Stat Comput (2013) 23:653–662
DOI 10.1007/s11222-012-9344-6

Zero variance Markov chain Monte Carlo for Bayesian estimators

Antonietta Mira · Reza Solgi · Daniele Imparato

Received: 11 January 2012 / Accepted: 11 July 2012 / Published online: 28 July 2012

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

Electronic supplementary material The online version of this article (doi:10.1007/s11222-012-9344-6) contains supplementary material, which is available to authorized users.

A. Mira · R. Solgi
Swiss Finance Institute, University of Lugano, via Buffi 13, 6904
Lugano, Switzerland

A. Mira
e-mail: antonietta.mira@usi.ch

R. Solgi
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

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, $\mathbf{x}^i, i = 1, \dots, N$, collected along the sample path of an ergodic Markov chain having π (normalized) as its unique stationary and limiting distribution. In this context, the MCMC estimate of μ_f is $\hat{\mu}_f = \sum_{i=1}^N f(\mathbf{x}^i)/N$.

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 rest of this section we briefly explain the zero-variance (ZV) principle introduced in Assaraf and Caffarel (1999), Assaraf and Caffarel (2003): 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 kinetic 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 Sect. 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:

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

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 (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 et al. 2001); by hybrid Monte Carlo (Duane et al. 2010; Neal 1994; Brewer et al. 1996; Fort et al. 2003; Ishwaran 1999); by exploiting nonreversible Markov chains (Diaconis et al. 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 et al. 2001; Mira and Tierney 2002). Up to our knowledge, the only other research line that uses control variates in MCMC estimation follows the PhD thesis by Henderson (1997) and has its most recent development in Dellaportas and Kontoyiannis (2012). 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 $U(\mathbf{x}) := g(\mathbf{x}) - \mathbb{E}[g(X^{n+1})|X^n = \mathbf{x}]$ 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 Dellaportas and Kontoyiannis (2012), the use of this type of control variates is further explored in the setting of reversible Markov chains.

In Assaraf and Caffarel (1999) and Assaraf and Caffarel (2003) unbiasedness and existence of a central limit theorem (CLT) for the ZV estimator are not discussed, neither in Leisen and Dalla Valle (2010), 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

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

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:

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

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 Dellaportas and Kontoyiannis (2012), 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

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

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 Assaraf and Caffarel (1999): 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 (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 Sect. 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:

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

where $\mathbf{z}=-\frac{1}{2}\nabla\ln\pi(\mathbf{x})$, $\nabla=(\frac{\partial}{\partial x_1},\dots,\frac{\partial}{\partial x_d})$ 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 Brooks and Gelman (1998) and Philippe and Robert (2001), 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 $P(\mathbf{x})=\sum_{j=1}^{d}a_jx_j$ (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:

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

For a more general approach to the choice of coefficients using control variates, reference should be made to Nelson (1989) and Loh (1994). We anticipate that conditions under which the ZV-MCMC estimator obeys a CLT (Sect. 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 order d^2. This makes the use of second order polynomials computationally expensive when dealing with high-dimensional sampling spaces.

5 Unbiasedness and central limit theorem

As remarked in Sect. 1, condition (1) may not be sufficient to ensure unbiasedness of the estimator when the Schrödinger operator (7) is used. In this section general conditions 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 $\mathbf{x} \in \partial\Omega$, $j = 1, \dots, d$.*

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

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

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*

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

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 Assaraf and Caffarel (1999) 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 $P(x) = a_0 + a_1x + a_2x^2$ 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 (Tierney 1994). 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 Sect. 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 $k \in \{0, \deg P - 1\}$.*

[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 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

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

The quantity m_j is known as the Linnik functional (if considered as a function of the target distribution, $I(\pi)$) since it was introduced by Linnik (1959). The quantity m_j is also interpretable as the Fisher information of a location family in a frequentest 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 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 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 $(k = 1, 2, \ldots, d)$. The Bayesian analysis of this problem is discussed in Marin and Robert (2007). In order to find the optimal vector of parameters a_k of the trial functions, a short Gibbs sampler, following (Albert and Chib 1993), (of length 2000, after 1000 burn in steps) is run, and the optimal coefficients are estimated: $\hat{\mathbf{a}}_k = -\hat{\Sigma}_{\mathbf{zz}}^{-1}\hat{\sigma}(\mathbf{z}, \beta_k)$. 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

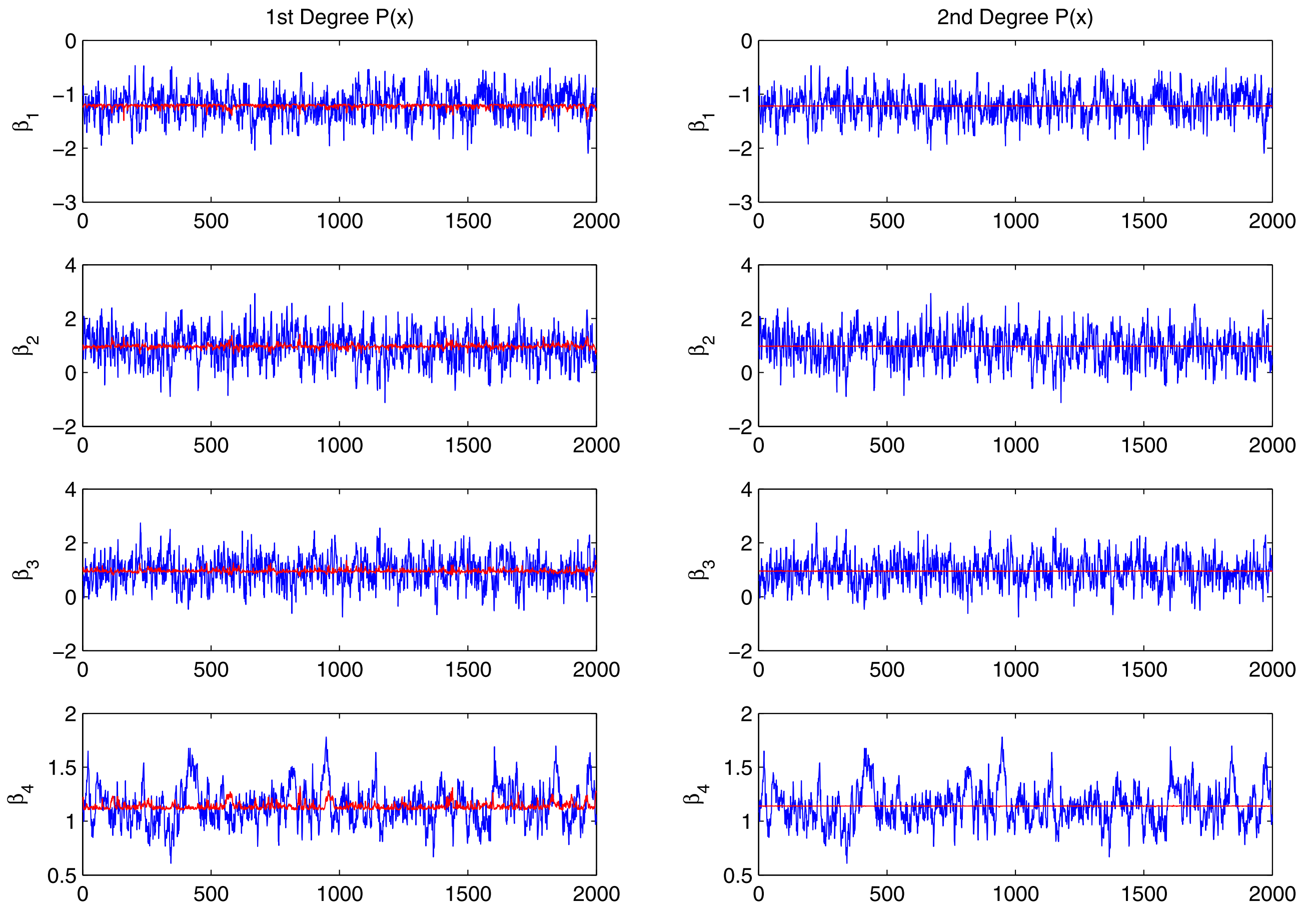

Fig. 1 Ordinary MCMC (*blue*) and ZV-MCMC (*red*) for probit model: rows are parameters, columns are degree polynomials (Color figure online)

(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.

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 ($k = 1, 2, \dots, d$). 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.

6.3 GARCH model

Generalized autoregressive conditional heteroskedasticity (GARCH) models (Bollerslev 1986) 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 $r(t) := [S(t) - S(t -$

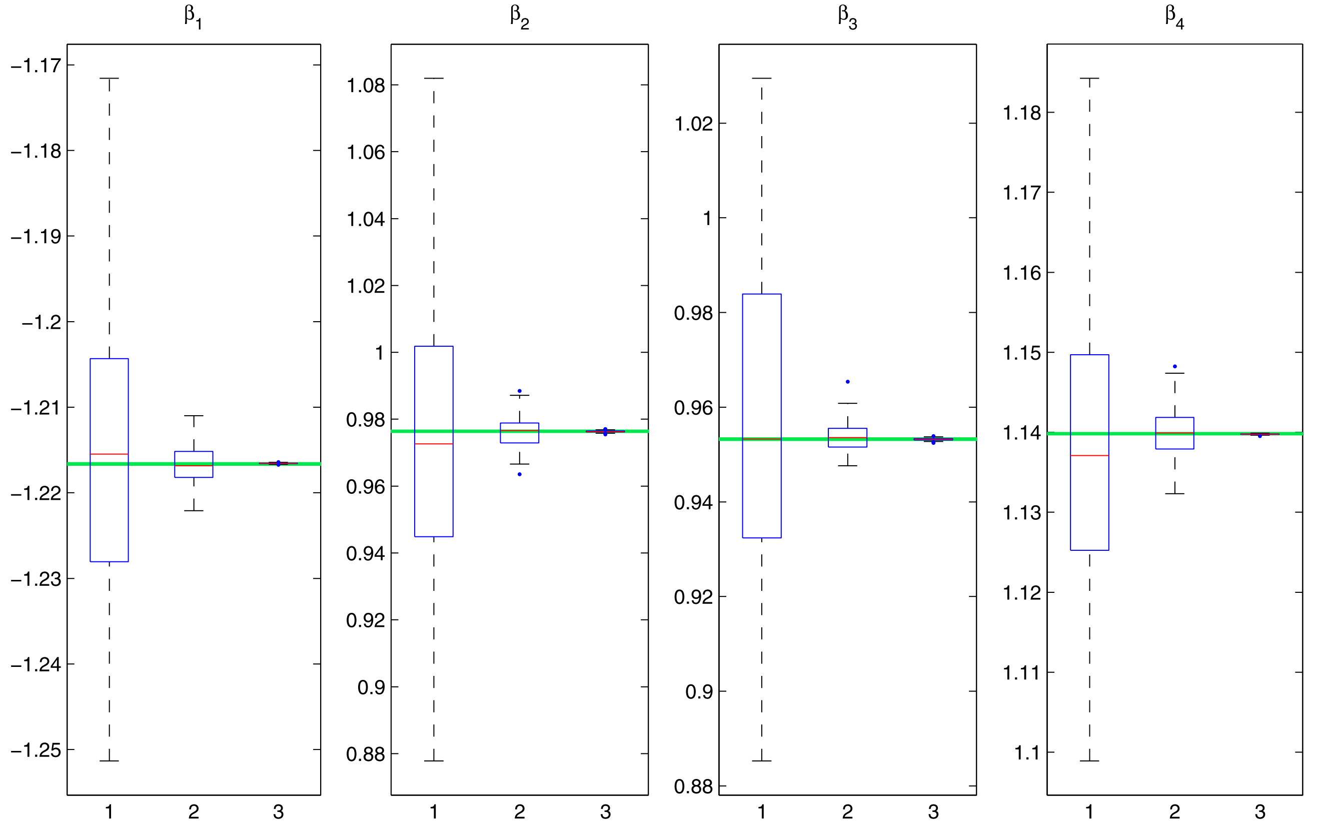

Fig. 2 Boxplots of ordinary MCMC estimates (1) and ZV-MCMC estimates (2 and 3) for the probit model, along with the 95 % confidence region obtained by an ordinary MCMC of length 10^8 (*green regions*) (Color figure online)

$1)]/S(t-1) \approx \ln(S(t)/S(t-1))$. 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 (Ardia 2008) 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 10,000) 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 10,000 (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.

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

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.

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, Bollerslev 2010), provided it is possible to analyt-

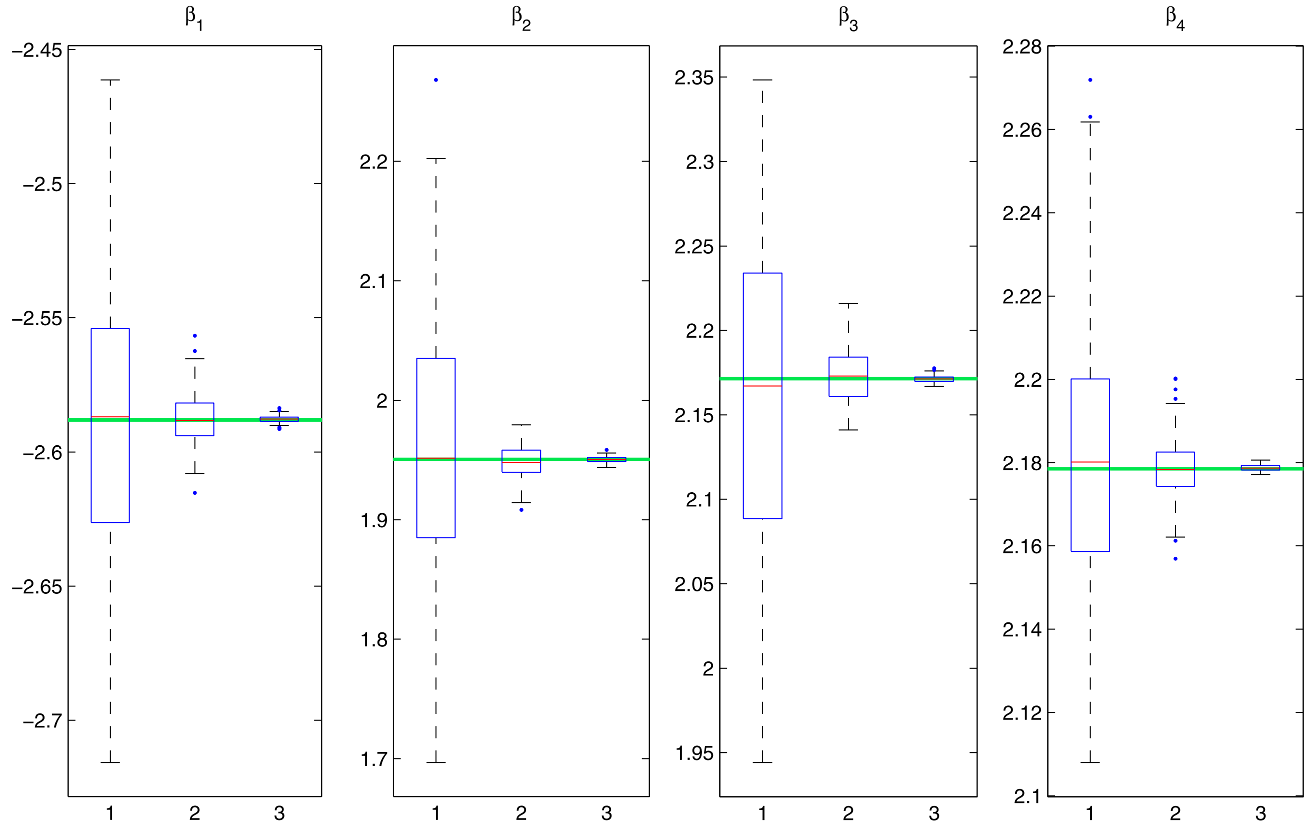

Fig. 3 Box-plots of ordinary MCMC estimates (1) and ZV-MCMC estimates (2 and 3) for the logit model, along with the 95 % confidence region obtained by an ordinary MCMC of length 10^8 (*green regions*) (Color figure online)

ically 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 Hastings (1970) first and Gelfand and Smith (1990) later on.

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. 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 Girolami and Calderhead (2011) 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 Girolami and Calderhead 2011) 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 Dellaportas and Kontoyiannis (2012) 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 vari-

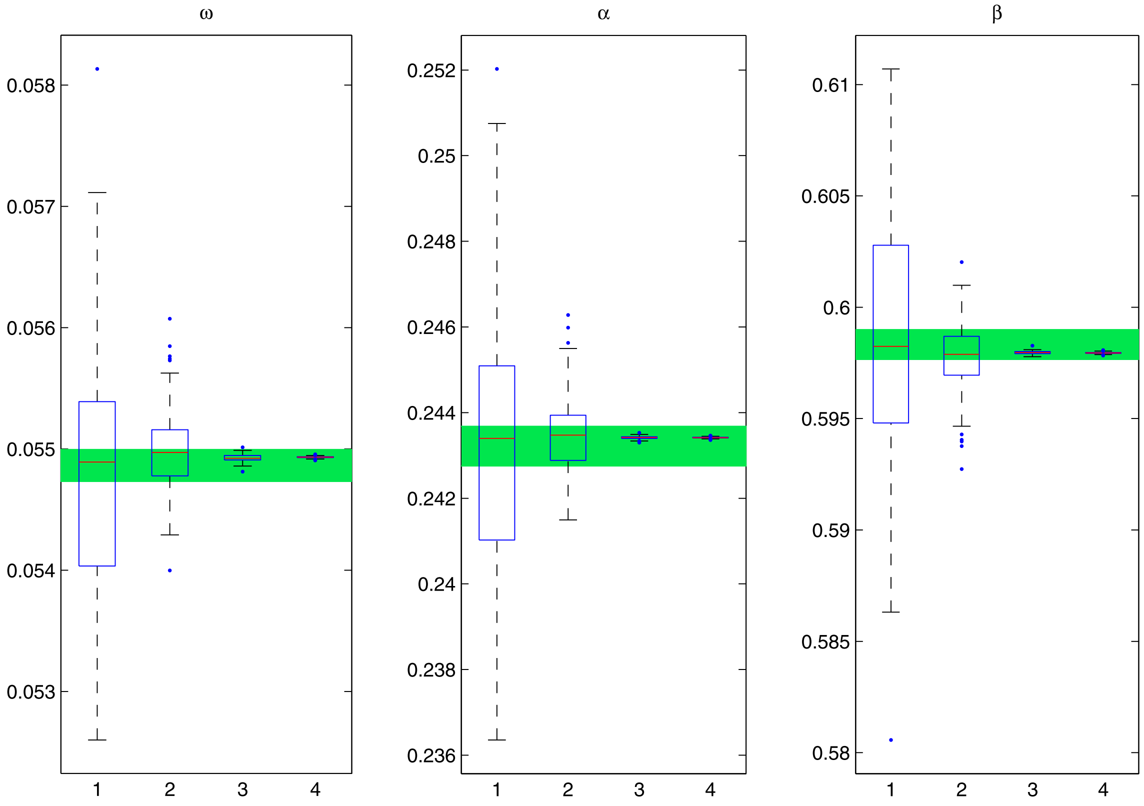

Fig. 4 Boxplots of ordinary MCMC estimates (1) and ZV-MCMC estimates (2, 3 and 4) for the GARCH model, along with the 95 % confidence region obtained by an ordinary MCMC of length 10^7 (*green regions*) (Color figure online)

ance reduction obtained in the examples in Sect. 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 Assaraf and Caffarel (2003), 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 Sect. 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.

8 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 Sect. 5 are reported and verified for the three examples.

Acknowledgements 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.

References

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

Albert, J., Chib, S.: Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88**(422), 669–679 (1993)

Ardia, D.: Financial Risk Management with Bayesian Estimation of GARCH Models: Theory and Applications. Lecture Notes in Economics and Mathematical Systems, vol. 612. Springer, Berlin (2008)

Assaraf, R., Caffarel, M.: Zero-Variance principle for Monte Carlo algorithms. Phys. Rev. Lett. **83**(23), 4682–4685 (1999)

Assaraf, R., Caffarel, M.: Zero-variance zero-bias principle for observables in quantum Monte Carlo: application to forces. J. Chem. Phys. **119**(20), 10,536–10,552 (2003)

Barone, P., Frigessi, A.: Improving stochastic relaxation for Gaussian random fields. Probab. Eng. Inf. Sci. **4**, 369–389 (1989)

Barone, P., Sebastiani, G., Stander, J.: General over-relaxation Markov chain Monte Carlo algorithms for Gaussian densities. Stat. Probab. Lett. **52**(2), 115–124 (2001)

Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. J. Econom. **31**(3), 307–327 (1986)

Bollerslev, T.: Glossary to ARCH (GARCH). In: Bollerslev, T., Russell, J., Watson, M. (eds.) Volatility and Time Series Econometrics, Essays in Honor of Robert Engle. Oxford University Press, Oxford (2010)

Brewer, M., Aitken, C., Talbot, M.: A comparison of hybrid strategies for Gibbs sampling in mixed graphical models. Comput. Stat. **21**, 343–365 (1996)

Brooks, S., Gelman, A.: Some issues in monitoring convergence of iterative simulations. In: Computing Science and Statistics (1998)

Craiu, R., Lemeieux, C.: Acceleration of the multiple-try Metropolis algorithm using antithetic and stratified sampling. J. Stat. Comput. **17**(2), 109–120 (2007)

Craiu, R., Meng, X.: Multiprocess parallel antithetic coupling for backward and forward Markov chain Monte Carlo. Ann. Stat. **33**(2), 661–697 (2005)

Dellaportas, P., Kontoyiannis, I.: Control variates for estimation based on reversible Markov chain Monte Carlo samplers. J. R. Stat. Soc. B **74**(1), 133–161 (2012)

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

Duane, S., Kennedy, A., Pendleton, B., Roweth, D.: Hybrid Monte Carlo Phys. Lett. B **195**, 216–222 (2010)

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

Fort, G., Moulines, E., Roberts, G., Rosenthal, S.: On the geometric ergodicity of hybrid samplers. J. Appl. Probab. **40**(1), 123–146 (2003)

Gelfand, A., Smith, A.: Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. **85**, 398–409 (1990)

Girolami, M., Calderhead, B.: Riemannian manifold Langevin and Hamiltonian Monte Carlo methods. J. R. Stat. Soc. B **73**(2), 1–37 (2011)

Green, P., Han, X.: Metropolis methods, Gaussian proposals, and antithetic variables. In: Barone, P., Frigessi, A., Piccioni, M. (eds.) Lecture Notes in Statistics, Stochastic Methods and Algorithms in Image Analysis, vol. 74, pp. 142–164. Springer, Berlin (1992)

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

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

Henderson, S.: Variance reduction via an approximating Markov process. Ph.D. thesis, Department of Operations Research, Stanford University, Stanford, CA (1997)

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

Higdon, D.: Auxiliary variable methods for Markov chain Monte Carlo with applications. J. Am. Stat. Assoc. **93**, 585–595 (1998)

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

Leisen, F., Dalla Valle, L.: A new multinomial model and a zero variance estimation. Commun. Stat., Simul. Comput. **39**(4), 846–859 (2010)

Linnik, Y.V.: An information-theoretic proof of the central limit theorem with Lindeberg conditions. Theory Probab. Appl. **4**, 288–299 (1959)

Loh, W.: Methods of control variates for discrete event simulation. Ph.D. thesis, Department of Operations Research, Stanford University, Stanford, CA (1994)

Marin, J.M., Robert, C.: Bayesian Core: A Practical Approach to Computational Bayesian Statistics. Springer, Berlin (2007)

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

Mira, A., Möller, J., Roberts, G.O.: Perfect slice samplers. J. R. Stat. Soc. B **63**(3), 593–606 (2001)

Mira, A., Tierney, L.: Efficiency and convergence properties of slice samplers. Scand. J. Stat. **29**, 1–12 (2002)

Neal, R.: An improved acceptance procedure for the hybrid Monte Carlo algorithm. J. Comput. Phys. **111**, 194–203 (1994)

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

Nelson, B.: Batch size effects on the efficiency of control variates in simulation. Eur. J. Oper. Res. **2**(27), 184–196 (1989)

Philippe, A., Robert, C.: Riemann sums for MCMC estimation and convergence monitoring. Stat. Comput. **11**, 103–105 (2001)

Ripley, B.: Stochastic Simulation. Wiley, New York (1987)

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

Swendsen, R., Wang, J.: Non universal critical dynamics in Monte Carlo simulations. Phys. Rev. Lett. **58**, 86–88 (1987)

Tierney, L.: Markov chains for exploring posterior distributions. Ann. Stat. **22**, 1701–1762 (1994)

Tierney, L., Mira, A.: Some adaptive Monte Carlo methods for Bayesian inference. Stat. Med. **18**, 2507–2515 (1999)

Van Dyk, D., Meng, X.: The art of data augmentation. J. Comput. Graph. Stat. **10**, 1–50 (2001)