New perspectives for Estimating Normalizing Constants via Posterior Simulation

Giovanni Petris
University of Arkansas

Luca Tardella
Università di Roma "La Sapienza"

December 31, 2007

Abstract

In this paper we propose a new effective tool for evaluating the normalizing constant of an arbitrary density function with the aid of an arbitrary MC or MCMC sampling scheme. The new original estimators proposed here stem from the idea of suitably perturbing the original target density function whose normalizing constant has to be evaluated in such a way that the perturbed density has the same original normalizing constant plus a known arbitrary positive mass. The proposed estimators can be easily implemented sharing the original simplicity of the harmonic mean estimator of Newton and Raftery (1994) yielding consistent MC or MCMC estimators based only on a simulated sample from the distribution proportional to the original density. However, under fairly general sufficient conditions, they avoid the infinite variance shortcoming. Effectiveness is illustrated through controlled simulated examples with distributions in dimension one up to one hundred, as well as on a more practical context of real data sets. Extensions to the ratio of constants is discussed together with relations of the new proposed approach with bridge sampling and path sampling.

Key Words: Normalizing constant; MCMC; Integrated Likelihood; Generalized Harmonic Mean; Bridge Sampling; Path Sampling; Bayesian inference.

1 Introduction

Evaluating the normalizing constant of a probability density function known only up to a multiplicative constant is a fundamental part of many statistical problems. Gelman and Meng (1998) gives an account of other intertwined fields where the problem of evaluating

$$c = \int_{\Theta} g(\theta) d\theta. \tag{1}$$

is also of interest, ranging from missing data problems to computational physics from spatial statistics to computational chemistry. The need of efficient and flexible tools for evaluating normalizing constants is particularly called for in Bayesian computations. When the integral cannot be solved analytically possible general approaches to tackle the integration problem can be reduced to two mainstreams: either numerical solutions by approximating the integrand function with sums of functions with analytical computable integrals or via simulation techniques. The interest in improving currently available tools and comparing their performances is still very high. Among the most recent review paper are Clyde et al. (2006), Han and Carlin (2001), Gelman and Meng (1998) and Chen and Shao (1997). More recent and less recent tools for Bayesian computations were often derived from basic standard Monte Carlo principles and suitable starting identities such as the Bridge Sampling of Meng and Wong (1996) the Harmonic Mean Estimator of Newton and Raftery (1994) and later generalizations (Gelfand and Dey, 1994; Chen, 1994; Raftery et al., 2006) or the Ratio Importance Sampling of Chen and Shao (1997). Other relevant recent ideas and techniques are illustrated Chib (1995) and his related series Chib and Jeliazkov (2001, 2005) the Nested Sampling Skilling (2006) and also Bartolucci et al. (2006) in trans-dimensional simulation.

2 Importance sampling and generalized harmonic mean estimators

Let us start reviewing two available Monte Carlo solutions for the problem of evaluating the integral quantity (1).

Importance Sampling (IS)

In the standard Monte Carlo approach the target constant c is viewed as an expected value with respect to some known probability density f with the usual Importance Sampling trick as follows

$$c = \int_{\Theta} \frac{g(\theta)}{f(\theta)} f(\theta) d\theta = E_f \left[\frac{g(\theta)}{f(\theta)} \right]. \tag{2}$$

The Importance Sampling (IS) is a Monte Carlo technique to estimate c based on a sample $\eta_1, ..., \eta_T$ i.i.d. from f through the empirical mean

$$\hat{c}_{IS} = \frac{1}{T} \sum_{t=1}^{T} \frac{g(\eta_t)}{f(\eta_t)}, \tag{3}$$

where one can choose the probability density f almost arbitrarily. In fact, in order for the IS to work the following minimal assumptions are imposed:

IS1 $g(\cdot)$ needs to be integrable

IS2 g must be a probability density dominated by f, i.e. $g << f$

In fact, when $g(\theta)$ is an almost everywhere non negative function on $\Theta \subset \mathbb{R}^n$ and the target constant is strictly positive, c can be regarded as the *normalizing constant* which turns the non negative density g into the probability density function

$$\tilde{g}(\theta) = \frac{g(\theta)}{c}.$$

This will be mostly the case in the rest of this paper.

Reciprocal Importance Sampling (RIS) or Generalized Harmonic Mean (GHM)

When $g(\theta)$ is positive almost everywhere and can be then interpreted as a density of a finite measure with total mass c another alternative (dual) idea is the Reciprocal Importance Sampling (RIS) or Generalized Harmonic Mean (GHM) strategy. It is in fact a disguised form of importance sampling based on recasting the problem in terms of the reciprocal of the unknown total mass c. One starts with a known quantity, say 1 without loss of generality, evaluated as an integral of a known suitable Lebesgue integrable function $h(\theta)$

$$1 = \int_{\Theta} h(\theta) d\theta \tag{4}$$

then divides both terms by the target c and performs a trick similar to the IS by dividing and multiplying the integrand function $h(\theta)$ in (4) by the original integrand function $g(\theta)$ in (1)

$$\frac{1}{c} = \int_{\Theta} \frac{h(\theta)}{c} d\theta = \int_{\Theta} \frac{h(\theta)}{g(\theta)} \frac{g(\theta)}{c} d\theta = E_{\tilde{g}} \left[\frac{h(\theta)}{g(\theta)} \right] \tag{5}$$

so that, taking the reciprocal terms, one gets

$$c = \frac{1}{E_{\tilde{g}} \left[\left(\frac{g(\theta)}{h(\theta)} \right)^{-1} \right]} = Har_{\tilde{g}} \left[\frac{g(\theta)}{h(\theta)} \right] \tag{6}$$

which redisplay the target quantity c as the harmonic mean with respect to the probability distribution $\tilde{g}$ heretoafter denoted with $Har_{\tilde{g}}$.

In order for (5) to hold the following minimal assumptions are needed

RIS1 $g(\cdot)$ needs to be integrable

RIS2 $h(\theta)$ is dominated by $g(\theta)$ i.e. $h << g$.

The advantage of this method is that in some instances, especially in Bayesian computations, where g represents the posterior distribution known only up to the unknown normalizing constant, one has usually readily available a sample $\theta_1, ..., \theta_T$ from the (normalized) probability density

$$\tilde{g}(\theta) = \frac{g(\theta)}{c}$$

and from that sample one can estimate c through the empirical harmonic mean as follows

$$\hat{c}_{RIS} = \hat{c}_{GHM} = \frac{T}{\sum_{t=1}^{T} \frac{h(\theta_t)}{g(\theta_t)}} = \frac{}{\frac{1}{T} \sum_{t=1}^{T} \left(\frac{g(\theta_t)}{h(\theta_t)} \right)^{-1}}. \tag{7}$$

In view of its derivation we think it would be better to refer to this approach as Reciprocal Importance Sampling (RIS) although, probably, the terminology Generalized Harmonic Mean (GHM) is nowadays more frequent in the literature. While the original IS estimator $\hat{c}_{IS}$ in (3) is unbiased the alternative $\hat{c}_{RIS}$ or $\hat{c}_{GHM}$ in (7) is biased although asymptotically unbiased. In fact they are consistent since in both cases conditions for the Strong Law of Large Numbers are met. We recall that one of the first particular formulation of $\hat{c}_{GHM}$ as a tool for Bayesian computations was made in Newton and Raftery (1994) where $g(\theta)$ was the posterior density up to the normalizing constant and $h(\theta)$ was taken to be the prior distribution of a Bayesian model so that the ratio $\frac{g(\theta)}{h(\theta)}$ turned out to reduce to the likelihood function and the estimator $\hat{c}_{GHM}$ was the harmonic mean of the sample likelihood function evaluations. The most general version $\hat{c}_{GHM}$ was proposed in Gelfand and Dey (1994) and also in Chen (1994) although under the different terminology $\hat{c}_{IWMDE}$ (Importance Weighted Marginal Density Estimator). One annoying drawback of the original proposal of Newton and Raftery (1994), with $h(\theta)$ being the prior density of a Bayesian model, was that the corresponding $\hat{c}_{GHM}$ can easily end up having very large variance and sometimes even infinite for simple common models. Hence we now review some fundamental results concerning appropriate conditions for deriving a suitable asymptotic control of of both $\hat{c}_{IS}$ and $\hat{c}_{GHM}$ estimators. Improved, stabilized version of the original proposal were successively proposed see for instance Raftery et al. (2006). We illustrate in the next section a new strategy to build up f and h so that the corresponding $\hat{c}_{IS}$ and $\hat{c}_{GHM}$ estimators may have a competitive performance and investigate their absolute and comparative effectiveness.

Remark 1. *We stress on the fact that it is usually not explicitly recognized that $h(\cdot)$ need not be a probability density. We will in fact consider such more general context.*

We will eventually highlight in Section 6 some interesting connections of our approach with other currently available techniques to estimate normalizing constants.

3 Error control of IS and GHM estimators and ideas for new implementations of GHM

In order to discuss later comparative performance of alternative IS or RIS-GHM estimators we review some fundamental results under a standard Monte Carlo setting. Simple standard modifications can be made to obtain analogous results under more general MCMC setting where either $f(\theta)$ or $\tilde{g}(\theta)$ represents the stationary distribution. To compactify notation we will use the following short-hands: when $\eta_t \sim f$ we denote with

$$Y_t = \frac{g(\eta_t)}{f(\eta_t)}$$

such that the expected value of the random variable Y_t, also denoted as $\mu_Y = E[Y_t]$, exists and coincides with the target quantity c.

Theorem 1. *Suppose $\eta_1, ..., \eta_T$ are i.i.d. according to f. The IS estimator*

$$\hat{c}_{IS} = \frac{1}{T}\sum_{t=1}^{T} Y_t$$

is an unbiased and consistent estimator of c. Moreover, if $\sigma_Y^2 = Var[Y_t] < \infty$ then

$$\sqrt{T}\,(\hat{c}_{IS} - c) \overset{T\to\infty}{\Longrightarrow} N(0, \sigma_Y^2). \tag{8}$$

It is routine to derive the asymptotic

$$RMSE_{\hat{c}_{IS}} = \sqrt{E\left[\left(\frac{\hat{c}_{IS} - c}{c}\right)^2\right]} = \sqrt{E\left[\left(\frac{\hat{c}_{IS}}{c} - 1\right)^2\right]} \approx \frac{\sigma_Y}{c\sqrt{T}}$$

and the asymptotic confidence interval for the target quantity c namely from

$$1 - \alpha \approx Pr\left\{ z_{\frac{\alpha}{2}} \leq \sqrt{T}\frac{(\hat{c}_{IS} - c)}{\sigma_Y} \leq z_{1-\frac{\alpha}{2}} \right\}.$$

On the other hand when $\theta_t \sim \tilde{g}$ we denote the ratio of densities evaluated at θ_t with

$$X_t = \frac{g(\theta_t)}{h(\theta_t)}$$

the corresponding random variable so that the mean of its reciprocal

$$\mu_{\frac{1}{X}} = E\left[X_t^{-1}\right]$$

exists and coincides with c^{-1} while (6) can be rewritten as

$$c = \frac{1}{\mu_{\frac{1}{X}}} = \alpha_X. \tag{9}$$

This amounts to say that c can be regarded as the harmonic mean α_X of the transformed random variable X_t. Moreover, the $\hat{c}_{GHM}$ estimator is nothing but the empirical harmonic mean of X_t for which a suitable asymptotic control can be guaranteed.

Theorem 2. *Suppose $\theta_1, ..., \theta_T$ are i.i.d. according to $\tilde{g}$. The generalized harmonic mean estimator*

$$\hat{c}_{GHM} = \frac{T}{\sum_{t=1}^{T} X_t^{-1}}$$

is a consistent estimator of c. Moreover, if $\sigma^2_{\frac{1}{X}} < \infty$ then

$$\sqrt{T}\left(\frac{\hat{c}_{GHM} - c}{c}\right) \stackrel{T\to\infty}{\Longrightarrow} N(0, A^2) \tag{10}$$

where

$$A = \frac{\sigma_{\frac{1}{X}}}{\mu_{\frac{1}{X}}}. \tag{11}$$

Proof - The proof immediately follows from the asymptotic behavior of H_n the sample harmonic mean of an i.i.d sample $X_1, ..., X_n$ as recently provided in Pakes (1999), although also previously investigated in Norris (1940).

$\diamond$

If we consider the Relative Mean Square Error of $\hat{c}_{GHM}$ as follows

$$RMSE_{\hat{c}_{GHM}} = \sqrt{E\left[\left(\frac{\hat{c}_{GHM}}{c} - 1\right)^2\right]} = \sqrt{E\left[\left(\frac{\hat{c}_{GHM} - c}{c}\right)^2\right]},$$

the last statement of the above theorem allows us to derive a suitable approximate expression of $RMSE_{\hat{c}_{GHM}}$, in terms of

$$\sigma^2_{\frac{1}{X}} = Var\left[\frac{h(\theta_t)}{g(\theta_t)}\right]$$

(provided it exists finite) using (9), (10) and (11), namely

$$RMSE_{\hat{c}_{GHM}} \approx RMSE_{\hat{c}_{GHM},\delta} = \frac{A}{\sqrt{T}} = c\,\frac{\sigma_{\frac{1}{X}}}{\sqrt{T}} = \frac{c}{\sqrt{T}}\sqrt{Var\left[\frac{h(\theta_t)}{g(\theta_t)}\right]}. \quad (12)$$

Notice that if we consider an arbitrary rescaling of the g density this does not affect the evaluation of the corresponding estimator in terms of $RMSE$.

In order to get a consistent estimator of the last approximated expression of the $RMSE$ it suffices to replace $c = \frac{1}{\mu_{\frac{1}{X}}}$ and $\sigma_{\frac{1}{X}}$ with their sample counterparts namely, respectively, $\hat{c}_{GHM}$ and the sample variance of X_t^{-1} denoted as $s^2_{\frac{1}{X}}$. These replacements yield

$$\widehat{RMSE}_{\hat{c}_{GHM},\delta} = \frac{1}{\sqrt{T}}\; s_{\frac{1}{X}}\hat{c}_{GHM}$$

Finally, using the CLT statement of Theorem 2 we can derive an asymptotic confidence interval for c from the following

$$\begin{aligned}
1-\alpha &\approx Pr\left\{\frac{A}{\sqrt{T}}z_{\frac{\alpha}{2}} \le \frac{\hat{c}_{GHM}-c}{c} \le \frac{A}{\sqrt{T}}z_{1-\frac{\alpha}{2}}\right\}\\
&= Pr\left\{\frac{\hat{c}_{GHM}}{1+\frac{A}{\sqrt{T}}z_{1-\frac{\alpha}{2}}} \le c \le \frac{\hat{c}_{GHM}}{1+\frac{A}{\sqrt{T}}z_{\frac{\alpha}{2}}}\right\}\\
&\approx Pr\left\{\frac{\hat{c}_{GHM}}{1+\frac{s_{\frac{1}{X}}\hat{c}_{GHM}}{\sqrt{T}}z_{1-\frac{\alpha}{2}}} \le c \le \frac{\hat{c}_{GHM}}{1+\frac{s_{\frac{1}{X}}\hat{c}_{GHM}}{\sqrt{T}}z_{\frac{\alpha}{2}}}\right\}\\
&= Pr\left\{\frac{1}{\frac{1}{\hat{c}_{GHM}}+\frac{s_{\frac{1}{X}}}{\sqrt{T}}z_{1-\frac{\alpha}{2}}} \le c \le \frac{1}{\frac{1}{\hat{c}_{GHM}}+\frac{s_{\frac{1}{X}}\hat{c}_{GHM}}{\sqrt{T}}z_{\frac{\alpha}{2}}}\right\}
\end{aligned}$$

where z_α is the α-quantile of a standard normal distribution.

In the next two sections we investigate theoretical properties and practical effectiveness of a particular version of the $\hat{c}_{GHM}$ estimator obtained with a particular choice of $h(\cdot)$ automatically constructed using only a sample from $\tilde{g}$ and the ability to compute of $g(\cdot)$.

In fact, looking at (12) one can immediately argue that an optimal choice of h which would lead to a perfect estimator of c by means $\hat{c}_{GHM}$ would be

$h \propto g$ which, unfortunately, in view of the starting constraint (4), requires that we should know c in advance.

Theoretically one could look for an ϵ-optimal solution of the following variational problem

$$\inf_{\{h(\cdot):\int h(\theta)d\theta=1\}} Var\left[\frac{h(\theta)}{g(\theta)}\right] \tag{13}$$

whose optimal solution would certainly give a zero variance estimator achievable in the limit, as already argued above, at the unaffordable price of knowing c in advance.

However, one can think of constructively defining a function h which exactly satisfies the constraint (4) and only approximately produce $h \propto g$ without requiring to know c in advance. In fact, although we derived our idea from a slightly different perspective we eventually realized that it fitted into the $\hat{c}_{GHM}$ framework.

Our original idea was to build up perturbations of $\tilde{g}$ without knowing the total mass of g perturbing g into g_P so that the total mass of g_P has some known functional relation to that of g. For instance one can *inflate* g parametrically so that g_{P_k} is such that

$$\int g_{P_k}(\theta)d\theta = c + k \tag{14}$$

where k is a known constant which can be arbitrarily fixed possibly producing (for small values of k) a perturbed g_P which closely resembles the original c.

When this is possible one can get the desired h satisfying (4) as follows

$$h_k(\theta) = \frac{g_{P_k}(\theta) - g(\theta)}{k} \tag{15}$$

In fact, given a completely known density $g(\theta)$, whose unknown total mass c must be evaluated, there exist several ways to obtain another completely known density such as $g_{P_k}(\theta)$. One naive idea would be that one can just add a known probability density $p(\theta)$ multiplied by a known constant factor k so that

$$g_{P_k}(\theta) = g(\theta) + k\, p(\theta). \tag{16}$$

However, this would lead to a trivial solution which gives back $h(\theta) = p(\theta)$ which does not depend on k at all.

An alternative effective idea when $\Theta = \mathbb{R}$ can be the Hyperplane Inflation method of Petris and Tardella (2003) where the perturbed density g_{P_k} can be constructed as follows:

$$g_{P_k}(\theta) = \begin{cases} g(\theta + r_k) & \theta < -r_k \\ g(0) & -r_k < \theta < r_k \\ g(\theta - r_k) & \theta < r_k \end{cases}$$

where

$$2r_k = \frac{k}{g(0)}$$

represents the length of the interval centered (w.l.o.g.) around the origin which allows to have a new density g_{P_k} with the same total mass of g plus the fixed quantity k which can be easily proved since the density g_{P_k}, outside the central interval $(-d_k, d_k)$, is nothing but a horizontal shift of the original g. In fact, in the general case when $\Theta = \mathbb{R}^n$ a similar trick can be implemented relying on the following inflating mapping which preserves the Lebesgue measure: for a fixed $k > 0$ the function $\psi_n(\cdot; r)$ is defined for any $\theta \in \mathbb{R}^n$, $\theta \neq 0$, by

$$\psi_n(\theta; r) = \frac{\theta}{|\theta|} \left(|\theta|^n + r^n \right)^{1/n}$$

where $B_r^{(n)} = \{\theta \in \mathbb{R}^n : |\theta| \leq r\}$ is the n-dimensional closed ball of radius r, centered at the origin. Since $\psi_n(\theta; r)$ preserves the Lebesgue measure $\lambda(\cdot)$, it can been shown (see Appendix A) that the following density function

$$g_{P_k}(\theta) \quad = \quad \begin{cases} g(0) & \text{if } \theta \in B_{r_k}^{(n)}, \\ g(\psi_n^{-1}(\theta; r_k)) & \text{if } \theta \notin B_{r_k}^{(n)} \end{cases} \tag{17}$$

has a total mass given by

$$\int_\Theta g_{P_k}(\theta)\, d\theta = c + k$$

where

$$k = g(0) \cdot \lambda(B_{r_k}^{(n)}). \tag{18}$$

In the above sense we can think of g_{P_k} as a parametrically inflated or perturbed version of the original target density.

With this perturbed g_{P_k} we can define the function $h_k(\theta)$ as in (15) so that (4) holds and one can consider the following GHM estimator as in (7)

$$\hat{c}_{GHM} = \frac{1}{\frac{1}{T}\sum_{t=1}^{T}\left[\left(\frac{k\,g(\theta_t)}{g_{P_k}(\theta_t)-g(\theta_t)}\right)^{-1}\right]} \tag{19}$$

Hence one can abandon the variational problem (20) on a function space and try to tackle a simpler unidimensional sub-problem

$$\inf_{\{h_k(\cdot)=\frac{g_{P_k}-g}{k},k>0\}} Var\left[\frac{h_k(\theta)}{g(\theta)}\right]. \tag{20}$$

We quickly mention that there is also the possibility of exploiting the perturbed density g_{P_k} as sampling density so that a simple suitable modification to the straight IS can be used to estimate c using an i.i.d. sample from the normalized distribution $\tilde{g}_{P_k}$. Starting from the following identity

$$c = \frac{k\ E_{\tilde{g}_I(\theta)}\left[\frac{g(\theta)}{g_{P_k}(\theta)}\right]}{1 - E_{\tilde{g}_I(\theta)}\left[\frac{g(\theta)}{g_{P_k}(\theta)}\right]}. \tag{21}$$

simulating $\eta_1, ..., \eta_T$ i.i.d. from $\tilde{g}_I$, a consistent estimate of the unknown quantity namely can be computed as follows

$$\hat{c}_{IDS} = \frac{\frac{k}{T}\sum_{t=1}^{T}\left[\frac{g(\eta_t)}{g_{P_k}(\eta_t)}\right]}{1 - \frac{1}{T}\sum_{t=1}^{T}\left[\frac{g(\eta_t)}{g_{P_k}(\eta_t)}\right]} \tag{22}$$

and we call it *Inflated Density Sampling* estimate just to highlight that it is based on a random sample from a different sampling distribution, namely the inflated g_{P_k}.

Let us compare the two alternative identities (6) and (21). The second identity on one hand has the advantage that the ratio appearing inside the expectation can be suitably controlled so that the ratio can bounded from above by 1 and hence the resulting estimator will be guaranteed (see Section 2) to have always finite variance. Also, if the inflated mass k is relatively small the corresponding inflated density g_{P_k} may resemble very closely (see Figure 4.1) the original g which is a nice prerequisite for a well behaved

Monte Carlo error. On the other hand, it is often the case, for instance in Bayesian computation, that a sample from g is readily available while obtaining a sample from g_{P_k} could require additional computational efforts. For this reason we will not investigate much further the properties of $\hat{c}_{IDS}$ in this paper.

4 Investigation of a new class of GHM estimators with HI perturbation

4.1 Theoretical investigation in one-dimensional problems

We now focus on implementing the new class of estimators (7) with $g_{P_k}(\theta)$ constructed via (15) with (14) implemented with the Hyperplane Inflation (HI) idea originally developed in (Petris and Tardella, 2003).

We first investigate theoretically two simple examples with Normal and Cauchy densities to show how the behavior of the ratio $\frac{g_{P_k}(\theta)}{g(\theta)}$ is suitably controlled and conditions for Theorem 2 are guaranteed. Moreover, practical implementation turned out particularly encouraging for supporting the method based on identity (21) and HI.

Example. Let us consider the very basic case of a univariate ($n = 1$) standard normal density up to a multiplicative constant

$$g(\theta) \;=\; exp\left\{-\frac{\theta^2}{2}\right\} \propto f_{N(0,1)}(\theta).$$

In this case, inflating with a constant bounded density over the interval $(-d_k, d_k)$ around the mode 0 one gets

$$\begin{aligned} g_{P_k}(\theta) &= \begin{cases} 1 & |\theta| \leq d_k \\ exp\left\{-\frac{1}{2}(|\theta| - d_k)^2\right\} & |\theta| > d_k \end{cases} \\ g(\theta) &= exp\left\{-\frac{\theta^2}{2}\right\} \propto f_{N(0,1)}(\theta) \end{aligned}$$

so that the inflated density has the same total mass corresponding to g plus a known quantity namely $k = 2d_k$. Unfortunately, in this case, we do not end up with a bounded ratio since, for $\theta > d_k$

$$\begin{aligned} \frac{g_{P_k}(\theta)}{g(\theta)} &= exp\left\{-\frac{1}{2}[(|\theta| - d_k)^2 - \theta^2]\right\} \\ &= exp\left\{\frac{1}{2}[-\theta^2 - d_k^2 + 2|\theta|d_k + \theta^2]\right\} = exp\left\{-\frac{d_k^2}{2} + |\theta|d_k\right\} \end{aligned}$$

which is increasing with $|\theta|$ and diverging at an exponential rate as $|\theta| \to +\infty$. Yet we end up achieving a positive result for our first proposed procedure. In fact, we eventually aim at having a finite variance of the estimator $\hat{c}_{GHM}$ and this is actually true even if in this case the ratio is unbounded. In fact, since

$$RMSE\left[\hat{c}_{GHM}\right] = \sqrt{Var_{\tilde{g}(\theta)}\left[\frac{g_{P_k}(\theta)}{g(\theta)}\right]} < \infty \Longleftrightarrow E_{\tilde{g}(\theta)}\left[\left(\frac{g_{P_k}(\theta)}{g(\theta)}\right)^2\right] < \infty \tag{23}$$

we have that the second moment of the ratio can be written as

$$\begin{aligned} E_{\tilde{g}(\theta)}\left[\left(\frac{g_{P_k}(\theta)}{g(\theta)}\right)^2\right] &= \frac{1}{\sqrt{2\pi}}\left[\int_{\{|\theta|<=d_k\}} \frac{1}{g(\theta)}d\theta + \int_{\{|\theta|>d_k\}} \frac{g(|\theta| - d_k)^2}{g(\theta)}d\theta\right] && (24) \\ &= \frac{1}{\sqrt{2\pi}}\left[\int_{\{|\theta|<=d_k\}} exp\left\{\theta^2\right\}d\theta + \right. && (25) \\ &\quad \left. + \int_{\{|\theta|>d_k\}} exp\left\{-\frac{\theta^2}{2} - d_k^2 + 2|\theta|d\right\}d\theta\right] && (26) \end{aligned}$$

where the first integral is finite since the integrand function is continuous over a bounded closed interval. However, one must immediately realize that the choice of k and hence d_k will affect the performance of the $\hat{c}_{GHM}$ estimator since the ratio $g_{P_k}^2/g$ yields very large values for very large d_k when a constant value of g_{P_k} is divided by a very little quantity corresponding to the tail of the Normal density (see Figure 4.1) .

In fact one can also show that in some particular cases such as the Cauchy density the original ratio $g_{P_k}(\theta)/g(\theta)$ itself is actually bounded.

Example. Let

$$g(\theta) = \frac{1}{1+\theta^2}$$

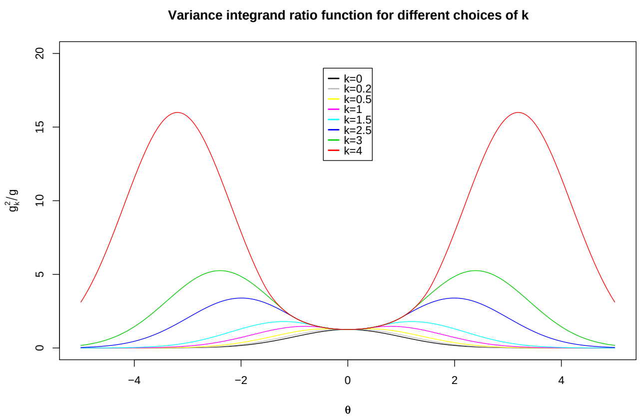

Figure 1: *Ratio $g^2_{P_k}(\theta)/g(\theta)$ for different choices of k in the Normal case*

then, inflating again with a constant bounded density over the interval $(-d_k, d_k)$ around the mode 0, one gets

$$g_{P_k}(\theta) \quad \propto \quad \begin{cases} 1 & |\theta| \leq d_k \\ \frac{1}{1+(|\theta|-d_k)^2} & |\theta| > d_k \end{cases}$$

and hence for $|\theta| > c$

$$\frac{g_{P_k}(\theta)}{g(\theta)} \quad = \quad \frac{1 + (|\theta| - d_k)^2}{1 + \theta^2}$$

which is in fact bounded (see Figure 4.1).

In fact, in the univariate case, one can derive some useful general sufficient conditions as in the following result.

Lemma 1. *Suppose $g(\theta)$ a continuous positive log-Lipschitz density of a finite measure on $\mathbb{R}$ and $g_{P_k}(\theta)$ is the transformed density obtained through the inflating transformation as in Petris and Tardella (2003). Then*

$$Var_{\tilde{g}(\theta)} \left[\frac{g_{P_k}(\theta)}{g(\theta)} \right] < \infty \tag{27}$$

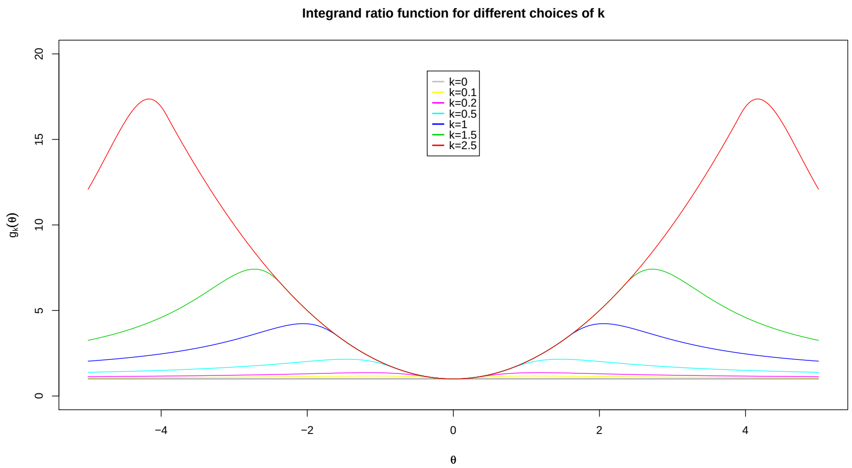

Figure 2: *Ratio $g_{P_k}(\theta)/g(\theta)$ for different choices of k in the Cauchy case*

Proof - It suffices to show that

$$E_{\tilde{g}(\theta)}\left[\left(\frac{g_{P_k}(\theta)}{g(\theta)}\right)^2\right] < \infty \tag{28}$$

which basically depends on the tail behavior of the ratio $\frac{g(|\theta|-d)}{g(\theta)}$ as $|\theta| \to \infty$. In particular, if $g(\theta)$ is log-Lipschitz then for sufficiently large θ

$$|\log g(\theta - d) - \log g(\theta)| \leq Cd$$

hence

$$\int_{\theta_0}^{\infty}\left(\frac{g(\theta-d)}{g(\theta)}\right)^2 g(\theta)d\theta \leq \int_{\theta_0}^{\infty} e^{2Cd}\tilde{g}(\theta)d\theta \leq Cd < \infty$$

$\diamond$

For instance the log-Lipschitz condition is met for the Laplace distribution. In cases where log-Lipschitz condition is not met as in the Gaussian case one can derive an alternative sufficient condition.

Lemma 2. *Suppose $g(\theta)$ a continuous positive log-concave density of a finite measure on $\mathbb{R}$ and $g_{P_k}(\theta)$ is the transformed density obtained through the inflating transformation as in Petris and Tardella (2003). Suppose that*

$$E_{\tilde{g}}\left[e^{2d\frac{d}{d\theta}\log g(\theta)}\right] < \infty$$

then

$$Var_{\tilde{g}(\theta)}\left[\frac{g_{P_k}(\theta)}{g(\theta)}\right] < \infty \tag{29}$$

Proof - It suffices to show that

$$E_{\tilde{g}(\theta)}\left[\left(\frac{g_{P_k}(\theta)}{g(\theta)}\right)^2\right] < \infty \tag{30}$$

which basically depends on the tail behavior of the ratios $\frac{g(|\theta|-d)^2}{g(\theta)}$ as $|\theta| \to \infty$. In particular, since log-concave densities are strongly unimodal and certainly decreasing for sufficiently large values $\theta > \theta_0$ w.l.o.g. consider the behavior of

$$\int_{\theta_0}^{\infty} \frac{g(\theta-d)^2}{g(\theta)} d\theta$$

Since g is eventually decreasing and from log-concavity, for any $\alpha \in (0,1)$

$$\log g(\theta - \alpha d) = \log g((1-\alpha)\theta + \alpha(\theta - d)) \quad \geq \quad (1-\alpha)\log g(\theta) + \alpha \log g(\theta - d)$$

which for $\alpha = \frac{1}{k}$ with arbitrary $k > 0$ gives

$$g\left(\theta - \frac{d}{k}\right) \quad \geq \quad g(\theta)^{\frac{k-1}{k}} \cdot g(\theta - d)^{\frac{1}{k}}$$

or, equivalently,

$$\frac{g\left(\theta - \frac{d}{k}\right)}{g(\theta)} \quad \geq \quad g(\theta)^{\frac{k-1}{k}-1} \cdot g(\theta-d)^{\frac{1}{k}}$$

that is

$$\left[\frac{g(\theta-\frac{d}{k})}{g(\theta)}\right]^{2k} \quad \geq \quad g(\theta)\left[\frac{g(\theta-d)}{g(\theta)}\right]^2$$

Then, taking the *lhs* to the limit with $k \to \infty$ one gets

$$\left[\frac{g(\theta-d)}{g(\theta)}\right]^2 \quad \leq \quad e^{-(2d)\frac{d}{d\theta}\log g(\theta)}$$

which establish our claim.

$\diamond$

As an aside we just mention that the advantage of $\hat{c}_{IDS}$ derived from the alternative identity (21) is that the integrand function obtained as a ratio could be more easily bounded above from 1 and hence the ensuing estimator can be guaranteed to have finite variance with no restricting assumption whatsoever. This is immediate to verify whenever the inflated $g_{P_k}(\theta)$ is constructed by summing $g(\theta)$ with a density function $h(\theta)$ which possesses known total mass k. If the inflated $g_{P_k}(\theta)$ is constructed via Hyperplane Inflation method of Petris and Tardella (2003) the ratio can be similarly bounded when g is unimodal simply recentering the the posterior mode around 0.

Another advantage of the second method derived from the alternative identity (21) with the inflated $g_{P_k}(\theta)$ constructed via the Hyperplane Inflation idea of Petris and Tardella (2003) is that one can exploit regeneration techniques to estimate the MC error more accurately when an MCMC simulating scheme is used for $\eta_1, ..., \eta_T$.

4.2 Theoretical and practical implementation with parametric tuning of the perturbation

We investigate in more depth the behavior of the $\hat{c}_{GHM}$ estimator implemented with the HI perturbation as a function of k. On one hand it is easy to understand that as $k \to 0$ the inflated density tends to approximate the original density in such a way that their ratio becomes approximately one. This implies that the variance of the empirical mean of the ratio of densities $W_t = \frac{g(\theta_t)}{g_{P_k}(\theta_t)}$ namely $\hat{\mu}_W = \frac{k}{T}\sum_{t=1}^{T}\left[\frac{g(\theta_t)}{g_{P_k}(\theta_t)}\right]$ becomes as small as possible. However, the same empirical mean is an unbiased estimator of $\mu_W = \frac{c+k}{c}$ which, in turn, approaches 1 as $k \to 0$ so that very little deviations from 1 at the denominator of

$$\hat{c}_{GHM} = \frac{k}{\hat{\mu}_W - 1}$$

might have a very large impact on the variability of the whole ratio even for a small numerator.

To understand this trade-off better one may focus on the last expression of the approximate relative mean square error of $\hat{c}_{GHM}$ obtained in (12) which can be now written as

$$RMSE_{\hat{c}_{GHM}} \approx \frac{c}{k}\sqrt{Var\left[\frac{g_{P_k}(\theta_t)}{g(\theta_t)}\right]}.$$

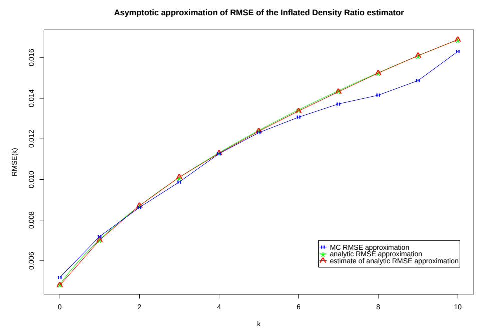

Figure 3: *Cauchy case: comparison of different evaluations of the Relative Mean Square Error of* $\hat{c}_{GHM}$ *as a function of* k
. The Monte Carlo approximation is obtained with sample size $T = 10000$

The approximate expression of the RMSE highlights explicitly which is the trade-off for the optimal choice of k. If on one hand $Var\left[\frac{g_{P_k}(\theta_t)}{g(\theta_t)}\right]$ would favor as little values of k as possible on the other hand $\frac{1}{k}$ acts in the opposite direction.

We now see how one can investigate simple cases such as the former Cauchy example and get an explicit expression for the approximate RMSE which would yield a better first understanding of the real performance of a possible general approximate optimal calibration of k to be used in practical applications.

$$\pi\sqrt{\frac{k\pi}{12}+\frac{\pi^2}{8}-1} \tag{C}$$

In fact, from the approximation in (C) we should choose k as close as possible to 0. We cannot of course choose k exactly equal to 0 for implementing $\hat{c}_{GHM}(k)$ because this would yield an indeterminate form in (6).

However, the following practical approach can be carried out: one can use the same sample to get different $\hat{c}_{GHM}(k)$ approximations and then choose the optimal k^* for which $RMSE_{\hat{c}_{GHM}}(k^*)$ is minimum.

We have verified in this Cauchy case (see Figure 4.2) that $RMSE_{\hat{c}_{GHM}}(k)$ as evaluated either by our estimate $\widehat{RMSE}^{MC}_{\hat{c}_{GHM}}(k)$ or by its analytic asymptotic approximation $RMSE_{\hat{c}_{GHM},\delta}(k)$ in (C) and or by the MC estimate

k	$RMSE_{\hat{c}_{GHM},\delta}(k)$	$\widehat{RMSE}_{\hat{c}_{GHM},\delta}(k)$	$\widehat{RMSE}$
10^{-4}	0.0052	0.0049	0.0048
1	0.0072	0.0071	0.0070
2	0.0086	0.0088	0.0087
3	0.0099	0.0102	0.0101
4	0.0113	0.0114	0.0113
5	0.0123	0.0125	0.0124
6	0.0131	0.0135	0.0134
7	0.0137	0.0144	0.0143
8	0.0142	0.0152	0.0152
9	0.0149	0.0160	0.0161
10	0.0163	0.0167	0.0169

Table 1: *Cauchy example where $\tilde{g}$ is $Cauchy(0,1)$ and $c = \pi$. Comparison for different values of k of RMSE approximation (theoretical and empirical) of formula (12), actual RMSE evaluated by 100 Monte Carlo replications using the true unknown value of $c = \pi$. Our $\widehat{RMSE}_{\hat{c}_{GHM},\delta}(k)$ estimate is obtained with sample size $T = 10000$.*

$\widehat{RMSE}$ of the exact RMSE estimated by 100 Monte Carlo replications of its original definition using the known true value of c are pretty close and the optimal

$$RMSE(k^*) = \inf_{k>0} RMSE(k) \approx \frac{1}{\sqrt{10000}} \sqrt{\frac{\pi^2}{8} - 1} = 0.004834$$

is actually obtained (and approximately achieved) for very little values of k. In fact, the behavior of $\widehat{RMSE}_\delta$ has been verified to be quite stable within the range $(10^{-2}, 10^{-13})$ and starts showing clear signs of numerical instability only for $k < 10^{-13}$ (see Table 4.2).

4.3 Performance in high-dimensional problems

We decided to check whether the proposed approach could be successfully implemented in very high-dimensional integration. In particular we have used as a benchmark the normalizing constant of a 100-dimensional multivariate normal with independent component was used.

The following table was obtained for different values of the inflated mass k. Surprisingly the optimal calibration of k led to a very large optimal k^* value. Despite that, we observe that a pretty stable behavior is shown within a reasonably wide range of so large values of k. Also, notice that the corresponding huge k values correspond to radii d_k in the range of 5 up to 10. Overall the simulation experiment confirms that the new method can perform well even in very high dimension.

	klist	chat	rmse.chat	chat.low	chat.upp	asy.sd	size
10^{5}		0.27	1.00	0.07	-0.13	316.23	100000.00
10^{6}		0.37	0.52	0.14	-0.65	165.28	100000.00
10^{7}		0.52	0.40	0.24	-2.80	125.19	100000.00
10^{8}		1.05	0.32	0.54	26.10	101.16	100000.00
10^{9}		1.04	0.20	0.65	2.61	63.48	100000.00
10^{10}		0.90	0.12	0.66	1.41	38.10	100000.00
10^{11}		1.07	0.08	0.85	1.42	26.28	100000.00
10^{12}		0.93	0.05	0.80	1.11	16.96	100000.00
10^{13}		1.04	0.04	0.93	1.17	12.05	100000.00
10^{14}		0.97	0.03	0.90	1.05	8.52	100000.00
10^{15}		0.97	0.02	0.92	1.04	6.41	100000.00
10^{16}		1.03	0.02	0.98	1.08	5.10	100000.00
10^{17}		1.01	0.01	0.98	1.05	4.08	100000.00
10^{18}		0.99	0.01	0.96	1.03	3.42	100000.00
10^{19}		1.01	0.01	0.98	1.04	3.07	100000.00
10^{20}		1.01	0.01	0.99	1.04	2.87	100000.00
10^{21}		1.01	0.01	0.98	1.04	2.83	100000.00
10^{22}		0.98	0.01	0.95	1.01	2.94	100000.00
10^{23}		1.00	0.01	0.97	1.03	3.29	100000.00
10^{24}		1.00	0.01	0.97	1.04	3.94	100000.00
10^{25}		1.01	0.02	0.97	1.06	4.98	100000.00
10^{26}		1.02	0.02	0.96	1.09	6.76	100000.00
10^{27}		1.03	0.03	0.94	1.14	9.81	100000.00
10^{28}		0.98	0.05	0.86	1.15	14.88	100000.00
10^{29}		1.01	0.08	0.81	1.32	24.86	100000.00
10^{30}		1.02	0.13	0.74	1.66	40.63	100000.00

5 Real data applications

In this section we have applied our approach to two literature examples which are used as illustrative applications in `MCMCpack` a contribute R package for Bayesian computation. The first one deals with the birth weight data in Hosmer and Lemeshow (1989) where an linear model is considered to verify the statistical relevance of factors like mother's age, race, weight at last menstrual period and smoking habits in determining birth weight of children. The second one is a data set from Ornstein (1976) where negative binomial regression analysis is considered to explain the number of interlocking director and executive positions shared with other major firms of different countries.

The `MCMCpack` readily provides two alternative methods to evaluate the marginal likelihood with either the Laplace method or the original Chib method (Chib, 1995) in the first case where Gibbs Sampling is implemented to simulate a MCMC sample from the posterior distribution while there is no method readily available in the package for the latter case where an all-purpose Metropolis Hastings technique is used to sample from the posterior.

5.1 Birth weight data

With the help of the `MCMCregress` function one immediately gets two posterior samples of size 50000 from two alternative models

```
model1 <- MCMCregress(bwt~age+lwt+as.factor(race) + smoke + ht, data=birthwt,
b0=c(2700, 0, 0, -500, -500, -500, -500),
B0=c(1e-6, .01, .01, 1.6e-5, 1.6e-5, 1.6e-5,1.6e-5), c0=10, d0=4500000,
marginal.likelihood="Chib95", mcmc=50000)
model2 <- MCMCregress(bwt~age+lwt+as.factor(race) + smoke,
data=birthwt,
b0=c(2700, 0, 0, -500, -500, -500),
B0=c(1e-6, .01, .01, 1.6e-5, 1.6e-5, 1.6e-5),
c0=10, d0=4500000,
marginal.likelihood="Chib95", mcmc=50000)
```

hence one can get an approximation of the Bayes Factor with the function `BayesFactor` as follows

```
> BF = BayesFactor(model1, model2)
> print(BF)
The matrix of Bayes Factors is:
           posterior1 posterior2
```

```
posterior1      1.000       14.1
posterior2      0.071        1.0

The matrix of the natural log Bayes Factors is:
           posterior1 posterior2
posterior1       0.00       2.64
posterior2      -2.64       0.00

 posterior1 :
   call =
MCMCregress(formula = bwt ~ age + lwt + as.factor(race) + smoke +
    ht, data = birthwt, mcmc = 50000, b0 = bb0, B0 = BB0, c0 = cc0,
    d0 = dd0, marginal.likelihood = "Chib95")

   log marginal likelihood =  -1505.270

 posterior2 :
   call =
MCMCregress(formula = bwt ~ age + lwt + as.factor(race) + smoke,
    data = birthwt, mcmc = 50000, b0 = bb0, B0 = BB0, c0 = cc0,
    d0 = dd0, marginal.likelihood = "Chib95")

   log marginal likelihood =  -1507.915
```

We then use our method to see how it behaves in this real data context. A simple pre-processing of the MCMC sample was carried out so that the origin coincides with the empirical posterior mode and the correlation matrix was standardized. We have then taken this into account in the logposterior evaluation with the further multiplication on the logposterior density for an appropriate constant so that the rescaled logposterior is equal to 0 at the origin. This preprocessing can be easily performed under any circumstances and improve the chances to get a better shaped starting density g with an optimal additional constant k^* (and corresponding radius of hyperspherical opening d_k) within reasonably computable ranges. The computational cost of this reshaping is quite moderate. From the posterior sample in `model1` and `model2` we get the figures in Table 5.1 and Table 5.1 from which one can derive a Bayes factor in favor of the largest model of 14.07 which is quite close to the one obtained with the standard methods implemented in `MCMCpack`.

	klist	log.chat	rmse.chat	log.chat.low	log.chat.upp	asy.sd	size
1	1e-05	−1505.05	0.04	−1505.16	−1504.91	9.20	50000.00
2	1e-04	−1505.05	0.04	−1505.16	−1504.91	9.20	50000.00
3	1e-03	−1505.05	0.04	−1505.16	−1504.91	9.22	50000.00
4	1e-02	−1505.05	0.04	−1505.17	−1504.91	9.40	50000.00
5	1e-01	−1505.13	0.10	−1505.39	−1504.78	22.15	50000.00
6	1e+00	−1505.21	0.06	−1505.38	−1505.01	13.60	50000.00
7	1e+01	−1505.25	0.04	−1505.36	−1505.14	7.94	50000.00
8	1e+02	−1505.27	0.02	−1505.33	−1505.21	4.47	50000.00
9	1e+03	−1505.28	0.01	−1505.31	−1505.24	2.53	50000.00
10	1e+04	−1505.27	0.01	−1505.29	−1505.25	1.81	50000.00
11	1e+05	−1505.29	0.01	−1505.33	−1505.25	2.64	50000.00

	klist	log.chat	rmse.chat	log.chat.low	log.chat.upp	asy.sd	size
1	1e-05	−1507.74	0.04	−1507.84	−1507.63	7.89	50000.00
2	1e-04	−1507.74	0.04	−1507.84	−1507.63	7.89	50000.00
3	1e-03	−1507.74	0.04	−1507.84	−1507.63	7.93	50000.00
4	1e-02	−1507.75	0.04	−1507.86	−1507.63	8.42	50000.00
5	1e-01	−1507.84	0.07	−1508.03	−1507.61	15.43	50000.00
6	1e+00	−1507.83	0.04	−1507.93	−1507.72	7.84	50000.00
7	1e+01	−1507.89	0.02	−1507.95	−1507.83	4.73	50000.00
8	1e+02	−1507.91	0.01	−1507.95	−1507.88	2.70	50000.00
9	1e+03	−1507.93	0.01	−1507.95	−1507.91	1.62	50000.00
10	1e+04	−1507.91	0.01	−1507.93	−1507.89	1.51	50000.00
11	1e+05	−1507.94	0.02	−1508.00	−1507.86	5.13	50000.00

5.2 Ornstein data

The data comes from an investigation (Hosmer and Lemeshow, 1989) on the boards and executives of the largest Canadian corporations and one objective is to explore factors affecting board interlocks. We can then fit a Bayesian negative binomial regression to the Ornstein data explaining the count of interlocking director and executive positions shared with other major firms. Within the `MCMCpack` one can easily derive with very few programming an MCMC sample with the help of the `MCMCmetrop1R()` function which implements a random walk Metropolis algorithm to sample from an arbitrary user-defined log-posterior density. We can use this as an example where, in the absence of available full-conditionals and not built-in resources to compute marginal likelihood and Bayes factors, one can exploit the same user-defined log-posterior density to implement our approach. In this case we use alternative design matrices to determine the relevance of the `sector` factor compared to the `nation` factor we get the following results which suggest to use the only `sector` factor as covariate. This is another example where our approach can be successfully implemented in a moderately large dimensional (up to 14 dimension) data set.

6 Connections with some more recent solutions

This section is devoted to explain how the new strategy can be related to some of the most popular recent techniques used in the context of estimating normalizing constants. As already highlighted earlier our inflated density ratio estimator $\hat{c}_{GHM}$ is just a particular instance of the generalized harmonic mean estimator originally developed in Newton and Raftery (1994). We will illustrate now how Bridge Sampling (BS) and Path Sampling can be related to our proposal.

Bridge Sampling (BS)

Although in their review paper Gelman and Meng (1998) claimed that in some particular case the $\hat{c}_{GHM}$ estimator can be seen as a particular case of Bridge sampling one can rather take the opposite perspective and argue that the BS estimator is a particular instance of a ratio of two GHM estimators.

In fact, if we set

$$\begin{aligned} c_0 &= \int g_0(\theta)d\theta \\ \tilde{g}_0(\theta) &= \frac{g_0(\theta)}{c_0} \\ c_1 &= \int g_1(\theta)d\theta \\ \tilde{g}_1(\theta) &= \frac{g_1(\theta)}{c_1} \end{aligned}$$

and use a *bridge* function $\alpha(\theta)$ satisfying

$$\int \alpha(\theta)g_0(\theta)g_1(\theta)d\theta < \infty$$

we get

$$r = \frac{c_1}{c_0} = \frac{\int g_1(\theta)\alpha(\theta)\tilde{g}_0(\theta)d\theta}{\int g_0(\theta)\alpha(\theta)\tilde{g}_1(\theta)d\theta}$$

so that one can estimate the ratio c_1/c_0 of the two normalizing constants as follows

$$\hat{r}^{BS}_{\frac{c_1}{c_0}} = \frac{\frac{1}{N_0}\sum_{i=1}^{N_0} g_1(\theta_{0,i})\alpha(\theta_{0,i})}{\frac{1}{N_1}\sum_{j=1}^{N_1} g_0(\theta_{1,j})\alpha(\theta_{1,j})}$$

with a suitable number N_0 of simulations $\theta_{0,i}$ from $\tilde{g}_0$ and number N_1 of simulations $\theta_{1,j}$ from $\tilde{g}_1$. Hence, to see how BS can be seen as a particular case of GHM we can use

$$h(\theta) = \alpha(\theta)g_0(\theta)g_1(\theta)$$

and combine two GHM estimators as in (7) with g_1 and g_0 respectively instead of g then the ratio

$$\frac{\hat{c}_{GHM,1}}{\hat{c}_{GHM,0}}$$

is nothing but the BS estimator $\hat{r}^{BS}$. Notice that in both $\hat{c}_{GHM,1}$ and $\hat{c}_{GHM,0}$ we are using the same function h satisfying (4) which is in fact unnecessary. With a different choice of h_1 and h_0 one could get some additional advantage in terms of efficiency of the resulting ratio estimator.

Notice also that as a matter of fact for the BS as well as for the GHM strategy the starting point is the existence of a finite integral for the h function although it need not be a known quantity as in (4) for the same finite quantity is used both at the numerator and at the denominator and hence it cancels out.

Path Sampling (PS)

In this section we will establish a formal rigorous limit connection between the PS estimator and our $\hat{c}_{GHM}$.

In fact, if we regard the $g_{P_k}(\theta) = g(\theta, k)$ in (17) as an explicit function of k and θ and notice that $g(\theta, 0) = g(\theta)$ and consider the limit of the function appearing in (15) when $k \to 0$ we discover that the resulting expression

$$\left.\frac{\partial}{\partial k} g(\theta, k)\right|_{k=0} = \lim_{k\to 0} h_k(\theta) = \lim_{k \to 0} \frac{g(\theta, k) - g(\theta, 0)}{k}$$

is nothing but the partial derivative of $g(\theta, k)$ evaluated at $k = 0$. Hence, taking the limit expression of (6) we get

$$c = Har_{\tilde{g}}\left[\frac{\frac{\partial}{\partial k} g(\theta, k)\big|_{k=0}}{g(\theta, 0)}\right]$$

or, equivalently

$$E_{\tilde{g}}\left[\frac{\partial}{\partial k}\log g(\theta,k)\bigg|_{k=0}\right]=\frac{1}{c}$$

which is in close connection to the corresponding PS identity for which

$$\log\left(\frac{c+k}{c}\right)=E_{\tilde{\tilde{g}}(\theta,\lambda)}\left[\frac{\partial}{\partial\lambda}\log g(\theta,\lambda)\right]$$

where the expectation is rather defined with respect to the joint distribution of (θ,λ)

$$\tilde{\tilde{g}}(\theta,\lambda)\propto g(\theta,\lambda)\pi(\lambda)$$

obtained normalizing the product of the original target and a uniform marginal density $\pi(\cdot)$ for $\lambda\in[0,k]$. In fact, taking the limit

$$\lim_{k\to 0}\frac{\log\left(\frac{c+k}{c}\right)}{k}=\lim_{k\to 0}\frac{\log(c+k)-\log c}{k}=\lim_{k\to 0}\frac{E_{\tilde{\tilde{g}}(\theta,\lambda)}\left[\frac{\partial}{\partial\lambda}\log g(\theta,\lambda)\right]}{k}=\frac{1}{c}$$

we can get our $\hat{c}_{GHM}$ as a limit case of the PS estimator.

References

Francesco Bartolucci, Luisa Scaccia, and Antonietta Mira. Efficient Bayes factor estimation from the reversible jump output. *Biometrika*, 93(1):41–52, 2006.

Ming-Hui Chen. Importance-weighted marginal bayesian posterior density estimation. *Journal of the American Statistical Association*, 89:818–824, 1994.

Ming-Hui Chen and Qi-Man Shao. On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25(4):1563–1594, 1997.

Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1995.

Siddhartha Chib and Ivan Jeliazkov. Accept-Reject Metropolis-Hastings sampling and marginal likelihood estimation. *Statistica Neerlandica*, 59(1): 30–44, 2005.

Siddhartha Chib and Ivan Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

M. A. Clyde, J. O. Berger, F. Bullard, W. H. Jeffreys, R. Luo, R. Paulo, and T. Loredo. Current challanges in bayesian model choice. In ., page ., 2006.

A. E. Gelfand and D. K. Dey. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B: Methodological*, 56:501–514, 1994.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, 13(2):163–185, 1998. ISSN 0883-4237.

Cong Han and Bradley P. Carlin. Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association*, 96(455):1122–1132, 2001.

D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, 1989.

Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.

Michael A. Newton and Adrian E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap (Disc: P26-48). *Journal of the Royal Statistical Society, Series B: Methodological*, 56:3–26, 1994.

Nilan Norris. The standard errors of the geometric and harmonic means and their application to index numbers. *Ann. Statist.*, 11(1):445–448, 1940.

M. Ornstein. The boards and executives of the largest canadian corporations. *Canadian Journal of Sociology*, 1:411–437, 1976.

A. G. Pakes. On the convergence of moments of geometric and harmonic means. *Statistica Neerlandica*, 53(1):96–110, 1999.

Giovanni Petris and Luca Tardella. A geometric approach to transdimensional Markov chain Monte Carlo. *Canad. J. Statist.*, 31(4):469–482, 2003. ISSN 0319-5724.

Adrian E. Raftery, Michael A. Newton, Jaya M. Satagopan, and Pavel N. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In J. M. Bernardo, Bayarry M. J., J. O. Berger, A. P. Dawid, Heckerman D., A. F. M. Smith, and West M., editors, *Bayesian Statistics 8 – Proceedings of the Eighth Valencia International Meeting*, pages 1–45. Oxford University Press, 2006.

John Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–860, 2006.