Bayesian Analysis (2012) **7**, Number 3, pp. 737–770

Computing the Bayes Factor from a Markov Chain Monte Carlo Simulation of the Posterior Distribution

Martin D. Weinberg[*]

Abstract. Determining the marginal likelihood from a simulated posterior distribution is central to Bayesian model selection but is computationally challenging. The often-used harmonic mean approximation (HMA) makes no prior assumptions about the character of the distribution but tends to be inconsistent. The Laplace approximation is stable but makes strong, and often inappropriate, assumptions about the shape of the posterior distribution. Here, I argue that the marginal likelihood can be reliably computed from a posterior sample using Lebesgue integration theory in one of two ways: 1) when the HMA integral exists, compute the measure function numerically and analyze the resulting quadrature to control error; 2) compute the measure function numerically for the marginal likelihood integral itself using a space-partitioning tree, followed by quadrature. The first algorithm automatically eliminates the part of the sample that contributes large truncation error in the HMA. Moreover, it provides a simple graphical test for the existence of the HMA integral. The second algorithm uses the posterior sample to assign probability to a partition of the sample space and performs the marginal likelihood integral directly. It uses the posterior sample to discover and tessellate the subset of the sample space that was explored and uses quantiles to compute a representative field value. When integrating directly, this space may be trimmed to remove regions with low probability density and thereby improve accuracy. This second algorithm is consistent for all proper distributions. Error analysis provides some diagnostics on the numerical condition of the results in both cases.

Keywords: Bayesian computation, marginal likelihood, algorithm, Bayes factors, model selection

1 Introduction

A Bayesian data analysis specifies joint probability distributions that describe the relationship between the prior information, the model or hypotheses, and the data. Using Bayes theorem, the posterior distribution is uniquely determined from the conditional probability distribution of the unknowns given the observed data:

$$P(\theta|\mathcal{M},\mathbf{D}) = \frac{P(\theta|\mathcal{M})P(\mathbf{D}|\theta,\mathcal{M})}{Z} \tag{1}$$

where $P(\theta|\mathcal{M})$ is the prior distribution, $P(\mathbf{D}|\theta,\mathcal{M})$ is the likelihood function, and

$$Z \equiv P(\mathbf{D}|\mathcal{M}) = \int d\theta\, P(\theta|\mathcal{M})P(\mathbf{D}|\theta,\mathcal{M}) \tag{2}$$

[*]Department of Astronomy, University of Massachusetts, Amherst, MA, weinberg@astro.umass.edu

DOI:10.1214/12-BA725

is the marginal likelihood. The symbol $\mathcal{M}$ denotes the assumption of a particular model and the parameter vector $\theta \in \Omega$. For physical models, the sample space Ω is most often a continuous space. The posterior may be used, for example, to infer the distribution of model parameters or to discriminate between competing hypotheses or models. The latter is particularly valuable given the wide variety of astronomical problems where diverse hypotheses describing heterogeneous physical systems are the norm (see Gelman et al. 2003, for a thorough discussion of Bayesian data analysis).

For parameter estimation, one often considers the marginal likelihood $P(\mathbf{D}|\mathcal{M})$ to be an uninteresting normalization constant. However, equation (2) admits a meaningful interpretation: it is the support for a model given the data. To see this, assume that the prior probability of some model, $\mathcal{M}_j$, is $P(\mathcal{M}_j)$. Then by Bayes theorem, the probability of the model given the data is $P(\mathcal{M}_j|\mathbf{D}) = P(\mathcal{M}_j)P(\mathbf{D}|\mathcal{M}_j)/P(\mathbf{D})$. The posterior odds of Model $j = 0$ relative to Model $j = 1$ becomes

$$\frac{P(\mathcal{M}_0|\mathbf{D})}{P(\mathcal{M}_1|\mathbf{D})} = \frac{P(\mathcal{M}_0)}{P(\mathcal{M}_1)} \frac{P(\mathbf{D}|\mathcal{M}_0)}{P(\mathbf{D}|\mathcal{M}_1)}. \tag{3}$$

If we have information about the ratio of prior odds, $P(\mathcal{M}_0)/P(\mathcal{M}_1)$, we should use it, but more often than not our lack of knowledge forces a choice of $P(\mathcal{M}_0)/P(\mathcal{M}_1) = 1$. Then, we estimate the relative probability of the models given $\mathbf{D}$ over their prior odds by the Bayes factor $P(\mathbf{D}|\mathcal{M}_0)/P(\mathbf{D}|\mathcal{M}_1)$ (see Lavine and Schervish 1999, for a discussion of additional concerns). When there is no ambiguity, we will omit the explicit dependence on $\mathcal{M}$ of the prior distribution, likelihood function, and marginal likelihood for notational convenience. The Bayes factor has a number of attractive advantages for model selection (Kass and Raftery 1995): (1) it is a consistent selector; that is, the ratio will increasingly favor the true model in the limit of large data; (2) Bayes factors act as Occam's razors, preferring simpler models if the fits are similar; (3) Bayes factors do not require the models to be nested in any way; that is, the models and their parameters need not be equivalent in any limit; and (4) once computed, a marginal likelihood value may be used for future model selection.

There is a catch: *direct* computation of the marginal likelihood (eq. 2) is intractable for most problems of practical interest. However, recent advances in computing technology together with developments in Markov chain Monte Carlo (MCMC) algorithms have the promise to compute the posterior distribution for problems that have been previously infeasible owing to dimensionality or complexity. Although dimension-switching algorithms, such as reversible-jump MCMC (Green 1995) incorporate model selection automatically without a need for Bayes factors, these simulations appear slow to converge for some real-world applications. Moreover, the marginal likelihood has archival value and may be used for a variety of tests, ex post facto.

Newton and Raftery (1994) presented a formula for estimating Z from a posterior distribution of parameters. They noted that an MCMC simulation of the posterior selects values of $\theta \in \Omega$ distributed as

$$Z \times P(\theta|\mathbf{D}) = P(\mathbf{D}|\theta)P(\theta)$$

and, therefore,

$$Z \times \int_{\Omega} d\theta \, \frac{P(\theta|\mathbf{D})}{P(\mathbf{D}|\theta)} = \int_{\Omega} d\theta \, P(\theta) = 1 \tag{4}$$

or

$$\frac{1}{Z} = \int_{\Omega} d\theta \, \frac{P(\theta|\mathbf{D})}{P(\mathbf{D}|\theta)} = E\left[\frac{1}{P(\theta|\mathbf{D})}\right]_{P(\theta|\mathbf{D})}. \tag{5}$$

This latter equation says that the marginal likelihood is the harmonic mean of the likelihood with respect to the posterior distribution. It follows that the harmonic mean computed from a sampled posterior distribution is an estimator for the marginal likelihood[1], e.g.:

$$\tilde{Z} = \left[\frac{1}{N}\sum_{i=1}^{N} \frac{1}{P(\mathbf{D}|\theta_i)}\right]^{-1}. \tag{6}$$

Unfortunately, this estimator is prone to domination by a few outlying terms with abnormally small values of $P(\mathbf{D}|\theta)$ (e.g. see Raftery et al. 2007, and references therein). Wolpert (2002) describes convergence criteria for equation (6) and Chib and Jeliazkov (2001) present augmented approaches with error estimates. This paper presents an easy-to-compute technique for assessing the existence of the integral in equation (5) using the posterior sample itself and estimating errors in its numerical evaluation (the Numerical Lebesgue Algorithm, NLA, see Section 2 and the Appendix). In essence, this algorithm defines a subset $\Omega_s \subset \Omega$ that decreases the error in $\tilde{Z}$. To be clear, the NLA does not and, indeed, cannot circumvent the limitations of the HMA but can easily diagnose them.

Alternative approaches to computing the marginal likelihood from the posterior distribution have been described at length by Kass and Raftery (1995). Of these, the Laplace approximation, which approximates the posterior distribution by a multidimensional Gaussian distribution and uses this approximation to compute equation (2) directly, is the most widely used. This approach does not suffer from the uncertainty of existence for the defining integrals, however, one must identify all the dominant modes. In addition, the modes may not be well-represented by a multidimensional Gaussian distribution for problems of practical interest, although many promising improvements have been suggested (e.g. DiCiccio et al. 1997). Alternatively, Trotta (2007) explored the use of the Savage-Dickey density ratio for cosmological model selection (see also Trotta 2008, for a full review of the model selection problem for cosmology), which is a good alternative when the distributions are appropriately separable and nested.

Finally, this paper considers evaluation of equation (2) directly. The MCMC simulation samples the posterior distribution by design, and therefore, this sample can be used to construct volume elements in k-dimensional parameter space, $d\theta$, e.g. when $\Omega \subset \mathbf{R}^k$. The volume will be sparsely sampled in regions of relatively low likelihood and the volume size must be chosen to minimize the bias and variance. The often-used approach from computational geometry, Delaunay triangulation, maximizes the

[1] We use $\tilde{A}$ to denote the computational estimate of A from a point set. This may be either a statistical estimator or a numerical estimate.

minimum angle of the facets and thereby yields the "roundest" volumes (e.g. de Berg et al. 2008). Unfortunately, the standard procedure scales as $\mathcal{O}(kN^2)$ for a sample of N points. This can be reduced to $\mathcal{O}(N \log N + N^{k/2})$ using the flip algorithm with iterative construction (Edelsbrunner and Shah 1966) but this scaling is prohibitive for large N and k typical of many problems. Rather, in this paper, we consider less optimal but tractable kd-trees and hyperoctrees for space partitioning (the Volume Tessellation Algorithm, VTA, see Section 2 and Appendix). However, we will see in Section 2 that the expression for Z may be reformulated rigorously with an appropriate choice $\Omega_s \subset \Omega$ to exclude the sparsely sampled regions that dominate the variance induced by the space partition. Together, the NLA and VTA provide robust alternatives to computing the the marginal likelihood with errors $|\delta \log Z| \lesssim 0.5$ for models with dimensionality $k \leq 20$. An in-depth report for an astronomically relevant case study—the inference galaxy image properties—is in preparation.

This paper is organized as follows. In Section 2, we apply Lebesgue integration to the marginal likelihood computation. This development explores the HMA and direct evaluation from the numerical standpoint and leads to an improved approach outlined in Section 3. In short, the proposed approach is motivated by methods of numerical quadrature rather than sample statistics. Examples in Section 4 compare the application of the new algorithms to the HMA and the direct integration of equation (2). The overall results are discussed in Section 5.

2 Numerical evaluation of the marginal likelihood integral

Consider the integral

$$I = \int_{\Omega} f(\theta)\, d\theta \tag{7}$$

where $f(\theta)$ is a density and θ has dimensionality k. Rewritten as a Lebesgue integral, I becomes

$$I = \int_0^{\sup\{f(\theta):\theta\in\Omega\}} M(y) dy \tag{8}$$

where

$$M(y) = \int_{f(\theta)>y} d\theta. \tag{9}$$

The Lebesgue integral (eq. 8) describes the measure $M(y)$ associated with each value of the density $f(\theta)$. We now apply this to the evaluation of the marginal likelihood from the sampled posterior distribution. We choose a subset Ω_s of the original domain Ω sampled by the Markov chain. The choice for the subset will be motivated by numerical analysis described below. The integral in equation (2) states that marginal likelihood is the expectation of the likelihood with respect to the prior distribution. This is the same as equation (8) with $P(\theta|\mathcal{M})P(\mathbf{D}|\theta, \mathcal{M})$ replacing $f(\theta)$. Alternatively, returning to equation (4), the integral $Z \equiv P(\mathbf{D})$ is implicitly defined by

$$P(\mathbf{D}) \int_{\Omega_s} \frac{d\theta\, P(\theta|\mathbf{D})}{P(\mathbf{D}|\theta)} = \int_{\Omega_s} d\theta\, P(\theta) \equiv J. \tag{10}$$

Since $\int_\Omega d\theta P(\theta) = 1$ and $\Omega_s \subset \Omega$, it follows that $J \leq 1$. Defining $Y \equiv 1/P(\mathbf{D}|\theta)$, the Lebesgue integral on the left-hand-side of equation (10) is

$$K \equiv \int_{\Omega_s} \frac{d\theta\, P(\theta|\mathbf{D})}{P(\mathbf{D}|\theta)} = \int_0^{Y_1} M(Y)\, dY = \int_{Y_0}^{Y_1} M(Y)\, dY + M(Y_0)Y_0 \tag{11}$$

with measure function

$$M(y) = \int_{Y(\mathbf{D}|\theta)>y} d\theta\, P(\theta|\mathbf{D}), \tag{12}$$

$Y_0 = \inf\{Y(\mathbf{D}|\theta) : \theta \in \Omega_s\}$, and $Y_1 = \sup\{Y(\mathbf{D}|\theta) : \theta \in \Omega_s\}$.

The Monte Carlo evaluation of K in equation (11) using a sample from the posterior distribution motivates the HMA. The HMA fails owing to domination by individual terms with small values of $P(\mathbf{D}|\theta)$ (i.e. large values of Y). Appendix 1 demonstrates that $M(Y) \propto Y^{-1-b}$ for normally-distributed $P(\mathbf{D}|\theta)$ and $P(\theta)$ where $b = \mathrm{Var}[P(\mathbf{D}|\theta)]/\mathrm{Var}[P(\theta)]$. That is, the integral expression for K will be consistent if $M(Y)$ decreases faster than Y^{-1} as $Y \to \infty$. We conclude that the NLA will be numerically well-conditioned only for problems with informative prior distributions. For these cases, K may be evaluated numerically with error estimates using the quadrature rule described in Appendix 2; this is the *Numerical Lebesgue Algorithm* (NLA). This appendix further shows that the standard expression for the HMA can be derived from the NLA.

Because the numerical procedure described in Appendix 2 trims the sample to eliminate intervals in Y with large error from equation (11), the evaluation of $Z \equiv P(\mathbf{D})$ using equation (10) requires an estimate of the integral J over the trimmed domain Ω_s. Appendix 3 describes a volume tessellation algorithm assigning a measure to each sample from the target distribution as follows: $\cup_j \omega_{js} \subseteq \Omega_s$ with $\omega_{js} \cap \omega_{ks} = \emptyset$, $j \neq k$. The resulting integral, then, may be evaluated using the Lebesgue or Riemann theory as described in Appendix 3; this is the *Volume Tessellation Algorithm* (VTA). The former is useful when the smallest cells in the tessellation contain more than a single sample. The two theories give identical results in the limit of a single sample per cell.

Note that Z may be evaluated directly from its definition using the VTA. Following the development from equation (10), we have

$$Z \times \int_{\Omega_s} d\theta\, P(\theta|\mathbf{D}) = \int_{\Omega_s} d\theta\, P(\theta) P(\theta|\mathbf{D}). \tag{13}$$

The subdomain Ω_s is arbitrary but should be chosen to minimize the variance of the integrals on the right- and left-hand sides of equation (10). The Monte Carlo evaluation of the integral on the left-hand side is simply the fraction of the posterior sample in Ω_s

relative to Ω: $F(\Omega_s) \equiv \sum_{i=1}^{N} \mathbf{1}_{\theta\in\Omega_s}/N$ where $\mathbf{1}_{\{\cdot\}}$ is an indicator function. If $\Omega_s = \Omega$, then $F(\Omega_s) = 1$. This allows us to write equation (13) in the convenient form:

$$Z = \frac{1}{F(\Omega_s)} \int_{\Omega_s} d\theta\, P(\theta) P(\theta|\mathbf{D}). \tag{14}$$

The factor $F(\Omega_s)$ corrects the normalization Z for the excluded part of the domain. As in the case of the NLA, the posterior sample may be trimmed to eliminate the very low density regions of the parameter space that contribute significant variance to the estimates of the right-hand side. The measure function for the integral on the right-hand side take the form

$$M(y) = \int_{P(\theta|\mathbf{D})>y} d\theta. \tag{15}$$

This approach to computing Z is consistent for all $M(y)$. For many problems of interest (e.g. with weakly informative prior distributions), $P(\theta)$ will be slowly varying over Ω while $P(\theta|\mathbf{D})$ will be large over a small subset of Ω. Therefore, the numerical evaluation of J only *weakly* depends on the details of the tessellation while the evaluation of Z may depend *strongly* on the tessellation. Tests suggest that the bias resulting from the tessellation details decreases slowly with sample size N for $\Omega_s = \Omega$. Therefore, the NLA often outperforms the VTA when $M(y)$ decreases fast enough for consistency. The performance of the VTA may be improved by careful choice of Ω_s.

3 The new algorithms

We now present the implementation details of two new algorithms, the *Numerical Lebesgue Algorithm* (NLA) and the *Volume Tessellation Algorithm* (VTA), that implement the strategies described in Section 2 and in Appendix 2 and Appendix 3. The NLA computes $\tilde{K}$ and VTA computes $\tilde{J}$ from equation (10) and $\tilde{Z}$ from equation (2). We will assume that the integral K exists and this can be checked empirically by plotting the partial quadrature sums as shown in Figures 8 and 9.

3.1 Description: NLA

Begin with an MCMC sample of size N from the posterior distribution. Sort the samples in order of increasing values of likelihood $L_j \equiv P(\mathbf{D}|\theta_j)$, and compute Δ_j (eq. 28) with $j = 1, \ldots, N$ to find the first value of $j = n$ satisfying $\Delta_j < \epsilon_*$. Then, compute the M_j for $j = n, \ldots, N$ using equation (25). This equation estimates M by counting the fraction of the MCMC sample satisfying the domain restriction in equation (9). One may estimate the effect of discreteness by computing M_j using both the restriction $P(\mathbf{D}|\theta) < L_j$ and $P(\mathbf{D}|\theta) \leq L_j$ (or, equivalently, $P^{-1}(\mathbf{D}|\theta) > Y_j$ and $P^{-1}(\mathbf{D}|\theta) \geq Y_j$) to obtain lower and upper estimates for $M(Y)$. Then, these yield Riemann-like upper and lower bounds on $\tilde{K}$. The criteria from equation (28) may be applied to identify and eliminate discontinuities in $M(Y)$. Excepting the numerical sort, the work required to implement this algorithm is no harder than the HMA. As described in Section 2 and Appendix 1, K (eq. 11) will not exist for many common problems, such as a Gaussian-process likelihood function with an uninformative uniform prior distribution (see Appendix 1). This algorithm will diagnose this condition directly from the posterior sample.

3.2 Description: VTA

The VTA uses a spatial partition to estimate the volume $\cup_j \omega_{sj} \subset \Omega_s$ associated with a sample size $m \leq c$. Any cell whose sample count exceeds a predefined value c is further subdivided (see Appendix 3). We have explored two easy-to-implement trees. The first, the kd-tree, splits $\mathbf{R}^k$ on planes perpendicular to one of the coordinate system axes. Our implementation splits at the median value along one of the coordinate axes (a *balanced* kd-tree). Traditionally, every node of a kd-tree, from the root to the leaves, stores a point. Here, the points are stored in leaf nodes only, although each splitting plane still goes through one of the sample points. This choice facilitates the computation of the volume spanned by the points for each node as follows. Let c_j be the number of parameter-space points $\theta^{[n]}, n = 1, \ldots, c_j$ in the j^{th} node. The volume for Node j, V_j, is geometrically determined by the splitting planes, or one may take

$$V_j = \prod_{i=1}^{k} \left[\max(\theta_i^{[1]}, \ldots, \theta_i^{[c_j]}) - \min(\theta_i^{[1]}, \ldots, \theta_i^{[c_j]}) \right] \qquad (16)$$

where θ_i is the i^{th} component of the vector θ. For simplicity, one may choose $c \equiv c_j = 2^q$ for some fixed integer q to determine an exclusive volume partition of the parameter space spanned by the point set, the *frontier*. One chooses the value of q to be large enough to limit the sampling bias of field quantities in the volume but small enough to resolve the posterior modes of interest. The values $q \in [2, \ldots, 6]$ seem to be good choices for many applications.

The second tree, the hyperoctree (or 2^k tree), partitions each node into 2^k children by bisecting the range of the parent node in *each* dimension. Unlike the kd-tree, all children need not be and generally will not be branches with leaves. We define a *terminal* branch to be a branch with leaves. Since a terminal branch will have between 0 and c leaves, one may estimate that the hyperoctree will have roughly twice the number of terminal branches than the kd-tree has cells. The volume for ω_s at each terminal branch is uniquely determined by its branching history, and its coordinate boundaries are uniquely specified by retaining the branching information at each node. It is convenient to encode the 'right' and 'left' branching history into a binary sequence stored as an integer key. Alternatively, one may use equation (16) to assign the volume for ω_s.

I discuss two alternative approaches in Appendix 3 for computing Z or J using the space partition. The first is based on the Lebesgue integral using equations (8), (31) and (32) to evaluate the measure function $M(Y)$. The second is based on the Riemann integral (eq. 7) using equation (33). Practically, both approaches may be computed for all evaluations without loss of efficiency since the computational work is dominated by the tree construction. For the integral J, Ω_s should be chosen to eliminate quadrature intervals that contribute large error to the result. For the integral in equation (14), Ω_s should be chosen to eliminate regions with very low probability density, if possible.

3.3 Computational performance

As with the NLA, the Lebesgue formulation facilitates performance and error analysis (see Section 4 for examples). In addition, the bias and variance can be reduced in some cases by performing a coordinate transformation. For volume tessellation, we can decrease the volume in the tails of the distribution, which is especially useful for dimensions with large or infinite domains. This may be accomplished with a careful choice of Ω_s. In most cases, it is sufficient to identify the region of the sample space with high posterior density and choose Ω_s by restricting the range in each dimension to include this region only. Alternatively, for irregular or complex volumes, I have had some success with the mapping

$$x \rightarrow y \equiv \frac{t}{\sqrt{1+t^2}} \quad \text{where} \quad t = \frac{x-\bar{x}}{s}, \tag{17}$$

$\bar{x}$ is the mean of the posterior distribution in x, and s^2 is the sample variance. Clearly $y \in (-1, 1)$ for $x \in (-\infty, \infty)$.

The NLA begins with a sort of the likelihood sequence $\{L_j\}$ and this scales as $\mathcal{O}(N \log N)$. The computation of the measure function, M_j, followed by the computation of $\tilde{K}$ is $\mathcal{O}(N)$. The sequence $\{M_j\}$ is also useful for diagnostics as we will explore in the next section. However, in many cases, we do not need the individual M_j but only the differential values $M_j - M_{j+1}$ to compute $\tilde{K}$, which contains a single term in equation (25). The values of likelihood might range over many orders of magnitude. Owing to the finite machine mantissa, the differential value may be necessary to achieve adequate precision for large N. The algorithm computes the lower, upper, and trapezoid-rule sums (eq. 26) for the final integral $\tilde{K}$. For large posterior samples, e.g. $N > 10000$, the differences between $K^{[l]}$ and $K^{[u]}$ (eq. 26) are small and the error is dominated by volume tessellation (see below and Appendix 3). Indeed, a more useful error estimate may be obtained by a random partitioning and subsampling of the original sequence $\{L_k\}$ to estimate the distribution of $\tilde{Z}$ (see the examples in Section 4). In practice, computing $\tilde{K}$ from a posterior sample with $N = 400000$ takes 0.2 CPU seconds on a single 2 GHz Opteron processor. Although NLA could easily be parallelized over n processors to reduce the total run time by $1/n$, this seems unnecessary.

For the VTA with a kd-tree, the computational complexity for building the tree from the N sampled points in parameter space scales as $\mathcal{O}(N \log^2 N)$ using the Quicksort algorithm at each successive level (this could be improved, see Cormen et al. 2001). The tree walk required to sum over differential node volumes to obtain the final integral scales as $\mathcal{O}(N \log N)$. Similarly, both constructing and walking the hyperoctree to obtain the final integral requires $\mathcal{O}(N \log N)$ operations. I confirmed this scaling empirically using the multidimensional example described in Section 4.3 with dimension $k \in [1, 40]$ and sample size $N \in [1000, 10000000]$. Computing $\tilde{J}$ and $\tilde{Z}$ directly from a posterior sample with $N = 400000$ and $k = 10$ takes 4.4 CPU seconds on a single 2 GHz Opteron processor and, therefore, the computation is unlikely to be a computational bottleneck in the inference overall, even when resampling to produce a variance estimate. The leading coefficient appears to vary weakly with the underlying distribution, although there could be undiscovered pathological cases that degrade the performance.

The kd-tree and hyperoctree have different and, in a sense, complementary partitioning strategies. The kd-tree ensures that all terminal branch nodes has the same number of leaves. Consequently, a particular ω_s containing points in the tail of the posterior distribution may have volumes with a large extent in one or more dimensions. The balanced kd-tree used here can make volumes whose aspect ratios are arbitrarily large. In addition, the kd-tree algorithm divides space into two subregions at each node. Therefore, a reliable volume partition requires $N \gg 2^k$ points to ensures that each dimension has been affected by the data distribution. With current hardware, this criterion limits the dimensionality to $k \log_2(10^8) \approx 25$. Future improvements in efficient partitioning of high-dimensional spaces may improve this limitation. Empirically, $N = 400000$ is sufficient for $k = 20$ in the tests below. Batch sample variance requires much larger chain lengths and, therefore, we resort to bootstrap resampling estimates in the tests below, especially for $k > 10$.

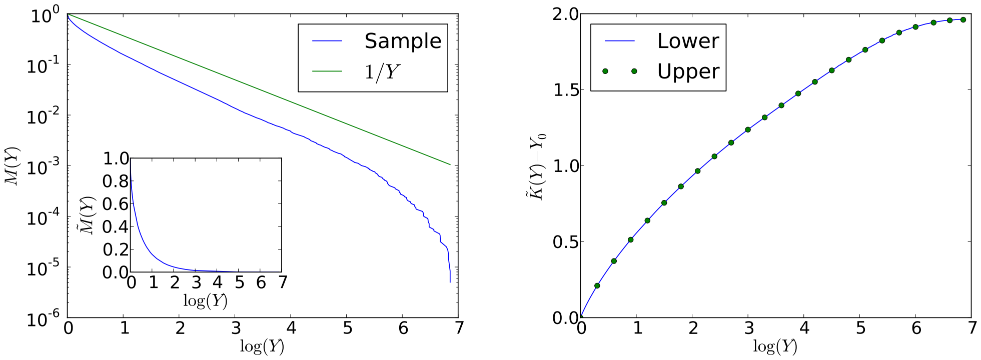
The hyperoctree algorithm, conversely, always creates self-similar subvolumes. This construction better adapts to multimodal distributions and is not adversely affected by large voids in the sampling of $\mathbf{R}^k$. On the other hand, the hyperoctree is recursively built with a stopping criterion of leaves per terminal node $m \leq c$. This may result in terminal branches with low volume filling factors which leads to oscillation in the integral value with sample size (see Appendix 3). This possibility motivates choosing the largest value c that still resolves the structure in the target distribution. The number of points required for the hyperoctree is clearly distribution dependent, but empirically, the criterion $N \gg 2^k$ seems to be a good guide. For both algorithms, the regions of low posterior probability density lead to large subvolumes. If Ω_s may be chosen to eliminate the low-density regions (e.g. by circumscribing the dominant mode), the error owing to tessellation variance will be greatly reduced.

4 Tests & Examples

To estimate the marginal likelihood using the methods of the previous section, we may use either the NLA to estimate K and the VTA to estimate J or use the VTA alone to estimate Z using equation (14). Examples below explore the performance of these strategies. The MCMC posterior simulations are all computed using the Bayesian Inference Engine (BIE, Weinberg and Moss 2010; Weinberg 2012), a general-purpose parallel software platform for Bayesian computation. All examples simulate the posterior distribution using the parallel tempering scheme (Geyer 1991) with $T = 128$ and 20 temperature levels or the differential evolution algorithm (Ter Braak 2006). MCMC convergence is assessed using the subsampling algorithm described in Giakoumatos et al. (1999), a generalization of the Gelman and Rubin (1992) test.

4.1 Fidelity of the NLA and the VTA

For a simple example, let us compute the marginal likelihood for a data sample $\mathbf{D}$ of 100 points $x \sim \mathcal{N}(0.5, 0.03)$ modeled by $\mathcal{N}(\theta, 0.03)$ with a uniform prior distribution $\theta \sim \mathcal{U}(-0.2, 1.2)$. The marginal likelihood Z can be computed analytically from $\mathbf{D}$. Application of the NLA for $\tilde{K}$ and the VTA for $\tilde{J}$ to 300,000 converged states using the parallel tempering algorithm gives a value of $\log \tilde{Z} = 31.31 \pm 0.04$ (95% CI), consistent with the analytic result: $\log Z = 31.36$. The autocorrelation for the posterior sample drops to zero at a lag of 8. Appendix 2 describes several numerical truncation criterion for the Riemann sum leading to $\tilde{K}$. Here, I truncate the sum when the logarithmic spacing in the likelihood exceeds the predefined limit h_*. A value of $h_* = 0.05$ seems appropriate from numerical considerations, although experiments suggest that the algorithm is not sensitive to this choice as long as h_* is not so small as to decimate the sample or so large that error-prone outliers are included. Computation of the batched standard deviation yields equivalent values and 95% CI. The VTA with $\Omega_s = \Omega$ yields $\log \tilde{Z} = 31.34 \pm 0.01$, consistent with the analytic result and with smaller bootstrap variance than the NLA. The 95% confidence intervals are produced by sampling ensembles of 75,000 states from the converged chain with replacement. Occasional extreme tail values result in some anomalously small values of $\log \tilde{Z}$ for the HMA: $29.24^{+4.51}_{-272.6}$.

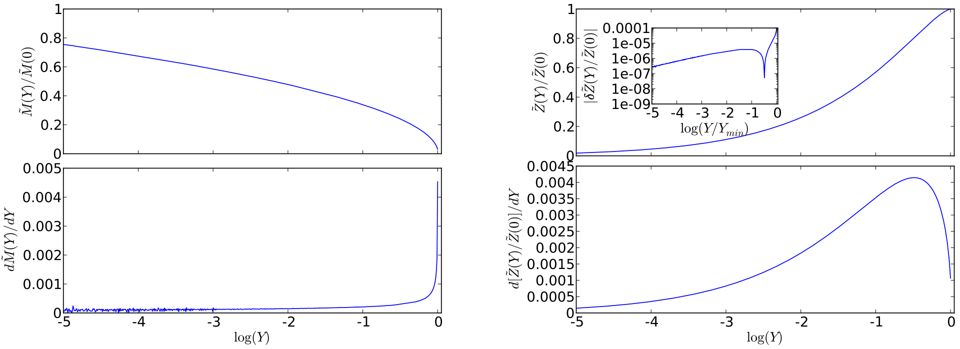


(a) $\tilde{M}(Y)$ (b) $\tilde{K}(Y) - Y_0$

Figure 1: Details of the marginal likelihood computation illustrating the numerical Lebesgue approach. Panel (a) compares the run of the measure $\tilde{M}$ function with $Y \equiv L_0/L$ computed from posterior simulation with $N = 400000$ elements using the NLA with the limiting case for existence: $M(Y) \propto 1/Y$. Panel (b) shows the Lebesgue *quadrature* term, $\tilde{K}(Y) - Y_0$ from eq. 22. $\tilde{K}_{lower/upper} - Y_0$ are the lower and upper Riemann sums. This illustrates the essence of the algorithm: anomalously small values of L degrade the fidelity of $\tilde{M}$ at large Y but these same values of $\tilde{M}$ make negligible contribution to $\tilde{K}$ and, therefore, can be truncated from the quadrature sums.

Figure 1 illustrates the details of the NLA applied to this computation. For ease of notation, let $L \equiv P(\mathbf{D}|\theta)$. Panel (a) plots $\tilde{M}$ from equation (25) where Y is the inverse scaled likelihood, $Y \equiv L_0/L$ and $L_0 = \sup\{L : \theta \in \mathbf{R}\}$. With this definition,

$Y_0 = 1$. The run of $\tilde{M}$ with Y drops more quickly than $1/Y$ near the posterior mode and again for very large values of Y (i.e. small likelihood values). The inset in this figure shows $\tilde{M}$ on a linear scale. The measure function $\tilde{M}$, and hence the integral $\tilde{K}$, is dominated by large values of L as long as M decreases sufficiently fast (see Appendix 1). The plot in Figure 1a readily reveals such failures. In this case, Figure 1a suggests that $M(Y) \propto 1/Y$ at moderate values of Y with a fall off at large Y owing to the cutoffs in the uniform prior distribution. Therefore, the variance of $\tilde{K}$ will be large although K formally exists (cf. Figs. 8 and 9). Figure 1b plots the cumulative sum defining the quadrature of $\tilde{Z}$ in equation (26), beginning with the largest values of likelihood first. The contribution to $\tilde{Z}$ is roughly uniform over the range of $\log Y$, rolling over near $\log Y \approx 6$. A more informative prior would increase the domination at smaller values of $\log Y$ as described in Appendix 1. In addition, NLA provides upper and lower bounds, and thereby some warning when the estimate is poorly conditioned, e.g. owing to an inappropriate choice for h_*.



(a) $\tilde{M}(Y)$

(b) $\tilde{Z}(Y)$

Figure 2: Details of the marginal likelihood computation using the Lebesgue version of the volume tessellation approach. Panel (a) compares the run of measure function $\tilde{M}(Y)$ and $d\tilde{M}(Y)/dY$ scaled by the maximum value, $\tilde{M}(0)$. Panel (b) shows the Lebesgue *quadrature* term, $\tilde{Z}(Y)$ and $d\tilde{Z}(Y)/dY$ from eq. 32. The inset panel shows the scale absolute error defined in Section 4.3.

The computational details for the VTA are shown in Figure 2. For ease of notation, let $P \equiv P(\theta|\mathbf{D})$. The measure function $M(Y)$ is computed from equation (32) where Y for the VTA is the scaled posterior probability: $Y \equiv P/P_0$. In comparison with Figure 1a, note that the ordinate scale is linear and that $dM(Y)/dY$ is peaked at large Y and not at small Y. This results in a straightforward, stable numerical quadrature owing to no existence issues. To illustrate the contribution to $\tilde{Z}$, we plot $\tilde{Z}(Y) \equiv \int_0^Y M(Y)\,dY$. The inset in Figure 2b plots the cumulative absolute error from the upper and lower Riemann estimates using equation (33) for each interval less than Y; i.e. the quadrature errors are tiny.

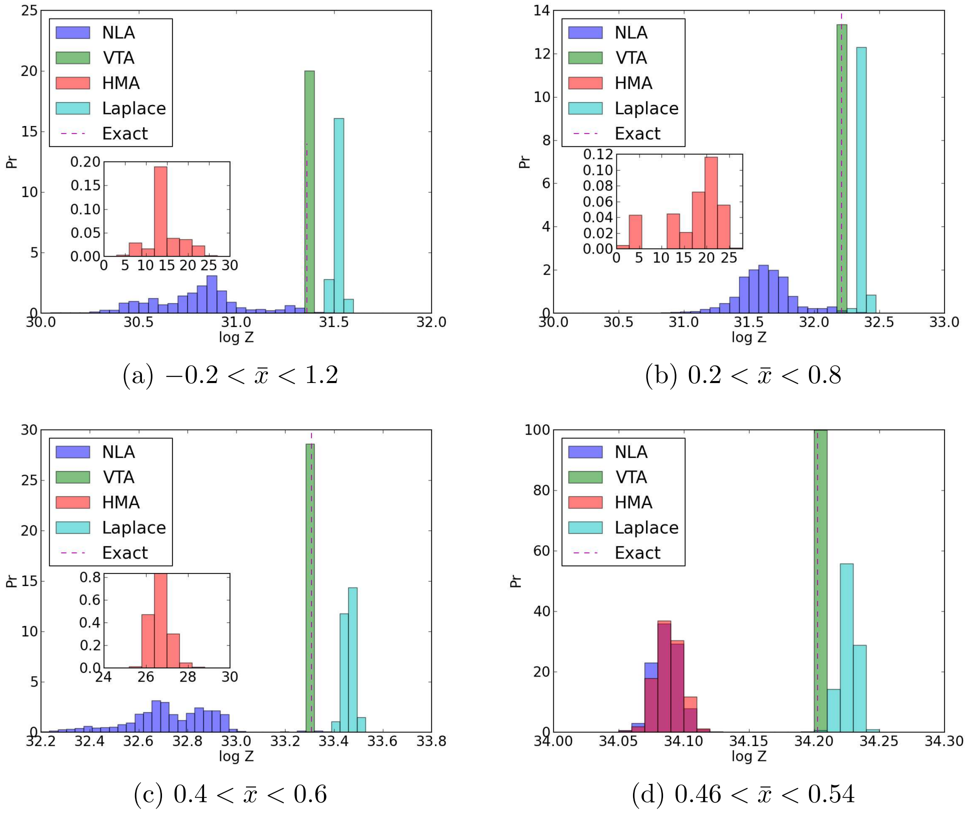
An MCMC posterior sample has $d\theta P(\theta|\mathbf{D}) =$ constant (asymptotically), and this may result in poorly explored tails. The effect of these tails may be reduced in several ways. First, as described in Section 2, we may choose a subdomain Ω_s that trims off the regions of low posterior density. If this can be done reliably by restricting parameter ranges or by a simple transformation, we may compute Z using equation (14) without further ado. Secondly, if an appropriate Ω_s cannot be found easily, the tails may be improved by sampling from an over-dispersed posterior distribution. For example, consider sampling from the "powered-up" distribution $P(\theta|\mathbf{D})^{1/T}$. Values of $T > 1$ better sample the tails of the distribution and provide better accuracy of the measure function $M(Y)$ at small Y but do so at the expense of requiring a larger sample to sample the modal region (see Appendix 3 for additional discussion). On the other hand, many advanced MCMC algorithms employ powered-up target distributions to improve mixing, and their output may be saved and reused for the computation of Z.

A more realistic assessment of the overall variance can be obtained by subsampling. The CPU time for these algorithms is sufficiently small that I recommend doing this in general. Consider the following experiment: (1) the posterior is simulated by MCMC to obtain a chain of 300,000 states; (2) the first half of the chain is discarded; (3) the second-half is randomly subsampled with replacement to obtain 128 samples of 10,000 states; (4) the marginal likelihood for each is computed using the NLA, VTA (kd-tree-based Riemann version with median value of the posterior probability in each subvolume), the Laplace approximation, and the HMA (approximately 2 CPU minutes in total). For all but the HMA, increasing the number of states decreases the variance for each distribution; samples with 10,000 states best revealed the differences between the algorithms on a single scale.

Figure 3 shows the distribution of $\tilde{Z}$ for the resampled ensembles for a sequence of four different prior distributions, each one successively more informative. Figure 3a is the model described at the beginning of this section; the range of the prior distribution is much larger than the values sampled from the posterior distribution. The colors are composited[2] with $\alpha = 0.5$ (e.g. HMA over VTA is brown, HMA over NLA is purple, Laplace over HMA is blue-gray, Laplace over VTA is blue-green). In Panel (d), the range is within the range of values sampled by the posterior in Panel (a). The analytic value for each panel is shown as a dashed vertical line.

The prior distributions for Panels (a) and (b) are nearly flat over the range of the posterior samples and the HMA performs poorly as expected (see insets); the truncation error criterion for the NLA removes the outliers from the HMA but slow decrease of $M(Y)$ at large Y results in a tail-heavy biased estimate. As the prior distribution becomes more informative, the outlier values in the likelihood are less extreme, the heavy tail is truncated, and the HMA becomes more accurate. Panel (d) shows the result of a strongly informative prior distribution: the VTA remains the most precise of the four algorithms, but the HMA and NLA become more precise with a small bias. The width of the prior distribution in Panel (c) is approximately six times the width of the likelihood function. This is sufficient to prevent extreme outliers and the wild estimates of $\tilde{Z}$ by the HMA, but the HMA bootstrap distribution for $\tilde{Z}$ is broad and

[2] For each color channel, value c_1 over c_0 yields the new value $c = (1 - \alpha)c_0 + \alpha c_1$.



(a) $-0.2 < \bar{x} < 1.2$

(b) $0.2 < \bar{x} < 0.8$

(c) $0.4 < \bar{x} < 0.6$

(d) $0.46 < \bar{x} < 0.54$

Figure 3: Histogrammed distributions of $\tilde{Z}$ for the NLA, VTA, HMA, and the Laplace approximation for 10,000 randomly resampled states of a converged posterior distribution of 200,000 states. The dashed line shows the true value computed by directly integrating Z from eq. 2. Each panel is labeled by the range of the flat prior distribution for the position of the normal distribution.

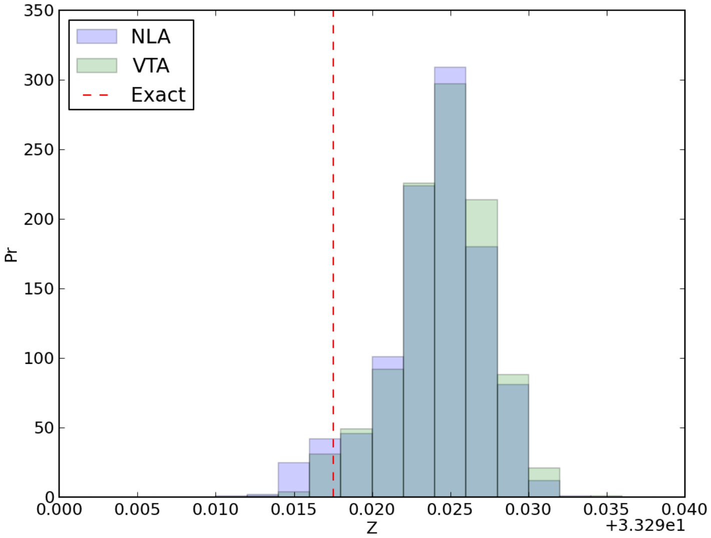


Figure 4: As in Fig. 3c for NLA and VTA with $\Omega_s \subset \Omega$ chosen to include half of the sample centered on the posterior mode.

biased.

In Panels (a)–(c), the prior distribution places little restriction on the resulting posterior sample, and since the likelihood function is normally distributed, the Laplace approximation is a close match to the true distribution. Indeed, it performs better than the HMA and NLA although it overestimates the marginal likelihood value. One should not expect such good performance in general. In the final case, Panel (d), the prior is strongly informative. The related HMA and NLA provide underestimates with similar overall distributions. The Laplace method performs adequately as expected, although it retains its upward bias. The VTA provies a narrow estimate with small bias in all cases.

Recall that the NLA uses the VTA for estimating the right-hand side of equation (10). These VTA-based estimates often may be further improved by restricting Ω_s as illustrated in Figure 4. This figure repeats the marginal likelihood computation from Figure 3c with Ω_s chosen to include half of the posterior sample by decreasing the parameter range symmetrically about the modal value. Although the estimate is biased, the variance is dramatically reduced and the range of the distribution includes the correct value.

4.2 Non-nested Linear Regression Models

Here, we test these algorithms on the radiata pine compressive strength data that were also analyzed by Han and Carlin (2001) and a number of previous authors. We use the data tabulated by Han and Carlin from Williams (1959). These data describe the maximum compressive strength parallel to the grain y_i, the density x_i, and the resin-adjusted density z_i for $N = 42$ specimens. Carlin and Chib (1995) use these data to compare the two linear regression models:

$$\begin{aligned} M = 1 : y_i &= \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \qquad i = 1, \ldots, N \\ M = 2 : y_i &= \gamma + \delta(z_i - \bar{z}) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \tau^2), \qquad i = 1, \ldots, N \end{aligned}$$

with $\mathcal{M} = \{1, 2\}, \theta_1 = \{\alpha, \beta, \sigma^2\}^T$, and $\theta_2 = \{\gamma, \delta, \tau^2\}^T$. We follow Han and Carlin (2001) and Carlin and Chib (1995), adopting $\mathcal{N}\left(\{3000, 185\}^T, \mathrm{Diag}\{10^6, 10^4\}\right)$ priors on $\{\alpha, \beta\}^T$ and $\{\gamma, \delta\}^T$, and $\mathrm{IG}\left(3, [2 * 300^2]^{-1}\right)$ priors on σ^2 and τ^2; $\mathrm{IG}(a, b)$ is the inverse gamma distribution with the density function

$$f(v) = \frac{e^{-1/(bv)}}{\Gamma(a) b^a v^{a+1}}$$

where $v > 0$ and $a, b > 0$. Han and Carlin point out that these priors are approximately centered on the least-squares solution but are otherwise rather vague. Using direct integration, Green and O'Hagan (1998) find a Bayes factor of about 4862 in favor of Model 2.

Table 1: Marginal likelihood for non-nested linear regression models

Model	$\log Z(M = 1)$	$\log Z(M = 2)$	B_{21}	$\Delta\%$
NLA	$-317.29^{+0.08}_{-0.07}$	$-308.77^{+0.06}_{-0.06}$	5014^{+101}_{-50}	+3.1
VTA (kd)	$-308.41^{+0.02}_{-0.02}$	$-300.06^{+0.02}_{-0.02}$	4675^{+173}_{-166}	-1.3
VTA (2^k)	$-309.31^{+0.01}_{-0.01}$	$-300.83^{+0.01}_{-0.01}$	4837^{+73}_{-110}	-0.5
Laplace	$-306.82^{+0.03}_{-0.03}$	$-298.3^{+0.04}_{-0.03}$	5014^{+310}_{-339}	3.1

The subscripted (superscripted) values show the offset from median for the 2.5% (97.5%) quantile.

Table 1 describes the results of applying the algorithms from the previous sections to a converged MCMC chain of 2.4 million states for both models using the differential evolution algorithm. The quoted value is the median and the bounds are the $p = 0.025$ and $p = 0.975$ quantiles computed from 1024 bootstrap subsamples of 200,000 states. Generally, the variance of the distribution in $\tilde{Z}$ decreases as the sample size increases. I chose a sample size of 200,000 to achieve a 95% CI of 0.1 or smaller in the logarithm of the marginal likelihood for both the NLA and the VTA. These estimates are based on bootstrap resampling. The sample batch variance method again yields comparable limits; e.g. $\log \tilde{Z}(M = 1)$ is 308.49 ± 0.016 (95% CI). The autocorrelation drops to zero quickly for the chain; the sample autocorrelation at lag 1 is 0.012. The second and third columns of the table are the values of marginal likelihood for Models 1 and 2 for each of the three models listed in the first column. The fourth column is the Bayes factor for Model 2 to Model 1, B_{21}, and the fifth column is the relative difference from the exact result. The NLA, the VTA with both the kd-tree and hyperoctree spatial partitions, and the Laplace approximation yield values within a few percent of the true value of B_{21}. The VTA presents the smallest variance, followed by Laplace and then NLA. The HMA was computed but the samples were too broadly distributed to be of use. A recomputation with a smaller volume chosen to reduce the sample by half yields comparable values for the Bayes factors. Figure 5 shows the distribution of B_{21} for the samples; counter to the trend from Section 4.1, both the VTA and Laplace approximation are more biased than the NLA here.

The value h_* used to compute the NLA will vary with the problem and the sample size. Therefore, some analysis of $\tilde{Z}$ is required to choose an appropriate value. As an example, Figure 6 plots the median and 95% confidence region for the bootstrap sampled marginal likelihood computation as a function of h_* for the regression problem. The value of the VTA is shown for reference. The values for $\tilde{Z}$ track each other closely for $0.001 \leq h_* \leq 0.008$. For $h_* < 0.001$, there are too few states for a reliable computation of $\tilde{Z}$. For $h_* > 0.008$, the NLA values are sensitive to the low-likelihood tail, resulting in an increasing variance with increasing h_*.

4.3 High-dimension parameter spaces

We adopt a 'data-free' likelihood function for the parameter vector θ with rank k:

$$P(\mathbf{D}|\theta) = L(\theta) = \left(2\pi\sigma^2\right)^{-k/2} e^{-\sum_{i=1}^{k} \theta_i^2/2\sigma^2}$$

with $\sigma^2 =$ constant. Further, we assume that each element θ_j of the vector θ is normally distributed with a mean of 0 and a variance of 1. The resulting expression for the marginal likelihood may be directly integrated, yielding $P(\sigma^2, k) = \left[2\pi(1+\sigma^2)\right]^{k/2}$. A straightforward analytic calculation (see Appendix 1) reveals that the integrand in the integral for J (eq. 10) will increasingly steepen towards large values of $1/L$ as k increases with $J \to \infty$ as $\sigma^2 \to 0$. This suggests that the quality of the NLA evaluations will degrade with increasing k.

For each model of dimension k, we compute a Markov chain sample of the posterior distribution using the Differential Evolution algorithm (DE, Ter Braak 2006). This algorithm evolves an ensemble of chains with initial conditions sampled from the prior distribution. A proposal is computed by randomly selecting pairs of states from the ensemble and using a multiple of their difference; this automatically 'tunes' the proposal width. We have further augmented this algorithm by including a tempered simulation step (Neal 1996) after every 20 DE steps (see Weinberg and Moss 2010; Weinberg 2012, for more details).

Each row in Table 2 describes the application of the NLA, VTA (both the Riemann and Lebesgue versions for the kd-tree), and Laplace approximation to a model

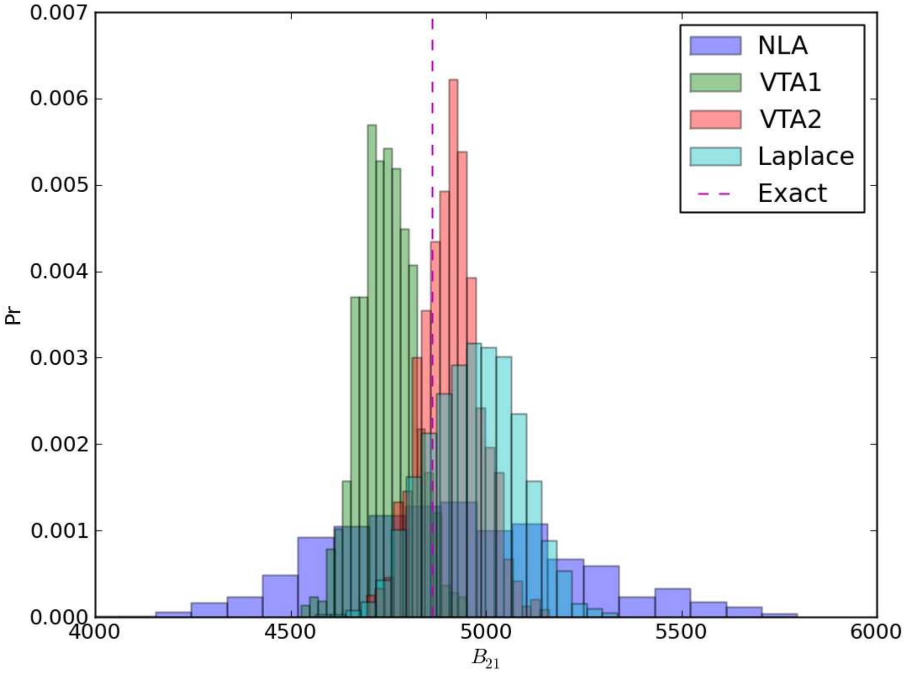


Figure 5: The histogrammed distribution of Bayes factors for the 1024 samples using the NLA, VTA, and Laplace approximation. Although the variance for the NLA is larger than the VTA or Laplace approximation, its bias is small.

of dimension k. The MCMC simulations produce approximately 1.4 million converged states. The convergence is tested using the Gelman-Rubin statistic. Each converged chain is resampled with replacement to provide 1024 subsamples of n states. The value $N \in [10000, 400000]$ is chosen to achieve 95% confidence intervals of approximately 1% of $\tilde{Z}$ or smaller. The 95% confidence intervals on $\tilde{Z}$ are indicated as sub- and superscripts. The Riemann VTA determines volume ω_s spanning c samples and approximates the integral by multiplying the volume by the median value of the sample. Finally, for each algorithm, the table presents the relative error: $\Delta\% \equiv |\log \tilde{Z} - \log Z|/|\log Z| \times 100$. The batched standard deviation was computed for $k = 1, 2, 5$, and 10 and yields bounds consistent with the bootstrap-resampled values listed in Table 2; the chain size is prohibitively large for the VTA at $k = 20$ for a reliable computation of the batched standard deviation.

As expected, the NLA performs poorly for large k owing to the increasingly larger dependence on small values of the likelihood. The VTA results are very encouraging: the relative error is within a few percent for $1 \leq k \leq 40$. For $k = 40$, I computed $\tilde{Z}$ with

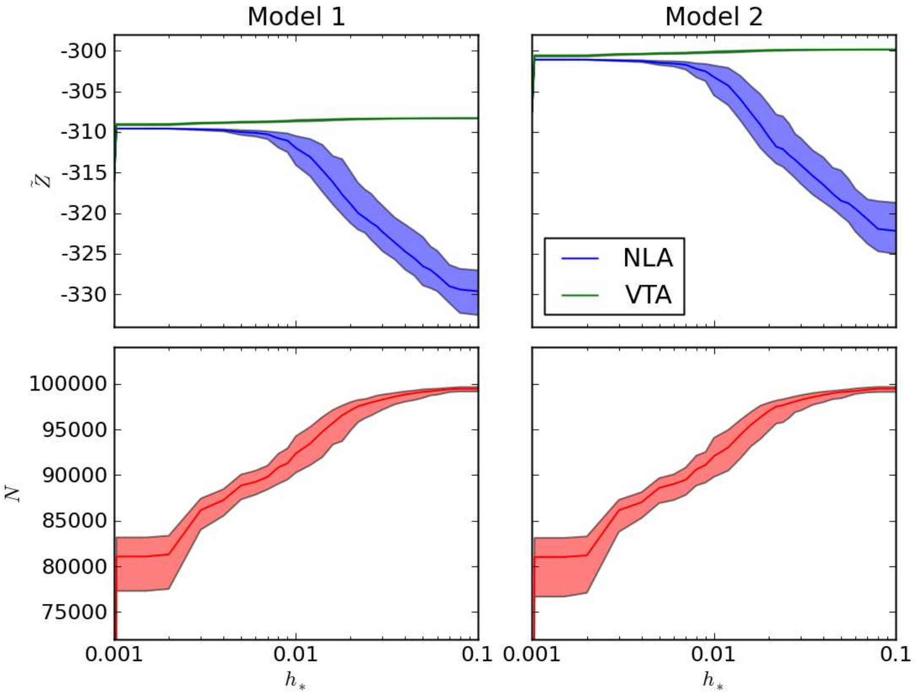


Figure 6: Comparison of the NLA and VTA as a function of h_* for Models 1 and 2. The upper panel shows the run of $\tilde{Z}$ for increasing h_*; the lower panel shows the number of states out of 100,000 that meet the h_* threshold criterion. The median (95% confidence region) is shown as a solid line (shaded band).

sample sizes of 400,000 states. Both the NLA and VTA tend to slightly overestimate Z for large k. It is surprising and encouraging that the accuracy of the VTA results exceeds the accuracy of the Laplace approximation for multivariate normal distributions.

Figure 7 depicts the intermediate numerical steps in computing $\tilde{Z}$ by the Lebesgue construction for the kd-tree and hyperoctree for the eight-dimensional normal distribution with varying sample size N. In each panel, the plotted quantities are defined as follows. $\tilde{M}(P)$ is the measure function defined by equation (9). The second two plots illustrate the contributions to the integral $\tilde{Z}$:

$$\tilde{Z}(P) = \int_{P_{min}\approx 0}^{P_{max}} dP\, \tilde{M}(P) = \int_{\log(P_{min})}^{\log(P_{max})} d(\log P)\, \tilde{M}(P)P.$$

Each panel summarizes sixteen uncorrelated samples for each value of N. The solid line and bands in each plot show the mean and extrema values, respectively, for each quantity. The top plot (blue) shows the volume associated with points with values of

Table 2: Test of high-dimensional marginal likelihood

Model		NLA		VTA				Laplace	
k	Exact	$\log\tilde{Z}$	$\Delta\%$	$\log\tilde{Z}_{Riemann}$	$\Delta\%$	$\log\tilde{Z}_{Lebesgue}$	$\Delta\%$	$\log\tilde{Z}$	$\Delta\%$
1	-1.468	$-1.54^{+0.01}_{-0.01}$	-5.0	$-1.49^{+0.06}_{-0.00}$	-1.4	$-1.49^{+0.00}_{-0.00}$	-1.4	$-1.39^{+0.00}_{-0.00}$	+5.1
2	-2.936	$-3.20^{+0.03}_{-0.03}$	-9.0	$-2.94^{+0.29}_{-0.00}$	-0.1	$-2.95^{+0.00}_{-0.00}$	-0.4	$-2.67^{+0.01}_{-0.01}$	+9.1
5	-7.34	$-7.75^{+0.01}_{-0.01}$	-5.5	$-7.29^{+0.47}_{-0.00}$	+0.6	$-7.40^{+0.01}_{-0.01}$	-0.8	$-6.88^{+0.02}_{-0.01}$	+6.3
10	-14.68	$-23.61^{+0.05}_{-0.03}$	-60.8	$-14.45^{+0.02}_{-0.01}$	+1.6	$-14.70^{+0.02}_{-0.01}$	+1.6	$-11.00^{+0.08}_{-0.11}$	+25.1
20	-29.36	$-39.63^{+0.06}_{-0.06}$	-35.0	$-32.40^{+0.32}_{-0.30}$	-10.4	$-28.51^{+0.05}_{-0.02}$	+2.9	$-18.15^{+0.09}_{-0.14}$	+38.2

probability greater than P, the middle plot (red) shows the contribution of that volume between probabilities P and $P + dP$ to the marginal likelihood integral, and the bottom plot (green) shows the contribution to the marginal likelihood integral for all probability values smaller than P. In all cases, the dominant contribution to $\tilde{Z}$ comes from neither tail but rather the mid-range values of probability. The computation with both trees exhibits shifts in $\tilde{Z}$ that are much larger than the sample variance. This is caused by the changing distribution of cell volumes for ω_s with changing sample size. The number of volume bisections for the kd-tree and the fraction of smaller volume cells for the hyperoctree increases with sample. A thresholding effect becomes increasingly apparent for small values of c. The mixture of differing volume sizes for the hyperoctree partition causes the multiple peaks in the run $\tilde{M}(P)P$. Despite these differences, both trees yield results of comparable accuracy.

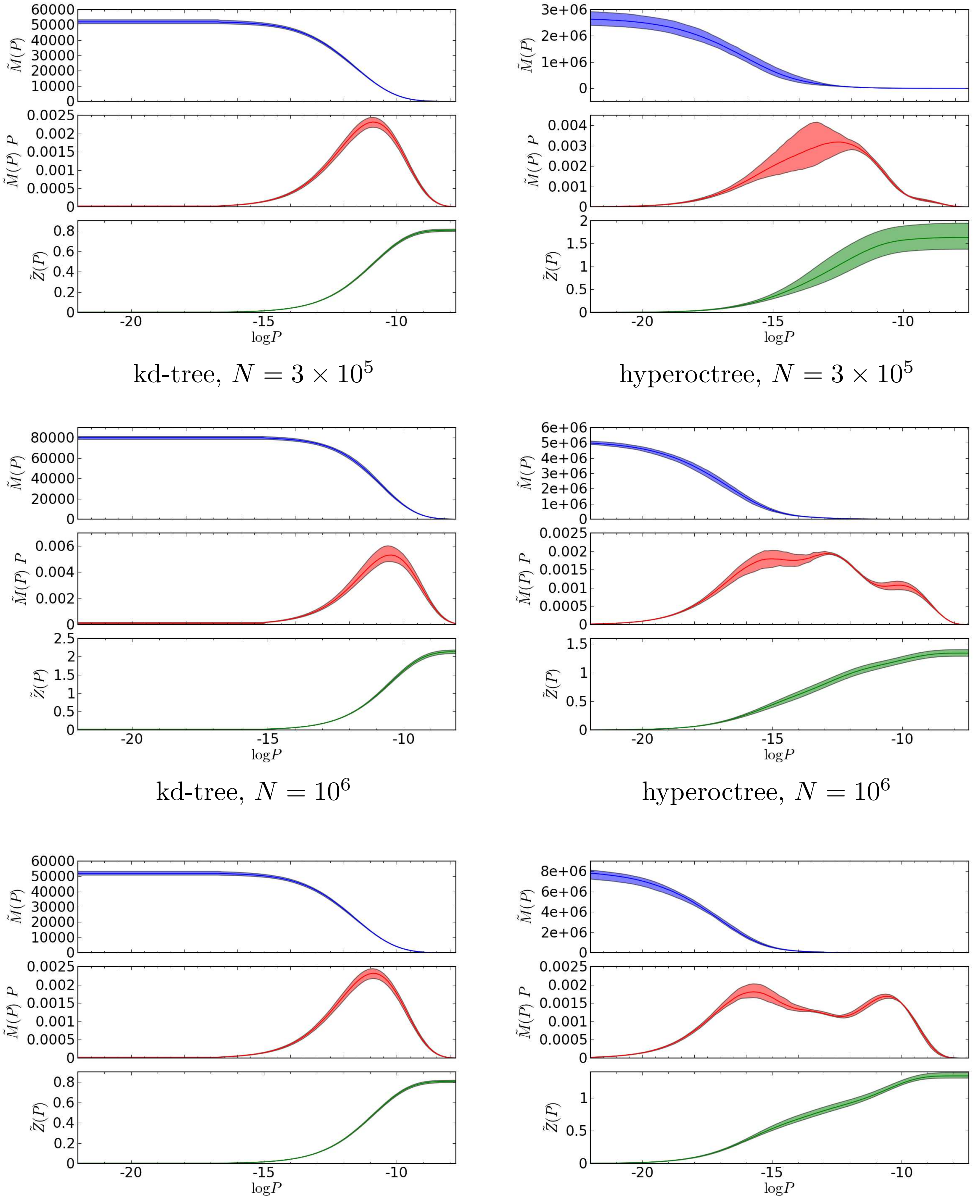
4.4 A multimodal example

All of the examples up to this point have been unimodal. We now explore a mixture of three normal distributions in four dimensions. Each component has equal probability: $w = 1/3$. The centers of each of the three components are $\mathbf{x}_c = [x_0, x_1, x_2, x_3] = [1/2, -1/2, 0, 0], [-1/2, -1/2, 0, 0], [0, 0, 1/2, 0]$ with $\sigma^2 = 0.03^2$ for each. We sample $N = 384$ points from this model and infer the three centers $(x)_c$ and weights $\mathbf{w}$. The likelihood function becomes

$$P(\mathbf{D}|\theta) = \prod_{i=1}^{N} \left[\sum_{j=1}^{3} w_j \frac{\exp(-(\mathbf{x}_i - \mathbf{x}_{c\,j})^2/2\sigma^2)}{(2\pi\sigma^2)^2} \right]. \tag{18}$$

Since $\sum_{j=1}^{3} w_j = 1$, there are 14 independent variables. We assume a Dirichlet prior with a shape parameter of $\alpha = 4$ suppresses domination by a single component. The prior distribution for the center is uniform in $[-1, 1]$ in each dimension x_j.

When the separation between the components is much larger than σ, as it is in this



kd-tree, $N = 3 \times 10^5$

hyperoctree, $N = 3 \times 10^5$

kd-tree, $N = 10^6$

hyperoctree, $N = 10^6$

kd-tree, $N = 3 \times 10^6$

hyperoctree, $N = 3 \times 10^6$

Figure 7: Illustration of the the measure function and the cumulative integral $\tilde{Z}(P)$ as a function of $\log P$ for the VTA algorithm using kd-trees and hyperoctrees for various sample sizes N as labeled. Although the hyperoctree is prone to oscillation owing to the correlation of occupation fraction with cell size, this tree yields more accurate results for the same eight dimensional normally distributed samples.

Table 3: Comparison of NLA and VTA for a multimodal model in 14 dimensions

c	NLA	Δ	VTA (Riemann)	Δ	VTA (Lebesgue)	Δ
8	$2783.4^{+1.0}_{-1.0}$	-3.0	$2780.1^{+0.8}_{-0.8}$	-6.3	$2782.4^{+1.0}_{-1.1}$	-4.0
16	$2785.7^{+0.7}_{-0.8}$	-0.7	$2781.8^{+0.7}_{-0.5}$	-4.6	$2784.2^{+0.8}_{-0.7}$	-2.2
32	$2787.3^{+0.6}_{-0.5}$	$+0.8$	$2783.1^{+10.3}_{-0.4}$	-3.3	$2785.6^{+0.7}_{-0.5}$	-0.8
64	$2788.6^{+7.9}_{-0.4}$	$+2.2$	$2784.1^{+10.1}_{-0.3}$	-2.3	$2786.9^{+7.6}_{-0.5}$	$+0.48$
128	$2789.6^{+8.6}_{-0.4}$	$+3.2$	$2784.9^{+10.0}_{-0.2}$	-1.5	$2788.0^{+8.0}_{-0.4}$	$+1.6$
256	$2790.5^{+8.7}_{-0.3}$	$+4.1$	$2785.5^{+10.0}_{-0.2}$	-0.92	$2789.0^{+8.3}_{-0.4}$	$+2.6$

case, the likelihood function is well-approximated by

$$
\begin{aligned}
P(\mathbf{D}|\theta) &\approx \prod_{j=1}^{3}\prod_{i_j=1}^{N_j} w_j \frac{e^{-(\mathbf{x}_i-\mathbf{x}_{c\,j})^2/2\sigma^2}}{(2\pi\sigma^2)^2} \\
&= (2\pi\sigma^2)^{-2N}\prod_{j=1}^{3} w_j^{N_j} e^{-(\bar{\mathbf{x}}_j-\mathbf{x}_{c\,j})^2/2\sigma^2} e^{-(N_j-1)s_j^2/2\sigma^2}, \qquad (19)
\end{aligned}
$$

where N_j is the number of points in component j and $\bar{x}_j$ is the mean for the N_j points with $j = 1, 2, 3$. Similarly, s_{jk}^2 is the sample variance for dimension k for each component j with $s_j^2 = \sum_{k=0}^{3} s_{jk}^2$. Using equation (19), the marginal likelihood may be evaluated analytically. For our particular data sample, $Z = 2786.42$.

The posterior distribution is sampled using the tempered differential evolution algorithm described in Section 4.3. From the approximately 2 million converged states, 400 samples of 100,000 are randomly selected and the NLA and the Riemann and Lebesgue variants of the kd-tree VTA are applied. The results for the NLA and the Riemann and Lebesgue variants of the VTA are as listed and compared with the analytic result in Table 3. The systematic bias in computing the volume from the tessellation is readily apparent. For small values of c and the smallest cell volumes, the volume is slightly underestimated and the resulting estimates of Z are underestimated. As c increases, the estimate of Z approaches the exact value, but the subsample variance increases as well. This is caused by the increasing inhomogeneity of the sampled points across the volume. As described in Appendix 3, this suggests that one should select c and the subvolume to be as large as one can to reduce the variance of the probability estimate within the subvolume but not so large so that the bootstrap variance is large owing to inhomogeneity. Referring to Table 3, this is $c = 16$ or 32. This strategy mildly prefers the Lebesgue variant of the VTA; it has the smallest difference from the analytic value on either side of the optimal choice, $c = 32$. Overall, both algorithms perform well here; the Lebesgue version of the VTA is slightly better on average.

5 Summary and Discussion

This paper presents two new algorithms for computing the marginal likelihood, Z, from an MCMC-sampled posterior distribution. Methods for sampling the posterior distribution with MCMC are well-developed, and robust techniques for evaluating Z will extend the utility of the often expensive-to-compute posterior sample. Standard techniques are either limited to low dimensionality, such as direct quadrature, or nearly normally-distributed samples, such as Laplace approximation. The harmonic mean approximation (HMA) has the further limitation of non-existence for a number of often used models (see Appendix 1). An obvious advantage of using the posterior sample is that the volume of the high-dimensional parameter space is naturally explored in regions where the posterior density is significant. The methods proposed here borrow techniques from numerical analysis and Lebesgue theory to compute this integral in high-dimensional parameter vectors θ with continuous values, e.g. $\theta \in \Omega \subset \mathbf{R}^k$.

The first algorithm follows from a Lebesgue reformulation of the HMA (Newton and Raftery 1994). This development reveals that the integral on the left-hand side of equation (4) will fail to exist if the measure function $M(Y)$ where $Y = L^{-1}$ from equation (12) decreases too slowly (Appendix 1). Most importantly, the numerical computation of $M(Y)$ for a posterior sample can be used to diagnose the lack of existence of the left-hand side of equation (10) and motivate the choice of a more informative prior or another evaluation method. Having verified existence, the truncation error can dominate the quadrature of the left-hand side of equation (4) unless the sample is appropriately truncated. The *Numerical Lebesgue Algorithm* (NLA) addresses this problem by determining a well-sampled subset $\Omega_s \subset \Omega$ from the MCMC sample, ex post facto (eqs. 10 and 11).

The second algorithm, the *Volume Tessellation Algorithm* (VTA), partitions the subset $\Omega_s \subseteq \Omega$ in parameter space for evaluating the measure function (eq. 9) for both J (eq. 10) and Z (eq. 2) using a space partitioning tree. Such trees recursively partition a k-dimensional parameter space into convex subspaces. The VTA is implemented with a kd-tree (Cormen et al. 2001) for simplicity. The proposed algorithms are a bit more difficult to implement and have a higher computational complexity than the simple HMA, but the overall CPU time is rather modest compared to the computational investment required to produce the MCMC-sampled posterior distribution itself. For a sample of size N, the sorting required by NLA and VTA has an approximate computational complexity of $\mathcal{O}(N \log N)$ rather than $\mathcal{O}(N)$ for the harmonic mean. Nonetheless, the computational time is a fraction of second to minutes for typical values of $10^5 < N < 10^8$ (see Section 3). Even with this directed sampling approach to numerical quadrature, the sample size required to resolve features of the same linear scale grows exponentially with dimensionality k. Nonetheless, these algorithms, and VTA in particular, promise estimates of the marginal likelihood with sufficient accuracy to compute useful Bayes factors for model spaces with dimensionality $k < 25$.

For dimensionality $k > 5$, our tests favor the VTA over the NLA unless the prior distribution is strongly informative. Similarly, the VTA is as good as or better than the Laplace approximation, even for multivariate Gaussian posterior distributions. Tests show that the Lebesgue variant of the VTA (using eqs. 31–32) has a smaller variance than the Riemann variant (eq. 33) owing to its use of sampled values within the smallest subvolumes. Conversely, because these algorithms exploit the additional structure implied by smooth, well-behaved likelihood and prior distribution functions, the algorithms developed here will be inaccurate and possibly fail miserably for *wild* density functions. So far, no such failures have appeared in practice. Based on current results, I tentatively recommend relying preferentially on VTA for the following reasons: 1) there is no concern over the existence of the integral defining the HMA; and 2) it appears to do well even in a high-dimensional parameter space. Additional real-world testing, especially on high-dimensional multimodal posteriors, will provide more insight. We are currently testing these algorithms for astronomical inference problems too complex for a simple example; the results will be reported in future papers. An implementation of these algorithms will be provided in the next release of the Bayesian Inference Engine (BIE, Weinberg and Moss 2010; Weinberg 2012). A stand-alone C++ implementation of these algorithms is available from the author upon request.

There are several natural algorithmic extensions and improvements not explored here. One may generalize the stepwise construction of M_i defined in Section 2 to a smoothed approximation for the computation of $M(Y)$. The direct integration of equation (2) using the Riemann variant of the VTA currently ignores the location of field values in each cell volume. At the expense of CPU time, the accuracy might be improved by fitting the sampled points with low-order multinomials and using the fits to derive a cubature algorithm for each cell. A similar generalization of equation (31) may benefit the Lebesgue version of the VTA. Philippe and Robert (2001) describe a number of useful techniques for computing MCMC-based Riemann sums and these may improve the VTA further. In addition, a more sophisticated tree structure may decrease the potential for bias by providing a tessellation with "rounder" cells with more adaptivity than the hyperoctree. Lastly, identifying primary modes and restricting Ω_s to the vicinity of these modes using equation (14) may eliminate the need for improved tree structures as long as the cardinality of Ω_s remains large.

References

Carlin, B. P. and Chib, S. (1995). "Bayesian model choice via Markov chain Monte Carlo methods." *Journal of the Royal Statistical Society, Series B*, 57: 473–484. 751

Chib, S. and Jeliazkov, I. (2001). "Marginal Likelihood From the Metropolis-Hastings Output." *Journal of the American Statistical Association*, 96(453): 270–281. 739

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms*. The MIT Press, 2nd edition. 744, 758

de Berg, M., Cheong, O., van Kreveld, M., and Overmars, M. (2008). *Computational Geometry: Algorithms and Applications*. Springer-Verlag. 740

DiCiccio, T., Kass, R., Raftery, A., and Wasserman, L. (1997). "Computing Bayes factors by combining simulation and asymptotic approximations." *American Statistical Association*, 92: 903–915. 739

Edelsbrunner, H. and Shah, N. (1966). "Incremental Topological Flipping Works for Regular Triangulations." *Algorithmica*, 15(3): 223–241. 740

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Texts in Statistical Science. Boca Raton, FL: CRC Press, 2nd edition. 738

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7: 457–472. 745

Geyer, C. (1991). "Markov chain Monte Carlo maximum likelihood." In *Computing Science and Statistics*, Proceedings of the 23rd Symposium on the Interface, 156. American Statistical Association. 745

Giakoumatos, S. G., Vrontos, I. D., Dellaportas, P., and Politis, D. N. (1999). "An MCMC Convergence Diagnostic using Subsampling." *Journal of Computational and Graphical Statistics*, 1: 431–451. 745

Green, P. and O'Hagan, A. (1998). "Model choice with MCMC on product spaces without using pseudo-priors." Technical report, Nottingham University. Statistics Research Report 98-01. 751

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82: 711–32. 738

Han, C. and Carlin, B. P. (2001). "MCMC Methods for Computing Bayes Factors: A Comparative Review." *Journal of the American Statistical Association*, 96: 1122–1132. 751

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90(430): 773–795. 738, 739

Lavine, M. and Schervish, M. (1999). "Bayes Factors: What they are and what they are not." *American Statistician*, 53: 119–122. 738

Neal, R. M. (1996). "Sampling from multimodal distributions using tempered transitions." *Statistics and Computing*, 6: 353–366. 752

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian inference by the weighted likelihood bootstrap." *Journal of the Royal Statistical Society, Series B*, 56: 3–48. 738, 758

Philippe, A. and Robert, C. P. (2001). "Riemann sums for MCMC estimation and convergence monitoring." *Statistics and Computing*, 11: 103–115. 759

Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. (2007). "Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity." *Bayesian Statistics*, 8: 1–45. 739

Ter Braak, C. J. F. (2006). "A Markov chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces." *Statistics and Computing*, 16: 239–249. 745, 752

Trotta, R. (2007). "Applications of Bayesian model selection to cosmological parameters." *Monthly Notices of the Royal Astronomical Society*, 378: 72–82. 739

— (2008). "Bayes in the sky: Bayesian inference and model selection in cosmology." *Contemporary Physics*, 49(2): 71–104. 739

Weinberg, M. D. (2012). "Computational statistics using the Bayesian Inference Engine." *Monthly Notices of the Royal Astronomical Society*. Submitted. 745, 752, 759

Weinberg, M. D. and Moss, J. E. B. (2010). "The UMass Bayesian Inference Engine." Technical report, University of Massachusetts/Amherst. `http://www.astro.umass.edu/BIE`. 745, 752, 759

Williams, E. (1959). *Regression Analysis*. New York: Wiley. 751

Wolpert, R. L. (2002). "Stable Limit Laws for Marginal Probabilities from MCMC Streams: Acceleration of Convergence." Discussion Paper 2002-22, Duke University ISDS. 739

Acknowledgments

I thank Michael Lavine for thoughtful discussion and both Neal Katz and Michael Lavine for comments on the original manuscript. Christian Robert provided many useful references and much encouragement. It is also a pleasure to acknowledge the thoughtful and helpful comments of the anonymous referees and the associate editor of the journal. This material is based upon work supported by the National Science Foundation under Grant Nos. 0611948 and 1109354 and by the National Aeronautics and Space Administration's AISR Program through Grant No. NNG06GF25G.

Appendix 1 Consistency of K

The evaluation of K (eq. 11) suffers from unacceptably large errors in common usage. As an example of the latter, consider the *textbook* inference of an unknown mean θ from a sample of N normally distributed points $x \sim \mathcal{N}(\theta, \sigma_x^2)$. The likelihood function is

$$L \equiv P(D|\theta) = \prod_{i=1}^{N} \frac{e^{-(x_i-\theta)^2/2\sigma_x^2}}{\sqrt{2\pi\sigma_x^2}} = L_0 e^{-(\bar{x}-\theta)^2 N/2\sigma_x^2} \tag{20}$$

where $L_0 = \sup\{L : \theta \in \mathbf{R}\}$ and $\bar{x}$ is the sample mean. Let the prior distribution for θ be $\mathcal{N}(\theta_0, \sigma_\theta^2)$. Assume that we have used an MCMC algorithm to sample the posterior distribution of θ.

Now, let us evaluate K using Lebesgue integration for this example. To compute the measure function, $M(Y)$, we solve equation (20) for $\theta = \theta(L)$, noting that the solution has two branches. After some algebra, one finds

$$M(Y) = 1 - \frac{1}{2}\left[\Phi\left(\frac{\bar{x} - \bar{y} + \Delta(Y)}{\sqrt{2\sigma_p^2}}\right) - \Phi\left(\frac{\bar{x} - \bar{y} - \Delta(Y)}{\sqrt{2\sigma_p^2}}\right)\right] \tag{21}$$

where $\Phi(\cdot)$ is the standard normal CDF and

$$\bar{y} \equiv \frac{\sigma_x^2\theta_0/N + \sigma_\theta^2\bar{x}}{\sigma_x^2/N + \sigma_\theta^2}, \qquad \sigma_p^2 \equiv \frac{\sigma_\theta^2\sigma_x^2/N}{\sigma_x^2/N + \sigma_\theta^2} = \left(\frac{1}{\sigma_\theta^2} + \frac{N}{\sigma_x^2}\right)^{-1},$$

$$Y \equiv L^{-1}, \qquad Y_0 \equiv L_0^{-1}, \qquad \Delta(Y) \equiv \sqrt{\frac{2\sigma_x^2}{N}\log(Y/Y_0)}.$$

The values $\bar{y}$ and σ_p^2 are the posterior mean and posterior variance of θ respectively. The value Y_0 is the minimum value of Y and $\Delta(\cdot)$ describes the offset of θ with increasing Y. Note that $\Delta(Y_0) = 0$ and $\Delta(Y)$ increases monotonically with Y.

Generally, the sample will not cover $[Y_0, \infty)$ but will be limited from above by the smallest sampled value of the likelihood $Y_{max} = L_{min}^{-1}$. We define this limited value of the integral K as

$$K(Y) \equiv \int_{Y_0}^{Y} dY M(Y) + M(Y_0)Y_0 = \int_{Y_0}^{Y} dY M(Y) + Y_0 \tag{22}$$

where the last equality uses $M(Y_0) = 1$. Clearly $K = K(\infty) > K(Y_{max})$. The magnitude of the truncation owing to finite sampling, $K(\infty) - K(Y_{max})$ depends on the width of the likelihood distribution relative to the prior distribution: $b \equiv \sigma_x^2/(N\sigma_\theta^2)$. The condition $\lim_{Y\to\infty}[K(\infty) - K(Y_{max})] \to 0$ requires that $M(Y)$ decreases faster than $Y^{-1-\epsilon}$ for some $\epsilon > 0$. For $b = 0$ and large Y, $\int^Y dY M(Y)$ increases as $\log(\log Y)$. For $b > 0$, $\int^Y dY M(Y)$ decreases at least as fast Y^{-b}. Figure 8 shows $K(\infty) - K(Y_0)$ as a function of b and suggests that $b > 0.1$ is sufficient to obtain sufficient numerical convergence for practical values of N. Qualitatively, a prior distribution that limits Y from above (or, equivalently, L from below) truncates the heavy tail that gives rise to the non-existence of K.

Similar asymptotic expressions for $K(Y)$ may be derived for multivariate normal distributions. Assume that data are i.i.d. in each of k dimensions and that true mean equals the sample mean. Then

$$M(Y) = \Gamma\left(\frac{k}{2}, (1+b)\log(Y/Y_0)\right)/\Gamma\left(\frac{k}{2}\right) \tag{23}$$

where $\Gamma(a, x)$ is the upper incomplete gamma function and $\Gamma(a)$ is the complete gamma function. Using the standard asymptotic formula for $\Gamma(a, x)$ in the limit of large Y, one

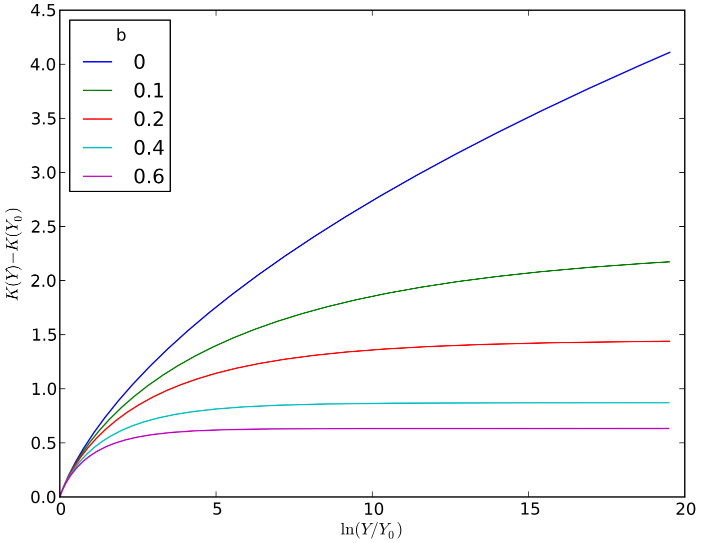


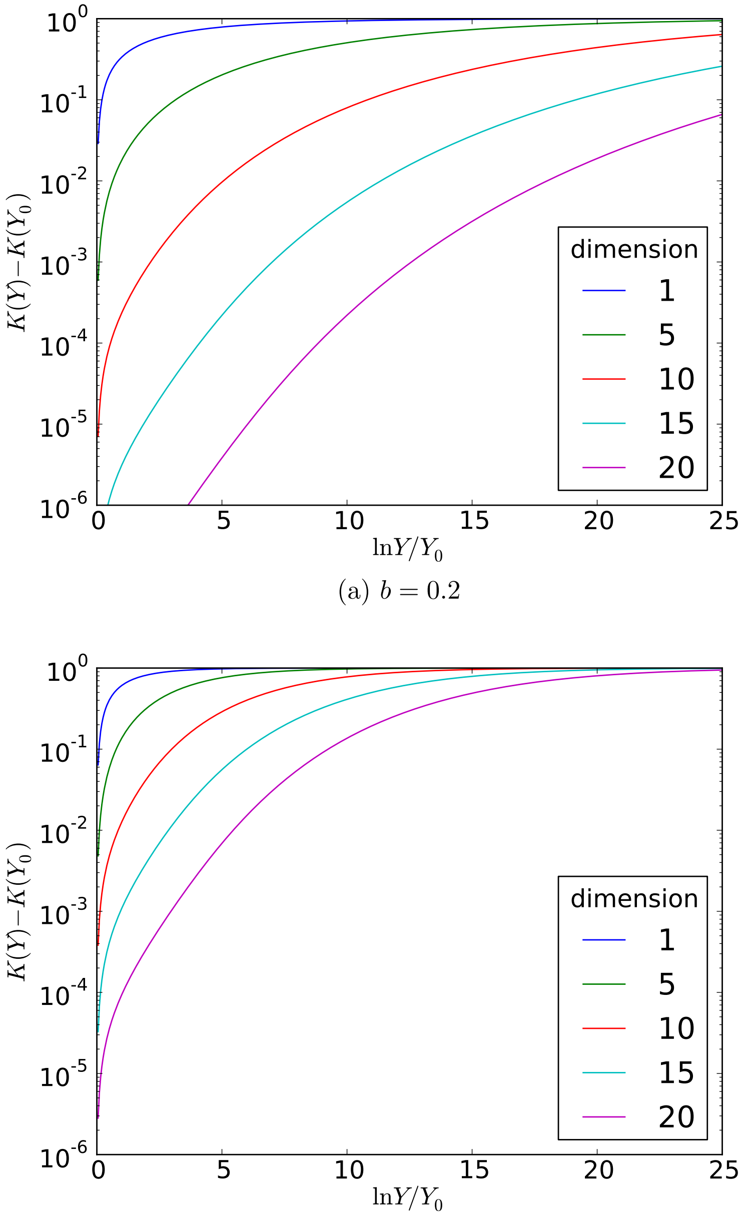
Figure 8: The integral $K(Y) - K(Y_0)$ is shown as a function of Y/Y_0 for various values of the variance ratio $b \equiv \sigma_x^2/(N\sigma_\theta^2)$. For an uninformative prior distribution $\sigma_\theta \gg \sigma_x/N$ and $b \to 0$ and $K(Y)$ increases without bound for increasing Y/Y_0. For an informative prior distribution $\mathcal{O}(b) \sim 1$ and $K(Y)$ approaches a constant value with Y/Y_0.

finds that

$$M(Y) \to \frac{1}{\Gamma(k/2)} \left[\sqrt{1+b^2} \log\left(\frac{Y}{Y_0}\right)\right]^{k/2-1} \left(\frac{Y}{Y_0}\right)^{-1-b} \qquad \text{for } Y \gg k/2. \tag{24}$$

This expression reduces to equation (21) when $k = 1$, but more importantly, this shows that the tail magnitude of $M(Y)$ increases with dimension k. Figure 9 illustrates this for various values of k and b.

In summary, $\lim_{\sigma_\theta^2 \to \infty} K \to \infty$ for a normally distributed likelihood function with an uninformative prior. Moreover, Figures 8 and 9 further demonstrate that $b = \mathcal{O}(1)$ for successful numerical evaluation. In other words, the HMA/NLA construction should not be used for problems with weakly informative priors. Intuitively, the cause is clear: if the Markov chain never samples the wings of the prior distribution which still make a significant contribution to K, then K will increase with sample size. As described at the beginning of Section 2, this failure is caused by the measure function decreasing too slowly as Y increases.



(a) $b = 0.2$

(b) $b = 0.6$

Figure 9: As in Figure 8, the integral $K(Y) - K(Y_0)$ is shown as a function of Y/Y_0 for the ratio $b = 0.2, 0.6$ for k-dimensional normal data distributions. The integral $K(Y)$ converges more rapidly with increasing b (as in Fig. 8) but increasingly slowly with k. The run of $K(Y) - K(Y_0)$ is normalized to 1 at $Y = \infty$ to facilitate comparison.

Appendix 2 Truncation error in estimating K from a Markov chain

We assume that $M(Y)$ approaches zero sufficiently fast as described in Appendix 1 and, therefore, that $\lim_{N\to\infty}\tilde{K} = L < \infty$. Then, the ranked values of $Y = 1/L$ for the MCMC-obtained posterior sample provide a natural, albeit irregular, grid for the numerical evaluation of K using standard quadrature algorithms. Even if $P(\theta|\mathbf{D})$ is smooth and continuous in θ, the mapping $Y \to \theta$ may be less well behaved. For example, consider $M(Y)$ for a one-dimensional unimodal distribution $P(\theta|\mathbf{D})$: it will be smooth and continuous. However, if we add secondary modes, discontinuities in the first derivative of $M(Y)$ will appear at the extremal values of each secondary mode. Fortunately, many practical problems are dominated by a small number of modes. Therefore, we expect the standard quadrature algorithms to be satisfactory.

To derive an algorithm, begin with the grid defined by the ordered set $\{Y_i\}$ induced by the MCMC sample. The Monte Carlo expression of equation (9) is

$$M_i^{[l]} \equiv \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{\{Y_j > Y_i\}}, \quad \text{or} \quad M_i^{[u]} \equiv \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{\{Y_j \geq Y_i\}}, \quad \text{or} \quad M_i \equiv \frac{M_i^{[l]} + M_i^{[u]}}{2}, \tag{25}$$

The indicator function includes a contribution from the index j only if $\{Y_j > Y_i\}$ or $\{Y_j \geq Y_i\}$ for the lower and upper form, respectively. Similarly, we may evaluate K from $\{M_i\}$ by upper and lower Riemann sums:

$$K^{[l]} \equiv \sum_{j=1}^{N}\left(h_j M_{j-1} + M(Y_0)Y_0 + \delta_j^{[l]}\right) \quad \text{or} \quad K^{[u]} \equiv \sum_{j=1}^{N}\left(h_j M_j + M(Y_0)Y_0 + \delta_j^{[u]}\right) \tag{26}$$

with $K \equiv (K^{[l]} + K^{[u]})/2$, $h_j \equiv Y_j - Y_{j-1}$, and

$$\delta_j^{[l]} \equiv \frac{h_j^2}{2}M'(Y_{j-1}) \approx \frac{h_j^2}{2}\left(\frac{M_j - M_{j-2}}{Y_j - Y_{j-2}}\right), \quad \delta_j^{[u]} \equiv \frac{h_j^2}{2}M'(Y_j) \approx \frac{h_j^2}{2}\left(\frac{M_{j+1} - M_{j-1}}{Y_{j+1} - Y_{j-1}}\right). \tag{27}$$

The term-by-term relative errors are then

$$\Delta_j^{[l]} = \delta_j^{[l]}/(h_j M_{j-1}), \quad \Delta_j^{[u]} = \delta_j^{[u]}/(h_j M_j). \tag{28}$$

The values of both δ_j and Δ_j can be used to eliminate terms in the sums of equation (26) with large errors, e.g. those with $\Delta_j \equiv \max(|\Delta_j^{[l]}|, |\Delta_j^{[l]}|) > \epsilon_*$ for some modest value of $\epsilon_* \ll 1$. Large errors will tend to occur for extremal values of Y or discontinuities from the multimodal nature of the posterior. In the large Y tail of the distribution, equation (28) is approximately $\ln Y_j - \ln Y_{j-1} = \ln L_{j-1} - \ln L_j$. This suggests an easy-to-apply stopping criterion for the sums in equation (26): $\ln Y_j - \ln Y_{j-1} > h_*$. In words, h_* is the maximum mesh spacing in log likelihood.

If we do not trim the sample using equation (27), we may recover the HMA from

$K^{[l]}$ (eq. 26) as follows:

$$\begin{aligned} K^{[l]} &\equiv \sum_{j=0}^{N} \left(\frac{1}{L_j} - \frac{1}{L_{j-1}} \right) M_{j-1} + \frac{M_0}{L_0} \\ &= \sum_{j=1}^{N} \frac{M_{j-1}}{L_j} - \sum_{j=1}^{N} \frac{M_j}{L_j} = \sum_{j=1}^{N} \frac{1}{L_i}(M_{i-1} - M_i) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{L_i}. \end{aligned} \tag{29}$$

In deriving equation (29), the term $M(Y_0)Y_0$ is absorbed into the sum and we use $M_N = M(Y_N) = 0$. Assuming that $J = 1$ in equation (10) and using equation (29) yields

$$\tilde{Z} \equiv \tilde{P}(\mathbf{D}) = \tilde{J}/\tilde{K} = \left(\frac{1}{N} \sum_j \frac{1}{L_j} \right)^{-1}. \tag{30}$$

Appendix 3 Evaluation of J and Z from the Markov chain

Consider the integral I described by equations (7) and (8) over the domain Ω_s. This expresses both integrals J and Z with $f(\theta) = \pi(\theta)$ and $f(\theta) = P(\theta)P(\mathbf{D}|\theta)$, respectively. We use the sampled posterior distribution to estimate the sampled volume for each subset needed to compute $M(Y)$ (eq. 9). This is then followed by a simple quadrature to compute $\int dY\, M(Y)$. The volume may be estimated straightforwardly using a space partitioning structure. A computationally efficient structure is a space partitioning tree, which divides a region of parameter space into some number of subregions at each node. The most easily implemented tree of this type for arbitrary dimension is the kd-tree (short for k-dimensional tree). Each leaf in the tree has zero volume. Each non-leaf node has the minimum volume enclosing the points in the node by coordinate planes. We partition Ω_s into exclusive subsets $\{\omega_s\}$ containing a fixed number of leaves c and let $v(\omega_s)$ be the volume enclosing each subset ω_s as determined by the partition. We also implemented the hyperoctree, a space partitioning tree structure which divides each parent volume into 2^k subvolumes by bisecting the coordinate domain in *each* dimension.

Assume that the tree has been constructed. Let $f(\theta_i)$ denote $P(\theta_i)$ or $P(\theta_i|\mathbf{D})$ for the evaluation of $M_J(y)$ or $M_Z(y)$, respectively for $\theta_i \in \omega_s$. Let $f_{min} = \inf \{f(\theta_i) : \theta_i \in \omega_s\}$ and $f_{max} = \sup \{f(\theta_i) : \theta_i \in \omega_s\}$. Order the set $\{f(\theta_i)\}$ from largest to smallest with $i \in [1, c]$. Then, the volume contribution from each ω_s to $M(y)$ is:

$$v(y, \omega_s) = v(\omega_s) \begin{cases} 0 & \text{if } y \geq f_{max} \\ 1 & \text{if } y \leq f_{min} \\ \frac{[f(\theta_j)-y](c-j)+[y-f(\theta_{j+1})](c-j-1)}{[f(\theta_j)-f(\theta_{j+1})][c-1]} & \text{s.t. } f(\theta_{j+1}) \leq y \leq f(\theta_j). \end{cases} \tag{31}$$

In equation (31), the index j takes the values $[1, c-1]$. Altogether, we have

$$M(y) \approx \sum_{\omega_s \in \Omega_s} v(y, \omega_s). \tag{32}$$

The density $f(\theta)$ and the value of y may be replaced by $g(f(\theta))$ and $g(y)$, respectively, where $g(\cdot)$ is a monotonic function. Ideally, $g(\cdot)$ is chosen to make $g(f(\theta))$ linear with sample number (e.g. $g(\cdot) = \Phi^{-1}(f)$ might be a good choice).

Alternatively, we can evaluate J or Z as a Riemann sum, multiplying $v(\omega_s)$ by some representative value of the prior probability, $f_*(\omega_s)$, for each element of ω_s (such as a p-quantile or mean value) and then sum the contributions for all ω_s:

$$I \approx \sum_{\omega_s \in \Omega_s} v(\omega_s) f_*(\omega_s). \tag{33}$$

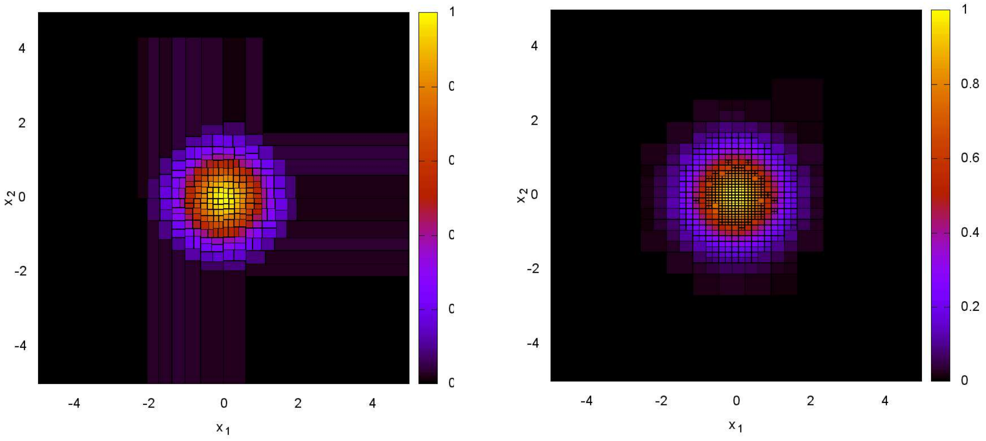
However, the error estimates resulting from the quadrature in the Lebesgue approach provide a useful check on the quality of the results. In addition, equation (31) provides additional precision since it exploits the information about the distribution of the values of $f(\omega)$ in each ω_s. Since the MCMC chain provides the values of $\pi(\theta)$ and $P(\theta|\mathbf{D}) = P(\theta)P(\mathbf{D}|\theta)$, we may use the same tree to evaluate both $\tilde{J}$ and $\tilde{Z}$ over the sampled volume Ω_s. The converged Markov chain samples the domain Ω proportional to the integrand of equation (2), and therefore, we expect

$$\lim_{N\to\infty} \int_{\Omega_s} d\theta \pi(\theta) P(\mathbf{D}|\theta) \gg \lim_{N\to\infty} \int_{\Omega\setminus\Omega_s} d\theta P(\theta) P(\mathbf{D}|\theta) \to 0$$

for large sample size by construction.

Figure 10 illustrates the tree construction for a single $k = 2$ sample. Each two-dimensional cell is colored by the median value of the posterior probability for the $c = 32$ points in each cell and scaled to the peak value of posterior probability P for the entire sample. A careful by-eye examination of the cell shape reveals a preponderance of large axis-ratio rectangles; this is a well-known artifact of the algorithm. Conversely, all cells in the hyperoctree are squares. For large values of P, the volume elements are small, and with a sufficiently large sample, the gradient in P across the volume is small. For small values of P, the volume elements are large, the gradients are large, and the large-axis ratio rectangles distort the reconstructed shape of the true posterior. The error from this feature of the partition can be reduced by restricting Ω_s to exclude the poorly sampled tails or by using an over-dispersed target distribution as described in Section 4.1. Alternatively, one may scale some or all of the dimensions using equation (17) to decrease the volume in the tails of the posterior distribution. The Jacobian of the transformation increases the weight of the probability value in the cells without improving the sampling, so this is not a guaranteed or generally recommended strategy.

There is a variance–bias trade off in choosing c, the number of leaves per subset ω_s. In tails of the sample, the variance (bias) in the volume estimate increases (decreases) as the number of sample points per subset ω_s increases (decreases). In other words, a different sample will have different outliers and lead to large changes in the volume of the partition owing to the sparse sampling. The converse applies to the estimation of the probability values in each subvolume. For the hyperoctree, the number of levels in the tree will increase with increasing sample size discretely. Therefore, the number of leaves per ω_s will tend to oscillate with increasing sample size. The oscillation amplitude will



kd-tree

hyperoctree

Figure 10: Two-dimensional illustration of the domain decomposition for the Gaussian likelihood example described in Section 4.3. The cells are colored according to posterior probability on a linear scale from 0 to $\sup\{P\}$.

decrease with increasing c. This suggests using c as large as possible while continuing to resolve the target distribution. For high-dimensional distributions, large c may be prohibitively expensive. Using small values of c introduces biases of unity order, but this may be suitable for many applications.

Tests to date suggest that the sensitivity to c is reduced by using the Lebesgue form for Z owing to the reduced bias from the posterior probability values. The prior probability value will be slowly varying over the posterior sample for a typical likelihood-dominated posterior distribution, so the bias in $\tilde{J}$ will be smaller than that in $\tilde{Z}$. This suggests that a smaller number of points per cell is better for the evaluation of J and that a larger number is better for Z. I recommend the sensitivity to c be investigated anew for each problem of interest. Some practical examples suggest that the resulting estimates are not strongly sensitive to the number of points per cell; $c = 8, 16$ or 32 appears to be a good compromise.

Finally, note that the VTA applies to importance sampled distributions as well. That is, suppose one samples the target distribution $P(\theta|\mathbf{D})$ using the sampling distribution $Q(\theta|\mathbf{D})$. Then, the marginal likelihood integral becomes

$$Z = \int dQ \int_{Q(\theta|\mathbf{D})>Q} d\theta \, \frac{P(\theta|\mathbf{D})}{Q(\theta|\mathbf{D})} \equiv \int dQ M(Q). \tag{34}$$

The measure function $M(Q)$ reduces to the standard expression (eq. 9) when $Q = P$.

For the powered-up sampling function discussed in Section 4.1, $Q = P^{1/T}$ for $T > 1$ and the measure function for this choice becomes

$$M(Q) = \int_{P(\theta|\mathbf{D})^{1/T} > Q} d\theta \, P(\theta|\mathbf{D})^{1-1/T}.$$

As the temperature T increases and the sampling function becomes increasingly broad in the parameter space, the measure function becomes a step function whose full value is the marginal likelihood. Presumably, each problem has an optimal value of T that minimizes the overall error in $\tilde{Z}$.