction, $\mu_{\tilde{f}} = \mu_f$ and $\sigma^2(\tilde{f}) = \sigma^2(f) + a^2\sigma^2(z) + 2a\sigma(f, z)$.

Minimizing $\sigma^2(\tilde{f})$ w.r.t. a gives the optimal choice of the parameter

$$a = -\frac{\sigma(f, z)}{\sigma^2(z)}, \tag{2}$$

that reduces the variance of $\sigma^2(\tilde{f})$ to $(1-\rho^2(f,z))\,\sigma^2(f)$. Therefore $\hat{\mu}_{\tilde{f}} := \sum_{i=1}^{N} \tilde{f}(\mathbf{x}_i)/N$ is a new unbiased MC estimator of μ_f, with variance

$$\sigma^2\left(\hat{\mu}_{\tilde{f}}\right) = \frac{1}{N}\sigma^2(\tilde{f}) = \frac{1}{N}\left(1-\rho^2(f,z)\right)\sigma^2(f) \quad \leq \quad \frac{1}{N}\sigma^2(f) = \sigma^2\left(\hat{\mu}_f\right).$$

This idea can be extended to more than one control variate.

In the rest of this section we briefly explain the zero-variance (ZV) principle introduced in (Assaraf and Caffarel 1999, 2003): an almost automatic method to construct control variates for Monte Carlo simulation. To this end, an operator, H, and a trial function, ψ, are introduced. H is required to be Hermitian (a self-adjoint operator, real in all practical applications) and

$$H\sqrt{\pi} = 0. \tag{3}$$

For $H = H(\mathbf{x},\mathbf{y})$, the weaker condition

$$\int H(\mathbf{x},\mathbf{y})\sqrt{\pi(\mathbf{y})}d\mathbf{y} = 0 \tag{4}$$

is needed. The trial function $\psi(\mathbf{x})$ is a rather arbitrary function, whose first and second derivatives are required to be continuous. The re-normalized function is defined to be

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{\int H(\mathbf{x},\mathbf{y})\psi(\mathbf{y})d\mathbf{y}}{\sqrt{\pi(\mathbf{x})}}. \tag{5}$$

As a consequence of (1) and (4) $\mu_f = \mu_{\tilde{f}}$, that is, both functions f and $\tilde{f}$ can be used to estimate the desired quantity via Monte Carlo or MCMC simulation. However, the statistical error of the two estimators can be very different. The optimal choice for (H,ψ), i.e. the one that leads to zero variance, can be obtained by imposing that $\tilde{f}$ is constant and equal to its average, $\tilde{f} = \mu_f$, which is equivalent to require that $\sigma(\tilde{f}) = 0$. The latter, together with (5), leads to the fundamental equation:

$$\int H(\mathbf{x},\mathbf{y})\psi(\mathbf{y})d\mathbf{y} = -\sqrt{\pi(\mathbf{x})}[f(\mathbf{x}) - \mu_f]. \tag{6}$$

In most practical applications equation (6) cannot be solved exactly, still, we propose to find an approximate solution in the following way. First choose H verifying (3). Second, parametrize ψ and derive the optimal parameters

by minimizing $\sigma^2(\tilde{f})$. The optimal parameters are then estimated using a first short MCMC simulation. Finally, a much longer MCMC simulation is performed using $\hat{\mu}_{\tilde{f}}$ instead of $\hat{\mu}_f$ as the estimator.

Previous research in the statistical literature aims at reducing the asymptotic variance of MCMC estimators by modifying the transition kernel of the Markov chain. These modifications have been achieved in many different ways, for example by trying to induce negative correlation along the chain path (Barone and Frigessi 1989; Green and Han 1992; Craiu and Meng 2005; So 2006; Craiu and Lemeieux 2007); by trying to avoid random walk behavior via successive over-relaxation (Adler 1981; Neal 1995; Barone, Sebastiani, and Stander 2001); by hybrid Monte Carlo (Duane, Kennedy, Pendleton, and Roweth 2010; Neal 1994; Brewer, Aitken, and Talbot 1996; Fort, Moulines, Roberts, and Rosenthal 2003; Ishwaran 1999); by exploiting non reversible Markov chains (Diaconis, Holmes, and Neal 2000; Mira and Geyer 2000), by delaying rejection in Metropolis-Hastings type algorithms (Tierney and Mira 1999; Green and Mira 2001), by data augmentation (Van Dyk and Meng 2001; Green and Mira 2001) and auxiliary variables (Swendsen and Wang 1987; Higdon 1998; Mira, Möller, and Roberts 2001; Mira and Tierney 2002). Up to our knowledge, the only other research line that uses control variates in MCMC simulation follows the seminal paper by (Henderson 1997) and has its most recent developement in (Dellaportas and Kontoyiannis 2010). In (Henderson and Glynn 2002) it is observed that, for any real-valued function g defined on the state space of a Markov chain $\{X^n\}$, the one-step conditional expectation $U(x) := g(x) - \mathbb{E}[g(X^{n+1})|X^n = x]$ has zero mean with respect to the stationary distribution of the chain and can thus be used as control variate. The Authors also note that the best choice for the function g is the solution of the associated Poisson equation which can rarely be obtained analytically but can be approximated in specific settings. A related technical report by (Dellaportas and Kontoyiannis 2010), further explores the use of this type of control variates in the setting of reversible Markov chains were a closed form expression for U is often available.

In (Assaraf and Caffarel 1999, 2003) unbiasedness and existence of a CLT for the ZV estimator are not discussed. The main contribution of this paper is to derive the rigorous conditions for unbiasedness and CLT for the ZV estimators in MCMC simulation. We also demonstrate that for some widely used models (probit, logit, and GARCH) under very mild condition (existence of MLE), the necessary conditions for unbiasedness and CLT are verified.

The paper is organized as follows. In Sections 2 and 3, different choices of the operator H and the trial function ψ are presented. In Section 4, expressions for the control variates are explicitely found, depending on the set of trial functions considered. The optimal control variates, that is, the optimal parameters which give maximal variance reduction for particular classes of ψ are discussed in Section 6. Sections 5 and 7 deal with the mathematical conditions which ensure that the optimal ZV-MCMC estimators are unbiased and obey a CLT. Sufficient conditions are given and verified, that will be verified in the final examples discussed in Section 8: Probit, Logit and Garch models in a Bayesian framework. The simulations show that, even by considering a low-dimensional parametric class of trial functions, a huge variance reduction can be achieved.

2 Choice of H

In this section guidelines to choose the operator H, both for discrete and continuos settings, are given. In a discrete state space, denote with $P(\mathbf{x}, \mathbf{y})$ a transition matrix reversible with respect to π (a Markov chain will be identified with the corresponding transition matrix or kernel). The following choice of H

$$H(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{\pi(\mathbf{x})}{\pi(\mathbf{y})}}[P(\mathbf{x}, \mathbf{y}) - \delta(\mathbf{x} - \mathbf{y})] \tag{7}$$

satisfies condition (4), where $\delta(\mathbf{x}-\mathbf{y})$ is the Dirac delta function: $\delta(\mathbf{x}-\mathbf{y}) = 1$ if $\mathbf{x} = \mathbf{y}$ and zero otherwise. It should be noted that the reversibility condition is essential in order to have a symmetric operator $H(\mathbf{x}, \mathbf{y})$, as required. With this choice of H, letting $\tilde{\psi} = \psi/\sqrt{\pi}$, equation (5) becomes:

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) - \sum_{\mathbf{y}} P(\mathbf{x}, \mathbf{y})[\tilde{\psi}(\mathbf{x}) - \tilde{\psi}(\mathbf{y})].$$

The same H can also be applied to continuous systems. In this case, P is the kernel of the Markov chain and equation (7) can be trivially extended to the continuous case. This choice of H is exploited in (Dellaportas and Kontoyiannis 2010), where the following fundamental equation is found for the optimal $\tilde{\psi}$: $\mathbb{E}[\tilde{\psi}(\mathbf{x}_1)|\mathbf{x}_0 = \mathbf{x}] = \mu_f - f(\mathbf{x})$. It is easy to prove that this equation coincides with our fundamental equation (6), with the choice of H

given in (7). The Authors observe that the optimal trial function is given by

$$\tilde{\psi}(\mathbf{x}) = \sum_{n=0}^{\infty} [\mathbb{E}[\hat{f}(\mathbf{x}_n)|\mathbf{x}_0 = \mathbf{x}] - \mu_f], \tag{8}$$

that is, $\tilde{\psi}$ is the solution to the Poisson equation for $f(\mathbf{x})$. However, an explicit solution cannot be obtained in general.

Another operator is proposed in (Assaraf and Caffarel 1999): if $\mathbf{x} \in \mathbb{R}^d$ consider the Schrödinger-type Hamiltonian operator:

$$H = -\frac{1}{2} \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} + V(\mathbf{x}), \tag{9}$$

where $V(\mathbf{x})$ is constructed to fulfill equation (3): $V = \frac{1}{2\sqrt{\pi}} \Delta \sqrt{\pi}$ and Δ denotes the Laplacian operator of second order derivatives. In this setting, $\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{H\psi(\mathbf{x})}{\sqrt{\pi(\mathbf{x})}}$. These are the operator and the re-normalized function that will be considered throughout this paper. Although it can be applied only to continuous state spaces, this Schrödinger-type operator shows several advantages with respect to the operator (7). First of all, in order to use (7) the conditional expectation appearing in (8) has to be available in closed form. Secondly, definition (9) does not require reversibility of the chain. Moreover, this definition is independent of the kernel $P(\mathbf{x}, \mathbf{y})$ and, therefore, also of the type of MCMC algorithm that is used in the simulation. Note that, for calculating $\tilde{f}$ both with the operator (9) and (7), the normalizing constant of π is not needed.

3 Choice of ψ

The optimal choice of ψ is the exact solution of the fundamental equation (6). In real applications, typically, only approximate solutions, obtained by minimizing $\sigma^2(\tilde{f})$, are available. In other words, we select a functional form for ψ, typically a polynomial, parameterized by some coefficients, and optimize those coefficients by minimizing the fluctuations of the resulting $\tilde{f}$. The particular form of ψ is very dependent on the problem at hand, that is on π, and on f. In the examples first, second and third order polynomials are considered. As one would expect, the higher is the degree of the polynomial, the higher is the number of control variates introduced and the higher is

the variance reduction of the estimators. It can be easily shown that in a d dimensional space, using polynomials of order p, provides $\binom{d+p}{d} - 1$ control variates.

4 Control Variates

In this section, general expressions for the control variates in the ZV method are found. Using the Schrödinger-type Hamiltonian H as given in (9) and trial function $\psi(\mathbf{x}) = P(\mathbf{x})\sqrt{\pi(\mathbf{x})}$, the re-normalized function is:

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) - \frac{1}{2}\Delta P(\mathbf{x}) + \nabla P(\mathbf{x}) \cdot \mathbf{z}, \tag{10}$$

where $\mathbf{z} = -\frac{1}{2}\nabla \ln \pi(\mathbf{x})$, $\nabla = \left(\frac{\partial}{\partial x_1}, ..., \frac{\partial}{\partial x_d}\right)$ denotes the gradient and $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ is the Laplacian operator of second derivatives.

Hereafter the function P is assumed to be a polynomial. Two special cases, that will be used in the examples discussed in the second part of the paper, are now considered. As a first case, for $P(\mathbf{x}) = \sum_{j=1}^{d} a_j x_j$ (1st degree polynomial), one gets:

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{H\psi(\mathbf{x})}{\sqrt{\pi(\mathbf{x})}} = f(\mathbf{x}) + \mathbf{a}^T\mathbf{z}.$$

Similarly for quadratic polynomials, $P(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T B\mathbf{x}$. The re-normalized $\tilde{f}$ is :

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

where $\mathbf{g}$ and $\mathbf{y}$ are column vectors with $\frac{1}{2}d(d+3)$ elements defined in the following way:

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

5 Unbiasedness

In this section general conditions on the target π are provided that guarantee that the ZV-MCMC estimator is (asymptotically) unbiased. Let π be a d-dimensional density defined on a bounded open set Ω with regular boundary $\partial\Omega$. Then, using integration by parts in d dimensions, we get

$$\left\langle \frac{H\psi}{\sqrt{\pi}} \right\rangle := \mathbb{E}_\pi \left[\frac{H\psi}{\sqrt{\pi}} \right] = \frac{1}{2} \int_{\partial\Omega} [\psi \nabla\sqrt{\pi} - \sqrt{\pi}\nabla\psi] \cdot \mathbf{n} d\sigma. \tag{11}$$

From this equality, it can be proved that, if $\psi = P\sqrt{\pi}$, a sufficient condition to get an unbiased estimator is $\pi(\mathbf{x})\dfrac{\partial P(\mathbf{x})}{\partial x_j} = 0$, for all $\mathbf{x} \in \partial\Omega$ and $j = 1, \ldots, d$. When π has unbounded support, the formula of integration by parts cannot be used directly. In this case, a sequence of bounded subsets $(B_r)_r$ is to be constructed, so that $B_r \nearrow \Omega$. In this case, a sufficient condition for unbiasedness is

$$\lim_{r \to +\infty} \int_{\partial B_r} \pi \nabla P \cdot \mathbf{n} d\sigma = 0.$$

By reducing all the previous computations for $d = 1$, a simple condition can be derived in the univariate case. If $\Omega = [l, u]$, where $u, l \in \overline{\mathbb{R}} := \mathbb{R} \cup \pm\infty$, it is sufficient that

$$\left. \frac{dP(x)}{dx} \right|_{x=l} \pi(l) = \left. \frac{dP(x)}{dx} \right|_{x=u} \pi(u), \tag{12}$$

which is true, for example, if $\frac{dP}{dx}\pi$ annihilates at the border of the support.

These results mean that, in order to get unbiasedness, one should consider trial functions ψ whose partial derivatives are zero on the set $\partial\Omega^* := \{\mathbf{x} \in \partial\Omega : \ \pi(\mathbf{x}) > 0\}$. For all the examples discussed in Section 8, the ZV-MCMC estimators have been found to be unbiased for any choice of (polynomial) P.

In the seminal paper by (Assaraf and Caffarel 1999) unbiasedness conditions are not explored since, typically, the target distribution the physicists are interested in, annihilate at the border of the domain with an exponential rate.

The following example shows how crucial the choice of trial functions is in order to have an unbiased estimator, even in trivial models.

Example 5.1 *Let $f(x) = x$ and π be exponential: $\pi(x) = \lambda e^{-\lambda x}\mathbb{I}_{\{x>0\}}$. If $P(x)$ is a first order polynomial, condition (12) does not hold. Moreover, this*

choice does not allow for a ZV-MCMC estimator, since the control variate $z = -\frac{1}{2}\frac{d}{dx}\ln\pi(\mathbf{x})$ *is constant and* $\sigma(x,z) = 0$. *However, to satisfy equation (12) it is sufficient to consider second order polynomials. Indeed, if* $P(x) = a_0 + a_1 x + a_2 x^2$ *equation (12) is satisfied provided that* $a_1 = 0$. *Therefore, the minimization of the variance of* $\tilde{f}$ *can be carried out within this special class. Note that higher order polynomials, whose derivative annihilate at zero, can provide ZV-MCMC estimators with greater variance reduction.*

6 Optimal coefficients

In this section variance reduction is discussed in the ZV context and the optimal choice of ψ is found for some special cases. Note that, if at least one of these hypotheses does not hold, $\sigma^2(\tilde{f})$ may be infinite or undefined:

A1: $\sigma^2(f) < \infty$; **A2**: $\sigma^2(\frac{H\psi}{\sqrt{\pi}}) < \infty$.

Therefore, from now on, both $A1$ and $A2$ are supposed to hold. As already observed, the ideal Zero Variance ψ cannot be usually explicitly found. Therefore, a particular subset of ψ is considered, which is typically a parametric class, and $\sigma^2(\tilde{f})$ is minimized within this class. However, we need to verify that this optimal solution still gives appreciable variance reduction. In the following proposition a useful criterion is stated in order to have variance reduction.

Proposition 6.1 *Under conditions* $A1$ *and* $A2$,

$$\sigma^2(\tilde{f}) \geq \sigma^2(f) - \frac{\left\langle \frac{f(x)H\psi}{\sqrt{\pi}} \right\rangle^2}{\left\langle \left(\frac{H\psi}{\sqrt{\pi}}\right)^2 \right\rangle}. \tag{13}$$

In particular, if $\psi(x) = P(x)\sqrt{\pi(x)}$ *and* $d = 1$, *equality holds if and only if*

$$\int f(x)[P''\pi + P'\pi'] = \frac{1}{2}\int [(P'')^2\pi + \frac{(P')^2(\pi')^2}{\pi} + 2P'P''\pi']. \tag{14}$$

Proof: By definition of $\tilde{f}$, it follows that

$$\sigma^2(\tilde{f}) = \sigma^2(f) + \left\langle \left(\frac{H\psi}{\sqrt{\pi}}\right)^2 \right\rangle + 2\sigma(f, \frac{H\psi}{\sqrt{\pi}}). \tag{15}$$

Fist of all, observe that we can always assume $\langle(\frac{H\psi}{\sqrt{\pi}})^2\rangle > 0$, since this second moment is zero if and only if $H\psi = 0$, but in this case $\tilde{f} \equiv f$. As a consequence, (13) easily follows from (15) because, for any x, y with $x \neq 0$, we have $x + 2y \geq -y^2/x$. Taking $x := \langle(\frac{H\psi}{\sqrt{\pi}})^2\rangle$ and $y := \sigma(f, \frac{H\psi}{\sqrt{\pi}})$ gives the result. Moreover, equality in (13) holds if and only if $x = -y$. Now, when $d = 1$, observe that

$$\begin{aligned} \left\langle \left(\frac{H\psi}{\sqrt{\pi}}\right)^2 \right\rangle &= \frac{1}{4}\int [(P'')^2\pi + \frac{(P')^2(\pi')^2}{\pi} + 2P'P''\pi'], \\ \sigma(f, \frac{H\psi}{\sqrt{\pi}}) &= -\frac{1}{2}\int f(x)[P''\pi + P'\pi']. \end{aligned}$$

Therefore, equality in (13) holds if and only if condition (14) is satisfied. ■
Whenever condition (14) is satisfied, $\sigma^2(\tilde{f})$ is certainly smaller than $\sigma^2(f)$. Of course, this is true if ψ solves the fundamental equation and therefore $\tilde{f}$ is constant and the ideal result of zero variance is achieved. However, a variance reduction is guaranteed even if $\sigma^2(\tilde{f})$ is minimized within a particular class of functions ψ or, equivalently, class of P.

6.1 Special case: polynomial trial functions

In the sequel, the case of polynomial trial functions is discussed.

1. *Univariate* π. For the sake of simplicity, the target π is first supposed to be univariate. Therefore, univariate polynomials are considered. Constant polynomials $P \in \mathcal{P} := \{P(x) \equiv c, \quad c \in \mathbb{R}\}$ are not interesting, since in this case $H\psi \equiv 0$, so that $\tilde{f} \equiv f$. Consider the class of first order polynomials $\mathcal{P} := \{P := a_0 + a_1 x, \ a_0, \ a_1 \in \mathbb{R}\}$. In this case, Equation (14) becomes $\int a_1 f(x)\pi' = a_1^2 I_0/2$, where $I_0 := \int(\pi')^2/\pi$ is the Fisher information of π with respect to the location parameter. I_0 is supposed to be finite, otherwise, $A2$ does not hold. If one introduces the control variate $z = -\frac{1}{2}\frac{d\log\pi}{dx}$, then the optimal parameter is $a_1 = -\sigma(f, z)/\sigma^2(z)$. This is just the solution (2), obtained by minimizing the variance of $\tilde{f}$ when only one control variate is considered. Moreover, if $f(x)\pi$ annihilates at infinity, integrating by parts the solution $a_1 = -2\langle f'\rangle/I_0$, for arbitrary a_0, is obtained. The solution $a_1 = 0$ is meaningless, because it reduces to the case of constant polynomials.

This example may be generalized for classes of higher-order polynomials. However, although theoretically it is not difficult to minimize the variance, stronger integrability conditions and computational problems appear. These issues are due to the integral $\int (P')^2(\pi')^2/\pi$ appearing in the right side of Equation (14). When P is a polynomial of order q, this integral involves, in turn, the computation of different integrals of the kind $I_n := \int x^n(\pi')^2/\pi$, for $n = 1, \ldots, q$, that are expected to be finite in order for $\langle (H\psi/\sqrt{\pi})^2 \rangle$ to be finite.

2. *Multivariate π, linear polynomials.* Generalizing the previous setting to d-dimensional target distributions, d control variates z_i $(1 \leq i \leq d)$ are needed and $\tilde{f}$ is equal to $\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + \mathbf{a}^T\mathbf{z}$, where $\mathbf{z} = [z_1, \cdots, z_d]^T$. The optimal choice of $\mathbf{a}$, that minimizes the variance of $\tilde{f}(x)$, is $\mathbf{a} = -\Sigma_{\mathbf{zz}}^{-1}\sigma(\mathbf{z}, f)$, where $\Sigma_{\mathbf{zz}} = \mathbb{E}(zz^T)$ and $\sigma(\mathbf{z}, f) = \mathbb{E}(zf)$. We anticipate that conditions under which the ZV-MCMC estimator obeys a CLT (Section 7) guarantee that the optimal $\mathbf{a}$ is well defined. In ZV-MCMC, the optimal $\mathbf{a}$ is estimated in a first stage, through a short MCMC simulation. When higher-degree polynomials are considered, the same optimal formula for the coefficients associated to the control variates is obtained, provided that an explicit formula for the control variate vector $\mathbf{z}$ has been found.

7 Central limit theorem

Conditions for existence of a CLT for $\hat{\mu}_f$ are well known in the literature ((Tierney 1994)):

Theorem 7.1 *Suppose an ergodic Markov chain $\{\mathbf{x}^n\}$, with stationary distribution π, and a real valued function f satisfy one of the following conditions:*

B1 : The chain is geometrically ergodic and $f(\mathbf{x}) \in L^{2+\delta}(\pi)$ for some $\delta > 0$.

B2 : The chain is uniformly ergodic and $f(\mathbf{x}) \in L^2(\pi)$.

Then

$$s_f^2 = \mathbb{E}_\pi\left[\left(f(\mathbf{x}^0) - \mu_f\right)^2\right] + 2\sum_{k=1}^{+\infty} \mathbb{E}_\pi\left[\left(f(\mathbf{x}^k - \mu_f\right)\left(f(\mathbf{x}^0) - \mu_f\right)\right]$$

is well defined, non-negative and finite, and

$$\sqrt{N}\,(\hat{\mu}_f - \mu_f) \xrightarrow{L} \mathcal{N}(0, s_f^2). \tag{16}$$

Therefore, the ZV-MCMC estimator obeys a CLT provided that the renormalized function $\tilde{f}$ satisfies one of the integrability conditions required in $B1$ and $B2$. By the definition of $\tilde{f}$, this implies, in turn, that the control variates $z_j = \frac{\partial \ln \pi}{\partial x_j}$ and the trial $P(x)$ satisfy some integrability conditions, which depend on the choice of $P(x)$. More precisely, from (10), ΔP and $\nabla P \cdot \mathbf{z}$ should belong to the space $L^{2+\delta}(\pi)$ when the chain is geometrically ergodic.

In the following corollary, the case of linear and quadratic polynomials P (used in the examples in Section 8) is considered.

Corollary 7.2 *Let $\psi(\mathbf{x}) = P(\mathbf{x})\sqrt{\pi}$, where $P(\mathbf{x})$ is a first or second degree polynomial. Then, the ZV-MCMC estimator $\hat{\mu}_{\tilde{f}}$ is a consistent estimator of μ_f which satisfies the CLT equation (16), provided one of the following conditions holds:*

- *C1 : The chain is geometrically ergodic and $x_i^k z_j \in L^{2+\delta}(\pi)$, for all $i, j \in \{1, ..., d\}$, for all $k = \{0, \deg P - 1\}$ and some $\delta > 0$.*
- *C2 : The chain is uniformly ergodic and $x_i^k z_j \in L^2(\pi)$, for all $i, j \in \{1, ..., d\}$ and for all $k = \{0, \deg P - 1\}$.*

It should be noted that, in the case of linear P, if $f \in L^2(\pi)$ and the Markov chain is uniformly ergodic, then a sufficient condition to get a CLT is

$$m_j = \mathbb{E}_\pi \left[\left(\frac{\partial}{\partial x_j} \ln(\pi(\mathbf{x})) \right)^2 \right] < \infty, \quad \forall j.$$

If the Markov chain is only geometrically ergodic, the stronger condition

$$\mathbb{E}_\pi \left[\left(\frac{\partial}{\partial x_j} \ln(\pi(\mathbf{x})) \right)^{2+\delta} \right] < \infty,$$

for some $\delta > 0$, is needed. The quantity m_j is known in the literature as Linnik functional (if considered as a function of the target distribution, $I(\pi)$) since it was introduced by (Linnik 1959) and is related to the Fisher

information and to the entropy. It has been shown in (McKean 1966) that, for univariate distributions, finiteness of $I[\pi]$ implies finiteness of the entropy of π. The Fisher information (in a frequentist setting with scalar parameter β) is:

$$I(\beta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\beta}\ln(\pi(x;\beta))\right)^2\right] = \int \frac{1}{\pi(x;\beta)}\left(\frac{\partial}{\partial\beta}\pi(x;\beta)\right)^2 \pi(x;\beta)dx.$$

If β is the location parameter, i.e., if $\pi(x;\beta) = h(x-\beta)$, the equality $\frac{\partial}{\partial\beta}h(x-\beta) = -\frac{\partial}{\partial x}h(x-\beta)$ implies

$$\int \frac{1}{\pi(x;\beta)}\left(\frac{\partial}{\partial\beta}\pi(x;\beta)\right)^2 dx \;=\; \int \frac{1}{\pi(x;\beta)}\left(\frac{\partial}{\partial x}\pi(x;\beta)\right)^2 dx.$$

Therefore, m_j is interpretable as the Fisher information of a location family in a frequentist setting. It has been proved that all the estimators discussed in the examples in Section 8 obey a CLT. In most cases, an explicit computation is the only way to discuss finiteness of m_j. However, for particular models, simpler conditions can be found. In the following section, the case of the exponential family is discussed.

7.1 Exponential family

Let π belong to a d-dimensional exponential family:

$$\pi(\mathbf{x}) \propto \exp(\beta \cdot \mathbf{T}(\mathbf{x}) - K_p(\beta))p(\mathbf{x}), \tag{17}$$

where $\beta \in \mathbb{R}^d$ is the vector of natural parameters, $\mathbf{T} = (T_1, T_2, \ldots, T_d)$ is the sufficient statistic, $K_p(\beta)$ is the cumulant generating function and $p(\mathbf{x})$ is a reference measure. The following theorem provides a sufficient condition for a CLT for ZV-MCMC estimators when the target belongs to the exponential family and a uniformly ergodic Markov Chain is considered. Similar results can be achieved when the Markov Chain is geometrically ergodic, by considering the $2+\delta$ moment.

Theorem 7.3 *Let π belong to an exponential family as in (17), with $p(\mathbf{x})$ such that $\dfrac{\partial \log p}{\partial x_j} \in L^2(\pi)$, for $j = 1, \ldots, d$. Then, the Linnik functional of π is finite if and only if $\frac{\partial T_k}{\partial x_j} \in L^2(\pi)$, for all $j, k = 1, 2, \ldots, d$.*

Proof: By a direct computation, we get

$$\frac{\partial \pi}{\partial x_j} = \exp(\beta \cdot \mathbf{T} - K_p(\beta)) \left[\frac{\partial p}{\partial x_j} + p(\mathbf{x}) \beta \cdot \frac{\partial \mathbf{T}}{\partial x_j} \right],$$

so that

$$\frac{\partial \log \pi(\mathbf{x})}{\partial x_j} = \frac{1}{\pi} \frac{\partial \pi}{\partial x_j} = \left(\beta \cdot \frac{\partial \mathbf{T}}{\partial x_j} + \frac{1}{p} \frac{\partial p}{\partial x_j} \right).$$

The thesis follows immediately since, by hypothesis, $\frac{\partial \log p}{\partial x_j} \in L^2(\pi)$. ■

Remark. In general, it can be hard to verify the hypothesis of Theorem 7.3 involving the reference measure $p(\mathbf{x})$. However, in most well known exponential models $p(\mathbf{x}) \equiv 1$, so that this condition is trivially satisfied.

Example 7.1 *The Gamma density* $\Gamma(\alpha, \theta)$ *can be written as an exponential family on* $(0, +\infty)$*, where* $p(x) \equiv 1$*,* $K_p(\beta) = -\alpha \log \theta$*, and the vector of parameters and the sufficient statistic are equal to:* $\beta = (\theta, \alpha - 1)$ *and* $\mathbf{T}(x) = (-x, \log x)$*, if* $\alpha \neq 1$*;* $\beta = \theta$*,* $T(x) = -x$ *if* $\alpha = 1$*. Since* $p \equiv 1$*, hypotheses of Theorem 7.3 are satisfied. Therefore, it is sufficient to study the gradient of the sufficient statistic. If* $\alpha = 1$*,* $T(x) = -x$*, whose derivative is constant and trivially belongs to* $L^2(\pi)$*. If* $\alpha \neq 1$*, we need to check wether* $1/x \in L^2(\pi)$*. To this end, we should study the finiteness of*

$$\int_0^{+\infty} \frac{1}{x^2} \pi(x) dx \propto \int_0^{+\infty} \frac{1}{x^2} x^{\alpha - 1} \exp(-\theta x) dx = \int_0^{+\infty} \frac{1}{x^{3-\alpha}} \exp(-\theta x) dx,$$

which is finite if and only if $\alpha > 2$*. Therefore, the Gamma density* $\Gamma(\alpha, \theta)$ *has finite Linnik functional for any* θ *and for any* $\alpha \in \{1\} \cup (2, +\infty)$*. Under these conditions, a CLT holds for the ZV-MCMC estimator.*

8 Examples

In the sequel standard statistical models are considered. For these models, the ZV-MCMC estimators are found in a Bayesian context; from now on, the target $\pi = \pi(\beta|\mathbf{x})$ is the posterior distribution in a Bayesian framework. Numerical simulations are provided, that confirm the effectiveness of variance reduction achieved, by minimizing the variance of $\tilde{f}$ among polynomial functions. Moreover, conditions for both unbiasedness and CLT for $\tilde{f}$ are verified for all the examples.

8.1 Probit Model

Let y_i be Bernoulli r.v.'s: $y_i|\mathbf{x}_i \sim \mathcal{B}(1, p_i)$, $p_i = \Phi(\mathbf{x}_i^T\beta)$, where $\beta \in \mathbb{R}^d$ is the vector of parameters of the model and Φ is the c.d.f. of a standard normal distribution. The likelihood function is:

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

As it can be seen by inspection, the likelihood function is invariant under the transformation $(\mathbf{x}_i, y_i) \to (-\mathbf{x}_i, 1 - y_i)$. Therefore, for the sake of simplicity, in the rest of the example we assume $y_i = 1$ for any i, so that the likelihood simplifies: $l(\beta|\mathbf{y}, \mathbf{x}) \propto \prod_{i=1}^{n} \Phi(\mathbf{x}_i^T\beta)$. This formula shows that the contribution of $\mathbf{x}_i = \mathbf{0}$ is just a constant $\Phi(\mathbf{x}_i^T\beta) = \Phi(0) = \frac{1}{2}$, therefore, without loss of generality, we assume for all i, $\mathbf{x}_i \neq \mathbf{0}$. Using flat priors, the posterior of the model is proportional to the likelihood, and the Bayesian estimator of each parameter, β_k, is the expected value of $f_k(\beta) = \beta_k$ under π ($k = 1, 2, \cdots, d$). Using Schrödinger-type Hamiltonians, H, defined in (9), and $\psi_k(\beta) = P_k(\beta)\sqrt{\pi(\beta)}$, as the trial functions, where $P_k(\beta) = \sum_{j=1}^{d} a_{j,k}\beta_j$ is a first degree polynomial, one gets:

$$\widetilde{f}_k(\beta) = f_k(\beta) + \frac{H\psi_k(\beta)}{\sqrt{\pi(\beta|\mathbf{y}, \mathbf{x})}} = f_k(\beta) + \sum_{j=1}^{d} a_{j,k} z_j,$$

where, for $j = 1, 2, \ldots, d$,

$$z_j = -\frac{1}{2} \sum_{i=1}^{n} \frac{x_{ij}\phi(\mathbf{x}_i^T\beta)}{\Phi(\mathbf{x}_i^T\beta)},$$

because of the assumption $y_i = 1$ for any i.

To demonstrate the effectiveness of ZV in this setting, a simple example, (Douc, Guillin, Marin, and Robert 2007), is presented. The bank dataset from (Flury and Riedwyl 1988) contains the measurements of four 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 probabilistic model as the regressors, and the type of the banknote y_i, as the response variable (0 for genuine and 1 for counterfeit). The model is the one outlined at the beginning of this section. Now, for $k = 1, \ldots, d$, the

optimal (in the sense of minimizing the asymptotic variance of the resulting MCMC estimators) vector of parameters a_k should be found. To this end, a short MCMC simulation (of length 2000, after 1000 burn in steps) is run, and the optimal coefficients are estimated: $\hat{\mathbf{a}}_k = -\hat{\Sigma}_{\mathbf{zz}}^{-1}\hat{\sigma}(\mathbf{z}, \beta_k)$. The Albert-Chib sampler (Albert and Chib 1993), that is a Gibbs sampler for GLM, is used to run the Markov chain. Then, another MCMC simulation (of length 2000, and independent of the first one) is run, along which $\widetilde{f}_d(\beta)$ is averaged. The MCMC traces have been depicted in the left plot of Fig. 1. The blue curves are the trace of f_k (ordinary MCMC), and the red curves are the trace of $\widetilde{f}_k$ (ZV-MCMC). It is clear from the figure that the variances of the estimator have substantially decreased. Indeed the ratio of the Sokal estimate of the asymptotic variances (Sokal 1996) of the two estimators (the ordinary MCMC and and ZV-MCMC estimates) are between 25 and 100. Even better performance can be achieved by a using second degree polynomial to define the trial function. In the right plot of Fig. 1 the traces of ZV-MCMC with second order $P(x)$ have been depicted along with the trace of the ordinary MCMC. As it can be seen from the figure, the variances of the ZV estimates are negligible. In this case the ratio of the Sokal variances of two estimators are between $25,000$ and $90,000$.

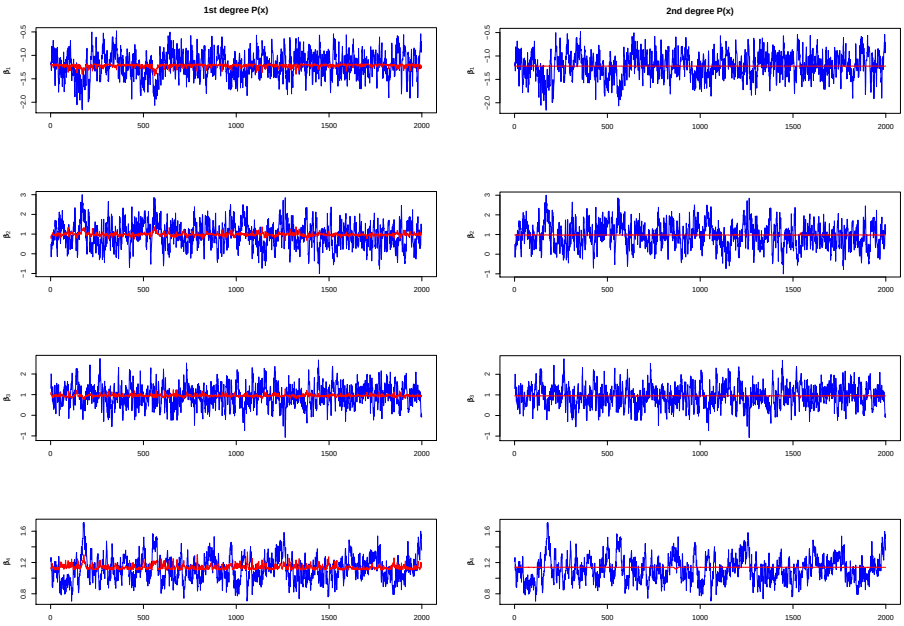
8.2 Logit Model

In the same setting as the probit model, let $p_i = \frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}$ where $\beta \in \mathbb{R}^d$ is the vector of parameters of the model. The likelihood function is:

$$l(\beta|\mathbf{y}, \mathbf{x}) \propto \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}\right)^{y_i} \left(\frac{1}{1+\exp(\mathbf{x}_i^T\beta)}\right)^{1-y_i}. \tag{18}$$

By inspection, it is easy to verify that the likelihood function is invariant under the transformation:$(\mathbf{x}_i, y_i) \to (-\mathbf{x}_i, 1-y_i)$. Therefore, for the sake of simplicity, in the sequel we assume $y_i = 0$ for any i, so that the likelihood simplifies as $l(\beta|\mathbf{y}, \mathbf{x}) \propto \prod_{i=1}^{n}[1+\exp(\mathbf{x}_i^T\beta)]^{-1}$. The contribution of $\mathbf{x}_i = \mathbf{0}$ to the likelihood is just a constant, therefore, without loss of generality, it is assumed that $\mathbf{x}_i \neq \mathbf{0}$ for all i. Using flat priors, the posterior distribution is proportional to (18) and the Bayesian estimator of each parameter, β_k, is the expected value of $f_k(\beta) = \beta_k$ under π ($k = 1, 2, \cdots, d$). Using the same

Figure 1: Traces of ordinary MCMC and ZV-MCMC, plotted in blue and red respectively for different parameters (in the rows) and different degree polynomials (in the columns).

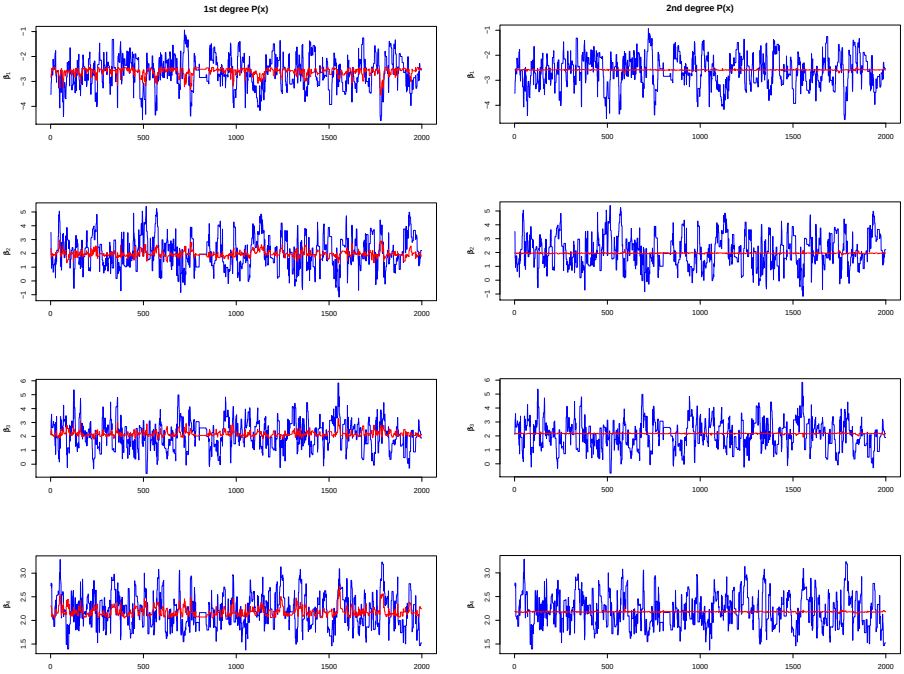


pair of operator H and test function ψ_k as before, the control variates are:

$$z_j = \frac{1}{2}\sum_{i=1}^{n} x_{ij}\frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}, \quad \text{for} \quad j = 1, 2, \ldots, d.$$

A logit model is fitted to the same dataset of Swiss banknotes, that has been introduced in the probit model example. Similar to the previous example, in the first stage a MCMC simulation is run, and the optimal parameters of $P(\beta)$ are estimated. Then, in the second stage an independent simulation is run, and $\tilde{f}_d$ is averaged, using the optimal trial function that has been estimated in the first stage. As shown in Fig. 2 for linear polynomial, the ratio of the Sokal's estimates of the asymptotic variances of the two estimators (the ordinary MCMC and and ZV-MCMC estimates) are between 10 and 40. Using the quadratic polynomial, the ratio of the Sokal's variances are between $2,000$ and $6,000$.

Figure 2: Traces of ordinary MCMC and ZV-MCMC plotted in blue and red respectively: different parameters in the rows and different degree polynomials in the columns.



8.3 GARCH Model

Generalized autoregressive conditional heteroskedasticity (GARCH) model (Bollerslev 1986) have become one of the most important building blocks of models in econometrics and financial econometrics. The widespread applications of GARCH models is due to its parameter parsimony and interpretability, and to some extent to the analytical tractability of the model. Using few parameters, these kinds of models can mimic some of the most important stylized features of financial time series, like volatility clustering, fat tails, and asymmetric volatility.

In financial applications, GARCH models have been widely used to model returns. Here it is shown how the ZV-MCMC principle can be exploited to estimate the parameters of a univariate GARCH model applied to daily returns of exchange rates in a Bayesian setting. In a Normal-GARCH model,

we assume the returns are conditionally Normally distributed, $r(t)|\mathcal{F}_t \sim \mathcal{N}(0, h_t)$, where h_t is a predictable ($\mathcal{F}_{t-1}$ measurable process): $h_t = \omega_1 + \omega_3 h_{t-1} + \omega_2 r_{t-1}^2$, where $\omega_1 > 0$, $\omega_2 \geq 0$, and $\omega_3 \geq 0$. Let $\mathbf{r} = (r_1, \ldots, r_T)$ be the observed time series. The likelihood function is equal to:

$$l\left(\omega_1, \omega_2, \omega_3 | \mathbf{r}\right) \quad \propto \quad \left(\prod_{t=1}^{T} h_t\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\frac{r_t^2}{h_t}\right)$$

and using independent truncated Normal priors for the parameters, the posterior is:

$$\pi\left(\omega_1, \omega_2, \omega_3 | \mathbf{r}\right) \quad \propto \quad \exp\left[-\frac{1}{2}\left(\frac{\omega_1^2}{\sigma^2(\omega_1)} + \frac{\omega_2^2}{\sigma^2(\omega_2)} + \frac{\omega_3^2}{\sigma^2(\omega_3)}\right)\right]\left(\prod_{t=1}^{T} h_t\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\frac{r_t^2}{h_t}\right).$$

The control variates (for the case of first degree polynomial in trial function) are:

$$\frac{\partial \ln \pi}{\partial \omega_i} = -\frac{\omega_i}{\sigma^2(\omega_i)} - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{h_t}\,\frac{\partial h_t}{\partial \omega_i} + \frac{1}{2}\sum_{t=1}^{T}\frac{r_t^2}{h_t^2}\,\frac{\partial h_t}{\partial \omega_i}, \quad i = 1, 2, 3,$$

where:

$$\frac{\partial h_t}{\partial \omega_1} = \frac{1-\omega_3^{t-1}}{1-\omega_3}, \quad \frac{\partial h_t}{\partial \omega_2} = \left(r_{t-1}^2 + \omega_3 \frac{\partial h_{t-1}}{\partial \omega_2}\right)\mathbb{I}_{t>1}, \quad \frac{\partial h_t}{\partial \omega_3} = \left(h_{t-1} + \omega_3 \frac{\partial h_{t-1}}{\partial \omega_3}\right)\mathbb{I}_{t>1}.$$

As an example, a Normal-GARCH(1, 1) is fitted to the daily returns of the Deutsche Mark vs British Pound (DEM/GBP) exchange rates from January 1985, to December 1987 (750 obs). In the first stage a short MCMC simulation, as proposed in (Ardia 2008), is used in order to estimate the optimal parameters of the trial function. Then in the second stage an independent simulation is run and $\tilde{f}_j(x)$ is averaged in order to efficiently estimate the posterior mean of each parameter. First, second and third degree polynomials in the trial function are used. As can be seen in Fig. 3 and Table 1, where the Sokal estimates of variance are reported, the ZV strategy reduces the variance of the estimators up to tens of thousands of times.

Table 1: Variance Reduction in GARCH Model Estimation:

Sokal estimate of variance of MC estimator /Sokal estimate of variance of ZV-MC estimator

	$\hat{\omega_1}$	$\hat{\omega_2}$	$\hat{\omega_3}$
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

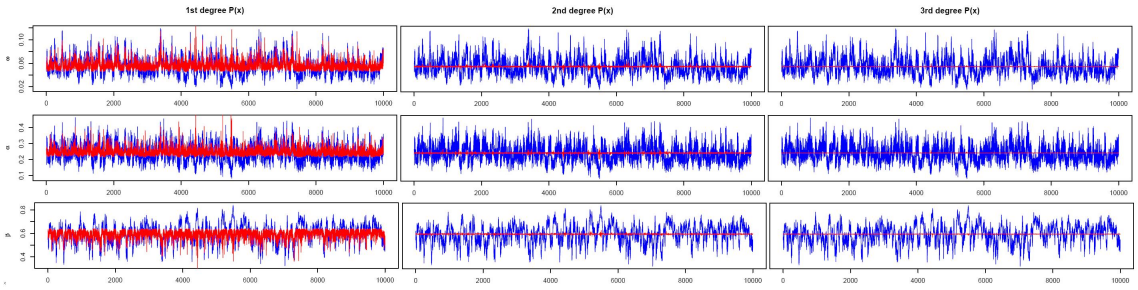


Figure 3: Traces of ordinary MCMC and ZV-MCMC plotted in blue and red respectively: different parameters in the rows and different degree polynomials in the columns

9 Possible generalizations

Two are the main ingredients to construct ZV-MCMC estimators; namely a trial function ψ and an operator H that are combined to define a re-normalized function $\tilde{f}$ as in (5). In this section two possible generalizations of the ZV principle, as illustrated so far, are proposed. The first one considers a bigger class of trial functions, the second one allows the use of a wider class of operators by defining a more general re-normalized function $\tilde{f}$.

9.1 Extended trial functions

Throughout this paper, a re-normalized $\tilde{f}$ as defined in (5) has been considered, where the trial function has been parametrized as $\psi(\mathbf{x}) = P(\mathbf{x})\sqrt{\pi(\mathbf{x})}$. This setting naturally leads to using the gradient of the log-target in the control variate formulation. In recent work by (Girolami and Calderhead 2011), the dynamcics of the classical Hamiltonian MCMC and of the MALA methods, in which first derivatives of the log-target appear, are efficiently improved by considering second and higher order derivatives of the same quantity. In a similar way, in our setting a finer definition of $\tilde{f}$ may lead to work with higher-order derivatives of the log-target. This can be easily achieved by considering a wider class of trial functions: $\psi(\mathbf{x}) = P(\mathbf{x})q(\mathbf{x})$, where, as before, $P(\mathbf{x})$ denotes a parametric class of polynomials, and $q(\mathbf{x})$ is an arbitrary (sufficiently regular) function. Then, by using the identity

$V = \frac{1}{2\sqrt{\pi}}\Delta\sqrt{\pi} = \frac{1}{2}\Delta \log \pi$, the re-normalization term of $\tilde{f}$ in (3) becomes:

$$\frac{H\psi}{\sqrt{\pi}} = \frac{1}{2}\left(-\frac{q}{\sqrt{\pi}}\Delta P - \frac{2}{\sqrt{\pi}}\nabla q \cdot \nabla P - \frac{P}{\sqrt{\pi}}\Delta q + \frac{P}{2\sqrt{\pi}} q \frac{\partial^2 \log \pi}{\partial x^2}\right).$$

Therefore, the second derivative of the log-target naturally arises due to the particular choice of the potential V. The formula obtained, which is quite involved, can be dramatically simplified for suitable q. For example, the choice $q \equiv 1$ gives

$$\frac{H\psi}{\sqrt{\pi}} = -\frac{1}{2\sqrt{\pi}}\left(\Delta P + \frac{P}{2}\Delta \log \pi\right).$$

However, it should be noted that, even in this simple case, unbiasedness conditions are not verified in general. In order to get unbiased estimators, one can use the following strategy: fix a certain number of parameters in P, so that the unbiasednes conditions are verified; next, find the optimal ZV-MCMC estimator by minimizing the variance of $\tilde{f}$ with respect to the remaining free parameters. In the univariate case, unbiasedeness conditions lead to fix two parameters, see (12), so we need to consider at least third degree polynomials to have at least one free parameter to minimize the variance of $\tilde{f}$. In this, more general setting, also CLT conditions should be carefully re-phrased.

9.2 Extended Hamiltonian operators

The Hamiltonian operator $H = -\frac{1}{2}\Delta + V$, where $V = \frac{1}{2\sqrt{\pi}}\Delta\sqrt{\pi}$, has been considered so far. In this setting, V is uniquely determined, because of the constrain (3), which is essetial to get unbiased estimators. In the paper by (Assaraf and Caffarel 2003), an alternative, more general renormalized function $\tilde{f}$ is defined:

$$\tilde{f} = f + \frac{H\psi}{\sqrt{\pi}} - \frac{\psi(H\sqrt{\pi})}{\pi}, \tag{19}$$

where, again, H is an Hamiltonian operator and ψ a quite arbitrary trial function. In this setting, if $H = -\frac{1}{2}\Delta + V$, under the same, mild conditions discussed in Section 5, $\tilde{f}$ has the same expectation as f under π. This is true without imposing condition (3), so that now V can be also chosen aribtrarly. Therefore, the re-normalization (19) allows for a more general class of Hamiltonians.

10 Discussion

As noted in the introduction, cross-fertilizations between the physics and the statistical literature have proved to be quite effective in the past, expecially in the MCMC framework. The first paradigmatic example is the paper by (Hastings 1970) first and (Gelfand and Smith 1990) later on, that brought to the attention of mainstream statisticians the Hastings algorithm - (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953) - that had been used to solve difficult problems in physics for over forty years before statisticians realized its potential. The paper by Gelfand e Smith has started very prolific new research lines in statistics (mostly Bayesian but also frequentist) both in theory and application and has sparked incredibly many interesting results that then become useful also to physicists.

In this paper a general variance reduction strategy, first introduced by the physicists (Assaraf and Caffarel 1999, 2003) is studied and applied to MCMC estimators in a Bayesian setting. Besides translating into statistical terms the paper by (Assaraf and Caffarel 1999), the main effort of our work has been the discussion of unbiasedness and convergence of the ZV-MCMC estimator. It should be noted that the study of CLT leads to the condition of finiteness for $\mathbb{E}_\pi[(\frac{\partial \log \pi(\mathbf{x})}{\partial \mathbf{x}})^2]$, where π is the target distribution of interest. This quantity coincides with the Fisher Information with respect to a location parameter. Fisher Information has also been used in the recent paper by (Girolami and Calderhead 2011) as a metric tensor in order to improve efficiency in both Langevin diffusion and Hamiltonian Monte Carlo methods. Their idea is to choose this metric as an optimal, local tuning of the dynamic, which is able to take into account the intrinsic anisotropy in the model considered. In our understanding, what makes ZV (introduced here) and RMHMC and RMALA (introduced in (Girolami and Calderhead 2011)) extremely efficient is the common strategy of exploiting information contained in the derivatives of the log-target. A combination of the two strategies could be explored: once the derivatives of the log-target are computed, they can be used both to boost the performance of the Markov chain (as suggested by (Girolami and Calderhead 2011)) and to achieve variance reduction by using them to design control variates. Combining ZV with clever samplers (as MMALA and RMHMC) is particularly easy since control variates can be constructed by simply post-processing the Markov chain and, thus, there is no need to re-run the simulation.

The second main contribution of this paper is the critical discussion of

the selection of H and ψ. A particular choice of H is proved to provide the same variance reduction framework exploited in (Dellaportas and Kontoyiannis 2010). In their work, control variates are derived for reversible MCMC and are related to the solution of the Poisson equation. In our context, their hypothesis of reversibility is implied by the simmetry of the particular H which is chosen. The solution to the Poisson equation depends on the transition kernel of the sampler and a closed analytical expression for the one-step ahead conditional expectations along the chain is needed to construct control variates in the setting of (Dellaportas and Kontoyiannis 2010). Moreover, the degree of variance reduction achieved depends on the MCMC implemented.

In this paper, the Schrödinger-type Hamiltonian H, introduced in the original article of the physicists, has been considered. This operator can be used only for continuous state spaces. However, it shows several advantages relative to the operator chosen by (Dellaportas and Kontoyiannis 2010). First, its definition does not depend on the kernel of the chain, so that it is simpler to evaluate. Moreover, the hypothesis of reversibility is not needed in our setting. Different choices of H and ψ could provide alternative efficient variance reduction strategies as discussed in Section 9. In the present research we have explored ψ based on first, second and thirs degree polynomials. Despite the use of this fairly restrictive class of trial functions, the degree of variance reduction obtained in the examples in Section 8 and in other simulation studies (not reported here) is impressive and of the order of tens of times (for first degree polynomials) and thousands of times (for higher degree polynomials), with practically no additional extra PCU time needed in the simulation.

11 Aknowledgements

Thanks are due to Dario Bressanini, for bringing to our attention the paper by Assaraf and Caffarel; to prof. Eugenio Regazzini, for discussing the CLT conditions for the examples; Paolo Tenconi, Filippo Carone and Fabrizio Leisen for comments and contributions to a preliminary version of this research.

References

Adler, S. (1981). Over-relaxation method for the monte carlo evaluation of the partition function for multiquadratic actions. *Phys. Rev. D 23*, 2901–2904.

Albert, J. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association 88, 422*, 669–679.

Ardia, D. (2008). Financial risk management with bayesian estimation of garch models: Theory and applications. In *Lecture Notes in Economics and Mathematical Systems 612*. Springer-Verlag.

Assaraf, R. and M. Caffarel (1999). Zero-Variance principle for Monte Carlo algorithms. *Physical Review letters 83, 23*, 4682–4685.

Assaraf, R. and M. Caffarel (2003). Zero-variance zero-bias principle for observables in quantum monte carlo: Application to forces. *The Journal of Chemical Physics 119, 20*, 10536–10552.

Barone, P. and A. Frigessi (1989). Improving stochastic relaxation for gaussian random fields. *Probability in the Engineering and Informational Sciences 4*, 369–389.

Barone, P., G. Sebastiani, and J. Stander (2001). General over-relaxation Markov chain Monte Carlo algorithms for gaussian densities. *Statistics & Probability Letters 52,2*, 115–124.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31, 3*, 307–327.

Brewer, M., C. Aitken, and M. Talbot (1996). A comparison of hybrid strategies for gibbs sampling in mixed graphical models. *Computational Statistics 21*, 343–365.

Craiu, R. and C. Lemeieux (2007). Acceleration of the multiple-try metropolis algorithm using antithetic and stratified sampling, journal statistics and computing. *Journal Statistics and Computing 17, 2*, 109–120.

Craiu, R. and X. Meng (2005). Multiprocess parallel antithetic coupling for backward and forward Markov chain Monte Carlo. *The Annals of Statistics 33, 2*, 661–697.

Dellaportas, P. and I. Kontoyiannis (2010). Control variates for reversible MCMC samplers. Technical report, arXiv:1008.1355v1.

Diaconis, P., S. Holmes, and R. F. Neal (2000). Analysis of a nonreversible Markov chain sampler. *Ann. Appl. Probab. 10,3*, 726–752.

Douc, R., A. Guillin, J. Marin, and C. Robert (2007). Minimum variance importance sampling via population monte carlo. *Probability and Statistics 11*, 427–447.

Duane, S., A. Kennedy, B. Pendleton, and D. Roweth (2010). Hybrid monte carlo. *Physics Letters B 195*, 216–222.

Flury, B. and H. Riedwyl (1988). *Multivariate Statistics*. Chapman and Hall.

Fort, G., E. Moulines, G. Roberts, and S. Rosenthal (2003). On the geometric ergodicity of hybrid samplers. *Journal of Applied Probability 40, 1*, 123–146.

Gelfand, A. and A. Smith (1990). Sampling-based approaches to calculating marginal densities. *J. American Statistical Association 85*, 398–409.

Girolami, M. and B. Calderhead (2011). Riemannian manifold langevin and hamiltonian monte carlo methods. *To appear on J. R. Statist. Soc. B 73, 2*, 1–37.

Green, P. and X. Han (1992). Metropolis methods, gaussian proposals, and antithetic variables. In P. Barone, A. Frigessi, and M. Piccioni (Eds.), *Lecture Notes in Statistics, Stochastic Methods and Algorithms in Image Analysis*, Volume 74, pp. 142–164. Springer Verlag.

Green, P. J. and A. Mira (2001). Delayed rejection in reversible jump Metropolis-Hastings. *Biometrika 88*, 1035–1053.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Henderson, S. (1997). *Variance Reduction Via an Approximating Markov Process.* Ph. D. thesis, Department of Operations Research, Stanford University, Stanford, CA.

Henderson, S. and P. Glynn (2002). Approximating martingales for variance reduction in Markov process simulation. *Math. Oper. Res. 27, 2*, 253–271.

Higdon, D. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association 93*, 585–595.

Ishwaran, H. (1999). Applications of hybrid Monte Carlo to Bayesian generalized linear models: quasicomplete separation and neural networks. *J. Comp. Graph. Statist. 8*, 779–799.

Linnik, Y. V. (1959). An information-theoretic proof of the central limit theorem with lindeberg conditions. *Theory of Probabiliy and its Applications 4*, 288–299.

McKean, H. P. (1966). Speed of approach to equilibrium for kac's caricature of a maxwellian gas. *Arch. Rat. Mech. Anal. 21*, 343367.

Metropolis, N., A. E. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics 21*, 1087–1092.

Mira, A. and C. J. Geyer (2000). On reversible Markov chains. *Fields Inst. Communic.: Monte Carlo Methods 26*, 93–108.

Mira, A., J. Möller, and G. O. Roberts (2001). Perfect slice samplers. *Journal of the Royal Statistical Soc. Ser. B 63, 3*, 593–606.

Mira, A. and L. Tierney (2002). Efficiency and convergence properties of slice samplers. *Scandinavian Journal of Statistics 29*, 1–12.

Neal, R. (1994). An improved acceptance procedure for the hybrid monte carlo algorithm. *Journal of Computational Physics 111*, 194–203.

Neal, R. M. (1995). Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation. Technical report, Learning in Graphical Models.

Ripley, B. (1987). *Stochastic Simulation*. John Wiley & Sons.

So, M. K. P. (2006). Bayesian analysis of nonlinear and non-gaussian state space models via multiple-try sampling methods. *Statistics and Computing 16*, 125–141.

Sokal, A. (1996). Monte carlo methods in statistical mechanics: Foundations and new algorithms. Lectures at the Cargese Summer School on Functional Integration: Basics and Applications.

Swendsen, R. and J. Wang (1987). Non universal critical dynamics in monte carlo simulations. *Phys. Rev. Lett. 58*, 86–88.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics 22*, 1701–1762.

Tierney, L. and A. Mira (1999). Some adaptive Monte Carlo methods for bayesian inference. *Statistics in Medicine 18*, 2507–2515.

Van Dyk, D. and X. Meng (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics 10*, 1–50.