Noname manuscript No.
(will be inserted by the editor)

Zero Variance Markov Chain Monte Carlo for Bayesian Estimators

Antonietta Mira · Reza Solgi · Daniele Imparato

Abstract Interest is in evaluating, by Markov chain Monte Carlo (MCMC) simulation, the expected value of a function with respect to a, possibly unnormalized, probability distribution. A general purpose variance reduction technique for the MCMC estimator, based on the zero-variance principle introduced in the physics literature, is proposed. Conditions for asymptotic unbiasedness of the zero-variance estimator are derived. A central limit theorem is also proved under regularity conditions. The potential of the idea is illustrated with real applications to probit, logit and GARCH Bayesian models. For all these models, a central limit theorem and unbiasedness for the zero-variance estimator are proved (see the supplementary material available on-line).

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,

$\mu_f = \int f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} / \int \pi(\mathbf{x})d\mathbf{x}$ 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

A. Mira
Swiss Finance Institute, University of Lugano, via Buffi 13, CH-6904 Lugano, Switzerland.
E-mail: antonietta.mira@usi.ch

R. Solgi
Swiss Finance Institute, University of Lugano, via Buffi 13, CH-6904 Lugano, Switzerland.
E-mail: reza.solgi@usi.ch

D. Imparato
Department of Economics, University of Insubria, via Monte Generoso 71, 21100 Varese, Italy.
E-mail: daniele.imparato@uninsubria.it

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 [39], is exploited.

In the rest of this section we briefly explain the zero-variance (ZV) principle introduced in [4,5]: an almost automatic method to construct control variates for MC simulation, in which an operator, H, acting as a map from functions to functions, and a trial function, ψ, are introduced.

In quantum mechanics, a commonly used operator H is the so-called Hamiltonian, which represents the total energy of the system, that is, the sum of the kynetic energy and the potential energy, where the kinetic energy is typically defined as a second-order differential operator. Such operator is Hermitian (that is, self-adjoint) if it acts on the restricted class of infinitely differentiable functions with compact support. If the trial function ψ belongs to this class, and if

$$H\sqrt{\pi} = 0 \tag{1}$$

the re-normalized function defined as

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{H\psi}{\sqrt{\pi(\mathbf{x})}} \tag{2}$$

satisfies $\mu_f = \mu_{\tilde{f}}$: thus both f and $\tilde{f}$ can be used to estimate the desired quantity via Monte Carlo or MCMC simulation. However, for general ψ the condition $\mu_f = \mu_{\tilde{f}}$ may not hold anymore and ad-hoc assumptions on the target π are necessary: this issue will be further discussed in Section 5.

Inspired by this physical setting, as a general framework H is supposed to be a Hermitian operator (self-adjoint and real in all practical applications) satisfying (1), and the re-normalized function is defined as in (2): depending on the specific choices of H and ψ, the condition $\mu_f = \mu_{\tilde{f}}$ has to be carefully verified.

Only a few operators will be considered in the paper, the key one being the Hamiltonian differential operator. An other important example discussed below is the Markov operator H acting as $H\psi(\mathbf{x}) = \int K(\mathbf{x},\mathbf{y})\psi(\mathbf{y})d\mathbf{y}$, where $K(\mathbf{x},\mathbf{y})$ needs to be symmetric. The re-normalized function, in this case, becomes

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{\int K(\mathbf{x},\mathbf{y})\psi(\mathbf{y})d\mathbf{y}}{\sqrt{\pi(\mathbf{x})}}. \tag{3}$$

and the condition $\mu_f = \mu_{\tilde{f}}$ holds as a simple consequence of (1).

Regardless of the specific choice of the operator and of the trial function, the optimal pair (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 $\sigma^2(\tilde{f}) = 0$, where $\sigma^2(\cdot)$ denotes the variance operator with respect to the target π. The latter, together with (2), leads to the fundamental equation:

$$H\psi = -\sqrt{\pi(\mathbf{x})}[f(\mathbf{x}) - \mu_f]. \tag{4}$$

In most practical applications equation (4) cannot be solved exactly, still, we propose to find an approximate solution in the following way. First choose a Hermitian operator H verifying (1). 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. This final estimator will be called Zero Variance (ZV) estimator through the paper.

Other research lines aim 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 ([6,21,13,41,12]); by trying to avoid random walk behavior via successive over-relaxation ([1,36,7]); by hybrid Monte Carlo ([16,35,10,18,27]); by exploiting non reversible Markov chains ([15,32]), by delaying rejection in Metropolis-Hastings type algorithms ([45,22]), by data augmentation ([46,22]) and auxiliary variables ([43,26,33, 34]). Up to our knowledge, the only other research line that uses control variates in MCMC estimation follows the PhD thesis by [24] and has its most recent developement in [14]. In [25] 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 $U(\mathbf{x}) := g(\mathbf{x}) - \mathbb{E}[g(X^{n+1})|X^n = \mathbf{x}]$ 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. In [14], the use of this type of control variates is further explored in the setting of reversible Markov chains were a closed form expression for U is often available.

In [4,5] unbiasedness and existence of a central limit theorem (CLT) for the ZV estimator are not discussed, neither in [28], where this estimator is applied to a toy example. The main contributions of this paper are, on the one hand, to derive the rigorous conditions for unbiasedness and CLT for the ZV estimators in MCMC simulation. On the other hand, we apply the ZV principle to some widely used models (probit, logit, and GARCH) and demonstrate that, under very mild restrictions, the necessary conditions for unbiasedness and CLT are verified.

2 Choice of H

In this section guidelines to choose the operator H, both for discrete and continuous 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). We restrict our attention in this section to operators H acting as $Hf := \sum_y K(\mathbf{x},\mathbf{y})f(\mathbf{y})$. The following choice

$$K(\mathbf{x},\mathbf{y}) = \sqrt{\frac{\pi(\mathbf{x})}{\pi(\mathbf{y})}}[P(\mathbf{x},\mathbf{y}) - \delta(\mathbf{x}-\mathbf{y})] \tag{5}$$

satisfies condition (1), 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 imposed on the Markov chain is essential in order to have a symmetric operator $K(\mathbf{x},\mathbf{y})$, as required.

With this choice of H, letting $\tilde{\psi} = \psi/\sqrt{\pi}$, equation (3) 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 in continuous settings. In this case, P is the kernel of the Markov chain and equation (5) can be trivially extended. This choice of H is exploited in [14], where the following fundamental equation is found for the optimal $\tilde{\psi}$: $\mathbb{E}[\tilde{\psi}(\mathbf{x}_1)|\mathbf{x}_0 = \mathbf{x}] - \tilde{\psi}(\mathbf{x}) = \mu_f - f(\mathbf{x})$. It is easy to prove that this equation coincides with our fundamental equation (4), with the choice of H given in (5). The Authors observe that the optimal trial function is given by

$$\tilde{\psi}(\mathbf{x}) = \sum_{n=0}^{\infty}[\mathbb{E}[f(\mathbf{x}_n)|\mathbf{x}_0 = \mathbf{x}] - \mu_f], \tag{6}$$

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 [4]: if $\mathbf{x} \in \mathbb{R}^d$ consider the Schrödinger-type Hamiltonian operator:

$$Hf = -\frac{1}{2}\sum_{i=1}^{d}\frac{\partial^2}{\partial x_i^2}f + V(\mathbf{x})f, \tag{7}$$

where $V(\mathbf{x})$ is constructed to fulfill equation (1): $V = \frac{1}{2\sqrt{\pi}}\Delta\sqrt{\pi}$ and Δ denotes the Laplacian operator of second order derivatives. In this setting, we obtain the general expression for $\tilde{f}$ reported in (2), where now H is the Schrödinger-type Hamiltonian. These are the operator and the re-normalized function that will be considered throughout this paper. Although it can only be applied to continuous state spaces, this Schrödinger-type operator shows several advantages with respect to the operator (5). First of all, in order to use (5) the conditional expectation appearing in (6) has to be available in closed form. Secondly, definition (7) does not require reversibility of the Markov 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 (7) and (5), 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 (4). 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, parameterized by some coefficients of a class of polynomials, 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 sequel it will be assumed that $\psi = P\sqrt{\pi}$, where P is a polynomial. 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 achieved. It can be easily shown that in a d dimensional space, using polynomials of order p, provides $\binom{d+p}{d} - 1$ control variates. However, some restrictions on the coefficients may occur in order to get an unbiased MCMC estimator. See Example 1 of Section 5 at this regard.

4 Control Variates and optimal coefficients

In this section, general expressions for the control variates in the ZV method are derived. Using the Schrödinger-type Hamiltonian H as given in (7) and trial function $\psi(\mathbf{x}) = P(\mathbf{x})\sqrt{\pi(\mathbf{x})}$, the re-normalized function is:

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) - \frac{1}{2}\Delta P(\mathbf{x}) + \nabla P(\mathbf{x}) \cdot \mathbf{z}, \qquad (8)$$

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 $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$. Like any other control variate (i.e. zero mean random variables under the distribution of interest), the variable $\mathbf{z}$ can be monitored to test convergence along the lines suggested by [11] and [38], where the same control variate $\mathbf{z} = \nabla \log \pi$ is used.

Hereafter the function P is assumed to be a polynomial. As a first case, for $P(\mathbf{x}) = \sum_{j=1}^{d} a_j x_j$ (1st degree polynomial), one gets:

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

The optimal choice of $\mathbf{a}$, that minimizes the variance of $\tilde{f}(x)$, is:

$$\mathbf{a} = -\Sigma_{\mathbf{zz}}^{-1}\sigma(\mathbf{z}, f), \qquad \text{where} \qquad \Sigma_{\mathbf{zz}} = \mathbb{E}(zz^T), \quad \sigma(\mathbf{z}, f) = \mathbb{E}(zf).$$

For a more general approach to the choice of coefficients using control variates, reference should be made to [37] and [30]. We anticipate that conditions under which the ZV-MCMC estimator obeys a CLT (Section 5) 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[1]. When higher-degree polynomials are considered, a similar formula for the coefficients associated to the control variates is obtained once an explicit formula for the control variate vector $\mathbf{z}$ has been found. As an example, 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 :

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

Using second order polynomials yields a vector of control variates of dimension $\frac{1}{2}d(d+3)$. Therefore, finding the optimal coefficients requires working with Σ_{zz} which is a matrix of dimension of orderd^2. This makes the use of second order polynomials computationally expensive when dealing with high-dimensional sampling spaces, say of the order of decades.

5 Unbiasedness and central limit theorem

As remarked in Section 1, condition (1) may not be sufficient to ensure unbiasedness of the estimator when the Schrödinger operator (7) is used. In this section general conditionys on the target π are provided that guarantee that the ZV-MCMC estimator is (asymptotically) unbiased for the class of trial functions discussed. Details can be found in the on-line supplementary material, Appendix D.

Proposition 1 *Let π be a d-dimensional density on a bounded open set Ω with regular boundary $\partial\Omega$, whose first and second derivatives are continuous. Then, if $\psi = P\sqrt{\pi}$, a sufficient condition for unbiasedness of the ZV-MCMC estimator is $\pi(\mathbf{x})\frac{\partial P(\mathbf{x})}{\partial x_j} = 0$, for all $\mathbf{x} \in \partial\Omega$, $j = 1, \ldots, d$.*

The previous proposition is a consequence of multidimensional integration by parts, from which one gets the equality

$$\mathbb{E}_\pi\left[\frac{H\psi}{\sqrt{\pi}}\right] = \frac{1}{2}\int_{\partial\Omega} [\psi\nabla\sqrt{\pi} - \sqrt{\pi}\nabla\psi] \cdot \mathbf{n} d\sigma, \tag{9}$$

where $\mathbf{n}$ denotes the versor orthogonal to $\partial\Omega$.

When π has unbounded support, integration by parts cannot be used directly. In this case, we can formulate the following result.

Proposition 2 *Let π be a d-dimensional density with unbounded support Ω, whose first and second derivatives are continuous, and let $(B_r)_r$ be a sequence of bounded subsets, so that $B_r \nearrow \Omega$. Then, a sufficient condition for unbiasedness of the ZV-MCMC estimator is*

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

[1] From a practical point of view there is no need to run two separate chains, one to get the control variates and one to get the final ZV estimator: everything can be done on a single Markov chain which is run once to estimate the optimal coefficients of the control variates and then post-processed to get the ZV estimator.

In the univariate case, if Ω is some interval of the real line, that is, $\Omega = (l, u)$, where $u, l \in \mathbb{R} \cup \pm\infty$, it is sufficient that

$$\left.\frac{dP(x)}{dx}\right|_{x=l} \pi(l) = \left.\frac{dP(x)}{dx}\right|_{x=u} \pi(u), \tag{10}$$

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

In the seminal paper by [4] unbiasedness conditions are not clearly 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 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, (10) does not hold and 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 (10) it is sufficient to consider second order polynomials. Indeed, if $P(x) = a_0 + a_1 x + a_2 x^2$ equation (10) is satisfied provided that $a_1 = 0$ and the minimization of the variance of $\tilde{f}$ can be carried out within this special class. The optimal choice $a_2 := \frac{1}{2\lambda}$ yields zero variance: $\sigma^2(\tilde{f}) \equiv 0$.

5.1 Central limit theorem

Conditions for existence of a CLT for $\hat{\mu}_f$ are well known in the literature ([44]). Using these classical results, from (8) we have that the ZV-MCMC estimator obeys a CLT provided f, ΔP and $\nabla P \cdot \mathbf{z}$ belong to $L^{2+\delta}(\pi)$ when the Markov chain run for the simulation is geometrically ergodic. In the next corollary, the case of linear and quadratic polynomials P (used in the examples in Section 6) is considered.

Corollary 1 *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, provided one of the following conditions holds:*

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

C2 : The Markov chain is uniformly ergodic and f, $x_i^k z_j \in L^2(\pi)$, $\forall i, j$ and for all $k \in \{0, \deg P - 1\}$.

In the case of linear P, using the definition of control variate, the statement of the previous corollary can be reformulated in this simple way: if $f \in L^2(\pi)$ and the chain is uniformly ergodic, then a sufficient condition to get a CLT is

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

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 [29]. The quantity m_j is also interpretable as the Fisher information of a location family in a frequentist setting.

5.2 Exponential family

Let π belong to a d-dimensional exponential family: $\pi(\mathbf{x}) \propto \exp(\beta \cdot \mathbf{T}(\mathbf{x}) - K_p(\beta))p(\mathbf{x})$, where $\beta \in \mathbb{R}^d$ is the vector of natural parameters. 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. This statement can be easily verified by a direct computation.

Theorem 1 *Let π belong to an exponential family, with p such that $\frac{\partial \log p}{\partial x_j} \in L^2(\pi)$, $\forall i, k$. Then, the Linnik functional of π is finite if and only if $\frac{\partial T_k}{\partial x_j} \in L^2(\pi)$, $\forall i, k$.*

Example 2 The Gamma density $\Gamma(\alpha, \theta)$ can be written as an exponential family on $(0, +\infty)$, where $p(x) \equiv 1$, so that hypotheses of Theorem 1 are satisfied. A direct computation shows that 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.

6 Examples

In the sequel standard statistical models are considered. For these models, the ZV-MCMC estimators are derived in a Bayesian context; from now on, the target $\pi = \pi(\beta|\mathbf{x})$ is the Bayesian posterior distribution: therefore, the argument associated with the state of the Markov chain is denoted by β instead of $\mathbf{x}$, which represents, now, the vector of data. The operator H considered is the Schrödinger-type Hamiltonian defined in (7), and $\psi = P\sqrt{\pi}$, where P is a polynomial.

Numerical simulations are provided, that confirm the effectiveness of variance reduction achieved, by minimizing the variance of $\tilde{f}$ within the class of trial functions considered. Moreover, conditions for both unbiasedness and CLT for $\tilde{f}$ are verified for all the examples. For the mathematical derivation of the zero-variance estimator and the proofs of unbiasedness and CLT for the models considered, we refer the reader to the appendices of the on-line supplementary material (Appendices A, B and C).

6.1 Probit Model

To demonstrate the effectiveness of ZV for probit models, a simple example is presented. The bank dataset from [17] 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 probit model as the regressors, and the type of the banknote y_i, is the response variable (0 for genuine and 1 for counterfeit). Using flat priors, the Bayesian estimator of each parameter, β_k, under squared error loss function, is the expected value of $f_k(\beta) = \beta_k$ under π $(k = 1, 2, \cdots, d)$. The Bayesian analysis of this problem is discussed in [31]. In order to find the optimal vector of parameters a_k of the trial functions, a short Gibbs sampler, following ([2]), (of length 2000, after 1000 burn in steps) is run, and the optimal coefficients are estimated: $\hat{\mathbf{a}}_k = -\hat{\Sigma}^{-1}_{\mathbf{zz}}\hat{\sigma}(\mathbf{z}, \beta_k)$. Finally another MCMC simulation of length 2000 is run (and using the estimated optimal values obtained in the previous step), along which $\widetilde{f}_k(\beta)$, for $k = 1, \ldots, 4$ is averaged. We have repeated this experiment 100 times. The MCMC traces of the ordinary MCMC and the ZV-MCMC in one of these MOnte Carlo experiments have been depicted in the left plot of Fig. 1. The blue curves are the traces of f_k (ordinary MCMC), and the red ones are the traces of $\widetilde{f}_k$ (ZV-MCMC). It is clear from the figure that the variances of the estimator have substantially decreased. Indeed for the linear trial functions, the ratios of the Monte Carlo estimates of the asymptotic variances of the two estimators (ordinary MCMC and ZV-MCMC) are between 25 and 100. Even better performance can be achieved using second degree polynomials to define the trial function. In the right column of Fig. 1 the traces of ZV-MCMC with second order $P(x)$ are reported along with the traces of the ordinary MCMC. As it can be seen from the figure, the variances of the ZV estimators are negligible: the ratio of the Monte Carlo estimates of the asymptotic variances of the two estimators are between $18,000$ and $90,000$. In this example (with the simulation length and burn-in reported above) the CPU time of ZV-MCMC is almost 3 times larger than the one of ordinary MCMC.

In order to study the unbiasedness of the ZV-estimators empirically, we have run a very long MCMC (of length 10^8) and obtained a very narrow 95% confidence region for each parameter. In Fig. 2 we have depicted the box-plot of the ordinary MCMC (first box-plot), and the ZV-estimators (second and third box-plot) along with these 95% confidence regions (the green regions). As it can be seen, the ZV-estimators are concentrated in the 95% confidence regions obtained from the very long chain.

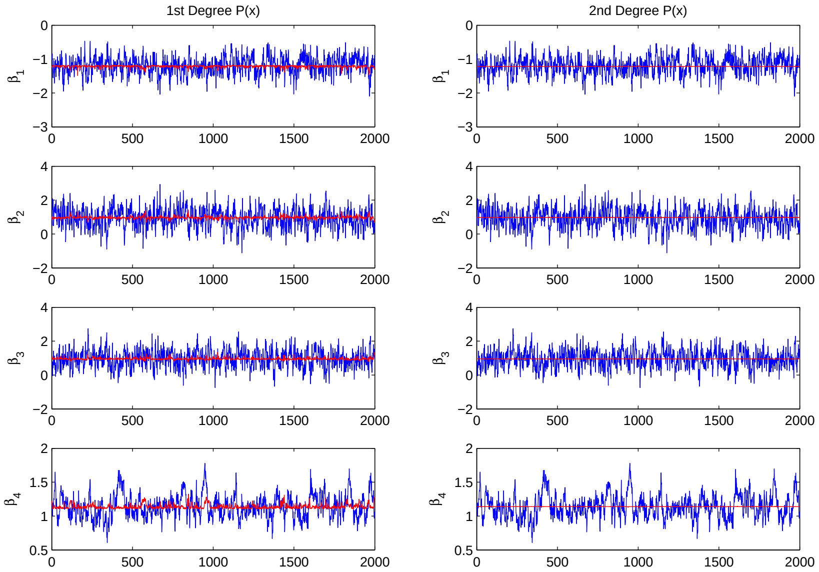

Fig. 1 Ordinary MCMC (blue) and ZV-MCMC (red) for probit model: rows are parameters, columns are degree polynomials.

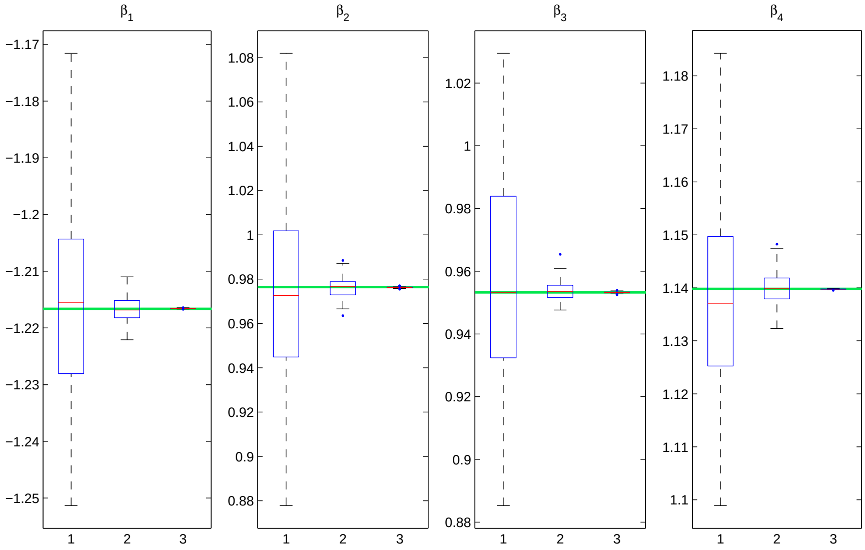

Fig. 2 Boxplots of ordinary MCMC estimates (1) and ZV-MCMC estimates (2 and 3) for the probit model, along with the 95% confidence region obtained by an ordinary MCMC of length 10^8 (green regions).

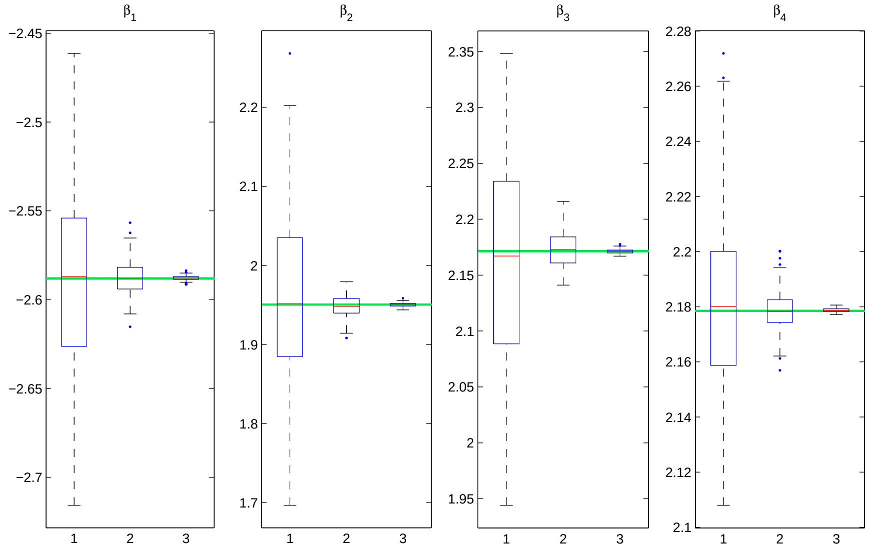

Fig. 3 Box-plots of ordinary MCMC estimates (1) and ZV-MCMC estimates (2 and 3) for the logit model, along with the 95% confidence region obtained by an ordinary MCMC of length 10^8 (green regions).

6.2 Logit Model:

A logit model is fitted to the same dataset of Swiss banknotes previously introduced. Flat priors are used and, as before, the Bayesian estimator of each parameter, β_k, (again under squared error loss functions) is the expected value of β_k under π $(k = 1, 2, \cdots, d)$. Similar to the probit 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 performed, and $\tilde{f}_k$ is averaged, using the optimal trial function estimated in the first stage (the same simulation length and burn-in, as in the probit example, have been used). For linear polynomial, the ratio of the Monte Carlo estimates of the asymptotic variances of the two estimators (ordinary MCMC and ZV-MCMC) are between 15 and 50. Using quadratic polynomials, these ratios are between $15,000$ and $20,000$. In this example the CPU time of the ZV-MCMC is almost 3 times higher than that of ordinary MCMC.

We have run a very long MCMC (of length 10^8) and obtained a very narrow 95% confidence region for each parameter. In Fig. 3 we have depicted the box-plot of the ordinary MCMC (first box-plot), and the ZV-estimators (second and third box-plot) along with these 95% confidence regions (the green regions). Again, as it can be seen, the ZV-estimators are concentrated in the 95% confidence regions obtained from the very long Markov chain.

Table 1 GARCH variance reduction: 95% Confidence interval for the ratio of the variances of ordinary MCMC estimators and ZV-MCMC estimator.

	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\omega}_3$
1st Degree $P(x)$	8-18	13-28	12-27
2nd Degree $P(x)$	1200-2700	6100-13500	6200-13800
3rd Degree $P(x)$	21000-47000	48000-107000	26000-58000

6.3 GARCH Model

Generalized autoregressive conditional heteroskedasticity (GARCH) models ([8]) have become one of the most important building blocks of models in financial econometrics, where they are 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. Let $S(t)$ be the exchange rate at time t. The daily returns are defined as $r(t) := [S(t) - S(t-1)]/S(t-1) \approx \ln(S(t)/S(t-1))$. 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 $h_t = \omega_1 + \omega_3 h_{t-1} + \omega_2 r_{t-1}^2$, and $\omega_1 > 0$, $\omega_2 \geq 0$, and $\omega_3 \geq 0$ are the parameters of the model. The aim is to estimate the expected value of ω_j under the posterior π, using independent truncated normal priors. As an example, a Normal-GARCH(1, 1) is fitted to the daily returns of the Deutsche Mark vs British Pound exchange rates from January 1985, to December 1987. In the first stage a short MCMC simulation ([3]) is used to estimate the optimal parameters of the trial function (2000 sweeps after 1000 burn-in). In the second stage an independent simulation is run (with length 10000) and $\tilde{f}_k(\omega)$ is averaged in order to efficiently estimate the posterior mean of each parameter. We compare this ZV-MCMC with an ordinary MCMC of length 10000 (after 1000 burn-in). First, second and third degree polynomials in the trial function are used. In order to study the effectiveness of ZV-MCMC, we have run these simulations (ordinary MCMC and ZV-MCMC) 100 times. As it can be seen in Table 1, where a 95% confidence interval for the variance reductions are reported, the ZV strategy reduces the variance of the estimators up to ten thousand times. In this example (with the simulation and burn-in lengths reported above) the CPU time of the ZV-MCMC is almost 20% higher than the CPU time of ordinary MCMC.

In order to study the unbiasedness of the ZV-estimators empirically, we have run a very long MCMC (of length 10^7) and obtained a narrow 95% confidence region for each parameter. In Fig. 4 we have depicted the box-plot of the ordinary MCMC (first box-plot), and the ZV-estimators (second, third and fourth box-plots) along with these 95% confidence regions (the green regions). As it can be seen the ZV-estimators lie in the range obtained by the very long MCMC.

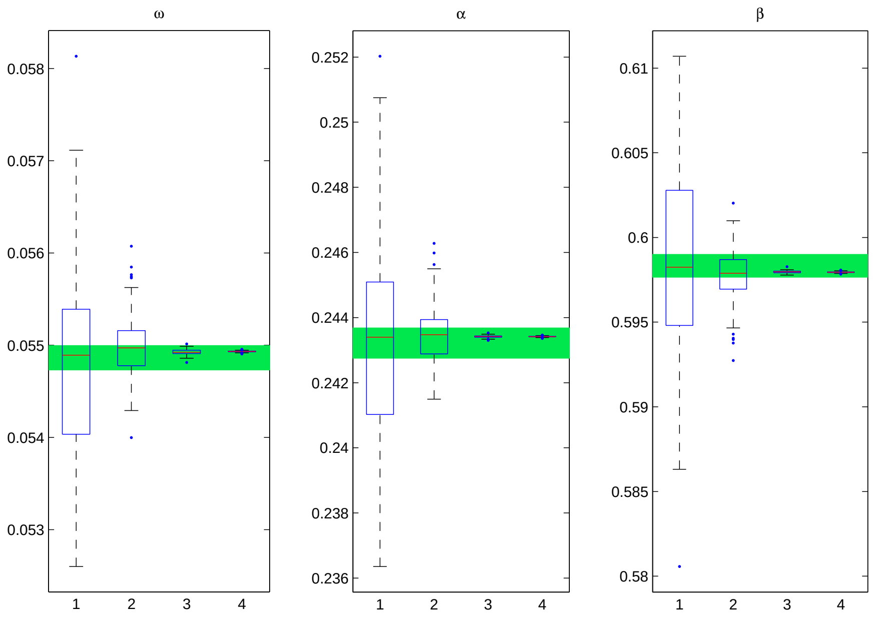

Fig. 4 Boxplots of ordinary MCMC estimates (1) and ZV-MCMC estimates (2, 3 and 4) for the GARCH model, along with the 95% confidence region obtained by an ordinary MCMC of length 10^7 (green regions).

Finally, note that the ZV strategy can be used in great generality and can be applied also to more complex GARCH models (such as E-GARCH, I-GARCH, Q-GARCH, GJR-GARCH, [9]), provided it is possible to analitically compute the necessary derivatives and verify the hypotheses needed for unbiasedness and CLT, in a way similar to the proof reported in Appendix C.

7 Discussion

Cross-fertilizations between physics and statistical literature have proved to be quite effective in the past, especially in the MCMC framework. The first paradigmatic example is the paper by [23] first and [19] later on.

Besides translating into statistical terms the paper by [4], the main effort of our work has been the discussion of unbiasedness and convergence of the ZV-MCMC estimator. The study of CLT leads to the condition of finiteness for $\mathbb{E}_\pi[(\frac{\partial \log \pi(\mathbf{x})}{\partial \mathbf{x}})^2]$. This quantity has also been used in the recent paper by [20] as a metric tensor to improve efficiency in Langevin diffusion and Hamiltonian MC 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 these methods and our 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 [20]) and to achieve variance reduction by using them to design control variates. This 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 comparison between the variance reduction framework exploited in [14] and the choice of different operators H in our context has remarked contras and benefits of the two approaches. Different choices of H and ψ could provide alternative efficient variance reduction strategies. 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.

In the present research we have explored ψ based on first, second and third degree polynomials. Despite the use of this fairly restrictive class of trial functions, the degree of variance reduction obtained in the examples in Section 6 and in other simulation studies (not reported here) is impressive and of the order of ten times (for first degree polynomials) and of thousand times (for higher degree polynomials), with practically small extra CPU time needed in the simulation.

Finally, mention should be made to an alternative, more general renormalized function $\tilde{f}$ reported in the paper by [5], defined as:

$$\tilde{f} = f + \frac{H\psi}{\sqrt{\pi}} - \frac{\psi(H\sqrt{\pi})}{\pi}, \tag{11}$$

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 (1), so that now V can be also chosen arbitrarily. Therefore, the re-normalization (11) allows for a more general class of Hamiltonians.

Supplementary Materials

Supplementary materials are available. In Appendices A, B and C the zero variance estimator and the proof of CLT are given for all the examples. In Appendix D computations of unbiasedness conditions discussed in Section 5 are reported and verified for the three examples.

Acknowledgement

Thanks are due to D. Bressanini, for bringing to our attention the paper by Assaraf and Caffarel and helping us translate it into statistical terms; to prof. E. Regazzini and F. Nicola, for discussing the CLT conditions for the examples; P. Tenconi, F. Carone and F. Leisen for comments and contributions to a preliminary version of this research. And finally, Assaraf and Caffarel themselves have given us interesting and useful comments that have greatly improved the paper.

Appendix A: Probit model

Mathematical formulation

Let y_i be Bernoulli r.v.'s: $y_i|\mathbf{x}_i \sim \mathcal{B}(1, p_i)$, $\quad p_i = \Phi(\mathbf{x}_i^T\beta)$, where $\beta \in \mathbb{R}^d$ is the vector of parameters of the model 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 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.

Central limit theorem

In the following, it is supposed that P is a linear polynomial. In the Probit model, the ZV-MCMC estimators obey a CLT if z_j have finite $2+\delta$ moment under π, for some $\delta > 0$:

$$\begin{aligned}\mathbb{E}_\pi\left[|z_j|^{2+\delta}\right] &= c_1 \mathbb{E}_\pi\left[\left|\sum_{i=1}^n \frac{x_{ij}\phi(\mathbf{x}_i^T\beta)}{\Phi(\mathbf{x}_i^T\beta)}\right|^{2+\delta}\right] \\ &= c_1 c_2 \int_{\mathbb{R}^d}\left|\sum_{i=1}^n \frac{x_{ij}\phi(\mathbf{x}_i^T\beta)}{\Phi(\mathbf{x}_i^T\beta)}\right|^{2+\delta}\prod_{i=1}^n \Phi(\mathbf{x}_i^T\beta)d\beta < \infty.\end{aligned}$$

where $c_1 = 2^{-2-\delta}$, and c_2 is the normalizing constant of π (the target posterior). Define:

$$\begin{aligned}K_1(\beta) &= \left|\sum_{i=1}^n \frac{x_{ij}\phi(\mathbf{x}_i^T\beta)}{\Phi(\mathbf{x}_i^T\beta)}\right|^{2+\delta}, \\ K_2(\beta) &= \prod_{i=1}^n \Phi(\mathbf{x}_i^T\beta), \\ K(\beta) &= K_1(\beta)K_2(\beta)\end{aligned}$$

and therefore:

$$\mathbb{E}_\pi\left[|z_j|^{2+\delta}\right] = c\int_{\mathbb{R}^d} K_1(\beta)K_2(\beta)d\beta.$$

where $c = c_1c_2$. Before studying the convergence of this integral, the following property of the likelihood for the probit model is needed.

Proposition 3 *Existence and uniqueness of MLE implies that, for any $\beta_0 \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, there exists i such that $\boldsymbol{x}_i^T\beta_0 < 0$.*

Proof (by contradiction). Uniqueness of MLE implies that $\mathbf{x}^T\mathbf{x}$ is full rank, that is, there is no β_0 orthogonal to all observations $\mathbf{x}_i$. This can be seen by contradiction: singularity of $\mathbf{x}^T\mathbf{x}$ implies existence of a non-zero β_0 orthogonal to all observations $\mathbf{x}_i$. This fact, in turn, implies $l(\beta|\mathbf{x},\mathbf{y}) = l(\beta + c\beta_0|\mathbf{x},\mathbf{y})$ and, therefore, $l(\bullet|\mathbf{x},\mathbf{y})$ does not have a unique global maximum.
Next, assume there exists some $\beta_0 \in \mathbb{R}^d$ such that, for any i, $\mathbf{x}_i^T\beta_0 > 0$. Then β_0 is a direction of recession for the negative log-likelihood function $-\sum_{i=1}^n \ln\Phi(\mathbf{x}_i^T\beta)$ (that is a proper closed convex function). This implies that this function does not have non-empty bounded minimum set ([40]), which means that the MLE does not exist. ■

Now, rewriting $\int_{\mathbb{R}^d} K(\beta)d\beta$ in hyper-spherical coordinates through the bijective transformation $(\rho, \theta_1, \ldots, \theta_{d-1}) := F(\beta)$, where F^{-1} is defined as

$$\begin{cases} \beta_1 &= \rho\cos(\theta_1) \\ \beta_l &= \rho\cos(\theta_l)\prod_{m=1}^{l-1}\sin(\theta_m), \qquad \text{for } l = 2, ..., d-1 \\ \beta_d &= \rho\prod_{m=1}^{d-1}\sin(\theta_m), \end{cases} \tag{12}$$

for $\theta \in \mathit{\Theta} := \{0 \leq \theta_i \leq \pi,\ i = 1, \ldots, d-2,\ 0 \leq \theta_{d-1} < 2\pi\}$ and $\rho > 0$, one gets

$$\begin{aligned} \int_{\mathbb{R}^d} K(\beta)d\beta &= \int_{\mathit{\Theta}}\int_0^{+\infty} K(F^{-1}(\rho,\theta))\rho^{d-1}\prod_{j=2}^{d-2}\sin^{d-j}(\theta_{j-1})\ d\rho d\theta. \\ &\leq \int_{\mathit{\Theta}}\int_0^{+\infty} K(F^{-1}(\rho,\theta))\rho^{d-1}\ d\rho d\theta \\ &:= \int_{\mathit{\Theta}} A(\theta)d\theta, \end{aligned}$$

Observe that the integrand is well defined for any (ρ, θ) on the domain of integration, so it is enough to study its asymptotic behaviour when ρ goes to infinity, and $\theta \in \mathit{\Theta}$.

First, analyze

$$K_1(F^{-1}(\rho,\theta)) = \left|\sum_{i=1}^{n}\frac{x_{ij}\phi(|\mathbf{x}_i|\rho\lambda_i(\theta)}{\Phi(|\mathbf{x}_i|\rho\lambda_i(\theta))}\right|^{2+\delta},$$

where, for any i, λ_i is a suitable function of the angles θ such that $\lambda_i \in [-1, 1]$, which takes into account the sign of the scalar product in the original coordinates system.

For any i, when $\rho \to \infty$

- if $\lambda_i < 0$, $\frac{x_{ij}\phi(|\mathbf{x}_i|\rho\lambda_i)}{\Phi(|\mathbf{x}_i|\rho\lambda_i)} \in \mathcal{O}(\rho)$;
- if $\lambda_i > 0$, $\frac{x_{ij}\phi(|\mathbf{x}_i|\rho\lambda_i)}{\Phi(|\mathbf{x}_i|\rho)\lambda_i} \in \mathcal{O}(\phi(\lambda_i\rho))$;
- if $\lambda_i = 0$, $\frac{x_{ij}\phi(|\mathbf{x}_i|\rho\lambda_i)}{\Phi(|\mathbf{x}_i|\rho\lambda_i)} = x_{ij}\sqrt{\frac{2}{\pi}} \in \mathcal{O}(1)$.

Therefore:

$$\sum_{i=1}^{n}\frac{x_{ij}\phi(|\mathbf{x}_i|\rho\lambda_i)}{\Phi(|\mathbf{x}_i|\rho\lambda_i)} \in \mathcal{O}(\rho)$$

and, for any $\theta \in \mathit{\Theta}$: $K_1(F^{-1}(\rho,\theta)) \in \mathcal{O}\left(\rho^{2+\delta}\right)$. Now, focus on $K_2(F^{-1}(\rho,\theta)) = \prod_{i=1}^{n}\Phi(|\mathbf{x}_i|\rho\lambda_i(\theta))$; existence of MLE for the probit model implies that, for any $\theta \in \mathit{\Theta}$, there exists some l $(1 \leq l \leq n)$, such that $\lambda_l(\theta) < 0$, and therefore:

$$K_2(F^{-1}(\rho,\theta)) < \Phi(|\mathbf{x}_l|\rho\lambda_l) \in \mathcal{O}(\phi(\lambda_l\rho)) \quad \rho \to \infty. \tag{13}$$

Putting these results together leads to

$$K(F^{-1}(\rho,\theta)) = K_1(F^{-1}(\rho,\theta))K_2(F^{-1}(\rho,\theta)) \in \mathcal{O}\left(\rho^{2+\delta}\phi(\lambda_l(\theta)\rho)\right)$$

so that, for any $\theta \in \Theta$,

$$K(F^{-1}(\rho,\theta))\rho^{d-1} \in \mathcal{O}\left(\rho^{1+\delta+d}\phi\left(\lambda_l(\theta)\rho\right)\right), \quad \rho \to +\infty.$$

Therefore, whenever the value $\theta \in \Theta$, its integrand converges to zero rapidly enough when $\rho \to +\infty$. This concludes the proof.

Note 1. In ([42]) it is shown that the existence of the posterior under flat priors for probit and logit models is equivalent to the existence and finiteness of MLE. This ensures us that the posterior is well defined in our context. In order to verify the existence of the posterior mean, we can use a simplified version of the proof given above. In other words we should show $\int \beta_j \prod_{i=1}^n \Phi(\mathbf{x}_i^T\beta)d\beta < +\infty$, that is, $K_1(\beta) = \beta_j \in \mathcal{O}(\rho)$. Therefore, a weaker version of the proof given above can be employed.

Note 2. In the proof given above we have used flat priors: although this assumption simplifies the proof, however a very similar proof can be applied for non-flat priors. Assume the prior is $\pi_0(\beta)$. Under this assumption the posterior is $\pi(\beta) = \pi_0(\beta)\ l(\beta|\mathbf{y},\mathbf{x}) \propto \pi_0(\beta)\ \prod_{i=1}^n \Phi(\mathbf{x}_i^T\beta)$ and the control variates are: $z_j = -\frac{1}{2}\frac{d\ln\pi_0(\beta)}{d\beta_j} - \frac{1}{2}\sum_{i=1}^n \frac{x_{ij}\phi(\mathbf{x}_i^T\beta)}{\Phi(\mathbf{x}_i^T\beta)}$. Therefore we need to prove the $2+\delta$-th moment of z_j under $\pi(\beta)$ is finite. A sufficient condition for this is the finiteness of $2+\delta$-th moments of $-\frac{1}{2}\frac{d\ln\pi_0(\beta)}{d\beta_j}$ and $-\frac{1}{2}\sum_{i=1}^n \frac{x_{ij}\phi(\mathbf{x}_i^T\beta)}{\Phi(\mathbf{x}_i^T\beta)}$ under $\pi(\beta)$. If we assume $\pi_0(\beta)$ is bounded above, the latter is a trivial consequence of the proof given above for the flat priors. Therefore we only need to prove the finiteness of the integral

$$\int_{\mathbb{R}^d}\left|\frac{d\ln\pi_0(\beta)}{d\beta_j}\right|^{2+\delta}\pi_0(\beta)\ \prod_{i=1}^n\Phi(\mathbf{x}_i^T\beta)d\beta.$$

Again if we assume the prior is bounded from above, a sufficient condition for the existence of this integral is the existence of the following integral:

$$\int_{\mathbb{R}^d}\left|\frac{d\ln\pi_0(\beta)}{d\beta_j}\right|^{2+\delta}\ \prod_{i=1}^n\Phi(\mathbf{x}_i^T\beta)d\beta$$

A proof very similar to the one given above will show that this integral is finite for common choices of priors $\pi_0(\beta)$ (such as Normal, Student's T, etc).

Appendix B: Logit model

Mathematical formulation

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:

$$l(\beta|\mathbf{y},\mathbf{x}) \propto \prod_{i=1}^n\left(\frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}\right)^{y_i}\left(\frac{1}{1+\exp(\mathbf{x}_i^T\beta)}\right)^{1-y_i}. \qquad (14)$$

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:

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

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 (14) 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 pair of operator H and test function ψ_k as in Appendinx A, the control variates are:

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

Central limit theorem

As for the probit model, the ZV-MCMC estimators obey a CLT if the control variates z_j have finite $2 + \delta$ moment under π, for some $\delta > 0$:

$$\begin{aligned} \mathbb{E}_\pi \left[z_j^{2+\delta} \right] &= c_1 \mathbb{E}_\pi \left[\left(\sum_{i=1}^{n} x_{ij} \frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)} \right)^{2+\delta} \right] \\ &= c_1 c_2 \int_{\mathbb{R}^d} \left(\sum_{i=1}^{n} x_{ij} \frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)} \right)^{2+\delta} \prod_{i=1}^{n} \frac{1}{1 + \exp(\mathbf{x}_i^T \beta)} d\beta \end{aligned}$$

where $c_1 = 2^{-2-\delta}$, and c_2 is the normalizing constant of π. The finitiness of the integral is, indeed, trivial. Observe that the function $e^y/(1 + e^y)$ is bounded by 1; therefore,

$$\mathbb{E}_\pi \left[z_j^{2+\delta} \right] \leq \left(\sum_{i=1}^{n} x_{ij} \right)^{2+\delta} < \infty.$$

As for probit model, note that it can easily be shown that the posterior means exist under flat priors. Moreover, a very similar proof (see Note 2 of Appendix A) can be used under a normal or Student's T prior distribution.

Appendix C: GARCH model

Mathematical formulation

We assume that the returns are conditionally normal 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:

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

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

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

Therefore, the control variates (when the trial function is a first degree polynomial) 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:

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

Central limit theorem

In order to prove the CLT for the ZV-MCMC estimator in the Garch model, we need:

$$\frac{\partial \ln \pi}{\partial \omega_2}, \quad \frac{\partial \ln \pi}{\partial \omega_3}, \quad \frac{\partial \ln \pi}{\partial \omega_1} \in L^{2+\delta}(\pi). \tag{15}$$

To this end, h_t and its partial derivatives should be expressed as a function of h_0 and $\mathbf{r}$:

$$h_t = \omega_1(\sum_{k=1}^{t-1} 1 + \omega_3^k) + \omega_3^t h_0 + \omega_2(\sum_{k=1}^{t-1} 1 + \omega_3^k r_{t-1-k}^2),$$

$$\begin{aligned}
\frac{\partial \ln h_t}{\partial \omega_1} &= \tfrac{1-\omega_3^{t-1}}{1-\omega_3}\mathbb{I}_{\{t>1\}}, \\
\frac{\partial \ln h_t}{\partial \omega_2} &= \left(r_{t-1}^2 + \sum_{j=0}^{t-2}\omega_3^{t-1-j} r_j^2\right)\mathbb{I}_{\{t>1\}}, \\
\frac{\partial \ln h_t}{\partial \omega_3} &= \left(h_{t-1} + \sum_{j=0}^{t-2}\omega_3^{t-1-j} h_j\right)\mathbb{I}_{\{t>1\}}.
\end{aligned}$$

Next, moving to spherical coordinates, the integral (15) can be written as

$$\int_{[0,\pi/2]^2}\int_0^\infty K_j(\rho;\theta,\phi)d\rho d\theta d\phi := \int_{[0,\pi/2]^2} A_j(\theta,\phi)d\theta d\phi,$$

where, for $j=1,2,3$, $K_j(\cdot;\theta,\phi)=|W_j|^{2+\delta}\times W$, with

$$\begin{aligned}
W_1 &= -\tfrac{1}{\sigma^2(\omega_1)}\rho\cos\theta\sin\phi - \frac{1}{2}\sum_{t=2}^{T}\left(\frac{1}{\tilde{h}_t}-\frac{r_t^2}{\tilde{h}_t^2}\right)\frac{1-\rho^{t-1}\cos^{t-1}\phi}{1-\rho\cos\phi},\\
W_2 &= -\tfrac{1}{\sigma^2(\omega_2)}\rho\sin\theta\sin\phi - \frac{1}{2}\sum_{t=2}^{T}\left(\frac{1}{\tilde{h}_t}-\frac{r_t^2}{\tilde{h}_t^2}\right)\left(r_{t-1}^2+\sum_{j=0}^{t-2}x_3^{t-1-j}r_j^2\right),\\
W_3 &= -\tfrac{1}{\sigma^2(\omega_3)}\rho\cos\phi - \frac{1}{2}\sum_{t=2}^{T}\left(\frac{1}{\tilde{h}_t}-\frac{r_t^2}{\tilde{h}_t^2}\right)\left(\tilde{h}_{t-1}+\sum_{j=0}^{t-2}x_3^{t-1-j}\tilde{h}_j\right),\\
W &= \exp\left(-\tfrac{1}{2}\rho^2\left(\tfrac{1}{\sigma^2(\omega_1)}\cos^2\theta\sin^2\phi+\tfrac{1}{\sigma^2(\omega_2)}\sin^2\theta\sin^2\phi+\tfrac{1}{\sigma^2(\omega_3)}\cos^2\phi\right)-\tfrac{1}{2}\sum_{t=1}^{T}\frac{r_t^2}{\tilde{h}_t^2}\right)\times\\
&\quad\times\rho^2\sin\theta\left(\prod_{t=1}^{T}\tilde{h}_t\right)^{-\frac{1}{2}}
\end{aligned} \tag{16}$$

and

$$\tilde{h}_t = -\rho\cos\theta\sin\phi\sum_{k=1}^{t-1}(1+\rho^k\cos^k\phi)+h_0\rho^t\cos^t\phi+\rho\sin\theta\sin\phi\sum_{k=1}^{t-1}(1+r_{t-1-k}^2\rho^k\cos^k\phi).$$

The aim is to prove that, for any $\theta,\phi\in[0,\pi/2]$ and for any j, $A_j(\theta,\phi)$ is finite. To this end, the convergence of A_j for any θ,ϕ should be discussed.

Let us study the proper domain of $K_j(\cdot;\theta,\phi)$. Observe that $K_j(\cdot;\theta,\phi)$ is not defined whenever $\tilde{h}_t=0$ and, if $j=1$ and $\phi\neq\pi/2$, also for $\rho=1/\cos\phi$. However, the discontinuity of K_3 at this point is removable, so that the domain of K_3 can be extended by continuity also at $\rho=1/\cos\phi$.

Since $\tilde{h}_t=0$ if and only if $\rho=0$, it can be concluded that, for any j and for any $\theta,\phi\in[0,\pi/2]$, the proper domain of $K_j(\cdot;\theta,\phi)$ is $\mathrm{dom}K_j(\cdot;\theta,\phi)=(0+\infty)$. By fixing the value of θ and ϕ, let us study the limits of K_j when $\rho\to0$ and $\rho\to+\infty$. Observe that, whatever the values of θ and ϕ are, W_j's are rationale functions of ρ. Therefore, for any j, $|W_j|^{2+\delta}$ cannot grow towards infinity more than polynomially at the boundary of the domain. On the other hand, W goes to zero with an exponential rate both when $\rho\to0$ and $\rho\to+\infty$, for any θ and ϕ. This is sufficient to conclude that, for any $\theta,\phi\in[0,\pi/2]$, the integral A_j is finite for $j=1,2,3$ and, therefore, condition (15) holds and the ZV estimators for the GARCH model obeys a CLT.

Appendix D: unbiasedness

In this Appendix, explicit computations are presented, which were omitted in Section 5. Moreover, it is proved that all the ZV-MCMC estimators discussed in Section 8 are unbiased.

Following the same notations as in Section 5, equation (9) follows because

$$\begin{aligned}
\left\langle \frac{H\psi}{\sqrt{\pi}} \right\rangle &:= \int_{\Omega} H\psi\sqrt{\pi} \\
&= \int_{\Omega} (V\psi\sqrt{\pi} - \frac{1}{2}\Delta\psi\sqrt{\pi}) \\
&= \int_{\Omega} V\sqrt{\pi}\psi - \frac{1}{2}\int_{\partial\Omega} \sqrt{\pi}\nabla\psi\cdot\mathbf{n}d\sigma + \frac{1}{2}\int_{\Omega}\nabla\sqrt{\pi}\cdot\nabla\psi \\
&= \int_{\Omega} V\sqrt{\pi}\psi - \frac{1}{2}\int_{\partial\Omega} \sqrt{\pi}\nabla\psi\cdot\mathbf{n}d\sigma + \frac{1}{2}\int_{\partial\Omega}\psi\nabla\sqrt{\pi}\cdot\mathbf{n}d\sigma - \frac{1}{2}\int_{\Omega}\psi\Delta\sqrt{\pi} \\
&= \int_{\Omega} (H\sqrt{\pi})\psi + \frac{1}{2}\int_{\partial\Omega}[\psi\nabla\sqrt{\pi} - \sqrt{\pi}\nabla\psi]\cdot\mathbf{n}d\sigma \\
&= \frac{1}{2}\int_{\partial\Omega}[\psi\nabla\sqrt{\pi} - \sqrt{\pi}\nabla\psi]\cdot\mathbf{n}d\sigma.
\end{aligned}$$

Therefore, $\left\langle \frac{H\psi}{\sqrt{\pi}} \right\rangle = 0$ if $\psi\nabla\sqrt{\pi} = \sqrt{\pi}\nabla\psi$ on $\partial\Omega$. Now, let $\psi = P\sqrt{\pi}$. Then,

$$\nabla\psi = \sqrt{\pi}\nabla P + \frac{P}{2\sqrt{\pi}}\nabla\pi,$$

so that $\left\langle \frac{H\psi}{\sqrt{\pi}} \right\rangle = 0$ if

$$\pi(\mathbf{x})\frac{\partial P(\mathbf{x})}{\partial x_j} = 0, \quad \forall\mathbf{x}\in\partial\Omega, \quad j = 1,\dots,d.$$

When π has unbounded support, following the previous computations integrating over the bounded set B_r and taking the limit for $r\to\infty$, one gets

$$\left\langle \frac{H\psi}{\sqrt{\pi}} \right\rangle = \frac{1}{2}\lim_{r\to+\infty}\int_{\partial B_r}\pi\nabla P\cdot\mathbf{n}d\sigma. \tag{17}$$

Therefore, unbiasedness in the unbounded case is reached if the limit appearing in the right-hand side of (17) is zero.

Now, the unbiasedness of the ZV-MCMC estimators exploited in Section 8 is discussed. To this end, condition (17) should be verified. Let B_ρ be a hyper-sphere of radius ρ and let $\mathbf{n} := \frac{1}{\rho}\beta$ be its normal versor. Then, for linear P, (17) equals zero if, for any $j = 1,\dots,d$,

$$\lim_{\rho\to+\infty}\frac{1}{\rho}\int_{B_\rho}\pi(\beta)\beta_j dS = 0. \tag{18}$$

The Probit model is first considered. By using the same notations as in Appendix A, the integral in (18) is proportional to

$$\lim_{\rho\to+\infty}\frac{1}{\rho}\int_{\Theta}K_2(F^{-1}(\rho,\theta))\rho^d d\theta. \tag{19}$$

Note that, because of (13), there exist ρ_0 and M such that

$$\begin{aligned}K_2(F^{-1}(\rho,\theta))\rho^d &\le M\phi(\lambda_{l(\theta)}\rho)\rho^d\\ &\le M\phi(\lambda_{l(\theta)}\rho_0)\rho_0^d := G(\theta) \ \ \forall\rho\ge\rho_0.\end{aligned}$$

Since $G(\theta)\in L^1$, by the dominated convergence theorem a sufficient condition to get unbiasedness is

$$\lim_{\rho\to+\infty}K_2(F^{-1}(\rho,\theta))\rho^{d-1}=0, \tag{20}$$

which is true, because of (13) for the Probit model.

We now consider the Logit model for which it is easy to prove that Proposition 3 holds. As done for the Probit model, one can write

$$\mathbb{E}_\pi\left[z_j^{2+\delta}\right]\propto\int_{\mathbb{R}^d}K_1(\beta)K_2(\beta)d\beta,$$

where

$$\begin{aligned}K_1(\beta)&=\left(\sum_{i=1}^n x_{ij}\frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}\right)^{2+\delta},\\ K_2(\beta)&=\prod_{i=1}^n\frac{1}{1+\exp(\mathbf{x}_i^T\beta)},\end{aligned}$$

By using the hyper-spherical change of variables in (12), we get

$$\mathbb{E}_\pi\left[z_j^{2+\delta}\right]\propto\int_{\Theta}\int_0^\infty K_1(F^{-1}(\rho,\theta))K_2(F^{-1}(\rho,\theta))\rho^{d-1}\ d\rho:$$

and, as for the Probit model, we must verify Equation (19). Now analyze $K_2(F^{-1}(\rho,\theta))$; for any θ, existence of MLE implies the existence of some l $(1\le l\le n)$, such that $\lambda_l(\theta)>0$, and therefore:

$$\begin{aligned}K_2(F^{-1}(\rho,\theta))&=\prod_{i=1}^n\frac{1}{1+\exp(|\mathbf{x}_i|\rho\lambda_i)}\\ &<\frac{1}{1+\exp(|\mathbf{x}_l|\rho\lambda_l)}\\ &\in\mathcal{O}\left(\exp(-\rho\lambda_l)\right).\end{aligned} \tag{21}$$

Therefore, there exist ρ_0, M such that

$$K_2(F^{-1}(\rho,\theta))\rho^d \leq M\exp(-\lambda_{l(\theta)}\rho)\rho^d$$
$$\leq M\exp(-\lambda_{l(\theta)}\rho_0)\rho_0^d := G(\theta) \ \ \forall\rho \geq \rho_0,$$

where $G(\theta) \in L^1$. These computations allow us to use Equation (20) as a sufficient condition to get unbiasedness, and its proof becomes trivial.

Finally consider the GARCH model. In this case, B_ρ is the portion of a sphere of radius ρ defined on the positive orthant. Then, the limit

$$\lim_{\rho\to+\infty}\frac{1}{\rho}\int_{[0,\pi/2]^2} W(F^{-1}(\rho,\theta))\rho^d d\theta,$$

where W was defined in (16), should be discussed. Again, an application of the dominated convergence theorem leads to the simpler condition

$$\lim_{\rho\to+\infty} W(F^{-1}(\rho,\theta))\rho^2 = 0,$$

which is true, since W decays with an exponential rate.

References

1. Adler, S.: Over-relaxation method for the Monte Carlo evaluation of the partition function for multiquadratic actions. Phys. Rev. D **23**, 2901–2904 (1981)
2. Albert, J., Chib, S.: Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association **88, 422**, 669–679 (1993)
3. Ardia, D.: Financial risk management with bayesian estimation of GARCH models: Theory and applications. In: Lecture Notes in Economics and Mathematical Systems 612. Springer-Verlag (2008)
4. Assaraf, R., Caffarel, M.: Zero-Variance principle for Monte Carlo algorithms. Physical Review letters **83, 23**, 4682–4685 (1999)
5. Assaraf, R., Caffarel, M.: Zero-variance zero-bias principle for observables in quantum Monte Carlo: Application to forces. The Journal of Chemical Physics **119, 20**, 10,536–10,552 (2003)
6. Barone, P., Frigessi, A.: Improving stochastic relaxation for Gaussian random fields. Probability in the Engineering and Informational Sciences **4**, 369–389 (1989)
7. Barone, P., Sebastiani, G., Stander, J.: General over-relaxation Markov chain Monte Carlo algorithms for Gaussian densities. Statistics & Probability Letters **52,2**, 115–124 (2001)
8. Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics **31, 3**, 307–327 (1986)
9. Bollerslev, T.: Glossary to ARCH (GARCH). In: Volatility and Time Series Econometrics, Essays in Honor of Robert Engle, Edited by Tim Bollerslev, Jeffrey Russell and Mark Watson. Oxford University Press, Oxford, UK (2010)
10. Brewer, M., Aitken, C., Talbot, M.: A comparison of hybrid strategies for Gibbs sampling in mixed graphical models. Computational Statistics **21**, 343–365 (1996)
11. Brooks, S., Gelman, A.: Some issues in monitoring convergence of iterative simulations. Computing Science and Statistics (1998)
12. Craiu, R., Lemeieux, C.: Acceleration of the multiple-try Metropolis algorithm using antithetic and stratified sampling. Journal Statistics and Computing **17, 2**, 109–120 (2007)

13. Craiu, R., Meng, X.: Multiprocess parallel antithetic coupling for backward and forward Markov chain Monte Carlo. The Annals of Statistics **33, 2**, 661–697 (2005)
14. Dellaportas, P., Kontoyiannis, I.: Control variates for estimation based on reversible Markov chain Monte Carlo samplers. Journal of the Royal Statistical Society, Series B. **74(1)**, 133–161 (2012)
15. Diaconis, P., Holmes, S., Neal, R.F.: Analysis of a nonreversible Markov chain sampler. Ann. Appl. Probab. **10,3**, 726–752 (2000)
16. Duane, S., Kennedy, A., Pendleton, B., Roweth, D.: Hybrid Monte Carlo. Physics Letters B **195**, 216–222 (2010)
17. Flury, B., Riedwyl, H.: Multivariate Statistics. Chapman and Hall (1988)
18. Fort, G., Moulines, E., Roberts, G., Rosenthal, S.: On the geometric ergodicity of hybrid samplers. Journal of Applied Probability **40, 1**, 123–146 (2003)
19. Gelfand, A., Smith, A.: Sampling-based approaches to calculating marginal densities. J. American Statistical Association **85**, 398–409 (1990)
20. Girolami, M., Calderhead, B.: Riemannian manifold Langevin and Hamiltonian Monte Carlo methods. J. R. Statist. Soc. B **73, 2**, 1–37 (2011)
21. Green, P., Han, X.: Metropolis methods, Gaussian proposals, and antithetic variables. In: P. Barone, A. Frigessi, M. Piccioni (eds.) Lecture Notes in Statistics, Stochastic Methods and Algorithms in Image Analysis, vol. 74, pp. 142–164. Springer Verlag (1992)
22. Green, P.J., Mira, A.: Delayed rejection in reversible jump Metropolis-Hastings. Biometrika **88**, 1035–1053 (2001)
23. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
24. Henderson, S.: Variance reduction via an approximating Markov process. Ph.D. thesis, Department of Operations Research, Stanford University, Stanford, CA (1997)
25. Henderson, S., Glynn, P.: Approximating martingales for variance reduction in Markov process simulation. Math. Oper. Res. **27, 2**, 253–271 (2002)
26. Higdon, D.: Auxiliary variable methods for Markov chain Monte Carlo with applications. Journal of the American Statistical Association **93**, 585–595 (1998)
27. Ishwaran, H.: Applications of hybrid Monte Carlo to Bayesian generalized linear models: quasicomplete separation and neural networks. J. Comp. Graph. Statist. **8**, 779–799 (1999)
28. Leisen, F., Dalla Valle, L.: A new multinomial model and a zero variance estimation. Communications in Statistics - Simulation and Computation **39(4)**, 846–859 (2010)
29. Linnik, Y.V.: An information-theoretic proof of the central limit theorem with Lindeberg conditions. Theory of Probability and its Applications **4**, 288–299 (1959)
30. Loh, W.: Methods of control variates for discrete event simulation. Ph.D. thesis, Department of Operations Research, Stanford University, Stanford, CA (1994)
31. Marin, J.M., Robert, C.: Bayesian Core: A Practical Approach to Computational Bayesian Statistics. Springer (2007)
32. Mira, A., Geyer, C.J.: On reversible Markov chains. Fields Inst. Communic.: Monte Carlo Methods **26**, 93–108 (2000)
33. Mira, A., Möller, J., Roberts, G.O.: Perfect slice samplers. Journal of the Royal Statistical Soc. Ser. B **63, 3**, 593–606 (2001)
34. Mira, A., Tierney, L.: Efficiency and convergence properties of slice samplers. Scandinavian Journal of Statistics **29**, 1–12 (2002)
35. Neal, R.: An improved acceptance procedure for the hybrid Monte Carlo algorithm. Journal of Computational Physics **111**, 194–203 (1994)
36. Neal, R.M.: Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation. Tech. rep., Learning in Graphical Models (1995)
37. Nelson, B.: Batch size effects on the efficiency of control variates in simulation. European Journal of Operational Research **2(27)**, 184–196 (1989)
38. Philippe, A., Robert, C.: Riemann sums for MCMC estimation and convergence monitoring. Statistics and Computing **11**, 103–105 (2001)
39. Ripley, B.: Stochastic Simulation. John Wiley & Sons (1987)
40. Rockafellar, R.: Convex analysis, pp. 264–265. Princeton University Press (1970)
41. So, M.K.P.: Bayesian analysis of nonlinear and non-Gaussian state space models via multiple-try sampling methods. Statistics and Computing **16**, 125–141 (2006)

42. Speckman P.L. Lee, J., Sun, D.: Existence of the mle and propriety of posteriors for a general multinomial choice model. Statistica Sinica **19**, 731–748 (2009)
43. Swendsen, R., Wang, J.: Non universal critical dynamics in Monte Carlo simulations. Phys. Rev. Lett. **58**, 86–88 (1987)
44. Tierney, L.: Markov chains for exploring posterior distributions. Annals of Statistics **22**, 1701–1762 (1994)
45. Tierney, L., Mira, A.: Some adaptive Monte Carlo methods for Bayesian inference. Statistics in Medicine **18**, 2507–2515 (1999)
46. Van Dyk, D., Meng, X.: The art of data augmentation. Journal of Computational and Graphical Statistics **10**, 1–50 (2001)