Marginal Likelihood Estimation via Arrogance Sampling

By Benedict Escoto

Abstract

This paper describes a method for estimating the marginal likelihood or Bayes factors of Bayesian models using non-parametric importance sampling ("arrogance sampling"). This method can also be used to compute the normalizing constant of probability distributions. Because the required inputs are samples from the distribution to be normalized and the scaled density at those samples, this method may be a convenient replacement for the harmonic mean estimator. The method has been implemented in the open source R package `margLikArrogance`.

1 Introduction

When a Bayesian evaluates two competing models or theories, T_1 and T_2, having observed a vector of observations $\boldsymbol{x}$, Bayes' Theorem determines the posterior ratio of the models' probabilities:

$$\frac{p(T_1|\boldsymbol{x})}{p(T_2|\boldsymbol{x})} = \frac{p(\boldsymbol{x}|T_1)}{p(\boldsymbol{x}|T_2)}\frac{p(T_1)}{p(T_2)}. \tag{1}$$

The quantity $\frac{p(\boldsymbol{x}|T_1)}{p(\boldsymbol{x}|T_2)}$ is called a *Bayes factor* and the quantities $p(\boldsymbol{x}|T_1)$ and $p(\boldsymbol{x}|T_2)$ are called the theories' *marginal likelihoods*.

The types of Bayesian models considered in this paper have a fixed finite number of parameters, each with their own probability function. If $\boldsymbol{\theta}$ are parameters for a model T, then

$$p(\boldsymbol{x}|T) = \int p(\boldsymbol{x}|\boldsymbol{\theta}, T)p(\boldsymbol{\theta}|T)\, d\boldsymbol{\theta} = \int p(\boldsymbol{x} \wedge \boldsymbol{\theta}|T)\, d\boldsymbol{\theta} \tag{2}$$

Unfortunately, this integral is difficult to compute in practice. The purpose of this paper is to describe one method for estimating it.

Evaluating integral (2) is sometimes called the problem of computing normalizing constants. The following formula shows how $p(\boldsymbol{x}|T)$ is a normalizing constant.

$$p(\boldsymbol{\theta}|\boldsymbol{x}, T) = \frac{p(\boldsymbol{\theta} \wedge \boldsymbol{x}|T)}{p(\boldsymbol{x}|T)} \tag{3}$$

Thus the marginal likelihood $p(\boldsymbol{x}|T)$ is also the normalizing constant of the posterior parameter distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, T)$ assuming we are given the density $p(\boldsymbol{\theta} \wedge \boldsymbol{x}|T)$ which is often easy to compute in Bayesian models. Furthermore, Bayesian statisticians typically produce samples from the posterior parameter distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, T)$ even when not concerned with theory choice. In these case, computing the marginal likelihood is equivalent to computing

the normalizing constant of a distribution from which samples and the scaled density at these samples are available. The method described in this paper takes this approach.

2 Review of Literature

Given how basic (1) is, it is perhaps surprising that there is no easy and definitive way of applying it, even for simple models. Furthermore, as the dimensionality and complexity of probability distributions increase, the difficulty of approximation also increases. The following three techniques for computing bayes factors or marginal likelihoods are important but will not be mentioned further here.

1. Analytic asymptotic approximations such as Laplace's method, see for instance Kass and Raftery (1995),
2. Bridge sampling/path sampling/thermodynamic integration (Gelman and Meng, 1998), and
3. Chib's MCMC approximation (Chib, 1995; Chib and Jeliazkov, 2005).

Kass and Raftery (1995) is a popular overview of the earlier literature on Bayes factor computation. All these methods can be very successful in the right circumstances, and can often handle problems too complex for the method described here. However, the method of this paper may still be useful due to its convenience.

The rest of section 2 describes three approaches that are relevant to this paper.

2.1 Importance Sampling

Importance sampling is a technique for reducing the variance of monte carlo integration. This section will note some general facts; see Owen and Zhou (1998) for more information.

Suppose we are trying to compute the (possibly multidimensional) integral I of a well-behaved function $f(\boldsymbol{\theta})$. Then

$$I = \int f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int \frac{f(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} g(\boldsymbol{\theta})\, d(\boldsymbol{\theta})$$

so if $g(\boldsymbol{\theta})$ is a probability density function and $\boldsymbol{\theta}_i$ are independent samples from it, then

$$I = \mathrm{E}_g[f(\boldsymbol{\theta})/g(\boldsymbol{\theta})] \approx \frac{1}{n} \sum_{i=1}^{n} \frac{f(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)} = I_n. \tag{4}$$

I_n is an unbiased approximation to I and by the central limit theorem will tend to a normal distribution. It has variance

$$\mathrm{Var}[I_n] = \frac{1}{n}\int \left(\frac{f(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} - I\right)^2 g(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \frac{1}{n}\int \frac{(f(\boldsymbol{\theta}) - Ig(\boldsymbol{\theta}))^2}{g(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \tag{5}$$

Sometimes f is called the *target* and g is called the *proposal* distribution.

Assuming that f is non-negative, then minimum variance (of 0!) is achieved when $g = f/I$—in other words when g is just the normalized version of f. This cannot be done in practice because normalizing f requires knowing the quantity I that we wanted to approximate; however (5) is still important because it means that the more similar the proposal is to the target, the better our estimator I_n becomes. In particular, f must go to 0 faster than g or the estimator will have infinite variance.

To summarize this section:

1. Importance sampling is a monte carlo integration technique which evaluates the target using samples from a proposal distribution.

2. The estimator is unbiased, normally distributed, and its variance (if not 0 or infinity) decreases as $O(n^{-1})$ (using big-O notation).

3. The closer the proposal is to the target, the better the estimator. The proposal also needs to have longer tails than the target.

2.2 Nonparametric Importance Sampling

A difficulty with importance sampling is that it is often difficult to choose a proposal distribution g. Not enough is known about f to choose an optimal distribution, and if a bad distribution is chosen the result can have large or even infinite variance. One approach to the selection of proposal g is to use non-parametric techniques to build g from samples of f. I call this class of techniques self-importance sampling, or **arrogance sampling** for short, because they attempt to sample f from itself without using any external information. (And also isn't it a bit arrogant to try to evaluate a complex, multidimensional integral using only the values at a few points?) The method of this paper falls into this class and particularly deserves the name because the target and proposal (when they are both non-zero) have exactly the same values up to a multiplicative constant.

Two papers which apply nonparametric importance sampling to the problem of marginal likelihood computation (or computation of normalizing constants) are Zhang (1996) and Neddermeyer (2009). Although both authors apply their methods to more general situations, here I will use the framework suggested by (3) and assume that we can compute $p(\boldsymbol{\theta} \wedge \boldsymbol{x}|T)$ for arbitrary $\boldsymbol{\theta}$ and also that we can sample from the posterior parameter distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, T)$. The goal is to estimate the normalizing constant, the marginal likelihood $p(\boldsymbol{x}|T)$.

Zhang's approach is to build the proposal g using traditional kernel density estimation. m samples are first drawn from $p(\boldsymbol{\theta}|\boldsymbol{x}, T)$ and used to construct g. Then n samples are drawn from g and used to evaluate $p(\boldsymbol{x}|T)$ as in traditional importance sampling. This approach

is quite intuitive because kernel estimation is a popular way of approximating an unknown function. Zhang proves that the variance of his estimator decreases as $O(m^{\frac{-4}{4+d}}n^{-1})$ where d is the dimensionality of $\boldsymbol{\theta}$, compared to $O(n^{-1})$ for standard (parametric) importance sampling.

There were, however, a few issues with Zhang's method:

1. A kernel density estimate is equal to 0 at points far from the points the kernel estimator was built on. This is a problem because importance sampling requires the proposal to have longer tails than the target. This fact forces Zhang to make the restrictive assumption that $p(\boldsymbol{\theta}|\boldsymbol{x}, T)$ has compact support.

2. It is hard to compute the optimal kernel bandwidth. Zhang recommends using a plug-in estimator because the function $p(\boldsymbol{\theta} \wedge \boldsymbol{x}|T)$ is available, which is unusual for kernel estimation problems. Still, bandwidth selection appears to require significant additional analysis.

3. Finally, although the variance may decrease as $O(m^{\frac{-4}{4+d}}n^{-1})$ as m increases, the difficulty of computing $g(\boldsymbol{\theta})$ also increases with m, because it requires searching through the m basis points to find all the points close to $\boldsymbol{\theta}$. In multiple dimensions, this problem is not trivial and may outweigh the $O(m^{\frac{-4}{4+d}})$ speedup (in the worst case, practical evaluation of $g(\boldsymbol{\theta})$ at a single point may be $O(m)$). See Zlochin and Baram (2002) for some discussion of these issues.

Neddermeyer (2009) uses a similar approach to Zhang and also achieves a variance of $O(m^{\frac{-4}{4+d}}n^{-1})$. It improves on Zhang's approach in two ways relevant to this paper:

1. The support of $p(\boldsymbol{\theta}|\boldsymbol{x}, T)$ is not required to be compact.

2. Instead of using kernel density estimators, linear blend frequency polynomials (LBFPs) are used instead. LBFPs are basically histograms whose density is interpolated between adjacent bins. As a result, the computation of $g(\boldsymbol{\theta})$ requires only finding which bin $\boldsymbol{\theta}$ is in, and looking up the histogram value at that and adjacent bins (2^d bins in total).

As we will see in section 3, the arrogance sampling described in this paper is similar to the methods of Zhang and Neddermeyer.

2.3 Harmonic Mean Estimator

The harmonic mean estimator is a simple and notorious method for calculating marginal likelihoods. It is a kind of importance sampling, except the proposal g is actually the distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, T) = p(\boldsymbol{\theta} \wedge \boldsymbol{x}|T)/p(\boldsymbol{x}|T)$ to be normalized and the target f is the known distribution $p(\boldsymbol{\theta}|T)$. Then if $\boldsymbol{\theta}_i$ are samples from $p(\boldsymbol{\theta}|x, T)$, we apparently have

$$1 \approx \frac{1}{n}\sum_{i=1}^{n}\frac{p(\boldsymbol{\theta}_i|T)}{p(\boldsymbol{\theta}_i|\boldsymbol{x},T)} = \frac{1}{n}\sum_{i=1}^{n}\frac{p(\boldsymbol{\theta}_i|T)}{p(\boldsymbol{x}|\boldsymbol{\theta}_i,T)p(\boldsymbol{\theta}_i|T)/p(\boldsymbol{x}|T)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p(\boldsymbol{x}|\boldsymbol{\theta}_i,T)/p(\boldsymbol{x}|T)}$$

hence

$$p(\boldsymbol{x}|T) \stackrel{?}{\approx} \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p(\boldsymbol{x}|\boldsymbol{\theta}_i,T)}\right)^{-1} \tag{6}$$

Two advantages of the harmonic mean estimator are that it is simple to compute and only depends on samples from $p(\boldsymbol{\theta}|x,T)$ and the likelihood $p(\boldsymbol{x}|\boldsymbol{\theta},T)$ at those samples. The main drawback of the harmonic mean estimator is that it doesn't work—as mentioned earlier the importance sampling proposal distribution needs to have longer tails than the target. In this case the target $p(\boldsymbol{\theta}|T)$ typically has longer tails than the proposal $p(\boldsymbol{\theta}|\boldsymbol{x},T)$ and thus (6) has infinite variance. Despite not working, the harmonic mean estimator continues to be popular (Neal, 2008).

3 Description of Technique

This paper's arrogance sampling technique is a simple method that applies the nonparametric importance techniques of Zhang and Neddermeyer in an attempt to develop a method almost as convenient as the harmonic mean estimator.

The only required inputs are samples $\boldsymbol{\theta}_i$ from $p(\boldsymbol{\theta}|\boldsymbol{x},T)$ and the values $p(\boldsymbol{\theta}_i \wedge \boldsymbol{x}|T) = p(x|\boldsymbol{\theta}_i,T)p(\boldsymbol{\theta}_i|T)$. This is similar to the harmonic mean estimator, but perhaps slightly less convenient because $p(\boldsymbol{\theta}_i \wedge \boldsymbol{x}|T)$ is required instead of $p(\boldsymbol{x}|\boldsymbol{\theta}_i,T)$.

There are two basic steps:

1. Take m samples from $p(\boldsymbol{\theta}|\boldsymbol{x},T)$ and using modified histogram density estimation, construct probability density function $f(\boldsymbol{\theta})$.

2. With n more samples from $p(\boldsymbol{\theta}|\boldsymbol{x},T)$, estimate $1/p(\boldsymbol{x}|T)$ via importance sampling with target f and proposal $p(\boldsymbol{\theta}|\boldsymbol{x},T)$.

These steps are described in more detail below.

3.1 Construction of the Histogram

Of the N total samples $\boldsymbol{\theta}_i$ from $p(\boldsymbol{\theta}|\boldsymbol{x},T)$, the first m will be used to make a histogram. The optimal choice of m will be discussed below, but in practice this seems difficult to determine. An arbitrary rule of $\min(0.2N, 2\sqrt{N})$ can be used in practice.

With a traditional histogram, the only available information is the location of the sampled points. In this case we also know the (scaled) heights $p(\boldsymbol{\theta} \wedge \boldsymbol{x}|T)$ at each sampled point. We can use this extra information to improve the fit.

Our "arrogant" histogram f is constructed the same as a regular histogram, except the bin heights are not determined by the number of points in each bin, but rather by the minimum density over all points in the bin. If a bin contains no sampled points, then $f(\boldsymbol{\theta}) = 0$ for $\boldsymbol{\theta}$ in that bin. Then f is normalized so that $\int f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 1$.

To determine our bin width, we can simply and somewhat arbitrarily set our bin width h so that the histogram is positive for 50% of the sampled points from the distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, T)$. To approximate h, we can use a small number of samples (say, 40) from $p(\boldsymbol{\theta}|\boldsymbol{x}, T)$ and set h so that $f(\boldsymbol{\theta}) > 0$ for exactly half of these samples.

Figure 1 compares the traditional and new histograms for a one dimensional normal distribution based on 50 samples. The green rug lines indicate the 50 sampled points which are the same for all. The arrogant histogram's bin width is chosen as above. The traditional histogram's optimal bin width was determined by Scott's rule to minimize mean squared error. As the figure shows, the modified histogram is much smoother for a given bin width, so a smaller bin width can be used. On the other hand, f will either equal 0 or have about twice the original density at each point, while the traditional histogram's density is numerically close to the original density.

3.2 Importance Sampling

The remaining $n = N - m - 40$ sampled points can be used for importance sampling. Using equation (4) with histogram f as our target and $p(\boldsymbol{\theta}|\boldsymbol{x}, T)$ as the proposal, we have

$$1 \approx I_n = \frac{1}{n} \sum_{i=1}^{n} \frac{f(\boldsymbol{\theta}_i)}{p(\boldsymbol{\theta}_i|\boldsymbol{x}, T)} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(\boldsymbol{\theta}_i)}{p(\boldsymbol{\theta}_i \wedge \boldsymbol{x}|T)/p(\boldsymbol{x}|T)}$$

hence

$$p(\boldsymbol{x}|T) \approx p(\boldsymbol{x}|T)/I_n = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{f(\boldsymbol{\theta}_i)}{p(\boldsymbol{\theta}_i \wedge \boldsymbol{x}|T)} \right)^{-1} = A_n \tag{7}$$

To underscore the self-important/arrogant nature of this approximation A_n, we can rewrite (7) as

$$p(\boldsymbol{x}|T) \approx H \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\min\{p(\boldsymbol{\theta}_j \wedge \boldsymbol{x}|T) : \boldsymbol{\theta}_j \text{ and } \boldsymbol{\theta}_j \text{ are in the same bin}\}}{p(\boldsymbol{\theta}_i \wedge \boldsymbol{x}|T)} \right)^{-1}$$

where H is the histogram normalizing constant. This equation shows that all the values in the numerator and the denominator of our importance sampling are from the same distribution $p(\boldsymbol{\theta} \wedge \boldsymbol{x}|T)$.

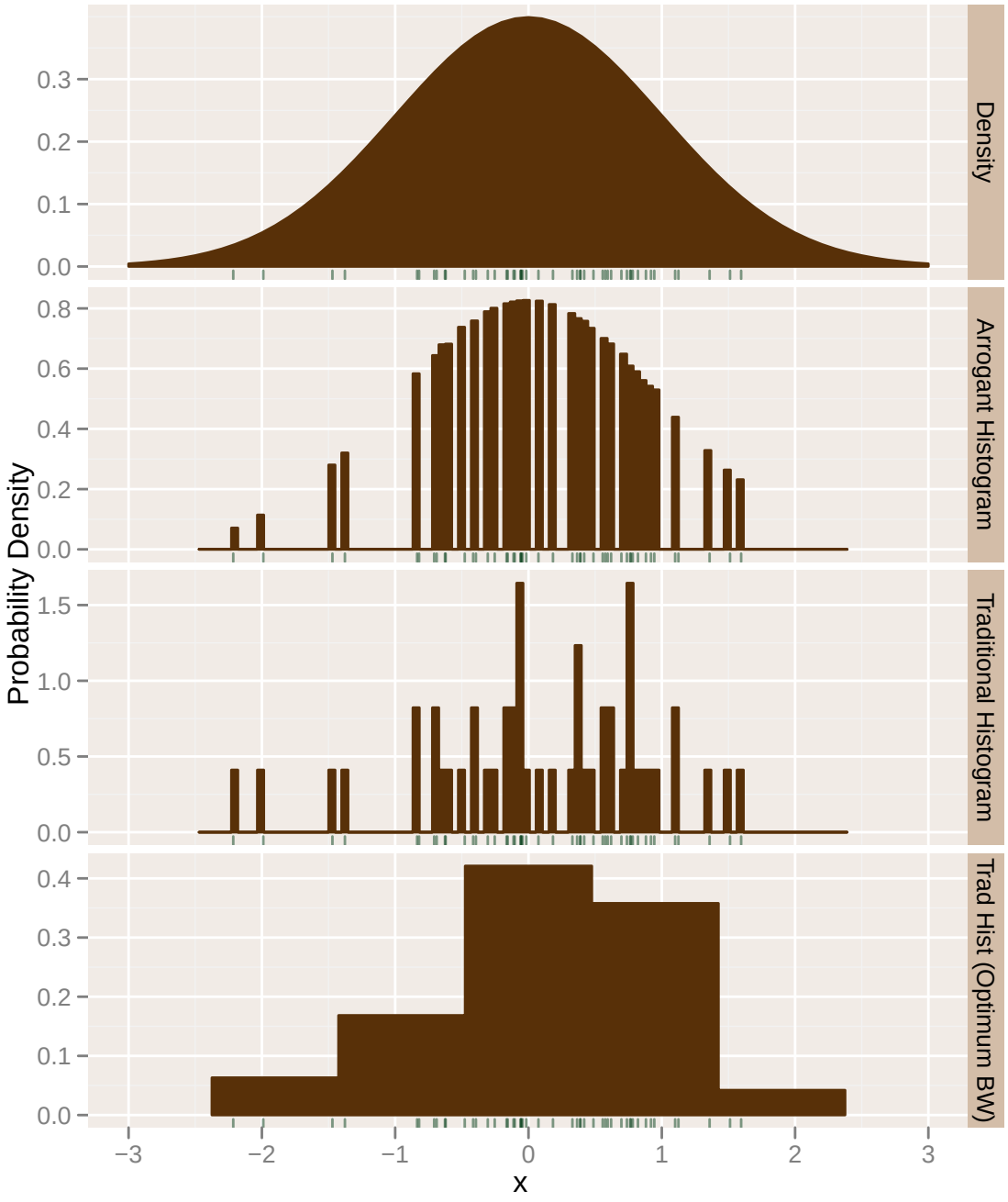

Figure 1: Histogram Comparison

Note that the histogram f is the target of the importance sampling and $p(\boldsymbol{\theta} \wedge \boldsymbol{x}|T)$ is the proposal. This is backwards from the usual scheme where the unknown distribution is the target and the known distribution is the proposal. Instead here the unknown distribution is the proposal, as in the harmonic mean estimator (see Robert and Wraith (2009) for another example of this.)

As in section 2.1, our approximation of $p(\boldsymbol{x}|T)^{-1}$ tends to a normal distribution as $n \to \infty$ by the central limit theorem. This fact can be used to estimate a confidence interval around $p(\boldsymbol{x}|T)$.

4 Validity of Method

This section will investigate the performance of the method. First, note that this method is just an implementation of importance sampling, so A_n^{-1} should converge to $p(\boldsymbol{x}|T)^{-1}$ with finite variance as long as the proposal density $p(\boldsymbol{\theta}|\boldsymbol{x}, T)$ exists and is finite and positive on the compact region where the target histogram density is positive.

To calculate the speed of convergence we will use equation (5) where f is the histogram, $g(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{x}, T)$, and $I = 1$ because the histogram has been normalized. Unless otherwise noted, we will assume below that $g : \mathbb{R}^d \to \mathbb{R}$ is finite, twice differentiable and positive, and that $\int \frac{\|\nabla \cdot g(\boldsymbol{\theta})\|^2}{g(\boldsymbol{\theta})} d\boldsymbol{\theta}$ is finite.

4.1 Histogram Bin Width

One important issue will be how quickly the d-dimensional histogram's selected bin width h goes to 0 as the number of samples $m \to \infty$. This section will only offer an intuitive argument. For any m, the histogram will enclose about the same probability ($\frac{1}{2}$) and will have about the same average density in a fixed region. Each bin has volume h^d, so if l is the number of bins then $lh^d = O(1)$ and $h \propto l^{-d}$.

Furthermore, the distribution of the sampled points converges to the actual distribution $g(\boldsymbol{\theta})$. If $m > O(l)$, an unbounded number of sampled points would end up in each bin. If $m < O(l)$, then some bins would have no points in them. Neither of these is possible because exactly one sampled point is necessary to establish each bin. Thus $m \propto l$ and $h \propto m^{-d}$.

4.2 Conditional Variance

Before estimating the convergence rate of A_n we will prove something about the conditional variance of importance sampling. Let $A = \{\boldsymbol{\theta} : f(\boldsymbol{\theta}) > 0\}$, $\mathbf{1}_A$ be the characteristic function of A, and $q = \int_A g(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. Define

$$g_A(\boldsymbol{\theta}) = \begin{cases} g(\boldsymbol{\theta})/q & \text{if } \boldsymbol{\theta} \in A \\ 0 & \text{otherwise} \end{cases}$$

Then g_A is the density of g conditional on $f > 0$. Define Var_A and E_A to mean the variance and expectation conditional on $f(\boldsymbol{\theta}) > 0$. Thus

$$\begin{aligned}
\mathrm{Var}(f(\boldsymbol{\theta})/g(\boldsymbol{\theta})) &= \mathrm{Var}(\mathrm{E}(f(\boldsymbol{\theta})/g(\boldsymbol{\theta})|\mathbf{1}_A)) + \mathrm{E}(\mathrm{Var}(f(\boldsymbol{\theta})/g(\boldsymbol{\theta})|\mathbf{1}_A)) \\
&= \mathrm{Var}\begin{pmatrix} \mathrm{E}_A(f(\boldsymbol{\theta})/g(\boldsymbol{\theta})) & \text{if } \boldsymbol{\theta} \in A \\ 0 & \text{otherwise} \end{pmatrix} \\
&\quad + \mathrm{E}\begin{pmatrix} \mathrm{Var}_A(f(\boldsymbol{\theta})/g(\boldsymbol{\theta})) & \text{if } \boldsymbol{\theta} \in A \\ 0 & \text{otherwise} \end{pmatrix} \\
&= \mathrm{Var}\begin{pmatrix} 1/q & \text{if } \boldsymbol{\theta} \in A \\ 0 & \text{otherwise} \end{pmatrix} + q\mathrm{Var}_A(f(\boldsymbol{\theta})/g(\boldsymbol{\theta})) \\
&= (1/q)^2 q(1-q) + \frac{1}{q}\mathrm{Var}_A(f(\boldsymbol{\theta})/qg(\boldsymbol{\theta})) \\
&= \frac{1-q}{q} + \frac{1}{q}\mathrm{Var}_A(f(\boldsymbol{\theta})/g_A(\boldsymbol{\theta}))
\end{aligned}$$

We will assume below that $q = \frac{1}{2}$, so that

$$\mathrm{Var}(f(\boldsymbol{\theta})/g(\boldsymbol{\theta})) = 1 + 2\mathrm{Var}_A(f(\boldsymbol{\theta})/g_A(\boldsymbol{\theta})) \tag{8}$$

4.3 Importance Sampling Convergence

With f, g, and A as defined above, f and g_A have the same domain. Assuming errors in estimating q and normalization errors are of a lesser order of magnitude, we can treat the histogram heights as being sampled from g_A. Suppose the histogram has l bins $\{B_j\}$, each with width h and based around the points $g_A(\boldsymbol{\theta}_j)$. Then by equation (5),

$$\begin{aligned}
\mathrm{Var}_A(f(\boldsymbol{\theta})/g_A(\boldsymbol{\theta})) &= \sum_{j=1}^{l} \int_{B_j} \frac{(f(\boldsymbol{\theta}) - g_A(\boldsymbol{\theta}))^2}{g_A(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \sum_{j=1}^{l} \int_{B_j} \frac{(g_A(\boldsymbol{\theta}) + \nabla g_A(\boldsymbol{\theta}) \cdot (\boldsymbol{\theta}_j - \boldsymbol{\theta}) + O((\boldsymbol{\theta}_j - \boldsymbol{\theta})^2) - g_A(\boldsymbol{\theta}))^2}{g_A(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \sum_{j=1}^{l} \int_{B_j} \frac{(\nabla g_A(\boldsymbol{\theta}) \cdot (\boldsymbol{\theta}_j - \boldsymbol{\theta}))^2 + O((\boldsymbol{\theta}_j - \boldsymbol{\theta})^3)}{g_A(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&\leq \sum_{j=1}^{l} \int_{B_j} \frac{\|\nabla \cdot g_A(\boldsymbol{\theta})\|^2 h^2}{g_A(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= h^2 \int \frac{\|\nabla \cdot g_A(\boldsymbol{\theta})\|^2}{g_A(\boldsymbol{\theta})} d\boldsymbol{\theta}
\end{aligned}$$

Because $h \propto m^{-d}$ where d is the number of dimensions, and m is the number of samples used to make the histogram,

$$\mathrm{Var}_A(f(\boldsymbol{\theta})/g_A(\boldsymbol{\theta})) \leq Cm^{-2/d}$$

where $C \propto \int \frac{\|\nabla \cdot g_A(\boldsymbol{\theta})\|^2}{g_A(\boldsymbol{\theta})} d\boldsymbol{\theta}$. Putting this together with (8), we get

$$\mathrm{Var}(I_n) = \mathrm{Var}(p(\boldsymbol{x}|T)/A_n) = n^{-1}(1 + O(Cm^{-2/d})) \tag{9}$$

5 Implementation Issues

5.1 Speed of Convergence

The variance of $n^{-1}(1 + O(Cm^{-2/d}))$ given by (9) is asymptotically equal to n^{-1}, which is the typical importance sampling rate. In practice however, the asymptotic results cannot distinguish useful from impractical estimators. If $Cm^{-2/d}$ is small and $\mathrm{Var}(p(\boldsymbol{x}|T)/A_n) \approx n^{-1}$, then $p(\boldsymbol{x}|T)$ can be approximated in only 1000 samples to about $6\% = \frac{1.96}{\sqrt{1000}}$ with 95% confidence. For many theory choice purposes, this is quite sufficient. Thus in typical problem cases the factor of $Cm^{-2/d}$ will be very significant. If $Cm^{-2/d} \gg 1$, then the convergence rate may in practice be similar to $n^{-1}m^{-2/d}$. Compare this to the rate of $n^{-1}m^{-4/(4+d)}$ for the methods proposed by Zhang and Neddermeyer.

This method also uses simple histograms, instead of a more sophisticated density estimation method (Zhang uses kernel estimation, Neddermeyer uses linear blend frequency polynomials). Although simple histograms converge slower for large d as shown above, they are much faster to compute for large d.

Neddermeyer's LBFP algorithm is quite efficient compared to Zhang's, but its running time is $O(2^d d^2 n^{\frac{d+5}{d+4}})$. d is a constant for any fixed problem, but if, say, $d = 10$, then the dimensionality constant multiplies the running time by $2^{10}10^2 \approx 10^5$.

By contrast, this paper's method takes only $O(dm\log(m))$ time to construct the initial histogram, and an additional $O(dn\log(m))$ time to do the importance sampling. The main reason for the difference is that querying a simple histogram can be done in $\log(m)$ time by computing the bin coordinates and looking up the bin's height in a tree structure. However, querying a LBFP requires blending all nearby bins and is thus exponential in d.

5.2 When $g = 0$

Our discussion assumed that $g(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{x}, T)$ was always positive. If g goes to 0 where the histogram is positive, the variance of A_n^{-1} will be infinite. However, this paper's method can still be used if $g(\boldsymbol{\theta})$ is 0 over some well-defined area.

For instance, suppose one dimension θ_k of $p(\boldsymbol{\theta}|T)$ is defined by a gamma distribution, so that $p(\theta_k|T) = 0$ if and only if $\theta_k \leq 0$. Then we can ensure the variance is not infinite by checking that the histogram is only defined where $\theta_k > \epsilon > 0$ for some fixed ϵ.

The `margLikArrogance` package contains a simple mechanism to do this. The user may specify a range along each dimension of $\boldsymbol{\theta}$ where it is known that $g > 0$. If the histogram is non-zero outside of this range, the method aborts with an error.

Note that the variance of the estimator increases with $\int \frac{\|\nabla \cdot g_A(\boldsymbol{\theta})\|^2}{g_A(\boldsymbol{\theta})} d\boldsymbol{\theta}$. In practice the estimator will work well only when g doesn't go to 0 too quickly where the histogram is positive. In these cases the histogram will be defined well away from any region where $g = 0$ and infinite variance won't be an issue even if $g = 0$ somewhere.

5.3 Bin Shape

Cubic histogram bins were used above—their widths were fixed at h in each dimension. Although the asymptotic results aren't affected by the shape of each bin, for usable convergence rates the bins' dimensions need to compatible with the shape of the high probability region of $p(\boldsymbol{\theta}|\boldsymbol{x}, T)$. Unfortunately, it is difficult to determine the best bin shapes.

The `margLikArrogance` package contains a simple workaround: by default the distribution is first scaled so that the sampled standard deviation along each dimension is constant. This is equivalent to setting each bin's width by dimension in proportion to that dimension's standard deviation. If this simple rule of thumb is insufficient, the user can scale the sampled values of $p(\boldsymbol{\theta}|\boldsymbol{x}, T)$ manually (and make the corresponding adjustment to the estimate A_n).

6 Conclusion

This paper has described an "arrogance sampling" technique for computing the marginal likelihood or Bayes factor of a Bayesian model. It involves using samples from the model's posterior parameter distribution along with the scaled values of the distribution's density at those points. These samples are divided into two main groups: m samples are used to build a histogram; n are used to importance sample the histogram using the posterior parameter distribution as the proposal.

This method is simple to implement and runs quickly in $O(d(m+n)\log(m))$ time. Its asymptotic convergence rate, $n^{-1}(1 + O(Cm^{-2/d}))$, is not remarkable, but in practice convergence is fast for many problems. Because the required inputs are similar to those of the harmonic mean estimator, it may be a convenient replacement for it.

7 References

1. S. Chib. "Marginal Likelihood from the Gibbs Output" *Journal of the American Statistical Association.* Vol 90, No 432. (1995)

2. S. Chib and I. Jeliazkov. "Accept-reject Metropolis-Hastings sampling and marginal likelihood estimation" *Statistica Neerlandica*. Vol 59, No 1. (2005)

3. A. Gelman and X. Meng. "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling" *Statistical Science*. Vol 13, No 2. (1998)

4. R. Kass and A. Raftery. "Bayes Factors" *Journal of the American Statistical Association*. Vol 90, No 430. (1995)

5. R. Neal. "The Harmonic Mean of the Likelihood: Worst Monte Carlo Method Ever". Blog post, `http://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/`. (2008)

6. J. Neddermeyer. "Computationally Efficient Nonparametric Importance Sampling" *Journal of the American Statistical Association*. Vol 104, No 486. (2009)

7. A. Owen and Y. Zhou. "Safe and effective importance sampling" *Journal of the American Statistical Association*. Vol 95, No 449. (2000)

8. C. Robert and D. Wraith. "Computational methods for Bayesian model choice" *arXiv* 0907.5123 `http://arxiv.org/abs/0907.5123`

9. P. Zhang. "Nonparametric Importance Sampling" *Journal of the American Statistical Association*. Vol 91, No 435. (1996)

10. M. Zlochin and Y. Baram. "Efficient Nonparametric Importance Sampling for Bayesian Inference" *Proceedings of the 2002 International Joint Conference on Neural Networks* 2498–2502. (2002)