D. Bressanini, F. Laisen, A. Mira, P. Tenconi

Zero variance in Markov chain Monte Carlo with an application to credit risk estimation

2008/4

UNIVERSITÀ DELL'INSUBRIA
FACOLTÀ DI ECONOMIA

http://eco.uninsubria.it

Printed in Italy in April 2008
Università degli Studi dell'Insubria
Via Monte Generoso, 71, 21100 Varese, Italy

Zero variance in Markov chain Monte Carlo with an application to credit risk estimation

Dario Bressanini[*], Fabrizio Leisen[†], Antonietta Mira[‡], Paolo Tenconi [§]

Abstract

We propose a general purpose variance reduction technique for Markov Chain Monte Carlo estimators based on the Zero-Variance principle introduced in the physics literature by Assaraf and Caffarel (1999). The potential of the new idea is illustrated with some toy examples and a real application to Bayesian inference for credit risk estimation.

Keywords: Markov chain Monte Carlo, Metropolis-Hastings algorithm, Variance reduction, Zero-Variance principle.

1 Main idea

We are interested in estimating the expected value of a function f with respect to a, possibly unnormalized, probability distribution π:

$$\mu_f = \frac{\int f(x)\pi(x)dx}{\int \pi(x)dx}. \tag{1}$$

Markov chain Monte Carlo methods (MCMC, Metropolis et al. 1953, Hastings 1970, Tierney, 1994), estimate integrals using a large but finite set of sample points, $x^i, i = 1, \cdots, N$,

[*] Dip. di Sc. Chimiche Fisiche e Matematiche, Università dell'Insubria, Como, Italy. (email: dario.bressanini@uninsubria.it)

[†] Dip. di Economia Università dell'Insubria, Via Monte Generoso 71, 21100 Varese, Italy. (email: fleisen@eco.uninsubria.it)

[‡] Dip. di Economia Università dell'Insubria, Via Monte Generoso 71, 21100 Varese, Italy. (email: amira@eco.uninsubria.it)

[§] Dip. di Economia Università dell'Insubria, Via Monte Generoso 71, 21100 Varese, Italy. (email: ptenconi@eco.uninsubria.it)

collected along the sample path of an ergodic Markov chain, P, having π (normalized) as its unique stationary and limiting distribution:

$$\hat{\mu}_f = \frac{1}{N} \sum_{i=1}^{N} f(x^i). \tag{2}$$

We have that

$$\mu_f = \hat{\mu}_f + \Delta\mu_f$$

where $\Delta\mu_f$ is the statistical error associated with the fact that the length of the simulated Markov chain path, N, is finite. For large enough N, standard statistical arguments lead to the following expression of the error:

$$\Delta\mu_f = K_f \frac{\sigma_f}{\sqrt{N}}$$

where the constant K_f is proportional to the amount of correlation along the sampled chain and σ_f is the standard deviation of f under π (assumed to be finite).

Recent literature (Peskun, 1973; Liu, 1996; Tierney, 1998; Tierney and Mira, 1999; Mira and Geyer, 2000; Green and Mira, 2001), aimed at reducing the statistical MCMC error, $\Delta\mu_f$, by reducing the correlation along the Markov chain, that is, by reducing K_f.

In this paper we suggest instead to reduce the 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 smaller (this is a standard variance reduction technique used in Monte Carlo simulation, see Ripley, 1987). To define $\tilde{f}$, an operator, H, and a trial function, ϕ, are introduced. We require that H is Hermitian (symmetric for finite state spaces, and real in all practical applications) and

$$\int H(x, y)\sqrt{\pi(y)}dy = 0. \tag{3}$$

The trial function $\phi(x)$ is a rather arbitrary function which is only required to be integrable. We define the renormalized function to be

$$\tilde{f}(x) = f(x) + \frac{\int H(x, y)\phi(y)dy}{\sqrt{\pi(x)}} = f(x) + \Delta f(x). \tag{4}$$

As a consequence of (1) and (3) we have that

$$\mu_f = \mu_{\tilde{f}} \tag{5}$$

that is, both functions f and $\tilde{f}$ can be used to estimate the desired quantity via Monte Carlo or MCMC simulation. However, the statistical error of the resulting estimator can be very different. The optimal choice for (H, ϕ) can be obtained by imposing that $\tilde{f}$ is constant and equal to its average, that is, by requiring

$$\sigma_{\tilde{f}} = 0,$$

which is equivalent to require that

$$\tilde{f} = \mu_f.$$

The latter, together with (4), leads to the fundamental equation:

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

In most practical applications equation (6) cannot be solved exactly, still, we propose to find an approximate solution in the following way. First choose H verifying (3) (in Section 2 we will suggest two general recipes to construct H). Second, parametrize ϕ and optimally choose the parameters by minimizing $\sigma_{\tilde{f}}$ over a finite set of points generated according to the Markov chain P. Finally, a much longer MCMC simulation is performed using $\hat{\mu}_{\tilde{f}}$ instead of $\hat{\mu}_f$ as the estimator. Note that the proposed approach can be used to obtain variance reduction also in Monte Carlo simulation if we can get i.i.d. draws from the target distribution π.

2 Choice of H

2.1 Discrete case

Denote with $P(x, y)$ a transition matrix reversible with respect to π (we identify a Markov chain with the corresponding transition matrix of kernel):

$$\pi(x)P(x, y) = \pi(y)P(y, x), \qquad \forall x, y.$$

The following choice of H

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

satisfies the requirements, where $\delta(x - y)$ is the Dirac delta function: $\delta(x - y) = 1$ if $x = y$ and zero otherwise. With this choice of H, letting $\tilde{\phi} = \frac{\phi}{\sqrt{\pi}}$, equation (4) becomes:

$$\tilde{f}(x) = f(x) - \int P(x, y)[\tilde{\phi}(x) - \tilde{\phi}(y)]dy. \tag{7}$$

The main difficulty with (7) is the evaluation of the integral.

2.2 Continuous case

If $x \in \Re^d$ we can consider the operator:

$$H = -\frac{1}{2}\sum_{i=1}^{d}\frac{\partial^2}{\partial x_i^2} + V(x) \tag{8}$$

where $V(x)$ is constructed to fulfill equation (3):

$$V(x) = \frac{1}{2\sqrt{\pi(x)}}\sum_{i=1}^{d}\frac{\partial^2\sqrt{\pi(x)}}{\partial x_i^2}. \tag{9}$$

In this setting we have that

$$\tilde{f}(x) = f(x) + \frac{H\phi(x)}{\sqrt{\pi(x)}}. \tag{10}$$

This is the function we will use in the examples considered in the following. To obtain the first and second order derivatives we used the R function "hessian" from the library "numDeriv" which evaluates an approximate Hessian of a scalar function using finite differences. Note that for calculating $\tilde{f}$ with the operator (8) the normalizing constant of $\pi(x)$ is not needed!

3 Choice of ϕ

The optimal choice of ϕ is the *exact solution* of the fundamental equation. In real applications, typically, only *approximate solutions*, obtained by minimizing $\sigma_{\tilde{f}}$, are available. The particular form of ϕ is very dependent on the problem at hand, that is on π, and on f. However an important point to notice is that, if we parametrize ϕ in terms of $c = \int \phi(x)dx$ and then minimize $\sigma_{\tilde{f}}$ with respect to c, the optimal choice of c is

$$c = -\frac{[E_\pi(f(x)\Delta f(x))]^2}{E_\pi(\Delta f(x))^2}$$

and, for this value of the parameter, from (4) we obtain

$$\sigma^2_{\tilde{f}} = \sigma^2_f - \frac{[E_\pi(f(x)\Delta f(x))]^2}{E_\pi(\Delta f(x))^2}. \tag{11}$$

Since the correction factor in (11) that leads from σ^2_f to $\sigma^2_{\tilde{f}}$ is always negative, regardless of the choice of ϕ, a variance reduction in the MCMC estimator is obtained by replacing f with $\tilde{f}$ in (2).

4 Examples of variance reduction in Monte Carlo case

In this section we present a few toy examples to demonstrate the power of the proposed technique. In particular we consider as target distributions:

1. Univariate and bivariate Gaussian distributions,
2. Univariate and bivariate standard Student-T distributions.

The functions of interest, f, are:

- $f(x) = x$ and $f(x) = x^2$ in the univariate case,
- $f(x_1, x_2) = x_1$, $f(x_1, x_2) = x_1^2$ and $f(x_1, x_2) = x_1 x_2$ in the bivariate case.

In the results presented we sample $T = 150$ values from π.

4.1 Univariate Gaussian distribution

Consider as target a normal distribution $N(\mu, \sigma^2)$ with non-normalized density $\pi(x) = \exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})$. In this case the theoretical functions ϕ that solve (6) are respectively for $f_1(x) = x$ and $f_2(x) = x^2$:

$$\phi_1(x) = (-2\sigma^2 x)\exp\left\{-\frac{1}{4}\frac{(x-\mu)^2}{\sigma^2}\right\};$$

and

$$\phi_2(x) = (-\sigma^2 x^2 - 2\mu\sigma^2 x)\exp\left\{-\frac{1}{4}\frac{(x-\mu)^2}{\sigma^2}\right\}.$$

In Table 1 we show simulation results for $f_1(x) = x$, $f_2(x) = x^2$ and the associated $\tilde{f}_1(x)$ and $\tilde{f}_2(x)$. Despite of the small sample, a great reduction in variance is achieved and the final estimated variance is nearly zero.

4.2 Univariate Student-T distribution

In this section we proceed as in the previous one but taking the univariate Student-T distribution with g degrees of freedom, $T(g)$, as the target. Suppose that $g > 2$. In this case the

Table 1: $N(1,2)$, $f_1(x) = x$, $f_2(x) = x^2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$
$\hat{\mu}_f$	0.912	1	2.824	3
$\hat{\sigma}^2_f$	2.013	2.28e-22	9.377	3.53e-21

Table 2: Univariate Student-T with df=5, $f_1(x) = x$, $f_2(x) = x^2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$
$\hat{\mu}_f$	-0.271	1.65e-12	1.834	1.666
$\hat{\sigma}^2_f$	1.778	5.19e-22	20.536	1.32e-23

Table 3: Bivariate Normal, $\underline{\mu}$=(2,1), (σ_1^2, σ_2^2)=(4,1), ρ=0.6 $f_1 = x_1$, $f_2 = x_1^2$, $f_3 = x_1 x_2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$	f_3	$\tilde{f}_3$
$\hat{\mu}_f$	1.703	2	6.518	8	2.582	3.2
$\hat{\sigma}^2_f$	3.654	3.48e-20	7.199	4.63e-18	13.053	5.76e-20

Table 4: Bivariate Student-T, df=7, $f_1 = x_1$, $f_2 = x_1^2$, $f_3 = x_1 x_2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$	f_3	$\tilde{f}_3$
$\hat{\mu}_f$	1.703	1.31e-11	6.518	1.4	2.582	-1.06e-12
$\hat{\sigma}^2_f$	3.654	8.85e-21	71.992	4.07e-19	13.053	2.03e-22

non-normalized density is $\pi(x) = \left(1+\frac{x^2}{g}\right)^{-\frac{g+1}{2}}$ and the theoretical functions ϕ that solve (6) are, respectively, for $f_1(x) = x$ and $f_2(x) = x^2$:

$$\phi_1(x) = \left(\frac{2}{3}\frac{1}{1-g}x^3 + 2\frac{g}{1-g}x\right)\left(1+\frac{x^2}{g}\right)^{-\frac{g+1}{4}}$$

and

$$\phi_2(x) = \left(\frac{1}{2}\frac{1}{2-g}x^4 + \frac{g}{2-g}x^2\right)\left(1+\frac{x^2}{g}\right)^{-\frac{g+1}{4}}.$$

Also in this case the simulation results displayed in Table 2 show an estimated variance close to zero.

4.3 Bivariate Gaussian

We consider here a two dimensional vector, $\underline{x} = (x_1, x_2)$, having a bivariate normal distribution with mean vector equal to $\underline{\mu} = (\mu_1, \mu_2)$, standard deviations equal to $\underline{\sigma} = (\sigma_1, \sigma_2)$, and correlation coefficient ρ. The theoretical ϕ functions for $f_1(x) = x_1$, $f_2(x) = x_1^2$ and $f_3(x) = x_1 x_2$, are, respectively:

$$\begin{aligned}
\phi_1(x_1, x_2) &= \left(-2\sigma_1^2 x_1 - 2\rho\sigma_1\sigma_2 x_2\right)\sqrt{\pi(x_1, x_2)};\\
\phi_2(x_1, x_2) &= \left\{\left[-\rho^2\frac{\sigma_1^4}{\sigma_1^2+\sigma_2^2} - \sigma_1^2\left(1-\rho^2\right)\right]x_1^2 + \left[-\rho^2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right]x_2^2 + \left[-\rho\frac{\sigma_1^3\sigma_2}{\sigma_1^2+\sigma_2^2}\right]x_1x_2\right.\\
&\quad + \left[-2\mu_1\sigma_1^2 + 2\rho\frac{\sigma_1^3\sigma_2}{\sigma_1^2+\sigma_2^2}\mu_2 - 2\rho^2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\mu_1\right]x_1\\
&\quad \left. + \left[-2\sigma_1\sigma_2\mu_1\rho - 2\rho\frac{\sigma_1\sigma_2^3}{\sigma_1^2+\sigma_2^2}\mu_1 + 2\rho^2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\mu_2\right]x_2\right\}\sqrt{\pi(x_1, x_2)};\\
\phi_3(x_1, x_2) &= \left\{\left[-\rho\frac{\sigma_1^3\sigma_2}{\sigma_1^2+\sigma_2^2}\right]x_1^2 + \left[-\rho\frac{\sigma_1\sigma_2^3}{\sigma_1^2+\sigma_2^2}\right]x_2^2 + \left[-2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right]x_1x_2\right.\\
&\quad \left. + \left[-2\frac{\sigma_1^4}{\sigma_1^2+\sigma_2^2}\mu_2 - 2\rho\frac{\sigma_1\sigma_2^3}{\sigma_1^2+\sigma_2^2}\mu_1\right]x_1 + \left[-2\frac{\sigma_2^4}{\sigma_1^2+\sigma_2^2}\mu_1 - 2\rho\frac{\sigma_1^3\sigma_2}{\sigma_1^2+\sigma_2^2}\mu_2\right]x_2\right\}\sqrt{\pi(x_1, x_2)}.
\end{aligned}$$

We consider first a standard bivariate Gaussian target and then move on to the case where $\underline{\mu} = (2,1), \underline{\sigma} = (2,1)$ and $\rho = 0.6$. In Table 3 we report the results obtained and we have again a near zero variance.

4.4 Bivariate Student-T

We conclude the Monte Carlo simulation with theoretical knowledge of the ϕ's functions, with the simulation of a bivariate T-Student distribution. The theoretical ϕ functions for $f_1(x) = x_1$, $f_2(x) = x_1^2$ and $f_3(x) = x_1 x_2$, are, respectively:

$$
\begin{aligned}
\phi_1(x_1, x_2) &= \left(\frac{2}{2-3g} x_1^3 + \frac{2}{2-3g} x_1 x_2^2 + \frac{6g}{2-3g} x_1 \right) \sqrt{\pi(x_1, x_2)};\\
\phi_2(x_1, x_2) &= \left\{ \left[\frac{1}{4} \frac{3-2g}{(2-g)(1-g)} \right] x_1^4 + \left[-\frac{1}{4} \frac{1}{(2-g)(1-g)} \right] x_2^4 + \left[\frac{1}{2} \frac{1}{2-g} \right] x_1^2 x_2^2 \right.\\
&\quad \left. + \left[\frac{1}{2} \frac{g(3-2g)}{(2-g)(1-g)} \right] x_1^2 + \left[-\frac{1}{2} \frac{g}{(2-g)(1-g)} \right] x_2^2 \right\} \sqrt{\pi(x_1, x_2)};\\
\phi_3(x_1, x_2) &= \left(\frac{1}{2} \frac{1}{1-g} x_1^3 x_2 + \frac{1}{2} \frac{1}{1-g} x_1 x_2^3 + \frac{g}{1-g} x_1 x_2 \right) \sqrt{\pi(x_1, x_2)}.
\end{aligned}
$$

The simulation results are reported in Table 4 and confirm the reduction of variance to zero.

4.5 A first discussion of the gained insight

As shown in the previous subsection, in the Monte Carlo framework this method works well when the theoretical ϕ is available. However, in most practical applications, two problems may arise:

1. The impossibility to sample directly from the target distribution;
2. The unavailability of the theoretical ϕ.

To overcome the first problem one could use MCMC simulation techniques, however it would be questionable if the machinery introduced at the beginning of this paper works properly also in a MCMC setting. The answer is affirmative, indeed it is straightforward to show that when the exact ϕ is available, i.e. the ϕ satisfying equation (6), $Cov(\tilde{f}(X_0), \tilde{f}(X_k))$ goes to zero, this is confirmed by our simulations.

The second problem is more delicate but, as pointed out in Section 3, any choice of ϕ reduces the variance. The choice of ϕ remains an open question we want to address here. In the previous examples we showed that the theoretical ϕ's take the form $P(x)\sqrt{\text{Target}}$ where $P(x)$ is a polynomial. In the examples we provide, we noticed the influence of the following two factors on the degree of the polynomial $P(x)$:

a) The degree of the function $f(x)$;

b) The structure of the target.

Regarding the first of the two, a simple suggestion would be to control it by imposing $P(x)$ to have the same degree of $f(x)$. The second factor varies strongly among problems faced, so it is difficult to give a general suggestion, however our experiments confirm some kind of robustness of results about a misspecification of the ϕ function. As an example we tried to impose a first order $P(x)$ for an univarite student target distribution, whose theoretical ϕ requires a third degree polynomial, when one is interested in $E_\pi(f)$. We obtained a promising 93% variance reduction, so a little decrease of performance with respect to the exact ϕ, confirmed even on MCMC samples.

5 Variance reduction in MCMC case and examples

We leave here the Monte Carlo framework, focusing henceforth on random draws obtained resorting to MCMC methods. We start referring to the examples studied in Section 4, if we estimate $E_\pi(f)$ by running a Markov chain by using the exact theoretical ϕ, we obtain results similar to the Monte Carlo case. In the Tables 5, 6, 7, 8 we report the simulation results. These are obtained by simulating 1000 points with a random walk Metropolis Hastings with an optimally scaled Normal proposal distribution (see Roberts and Rosenthal, 2001) and then discarding the first 850 points so that the number of MCMC actual points compare to the number of MC draws used in Section 4 (i.e. $T = 150$).

5.1 Gaussian-Gaussian model

Consider the following model for s iid observations y_i:

$$l(y_i|\theta) \sim N(\theta, \sigma_y^2) \qquad i = 1, \cdots, s;$$

where σ_y^2 is the known variance and θ is the parameter of interest. We assume a conjugate Normal prior:

$$h(\theta) \sim N(\mu_\theta, \tau_\theta^2)$$

where μ_θ and τ_θ^2 are known hyperparameters. It is well known that posterior distribution of the parameter of interest is

$$\pi(\theta|y_1, \cdots, y_s) = N(\mu_\pi, \sigma_\pi^2)$$

where

$$\mu_\pi = \frac{\mu_\theta \sigma_y^2 + s\tau_\theta^2 \overline{y}}{\sigma_y^2 + s\tau_\theta^2}$$

Table 5: $N(1,2)$, $f_1(x) = x$, $f_2(x) = x^2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$
$\hat{\mu}_f$	0.080	1	3,193	3
$\hat{\sigma}^2_f$	2.563	4.8e-20	13.209	1.31e-19

Table 6: Univariate Student-T with df=5, $f_1(x) = x$, $f_2(x) = x^2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$
$\hat{\mu}_f$	0.095	2.08e-12	1.55	1.666
$\hat{\sigma}^2_f$	1.551	1.08e-22	4.077	6.51e-24

Table 7: Bivariate Normal, $\underline{\mu}$=(2,1), (σ_1^2,σ_2^2)=(4,1), ρ=0.6 $f_1 = x_1$, $f_2 = x_1^2$, $f_3 = x_1 x_2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$	f_3	$\tilde{f}_3$
$\hat{\mu}_f$	1,683	2,549	5.366	8	2.136	3.2
$\hat{\sigma}^2_f$	2	2,01e-16	33.937	1.193e-14	7.14	7.11e-17

Table 8: Bivariate Student-T, df=7, $f_1 = x_1$, $f_2 = x_1^2$, $f_3 = x_1 x_2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$	f_3	$\tilde{f}_3$
$\hat{\mu}_f$	-0.09	7.29e-10	1.049	1.4	-0.038	-4,31e-12
$\hat{\sigma}^2_f$	1.04	1.02e-17	5.44	1.92e-17	1.254	1.95e-21

and

$$\sigma_\pi^2 = \frac{\sigma_y^2 \tau_\theta^2}{\sigma_y^2 + s\tau_\theta^2}$$

here $\overline{y}$ is the sample mean. In this setting we considered $f(\theta) = \theta$ and:

$$\phi(\theta) = \phi_1(\theta).$$

As a concrete example we used $\sigma_y = 3, \mu_\theta = 0, \tau_\theta = 3$ and generated the actual sample of size $s = 10$, from a Gaussian distribution with mean equal to one and standard deviation equal to 3. The posterior distribution has $\mu_\pi = 1.7487$ and $\sigma_\pi = 0.904$. The estimated mean and standard deviations of these distributions are presented in Table 9. Again, the advantage in terms of variance reduction of the method proposed is clear.

Table 9: Bayesian model, $f(\theta) = \theta$, $N = 500$.

	f	$\tilde{f}$
$\hat{\mu}_f$	1.7736	1.7399
$\hat{\sigma}_f^2$	0.8838	0.0362 (96% reduction)

5.2 Poisson-Gamma model

As a second model we consider the well known *Poisson-Gamma* model where:

$$l(y_i|\theta) \sim Po(\theta) \qquad i = 1, \cdots, s;$$
$$h(\theta) \sim Ga(\alpha = 4, \beta = 4).$$

We extract $s = 30$ values from a $Po(\theta = 4)$ distribution, we then

1. run a first MCMC simulation of length 1000 with a burn-in of 100;

2. minimize the variance on this first simulation and save the parameters;

3. run 100 parallel MCMC chains, each of length 10000 (after a burn-in of 150 steps);

4. compute, on each chain, f and $\tilde{f}$.

We are interested in the first moment of the posterior distribution, in this case we have the exact solution:

$$\frac{\beta + \sum_{i=1}^{s} y_i}{\alpha + s} = 4.058824$$

The inspection of parallel chains, for example at 500 iterations, shows that $\bar{f}_{500} = 4.060625$. $\bar{\tilde{f}}_{500} = 4.058843$ and $\sigma(f_{500}) = 0.0150$ while $\sigma(\tilde{f}_{500}) = 0.001777$.

A variance reduction by a factor of nearly ten is achieved. Figure 1 and Figure 2 depict the results obtained.

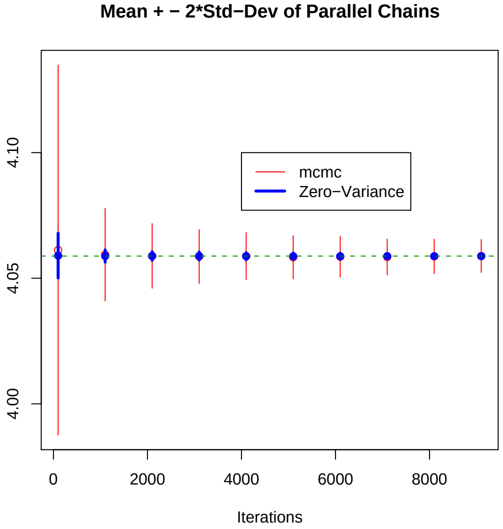

Figure 1: Poisson-Gamma: Parallel Chains

5.3 Logistic regression

We now consider a logistic regression model, commonly used in statistical practice. We simulate dependent binary data as follows:

$$l(y_i|\theta) \sim Be(\theta_i) \quad i = 1, ..., 100;$$

$$\theta_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad , \quad x_i \sim N(0, 1);$$

setting $\beta_0 = 0.5$, $\beta_1 = 1.5$, while assuming of an improper prior on each parameter.

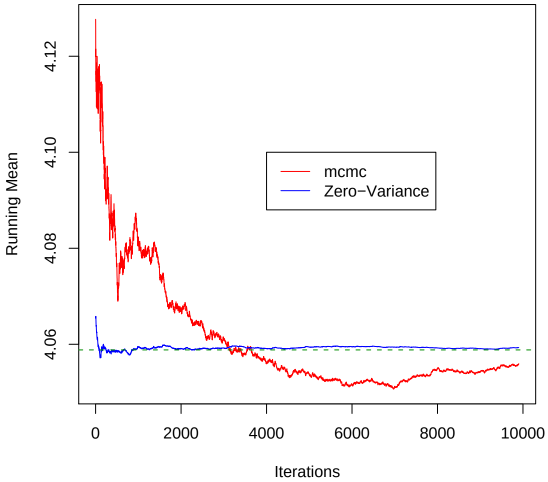

Figure 2: Poisson-Gamma: Single Chain

The posterior distribution does not have a closed form, however we resort to its normal approximation and therefore choose a ϕ with a structure similar to the optimal ϕ for the normal case. In the Tables 10 and 11, the resulting variance reductions are reported.

Table 10: Logit Model, β_0 $f(\beta_0) = \beta_0$, $N = 300$.

	f	$\tilde{f}$
$\hat{\mu}_f$	0.5676	0.5629
$\hat{\sigma}_f^2$	0.0923	0.0018 (80% reduction)

6 A simplified credit risk model

We now estimate the parameters of a logistic regression for creditworthiness, using a sample of 124 firms that gave rise to problematic credit and a sample of 200 healthy firms (so that $n = 324$). The models proposed is the following

$$\pi\left(\underline{\beta}|y,x\right) \propto \prod_{i=1}^{n} \theta_i^{y_i}\left(1-\theta_i\right)^{1-y_i} p\left(\underline{\beta}\right); \qquad (12)$$

Table 11: Logit Model, β_1 $f(\beta_1) = \beta_1$, $N = 300$.

	f	$\tilde{f}$
$\hat{\mu}_f$	2.0089	1.9758
$\hat{\sigma}^2_f$	0.1839	0.0122 (93% reduction)

$$Y_i \sim Be(\theta_i) \qquad \theta_i = \frac{\exp\left(\underline{x}_i^T \underline{\beta}\right)}{1 + \exp\left(\underline{x}_i^T \underline{\beta}\right)}, \qquad i = 1, \cdots, n;$$

where $\underline{\mathrm{x}}_i$ is a vector of four balance sheet indicators, including the intercept. We use a non informative improper prior distribution on $\underline{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$. This real data set has already been analyzed in Mira and Tenconi (2004), where a random effects model in the intercept was assumed.

We run an initial Markov Chain using a canonical Metropolis Hastings of length 300 (after a burn in of 700) and over this initial sample we estimate the optimal parameters of the ϕ function for each j dimension and for $f_j(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = \beta_j$, $j = 1, \ldots, 5$.

$$\phi^j(\beta) = \left(\gamma_1^j \beta_1 + \gamma_2^j \beta_2 + \gamma_3^j \beta_3 + \gamma_4^j \beta_4 + \gamma_5^j \beta_5\right) \sqrt{\pi\left(\underline{\beta}|y, x\right)} \qquad j = 1...5.$$

The optimization gave these estimates:

j	$\hat{\gamma}_1^j$	$\hat{\gamma}_2^j$	$\hat{\gamma}_3^j$	$\hat{\gamma}_4^j$	$\hat{\gamma}_5^j$
1	-0.09457704	-0.01333198	-0.05751499	-0.04640937	0.01208364
2	-0.01507528	-0.15816491	0.05934955	0.01612018	0.05508849
3	-0.05629736	0.06052546	-0.19269449	0.01473065	-0.03554821
4	-0.046095866	0.019266392	0.014117965	-0.101136218	0.003513808
5	0.0105972810	0.0597459884	-0.0345264631	0.0001133164	-0.0624642257

while the mean and variance for each parameter are reported in Table 12, where $f_j(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = \beta_j$, $j = 1, \ldots, 5$.

After performing a longer MCMC simulation of length 6000 (with a burn in of 1000 points), we obtain the results reported in Table 13:

While after 600000 MCMC iterations (with a burn in of 100000) we achieve the results reported in Table 14.

So with 5000 iterations only, the zero-variance estimators is close to the 500 000 MCMC results. This means that, to have results similar to the variance reduction technique, we must run a 100 times wider sample.

Table 12: Credit Risk Model, initial sample estimation

j	$\hat{\mu}_{f_j}$	$\hat{\mu}_{\tilde{f}_j}$	$\hat{\sigma}^2_{f_j}$	$\hat{\sigma}^2_{\tilde{f}_j}$	% variance reduction
1	-1.4761	-1.4339	0.0507	0.0015	97.04
2	-1.0337	-1.0138	0.0664	0.0018	97.28
3	-0.2858	-0.2830	0.0825	0.0043	94.78
4	-0.9687	-0.9746	0.0630	0.0007	98.88
5	0.8279	0.7756	0.0317	0.0012	96.21

Table 13: Credit Risk Model, N=6000 points

j	$\hat{\mu}_{f_j}$	$\hat{\mu}_{\tilde{f}_j}$	$\hat{\sigma}^2_{f_j}$	$\hat{\sigma}^2_{\tilde{f}_j}$	% variance reduction
1	-1.4045	-1.4431	0.0435	0.0032	92.64
2	-0.9831	-1.0122	0.0795	0.0028	96.47
3	-0.2810	-0.3078	0.1081	0.0097	91.02
4	-0.9466	-0.9716	0.0523	0.0007	98.66
5	0.7737	0.7762	0.0323	0.0019	94.11

Table 14: Credit Risk Model, N=600000 points

j	$\hat{\mu}_{f_j}$	$\hat{\sigma}^2_{f_j}$
1	-1.4354	0.0450
2	-1,0138	0.0820
3	-0.2941	0.0950
4	-0.9709	0.0510
5	0.7778	0.0310

7 Some tricks to speed up the simulation

When the operator defined in (8) is used, the function $\tilde{f}$ takes the form

$$\tilde{f}(x) = f(x) + \frac{(H\phi)(x)}{\sqrt{\pi(x)}}.$$

$H\phi(x)$ has to be computed on each point in the sample path, therefore when unavailable analytically, we must compute numerically the second order derivative that appears in the H operator. This is a time-consuming operation, however, by using some tricks we are able to speed up the necessary computations.

We have suggested to use functions having the form $\phi(x) = P(x)\sqrt{\pi(x)}$ where $P(x)$ is a polynomial. As the following theorem shows, this choice reduces the calculation of $H\phi(x)$ to a first order derivative.

Theorem 1. *Assume*

$$\phi(x) = P(x)\sqrt{\pi(x)}$$

where $P(x)$ is a polynomial. Then

$$(H\phi)(x) = -\frac{1}{2}\sum_{i=1}^{d}\left[\sqrt{\pi(x)}\frac{\partial^2}{\partial x_i^2}P(x) + 2\left(\frac{\partial}{\partial x_i}P(x)\right)\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right)\right]. \tag{13}$$

Before giving a proof of the proposition, some comments are required. Indeed, in (13) a second order derivative still appears but it is applied to a polynomial function and can thus be computed analytically. This theorem therefore reduces the computation to the first order derivative of the square root of the target. Also recall that the target has been already evaluated over all possible values x during the MCMC simulation: these values can thus be stored and re-used in the evaluation of $\tilde{f}$.

Proof. We must take the derivative of $\phi(x)$ twice with respect to a generic coordinate i:

$$\frac{\partial^2}{\partial x_i^2}\phi(x) = P(x)\frac{\partial^2}{\partial x_i^2}\sqrt{\pi(x)} + \sqrt{\pi(x)}\frac{\partial^2}{\partial x_i^2}P(x) + 2\left(\frac{\partial}{\partial x_i}P(x)\right)\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right).$$

Then

$$
\begin{aligned}
(H\phi)(x) &= \left(-\frac{1}{2}\sum_{i=1}^{d}\frac{\partial^2}{\partial x_i^2}\phi(x)\right) + \phi(x)V(x) \\
&= -\frac{1}{2}\sum_{i=1}^{d}\left[P(x)\frac{\partial^2}{\partial x_i^2}\sqrt{\pi(x)} + \sqrt{\pi(x)}\frac{\partial^2}{\partial x_i^2}P(x) + 2\left(\frac{\partial}{\partial x_i}P(x)\right)\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right)\right] \\
&\qquad + \frac{1}{2}\sum_{i=1}^{d}P(x)\frac{\partial^2}{\partial x_i^2}\sqrt{\pi(x)} \\
&= -\frac{1}{2}\sum_{i=1}^{d}\left[\sqrt{\pi(x)}\frac{\partial^2}{\partial x_i^2}P(x) + 2\left(\frac{\partial}{\partial x_i}P(x)\right)\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right)\right]
\end{aligned}
$$

□

Corollary 2. *Suppose that* $\phi(x) = P(x)\sqrt{\pi(x)}$ *and* $P(x) = P(x_1, \ldots, x_d)$ *is a first degree polynomial in* $\mathbb{R}^d$, *i.e.*

$$
P(x) = \sum_{i=1}^{d} a_i x_i \qquad a_i \in \mathbb{R}.
$$

Then

$$
(H\phi)(x) = -\sum_{i=1}^{d}\left[a_i\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right)\right]
$$

Proof. It follows from Theorem 1 by noting that

$$
\frac{\partial}{\partial x_i}P(x) = a_i \qquad \frac{\partial^2}{\partial x_i^2}P(x) = 0
$$

□

Remark 3. With the previous theorem $\tilde{f}$ becomes

$$
\tilde{f}(x) = f(x) - \frac{1}{\sqrt{\pi(x)}}\sum_{i=1}^{d}\left[a_i\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right)\right].
$$

A “logarithmic” version of the previous formula is also available:

$$
\tilde{f}(x) = f(x) - \frac{1}{2}\sum_{i=1}^{d}\left[a_i\left(\frac{\partial}{\partial x_i}\ln\pi(x)\right)\right].
$$

Corollary 4. *In the Credit Risk model of section 6*

$$(H\phi^j)(\beta) = -\sum_{i=1}^{5}\left[\gamma_i\left(\frac{\partial}{\partial\beta_i}\sqrt{\pi\left(\underline{\beta}|y,x\right)}\right)\right]$$

Proof. It follows from Theorem 2. □

We conclude this section with an intuition useful to avoid the numerical optimization, necessary to find a ϕ close to the optimal one. In the Credit Risk Model we noticed a great similarity of the matrix $\Gamma = \{\gamma_i^j\}_{i,j=1,\ldots,5}$ to the matrix -2Σ, where Σ is the estimated (from the MCMC output) covariance matrix of the target distribution.

This intuition is confirmed by the theoretical ϕ we obtained for the normal case when $f(x) = x_i$, described in section 4.1 and 4.3, as the coefficients of the polynomial are the elements of the i-th row of the target covariance matrix multiplied by -2. In the next subsection we will use the above mentioned tricks to reduce the variance in a complex credit risk model that builds on the simplified one introduced in Section 6.

7.1 An extended credit risk model

It is commonly accepted that the amount of credit risk is different among sectors. In Mira and Tenconi (2004) a hierarchical logistic regression model was proposed with the purpose to capture the sector specific baseline risks and to obtain a best fit of data. This model is reproposed here to investigate the zero variance principle on a highly parametrized model. The data contains 7513 firms allocated among $j = 7$ sectors, firm specific balance sheet indicators, x_{ij}, and default events, y_{ij} . The model presents a hierarchical structure in the intercepts α_j, allowing for greater variation among sectors, overcoming at the same time overfitting issues:

$$\pi\left(\underline{\alpha},\underline{\beta},\mu_a,\sigma_\alpha^2|y,x\right) \propto \prod_{j=1}^{7}\prod_{i=1}^{n_j}\theta_{ij}^{y_{ij}}\left(1-\theta_{ij}\right)^{1-y_{ij}}\prod_{j=1}^{7}p\left(\alpha_j|\mu_\alpha,\sigma_\alpha^2\right)p\left(\mu_\alpha\right)p\left(\sigma_\alpha^2\right)p\left(\underline{\beta}\right)$$

$$\theta_{ij} = \frac{\exp\left(\alpha_j + \underline{x}_{ij}^T\underline{\beta}\right)}{1+\exp\left(\alpha_j + \underline{x}_{ij}^T\underline{\beta}\right)}$$

with

$$\begin{aligned} \beta &\sim MN\left(0,\sigma^2 I_4 = 64\right) \\ \alpha_j|\mu_\alpha,\sigma_\alpha^2 &\sim N\left(\mu_\alpha,\sigma_\alpha^2\right) \\ \mu_\alpha &\sim N\left(0,\sigma^2 = 64\right) \\ \sigma_\alpha^2 &\sim Ga(\alpha = \frac{9}{5}, r = \frac{25}{9}). \end{aligned}$$

We focus on the functionals $f_k\left(\underline{\eta}\right) = \eta_k$ where $\underline{\eta}$ is the vector of all parameters, i.e $\underline{\eta} = (\underline{\alpha}, \underline{\beta}, \mu_\alpha, \sigma_\alpha)$. The ϕ functions are choosen as in Section 6 and the following steps are taken:

1. a Markov chain of lenght 50000 is run, discarding the first 10000 steps as burn-in, to obtain a sample from $\pi\left(\underline{\eta}|y, x\right)$;
2. the target variance-covariance matrix of $\underline{\eta}$, Σ_π, is estimated along the chain simulated at step 1. This estimate, $\hat{\Sigma}$, it used to parametrize the ϕ functions to compute $\tilde{f}$ with the "fast version" of our algorithm, i.e.
$$\tilde{f}_k\left(\underline{\eta}\right) = f_k\left(\underline{\eta}\right) - 2\hat{\Sigma} \times \nabla \ln\left(\pi\left(\underline{\eta}|y, x\right)\right);$$
3. We evaluate $\tilde{f}_k\left(\underline{\eta}\right)$ on a second MCMC sample of length 3000.

The results, in terms of variance reduction, for all parameters of interest, are presented in Table 15 which shows an average variance reduction of 78,95%. If we exclude the hyper parameters, η_{12} and η_{13}, which are of little interest for credit risk estimation, the variance reduction goes up to 85,49%.

Table 15: Variance reduction for complex credit risk model

k	η_k	$\hat{\mu}_{f_k}$	$\hat{\mu}_{\tilde{f}_k}$	$\hat{\sigma}^2_{f_k}$	$\hat{\sigma}^2_{\tilde{f}_k}$	% variance reduction
1	$\eta_1 = \alpha_1$	-6.5122	-6.4548	1.8261	0.7731	57.67
2	$\eta_2 = \alpha_2$	-5.3699	-6.5122	0.1546	0.0166	89.24
3	$\eta_3 = \alpha_3$	-5.1055	-5.1296	0.0884	0.0113	87.21
4	$\eta_4 = \alpha_4$	-4.8881	-4.9179	0.0876	0.0086	90.16
5	$\eta_5 = \alpha_5$	-5.2247	-5.2446	0.0869	0.0112	87.14
6	$\eta_6 = \alpha_6$	-3.9072	-3.9560	0.1057	0.0170	83.91
7	$\eta_7 = \alpha_7$	-6.3274	-6.3539	0.1097	0.0131	88.06
8	$\eta_8 = \beta_1$	-0.0942	-0.0901	0.0032	0.0005	83.83
9	$\eta_9 = \beta_2$	-1.2452	-1.2649	0.0999	0.0078	92.23
10	$\eta_{10} = \beta_3$	-1.4105	-1.4295	0.0415	0.0049	88.26
11	$\eta_{11} = \beta_4$	0.0870	0.0868	0.0027	0.0002	92.73
12	$\eta_{12} = \mu_\alpha$	-5.2806	-5.3548	0.3840	0.1114	70.98
13	$\eta_{13} = \sigma_\alpha$	1.3738	1.4248	0.1883	0.1601	15.00

8 Rao-Blackwellization

Rao-Blackwellization (Casella and Robert, 1996) can be seen as a special case of the variance reduction technique proposed in this paper. The Rao-Blackwellization idea is to replace $f(x^i)$ in $\hat{\mu}$ by a conditional expectation, $E_\pi[f(x^i)|h(x^i)]$, for some function h or to condition on the previous value of the chain thus using $E[f(x^i)|x^{i-1} = x]$ instead. Changing an expectation with a conditional expectation naturally reduces the variance of the resulting MCMC estimator. The functions $E_\pi[f(x^i)|h(x^i)]$ and $E[f(x^i)|x^{i-1} = x]$ can be considered as special instances of $\tilde{f}$ which do not minimize $\sigma_{\tilde{f}}$ but certainly reduce it. This suggests general guidelines that can be adopted to construct ϕ based on which we obtain $\tilde{f}$. In real applications, typically $E_\pi[f(x^i)|h(x^i)]$ or $E[f(x^i)|x^{i-1} = x]$ are not available in closed form, still, the researcher may have some intuition on the parametric form of such functions (or estimate them via pilot runs of the Markov chain). This intuition might aid the design of ϕ.

9 Conclusions

We have presented the advantages, in a statistical setting, of a general purpose variance reduction technique which has been originally suggested in the physics literature (Assaraf and Caffarel, 1999). Not only the zero variance physics principle has been adapted to the statistical framework but it has also been extended from Monte Carlo to Markov chain Monte Carlo simulation. The extent by which the variance of Monte Carlo and MCMC estimators can be reduced, is illustrated via some toy examples and a complex credit risk Bayesian model, fitted to a real dataset. The overall performance of the proposed technique is quite astonishing: in simple cases zero variance is indeed achieved, while in more complicated models, when the exact solution to the fundamental equation cannot be obtained analytically, a variance reduction between 80% and 95% is obtained. Moreover, useful tricks are proposed to dramatically speed up the application of the method to statistical modelling. Connections with the Rao-Blackwellization principle known in the MCMC literature are explored and exploited to better apply the zero-variance technique in a Bayesian setting.

References

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

Casella, G. and Robert, C. P. (1996), “Rao-Blackwellization of Sampling Schemes”, Biometrika, 83 (1), 81—94.

Delmas, J. F. and Jourdain, B. (2006), "Does waste recycling really improve Metropolis Hastings Monte Carlo algorithm?", Rapport du Recherce du CERMICS 2006-331

Green, P. J. and Mira, A. (2001), "Delayed Rejection in Reversible Jump Metropolis-Hastings", Biometrika, 88, 1035–1053.

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

Liu, J. S. (1996), "Peskun theorem and a modified discrete-state Gibbs sampler", Biometrika, 83 681–682.

Metropolis, N. and Rosenblutt, N. and Rosenblutt, A. W. and Teller, M. N. and Teller, A. H. (1953), "Equations of State Calculations by Fast Computing Machines", Journal of Chemical Physics, 21, 1087-1092.

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

Mira, A. and Tenconi, P. (2004), "Bayesian estimate via credit risk via MCMC with delayed rejection", Stochastic Analysis, *Random Fields and Applications IV in Progress in Probability*, Birkhauser Verlag, Basel, 2004, pp. 277-291, 26 93–108.

Peskun, P. H. (1973), "Optimum Monte Carlo sampling using Markov chains", Biometrika, 60, 607–612.

Ripley, B. D. (1987), " Stochastic Simulation", John Wiley and Sons, New York,

Roberts, G. O. and Rosenthal J. S. (2001), "Optimal Scaling for Various Metropolis-Hastings Algorithms", Statistical Science, Vol. 16, n. 4, 351-367

Sokal, A. D. (1989) "Monte Carlo methods in statistical mechanics: foundations and new algorithms", Cours de Troisième Cycle de la Physique en Suisse Romande, Lausanne.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions", Annals of Statistics, 22, 1701-1762

Tierney, L. (1998), " A note on Metropolis-Hastings kernels for general state spaces", Annals of Applied Probability, 8, 1–9.

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