A. Mira – D. Imparato

Density estimators through Zero Variance Markov Chain Monte Carlo

2011/8

UNIVERSITÀ DELL'INSUBRIA
FACOLTÀ DI ECONOMIA

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

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

Antonietta Mira*, Daniele Imparato†

March 10, 2011

Abstract

A Markov Chain Monte Carlo method is proposed for the point-wise evaluation of a density whose normalizing constant is not known. This method was introduced in the physics literature by Assaraf et al. (2007). Conditions for unbiasedness of the estimator are derived. A central limit theorem is also proved under regularity conditions. The new idea is tested on some toy-examples.

Keywords: Density estimator; Fundamental solution; MCMC simulation.

1 Introduction

Markov Chain Monte Carlo (MCMC) methods are able to gather information on a target distribution, known up to a normalizing constant. They are generally used for numerical evaluations of quantities related to the target itself, such as the computation of involved expectations and confidence intervals. MCMC methods are widely used in the Bayesian setting. In this context, the distribution of interest is the posterior of the model, for which the evaluation of the normalizing constant is not possible analytically, up to very simple cases. The combined simplicity and efficiency of MCMC algorithms, such as Metropolis-Hastings and Gibbs sampler, has lead to a renewed popularity of Bayesian Statistics.

*Università dell’Insubria, Varese. E-mail: antonietta.mira@uninsubria.it
†Università dell’Insubria, Varese. E-mail: daniele.imparato@uninsubria.it
When the normalizing constant is not known, a particular task arising in this framework is the evaluation of the target distribution at a given point. In this case, with the classical MCMC strategy a naive estimation of the unknown value is obtained by constructing the histogram of the distribution. This approach strongly depends on the choice of the length of the bins and fails along the tails of the target, where the simulated values of the Markov chain may rarely or, even never, occur if the bin width is not sufficiently large. Conversely, choosing too wide bins typically leads to unbiased estimations.

In this paper an alternative MCMC strategy to evaluate the target distribution at a given point is developed. The idea, which has been first proposed in the physical literature by [2], provides a continuous estimation of the density, which avoids the above mentioned computational issues involved in the MCMC histogram-like estimation.

2 Main idea

Let $\rho_d$ be a unnormalized, $d$-dimensional probability distribution, only known up to a normalizing constant: $\rho_d^* = c\rho_d$. Interest is in evaluating $\rho_d$ or any of its marginal distributions $\rho_r$, for $r \leq d$, at a given point in the support. Evaluating a marginal analytically is often a hard task, since it involves the computation of multidimensional integrals. The aim of this paper is to provide an MCMC estimation of the value of the marginal at a given point without the need of the classical histogram-like approximation. We first notice that $\rho_d(x_d)$ can be written as the expected value $\mu_f$ of a specific function $f$ and similarly for the marginals $\rho_r(x_r)$. Indeed, let $f(y_r) = \delta_0(y_r - x_r)$ and $\delta_0(x)$ denote the Dirac function:

$$
\mu_f = \int \delta_0(y_r - x_r)\rho_d(y_d)dy_d
= \int_{\mathbb{R}^{d-r}}\int_{\mathbb{R}^r} \delta_0(y_r - x_r)\rho_d(y_r, y_{r+1}, \ldots, y_d)dy_rdy_{r+1}\ldots dy_d \\
= \int_{\mathbb{R}^{d-r}} \rho_d(x_r, y_{r+1}, \ldots, y_d)dy_{r+1}\ldots dy_d
= \rho_r(x_r).$$

Starting from this equality, [3] developed a MCMC method to evaluate marginal distributions, intended to reduce the variance of the estimator. In analogy with previous work (see e.g. [1], [2], [5], [4]) the authors proposed
a MCMC estimator by modifying the function \( f(y_r) = \delta_0(y_r - x_r) \), whose expectation \( \mu_f \) we are interested in, and defining a new function \( \tilde{f} \), such that \( \mu_f = \mu_{\tilde{f}} \) and \( \sigma(\tilde{f}) \ll \sigma(f) \). Ideally, this re-normalized function could be constructed so that \( \sigma(\tilde{f}) \equiv 0 \). The corresponding estimator is, therefore, also called Zero Variance MCMC (ZV-MCMC) estimator (see [1] and [5]), although in practice the case of zero variance can be reached in, very rare, trivial cases.

The re-normalized function \( \tilde{f}_{x_r} \), which is proposed for estimating the marginal density \( \rho_r(x_r) \), is

\[
\tilde{f}_{x_r} = -\frac{1}{4\pi} \left[ \frac{1}{|y_r - x_r|} - g(x_r) \right] \frac{\Delta[\psi(y_r - x_r)\rho_d(y_d)]}{\rho_d(y_d)}. \tag{2}
\]

Here, \( \Delta \) denotes the Laplacian operator of partial second derivatives; \( \psi \) is a smooth function, satisfying \( \psi(0) = 1 \) and \( g \) is an other arbitrary smooth function. The proof that \( \mu_f = \mu_{\tilde{f}} \) is not trivial, since it is based on the functional equality

\[
\delta_0(y_r - x_r) = -\frac{\psi(y_r - x_r)}{4\pi} \frac{\Delta}{|y_r - x_r|}. \tag{3}
\]

Substituting (3) in (2) and integrating by parts twice gives the result, provided that \( h \) and \( g \) satisfy further suitable conditions, detailed later in the paper. In the original paper [3] these further hypotheses were not mentioned, since trivially satisfied in their specific application, but are, indeed, needed in more general situations. Moreover, it should be noted that equation (3) is true only when \( y_r, x_r \in \mathbb{R}^3 \). This fact was not stressed by the authors in [3], since they considered 3D coordinates, which represent the position of atoms in space. Generalizing [3], what is really needed to get a ZV-MCMC estimation of the target at a point is a function, say \( u(x) \), and a differential operator, say \( H \), such that \( Hu = \delta_0 \). Then, one can find a suitable re-normalized \( \tilde{f} \), in analogy with (2), where the Laplacian operator \( H = \Delta \) was considered for three-dimensional problems.

Starting from this setting, in this paper aim is at discussing generalizations of equation (2) in different dimensions and considering different operators \( H \). Secondly, the statistical properties of the resulting (ZV)-MCMC estimator will be studied, by carefully discussing the general hypotheses under which unbiasedness is reached and asymptotic results, such as Central Limit Theorem, hold. Moreover, optimal choices for the trial functions \( \psi \) and...
are studied in order to get minimum variance estimators. The classical, histogram-like density estimator will be compared with the new estimator we introduce through toy-examples (multivariate normal and multivariate t distribution), in order to show the variance reduction obtained with our method.

3 General framework

In this section we generalize the framework of [3] to more dimensions and consider different differential operators. To this end, we recall some functional properties of differential operators (see [6]).

Definition 3.1 Let $H$ be a differential operator. The function $u(x)$ is called a fundamental solution for $H$ if

$$Hu = \delta_0,$$  
(4)

where $\delta_0$ is the Dirac function, which acts as:

$$\int \delta_0(x)g(x)dx = g(0),$$

for any test function $g$.

It should be noted that equation (4) holds as an equality between distributions, that is, it is equivalent to

$$\int Hu(x)f(x)dx = \int \delta_0(x)f(x)dx = f(0).$$

Moreover, observe that, if $u$ is a fundamental solution for $H$, it holds

$$\psi Hu = \delta_0$$  
(5)

for any function $\psi$ such that $\psi(0) = 1$, since, for any $g$,

$$\int \psi(x)Hu(x)g(x)dx = \int \psi(x)\delta_0(x)g(x)dx = \psi(0)g(0) = g(0).$$

The analytical form of the fundamental solution is well known for some particular operators $H$. In the sequel the case of Laplacian operator is considered ([6], p. 80).
Proposition 3.1 The fundamental solution for the Laplacian operator \( H = \Delta \) is
\[
u_n(x) = \begin{cases} 
1 + x 1_{\{x > 0\}}, & \text{if } x \in \mathbb{R}, \\
\frac{1}{2\pi} \log |x|, & \text{if } x \in \mathbb{R}^2, \\
-\frac{1}{(n-2)c_n|x|^{n-2}} & \text{if } x \in \mathbb{R}^n, \quad n > 2
\end{cases}
\]
(6)
where \( 1_{\{x > 0\}} \) denotes the characteristic function of the set \( \{x > 0\} \) and \( c_n \) is the area of the \( n \)-th dimensional unitary sphere.

3.1 Core of the method

By using the mathematical formalism outlined at the beginning of Section 3, we are ready to define our MCMC strategy to evaluate densities at a given point with, often huge, variance reduction with respect to the classical MCMC density estimation. To this aim, if interest is in evaluating the marginal \( \rho_r \) at point \( x_r \), we first need to define a suitable re-normalized function, \( \tilde{f}_{x_r} \), such that
\[
\rho_r(x_r) = \mathbb{E}[\tilde{f}_{x_r}].
\]
This function is based on the fundamental solution of the Laplacian and is defined as:
\[
\tilde{f}_{x_r}(y_d) := u_r(y_r - x_r) \frac{\Delta[\psi(y_r - x_r)\rho_d(y_d)]}{\rho_d(y_d)};
\]
(7)
where \( u_r \) is defined in (6) and \( \psi \) is an arbitrary trial function such that \( \psi(0) = 1 \). When \( d = 3 \), we get the same result as in (2). Integrating by parts twice and using the definition of \( u_d \) gives the expected result \( \pi(x_r) = \mu_{\tilde{f}_{x_r}} \), under some ad hoc assumptions on \( \psi \) or the target \( \rho \).

Our computational strategy will thus be the following:

1. choose a class of test functions \( \psi(\cdot; \lambda) \), parametrized by a real vector \( \lambda \), such that \( \psi(0; \lambda) = 1 \), for any \( \lambda \);
2. construct the re-normalized function \( \tilde{f}_{x_r} = \tilde{f}_{x_r}(\cdot; \lambda) \) as in (7);
3. determine the optimal parameter value, \( \lambda^* \), which minimizes the variance of \( \tilde{f}_{x_r} \):
\[
\lambda^* = \arg\min_{\lambda} \sigma^2(\tilde{f}_{x_r});
\]
4. find the MCMC estimation of $\mu_{\tilde{f}_{X^n}}$: if $(X^n)$ is a Markov Chain with $\rho_r$ as its stationary distribution evaluate

$$\hat{\mu}_{\tilde{f}_{X^n}} = \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_{X^n}(X^i).$$

It should be noted that the variance depends on the choice of $\psi$ and on the target density which is considered. In general, the optimum $\lambda$, which minimizes $\sigma^2(\tilde{f}_{X^n})$, cannot be found analytically, so that numerical methods are to be introduced. Although the choice of the trial function $\psi$ is quite arbitrary, this function should satisfy, at least, two practical requirements: to be analytically simple, in order to easily compute its derivatives; to ensure the convergence of the variance of the modified function $\tilde{f}$, specially when the target itself does not converge rapidly enough to zero at the boundary of its proper domain. This second requirements ensures that a CLT holds for the MCMC estimator, as discussed in Section 4. It is also intuitive that a good choice for the class of $\psi$’s should allow for small $\sigma^2(\tilde{f}_{X^n})$, that is, for, possibly high, variance reduction in the resulting MCMC estimation.

As in [3], we use the following class of trial functions:

$$\psi(x; \lambda) = (1 + \lambda ||x||) \exp(-\lambda ||x||), \quad (8)$$

where $|| \cdot ||$ denotes the Euclidean norm and $\lambda$ is the parameter of the class. Easy computations show that

$$\nabla \psi(x) = -\lambda^2 x \exp(-\lambda ||x||),$$

$$\Delta \psi(x) = -\lambda^2 (d - \lambda ||x||) \exp(-\lambda ||x||).$$

With this choice, the renormalized function $\tilde{f}$ is proportional to $\exp(-\lambda ||x||)$. We, therefore, expect to achieve big variance reduction within this class, since the integrand rapidly converges to zero along the boundary, for any positive $\lambda$. This property, jointly with the minimization step to find the optimal $\lambda$, explains why our method generally gives more stable estimations than the classical MCMC procedure for density point evaluation. Of course, depending on the target density, different classes of trial functions may be considered, which are able to control the fluctuations not only at the boundary, but also at given critical points.
4 Unbiasedness and CLT conditions

In this section technical details are given about the statistical properties of our estimator. Mathematical conditions on the target densities are found in order to ensure the unbiasedness of the estimator. Moreover, asymptotical results regarding CLT are discussed: to this end, reference should be made to [7] and [8]. By using the following results, it is easy to verify unbiasedness and CLT of our estimator in all the examples shown.

Using the same notations as in Section 3.1, in the following proposition unbiasedness is discussed for targets with bounded support. The generalization to unbounded domains is straightforward and cited below. The proof is a simple application of the multiple integration by parts and Definition 3.1 and, therefore, is not reported here.

Proposition 4.1 Let \( \rho_d \) be a density with bounded support \( \Omega \): then, \( \mu_{\tilde{f}_x} = \pi(x_r) \) provided that

\[
\int_{\partial \Omega} [u_r((y_r-x_r)) \nabla \psi(y_r-x_r) \rho_d(y_d) - \rho_d(y_d) \psi(y_r-x_r) \nabla u_r(y_r-x_r)] \cdot n \, d\sigma = 0,
\]

(9)

where \( n \) is the versor normal to \( \sigma \).

When \( \rho_d \) has unbounded support, if \( B_r \) denotes the hypersphere of radius \( r \), using again integration by parts and taking the limits for \( r \to \infty \), the same condition (9) must be verified, with \( \partial \Omega = B_r \) and \( r \to \infty \). In most cases, condition (9) can be easily proved by considering hyperspherical coordinates and by studying if the integrand in (9) is (asymptotically) zero. This is a trivial task if the target considered is a function of the eucledean norm or scalar product. This is the case of the toy-examples of Section 5, for which Equation (9) is easily verified.

According to [8], a CLT holds for an ergodic Markov chain \( \{X^n\} \) with stationary distribution \( \pi \) provided that finiteness of second moment with respect to \( \pi \) is verified. More precisely, the following theorem holds.

Theorem 4.2 Suppose an ergodic Markov chain \( \{X^n\} \), with stationary distribution \( \pi \) and a real valued function \( f \), satisfies one of the following conditions:

B1 : The chain is geometrically ergodic and \( f \in L^{2+k}(\pi) \) for some \( \delta > 0 \).
B2 : The chain is uniformly ergodic and \( f \in L^2(\pi) \).

Then

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

is well defined, non-negative and finite, and

\[
\sqrt{N}(\hat{\mu}_f - \mu_f) \overset{L}{\to} \mathcal{N}(0, s_f^2).
\]  

(10)

It should be noted that, when the chain is uniformly ergodic, our estimator satisfies a CLT provided that \( \sigma(\tilde{f}_{x_r}) \) is finite. This means that the CLT holds whenever our method can be used. If this is not the case, a different choice of \( \psi \) has to be considered. Using trial function as in (8), the CLT condition is straightforward for the examples discussed below.

## 5 Examples

In this Section the MCMC density estimator described above is applied to few simple toy-examples, for which the true value of the target density to be evaluated is known. For the sake of simplicity, the Monte Carlo simulation is considered. A comparison is made of our method with the classical (MC)MC estimator; for the latter, the length of the bin has been selected empirically in order to avoid meaningless estimations. As it can be seen by considering both mean squared error (MSE) and mean absolute error (MAE), our method gives more stable estimations. Moreover, the bias of the MCMC histogram estimator is evident, as shown in the figures reported.

### 5.1 The multivariate normal distribution

Let us consider a multivariate normal target \( \rho_d \sim \mathcal{N}_d(\mu, \Sigma) \). In this case,

\[
\nabla \rho_d = -\Sigma^{-1}(x - \mu)\rho_d,
\]

\[
\Delta \rho_d = \rho_d (||\Sigma^{-1}(y - \mu)||^2 - \text{Tr}(\Sigma^{-1})),
\]

where \( \text{Tr}(A) \) is the trace of the matrix \( A \). Therefore, since \( \Delta(\rho_d \psi) = \psi \Delta \rho_d + 2 \nabla \rho_d \cdot \nabla \psi + \rho_d \nabla \psi \), the modified function \( \tilde{f}_{x_r} \), defined in (7), is equal to
\[ \tilde{f}_{x_r}(y_d) = u_r(y_r-x_r) \left[ \psi(||\Sigma^{-1}(y-\mu)||^2 - \text{Tr}(\Sigma^{-1})) - 2\Sigma^{-1}(y-\mu) \cdot \nabla \psi + \Delta \psi \right], \]  

(11)

where \( \psi \) and its derivatives were defined in (8).

As an example, consider the case of a 3D normal distribution with mean \( \mu = 0 \), and covariance matrix \( \Sigma \) equal to

\[ \Sigma = \begin{pmatrix} 1 & 1/5 & 1/2 \\ 1/5 & 1 & 1/3 \\ 1/2 & 1/3 & 3/2 \end{pmatrix}. \]

Different simulations of length \( n = 10^6 \) have been executed to simulate the joint density \( \rho_d \); the fluctuations of classical MC were compared with those arising with our method.

In Figure 1 the results are shown regarding the estimation of the value of \( \rho_d \) at points \( x_r = \mathbf{0} \) and \( x_r = \mathbf{2} = (2, 2, 2) \), respectively.

Figure 1: 30 simulations of the value of a 3D normal distribution at \( \mathbf{0} \) (left plot) and \( \mathbf{2} \) (right plot): black line is the classical MC counting method, red line is our MC smoothing method, green line is the true value.
Table 1: Variance Reduction with ZV-MCMC:
Comparison of Mean Squared Error (MSE) and Absolute Mean Error (AME) between MC and ZV-MC.

<table>
<thead>
<tr>
<th></th>
<th>x = 0</th>
<th>x = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td>2.0 10^-4</td>
<td>4.7 10^-7</td>
</tr>
<tr>
<td>ZV-MC</td>
<td>8.8 10^-7</td>
<td>8.6 10^-8</td>
</tr>
<tr>
<td>MAE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td>5.9 10^-2</td>
<td>3.1 10^-3</td>
</tr>
<tr>
<td>ZV-MC</td>
<td>3.9 10^-3</td>
<td>1.3 10^-3</td>
</tr>
</tbody>
</table>

5.2 The multivariate t distribution

Let us consider a $d$-dimensional t distribution with parameters $\mu$, $\Sigma$ and $k$:

$$
\rho_d(x) \propto \left[ 1 + \frac{1}{k} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]^{-\frac{k+d}{2}}
$$

In this case, the modified function $\tilde{f}_{x_r}$, defined in (7), is equal to

$$
\tilde{f}_{x_r}(y_d) := u_r(y_r - x_r) \left( \Delta \psi(y_r, x_r) + 2 \nabla \psi(y_r, x_r) \cdot \frac{\nabla \rho_d(y_d)}{\rho_d} + \psi(y_r, x_r) \frac{\Delta \rho_d(y_d)}{\rho_d} \right),
$$

where

$$
\nabla \rho_d(y_d) = \frac{-k + d}{d} \frac{\Sigma^{-1}(y_d - \mu)}{\Sigma^{-1}(y_d - \mu)^T (y_d - \mu)},
$$

$$
\frac{\Delta \rho_d(y_d)}{\rho_d} = \frac{(k + d)(k + d + 2)||\Sigma^{-1}(y_d - \mu)||^2}{k^2(1 + \frac{1}{n}(y_d - \mu)^T \Sigma^{-1}(y_d - \mu))^2} - \frac{(k + d) Tr(\Sigma^{-1})}{n^2(1 + \frac{1}{k}(y_d - \mu)^T \Sigma^{-1}(y_d - \mu))^2}.
$$

As an example, consider the case of a 3D t-distribution with parameters $\mu = 0$, $\Sigma = I$ and $k = 1$. In Figure 2 the results are shown regarding the estimation of the value of $\rho_d$ at points $x_r = 0$ and $x_r = 2 = (2, 2, 2)$, respectively.

Table 2: Variance Reduction with ZV-MCMC:
Comparison of Mean Squared Error (MSE) and Absolute Mean Error (AME) between MC and ZV-MC.

<table>
<thead>
<tr>
<th></th>
<th>x = 0</th>
<th>x = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td>4.3 10^-4</td>
<td>2.0 10^-7</td>
</tr>
<tr>
<td>ZV-MC</td>
<td>5.6 10^-6</td>
<td>2.5 10^-8</td>
</tr>
<tr>
<td>MAE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td>9.3 10^-2</td>
<td>1.9 10^-3</td>
</tr>
<tr>
<td>ZV-MC</td>
<td>1.1 10^-2</td>
<td>6.2 10^-4</td>
</tr>
</tbody>
</table>

10
Figure 2: 30 simulations of the value of a 3D normal distribution at 0 (left plot) and 2 (right plot): black line is the classical MC counting method, red line is our MC smoothing method, green line is the true value.

References


