# Automated Model Selection with Bayesian Quadrature

**Henry Chai** [1] **Jean-François Ton** [2] **Michael A. Osborne** [3 4] **Roman Garnett** [1]

## Abstract

We present a novel technique for tailoring Bayesian quadrature (BQ) to model selection. The state-of-the-art for comparing the evidence of multiple models relies on Monte Carlo methods, which converge slowly and are unreliable for computationally expensive models. Although previous research has shown that BQ offers sample efficiency superior to Monte Carlo in computing the evidence of an individual model, applying BQ directly to model comparison may waste computation producing an overly-accurate estimate for the evidence of a clearly poor model. We propose an *automated and efficient* algorithm for computing the most-relevant quantity for model selection: the posterior model probability. Our technique maximizes the mutual information between this quantity and observations of the models' likelihoods, yielding efficient sample acquisition across disparate model spaces when likelihood observations are limited. Our method produces more-accurate posterior estimates using fewer likelihood evaluations than standard Bayesian quadrature and Monte Carlo estimators, as we demonstrate on synthetic and real-world examples.

## 1. Introduction

Model selection is a fundamental problem that arises in the course of scientific inquiry: which of several candidate models best explains an observed dataset $\mathcal{D}$? The Bayesian approach to model selection involves computing *posterior model probabilities,* the probability that each model generated the observations. This approach requires computing *model evidences,* which can be expressed as integrals of the form $Z = \int f(\mathcal{D} \mid \theta)\, \pi(\theta)\, \mathrm{d}\theta$, where $\theta$ is a vector of model parameters, $f(\mathcal{D} \mid \theta)$ is a likelihood, and $\pi(\theta)$ is a prior.

Unfortunately, for many real-world model selection tasks, these integrals are computationally intractable and must be estimated numerically. Numerous commonly-used techniques to estimate such integrals rely on *Monte Carlo* estimators (Metropolis et al., 1953; Hastings, 1970). These methods converge slowly in terms of the number of required integrand samples as they do not incorporate knowledge about sample locations. This makes such methods ill-suited for settings where the integrand is expensive to evaluate.

One alternative is *Bayesian quadrature* (BQ) (Larkin, 1972; Diaconis, 1988; O'Hagan, 1991; Rasmussen & Ghahramani, 2003), which relies on a probabilistic belief on the integrand that can be conditioned on observations to derive a posterior belief about the value of the integral. The theoretical properties of kernel quadrature methods (including BQ) have been studied at length: these methods can achieve faster convergence rates than Monte Carlo estimators (Briol et al., 2015; Bach, 2017; Karvonen et al., 2018), even when the underlying model is misspecified (Kanagawa et al., 2016; 2017), a commonly-cited pitfall of kernel-based methods.

There has been significant work investigating how traditional, Monte Carlo based methods can be adapted to efficiently estimate posterior model probabilities for model selection (Neal, 2001; Green, 1995; Carlin & Chib, 1995; Meng & Wong, 1996; Godsill, 2001; Skilling, 2004). However, no analogous work has appeared to adapt BQ for this important task. That is our goal in this work.

We propose a principled adaptation of BQ designed to automate model selection. Specifically, we define a novel acquisition function for active selection of locations to observe model likelihoods. This acquisition function corresponds to the mutual information between observations of the model likelihood and a quantity specifically relevant to the task of model selection: the posterior model probabilities. This allows our method to automatically select informative sample locations across multiple model parameter spaces unlike previous active BQ approaches to model selection (Osborne et al., 2012; Gunter et al., 2014; Chai & Garnett, 2019), which focused on accurately estimating individual model evidences. We illustrate the shortcomings of such approaches using a toy motivating example. Experiments conducted on

[1]Department of Computer Science and Engineering, Washington University in St. Louis, Saint Louis, MO, USA [2]Department of Statistics, University of Oxford, Oxford, United Kingdom [3]Department of Engineering Science, University of Oxford, Oxford, United Kingdom [4]Mind Foundry, Oxford, United Kingdom. Correspondence to: Henry Chai <hchai@wustl.edu>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

real-world and synthetic data demonstrate that our proposed method can outperform existing BQ techniques and specialized Monte Carlo methods in terms of efficiently reaching accurate estimates of posterior model probabilities.

## 2. Related Work

Much work has been devoted to developing Monte Carlo methods specifically designed for model selection. Broadly speaking, these methods can be broken down into two groups: within-model approaches, such as annealed importance sampling (AIS) (Neal, 2001), nested sampling (Skilling, 2004), and bridge sampling (Bennett, 1976; Meng & Wong, 1996), and trans-dimensional approaches such as Green's (1995) reversible jump MCMC and Godsill's (2001) composite model space framework, a generalization of the product-space approach proposed by Carlin & Chib (1995). Within-model approaches estimate each model's model evidence separately, whereas trans-dimensional approaches directly estimate posterior model probabilities.

In our experiments we compare our method against one prominent Monte Carlo method from each category: bridge sampling (within-model) and reversible jump MCMC (trans-dimensional). All commonly used Monte Carlo methods for model selection have pros and cons, and their performance on specific tasks can be greatly affected by open-ended modeling choices such as the choice of intermediate densities for AIS or the choice of pseudo-priors for the composite model space framework of Godsill (2001). It is beyond the scope of this work to comprehensively analyze all widely used Monte Carlo model selection methods; however, we believe the chosen benchmarks to be reasonable and competitive. In particular, the design choices associated with bridge sampling are easily justified and give rise to greater transparency in our experimental design as opposed to many possible alternatives.

Among the commonly used trans-dimensional methods, the original product space approach of Carlin & Chib (1995) made use of a Gibbs sampler, which requires conjugate conditional likelihoods that do not exist in many model selection settings. Godsill (2001) showed that replacing the Gibbs sampler in their composite space model with a Metropolis–Hastings proposal mechanism gives rise to Green's (1995) reversible jump MCMC. Godsill (2001) also claimed that the use of such a Metropolis–Hastings proposal mechanism is preferable to the use of a Gibbs sampler in nested model settings, that is, settings where there is an overlap in model parameter spaces. As our real-world experimental setting is a model selection task between nested models, the choice to compare against reversible jump MCMC is well justified.

Previous work has been done on adapting BQ to situations where the integrand of an intractable integral is known to be nonnegative *a priori* (Osborne et al., 2012; Gunter et al., 2014; Chai & Garnett, 2019). Such integrals occur frequently in machine learning tasks, including model selection: the model evidence is an integral of probability distributions, which are nonnegative everywhere. These methods make use of warped Gaussian processes (GPs) (Snelson et al., 2004) to weakly enforce the nonnegativity constraint. They have been shown to outperform BQ algorithms that are agnostic to *a priori* information on a variety of model selection tasks. However, we will show that our method can lead to even greater improvements. Furthermore, our methodology is compatible with the use of warped GPs; in fact, our proposed method can be seen as an instantiation of the framework laid out by Chai & Garnett (2019) with a novel acquisition function.

As a final note, our focus in this manuscript is on the traditional Bayesian approach to model selection, which involves the computation of model posteriors. We acknowledge the existence of several alternative approaches to Bayesian model selection (Bernardo & Smith, 1994; Vehtari, 2001; Watanabe, 2010). An extension of our proposed method to these alternatives is a potential line of future inquiry.

## 3. Background

### 3.1. Bayesian Model Selection

For the purposes of this work, a model is defined as a parametric family of probability distributions that can be used to explain some observed dataset, $\mathcal{D}$. Given a finite set of candidate models $\{\mathcal{M}_1, \ldots, \mathcal{M}_k\}$, the Bayesian approach to inference in this setting is to reason about the conditional or posterior distribution over models via Bayes theorem:

$$\begin{aligned}\Pr(\mathcal{M}_i \mid \mathcal{D}) &= \frac{\Pr(\mathcal{D} \mid \mathcal{M}_i)\Pr(\mathcal{M}_i)}{\Pr(\mathcal{D})} \\ &= \frac{\Pr(\mathcal{D} \mid \mathcal{M}_i)\Pr(\mathcal{M}_i)}{\sum_{j=1}^{k}\Pr(\mathcal{D} \mid \mathcal{M}_j)\Pr(\mathcal{M}_j)} \end{aligned} \tag{1}$$

where $\Pr(\mathcal{M}_i \mid \mathcal{D})$ is known as the posterior probability of model $\mathcal{M}_i$, $\Pr(\mathcal{D} \mid \mathcal{M}_i)$ is the model evidence of model $\mathcal{M}_i$ and $\Pr(\mathcal{M}_i)$ is the prior probability of model $\mathcal{M}_i$. The computation of model evidences requires integrating out the model parameters that control the likelihood of a given model generating the observed data:

$$\Pr(\mathcal{D} \mid \mathcal{M}_i) = \int \Pr(\mathcal{D} \mid \mathcal{M}_i, \theta_i)\Pr(\theta_i)\,\mathrm{d}\theta_i \tag{2}$$

where $\theta_i$ is the vector of model parameters corresponding to model $\mathcal{M}_i$, $\Pr(\mathcal{D} \mid \mathcal{M}_i, \theta_i)$ is the likelihood of $\mathcal{D}$ under $\mathcal{M}_i$ parameterized by $\theta_i$, and $\Pr(\theta_i)$ is the prior probability of the model parameters $\theta_i$.

Given posterior model probabilities, one common practice is to find $\mathcal{M}^* = \arg\max_{\mathcal{M}} \Pr(\mathcal{M}_i \mid \mathcal{D})$ and then treat $\mathcal{M}^*$

as the true, data-generating model for the purpose of future inference tasks. We will refer to this variant of the model selection task as *model choice*. An alternative to model choice is *model averaging*: instead of simply using the single, most-likely candidate model, model averaging takes a fully-Bayesian viewpoint by using the posterior model probabilities to marginalize out the choice of model for subsequent inference tasks.

### 3.2. Mutual Information

The *mutual information* between two random variables is a measure of how much information observing the value of one provides about the other. Formally, given two random variables $X$ and $Y$, the mutual information of $X$ and $Y$ is

$$I(X;Y) = H(X) - H(X \mid Y), \tag{3}$$

where $H(X)$ is the *entropy* of $X$, and $H(X \mid Y)$ is the *conditional entropy* of $X$ given $Y$. If $X$ is discrete with PMF $p$ and domain $\mathcal{X}$, then the entropy is defined to be

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{4}$$

and if $X$ is continuous with PDF $p$ and domain $\mathcal{X}$, then we instead use the *differential entropy:*

$$H(X) = -\int_{\mathcal{X}} p(x) \log p(x) \, \mathrm{d}x. \tag{5}$$

The conditional entropy $H(X \mid Y)$ is defined to be the expected (differential) entropy of the posterior distribution $p(X \mid Y)$, where the expectation is taken with respect to $Y$. Therefore the mutual information can be interpreted as the expected information gained about $X$ (that is, the expected reduction in entropy) when measuring $Y$.

### 3.3. Bayesian Quadrature

Given some intractable integral $Z = \int f(\theta)\, \pi(\theta) \, \mathrm{d}\theta$, Bayesian quadrature places a Gaussian process (GP) prior belief on the function $f(\theta)$ (or occasionally on the product $f(\theta)\, \pi(\theta)$ directly). A GP specifies a probability distribution over functions, where the joint distribution of the function's value at finitely many locations is multivariate normal. A GP is fully specified by its first two moments: a mean function $\mu(x)$ and a covariance function $\Sigma(x, x')$. Given a set of observations at locations $x_D = \{x_1, \ldots, x_n\}$ with corresponding function values $f(x_D)$, a GP can be conditioned on these observations to arrive at a posterior GP with mean $\mu_D(x) = \mu(x) + \Sigma(x, x_D)\Sigma(x_D, x_D)^{-1}(f(x_D) - \mu(x_D))$ and covariance $\Sigma_D(x, x') = \Sigma(x, x') - \Sigma(x, x_D)\Sigma(x_D, x_D)^{-1}\Sigma(x_D, x')$. For a comprehensive overview of GPs, see (Rasmussen & Williams, 2006).

Bayesian quadrature makes use of the fact that GPs are closed under linear functionals (Rasmussen & Ghahramani, 2003), meaning that a GP belief on $f$ induces a Gaussian belief on $L[f]$, where $L$ is any linear functional. As integration against a probability distribution is such a functional, if $p(f) = \mathcal{GP}(f; \mu, \Sigma)$, then $p(Z) = \mathcal{N}(Z; m, K)$, where

$$m = \int \mu(x)\, \pi(x) \, \mathrm{d}x \,; \tag{6}$$

$$K = \iint \Sigma(x, x')\, \pi(x)\, \pi(x') \, \mathrm{d}x \, \mathrm{d}x' \,. \tag{7}$$

A design choice that must be addressed when using BQ is where to observe the integrand $f$. One natural approach is to select sample locations so as to minimize one's uncertainty about $Z$ which is equivalent to minimizing the entropy of $Z$. Because the posterior variance of a GP does not depend on the observed function values (see above), if a GP prior is placed directly on $f$, then an optimal sampling design (w.r.t. this objective) can be specified in advance (Minka, 2000). However, it is often appropriate to specify a GP prior not on $f$ but on an affine transformation of $f$ in order to incorporate some *a priori* information. Doing so introduces a dependency between the posterior variance and observed function values. Thus, the optimal sampling sequence cannot be precomputed. One option in this setting is to sequentially select sample locations in order to minimize the expected entropy of $Z$ as proposed by Osborne et al. (2012). As an alternative, Gunter et al. (2014) propose an active sampling mechanism that iteratively minimizes the entropy of the integrand instead of the value of the integral as doing so is more computationally efficient and numerically stable.

## 4. Motivation

Consider the task of selecting between two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, given data $\mathcal{D}$. Suppose that BQ is used to estimate the model evidences for both models. After some number of iterations of BQ, the posterior beliefs (implicitly conditioned on BQ evaluations) on the model evidences are plotted on the same axis; as an example, see Figure 1.

Figure 1 depicts a situation where there is high uncertainty about both model evidences; however, the uncertainty in the posterior model probabilities is low. In particular, for this toy example, $\Pr(\mathcal{M}_1 \mid \mathcal{D})$ is almost certainly close to one, while $\Pr(\mathcal{M}_2 \mid \mathcal{D})$ is almost certainly close to zero. This example illustrates the fact that it is not necessary to have low-entropy estimates of model evidences to have low-entropy estimates of posterior model probabilities and thus, methods that aim to achieve accurate estimates of model evidences may be inefficiently sampling observations when the goal to is to achieve accurate estimates of the posterior model probabilities.

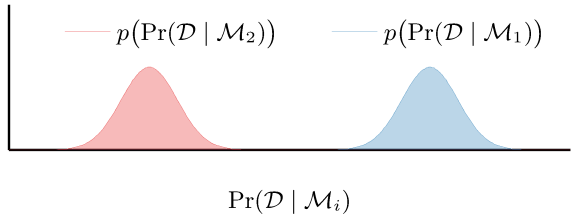



*Figure 1.* A toy example of the posteriors for two model evidences; observe that both posterior beliefs are Gaussian, a direct result of the closure of GPs under linear functionals (see (6) and (7))

## 5. Methods

We propose an adaptation of the standard BQ algorithm to the task of model selection. The principal novelty of our proposed method is in selecting where to observe the likelihood functions of the models involved. Rather than selecting locations with the goal of achieving accurate estimates of the model evidences, as previous work has considered at length (Osborne et al., 2012; Gunter et al., 2014), our method seeks to achieve accurate estimates of posterior model probabilities, a more essential quantity to model selection. Our method makes use of the mutual information between model likelihood observations and posterior model probabilities, allowing them to choose informative sample locations across multiple model parameter spaces simultaneously.

Suppose that we have observed some dataset $\mathcal{D}$ and have $k$ candidate models $\{\mathcal{M}_1, \ldots, \mathcal{M}_k\}$, to explain $\mathcal{D}$. Let $\ell_i(\theta_i) = p(\mathcal{D} \mid \theta_i, \mathcal{M}_i)$ be the likelihood for $\mathcal{M}_i$. We assume that these likelihoods have mutually independent Gaussian process priors: $p(\ell_i) = \mathcal{GP}(\ell_i; \mu_i, \Sigma_i)$. Now the model evidence of $\mathcal{M}_i$:

$$a_i = \int_{\Theta_i} \ell_i(\theta_i)\, \pi_i(\theta_i)\, \mathrm{d}\theta_i \tag{8}$$

is normally distributed as $p(a_i) = \mathcal{N}(a_i; m_i, K_i)$, where $m_i$ and $K_i$ are given by the BQ identities (6) and (7). The posterior probability of $\mathcal{M}_i$ can be expressed as:

$$z_i = \frac{a_i}{\sum_{j=1}^{k} a_j}. \tag{9}$$

Formally, we consider the mutual information between $\ell_i(\theta_i)$ and $\mathbf{z} = [z_1, \ldots, z_k]$:

$$I\big(\ell_i(\theta_i); \mathbf{z}\big) = H\big(\ell_i(\theta_i)\big) - H\big(\ell_i(\theta_i) \mid \mathbf{z}\big)\,. \tag{10}$$

Here $\ell_i(\theta_i)$ is just a univariate Gaussian and its entropy is $H\big(\ell_i(\theta_i)\big) = \frac{1}{2}\log 2\pi e\, \Sigma_i(\theta_i, \theta_i)$. Interestingly, the conditional random variable $\ell_i(\theta_i) \mid \mathbf{z}$ is also a univariate Gaussian. To see this, consider the joint density between $\ell_i(\theta_i)$ and $\mathbf{b_{-i}} = [b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_k]$ where

$$b_j = \bar{z}_j a_j + z_j \sum_{j' \in [k]\,, j' \neq j} a_{j'}\,, \tag{11}$$
$$\bar{z}_j = z_j - 1\,,\ [k] = \{1, \ldots, k\}\,.$$

As all $a_i$ are independent and Gaussian, $\mathbf{b_{-i}}$ follows a multivariate Gaussian distribution. We will denote the mean of $b_j$ by $\beta_j$, the variance of $b_j$ by $s_{j,j}$ and the covariance between $b_j$ and $b_{j'}$ by $s_{j,j'}$ where

$$\beta_j = \bar{z}_j m_j + z_j \sum_{j' \in [k]\,, j' \neq j} m_{j'}\,, \tag{12}$$

$$s_{j,j} = \bar{z}_j^2 K_j + z_j^2 \sum_{j' \in [k]\,, j' \neq j} K_{j'}\,, \tag{13}$$

$$s_{j,j'} = \bar{z}_j z_{j'} K_j + z_j \bar{z}_{j'} K_{j'} + z_j z_{j'} \sum_{j'' \in [k]\,, j'' \neq j, j'} K_{j''}\,. \tag{14}$$

Furthermore, $\ell_i(\theta_i)$ is jointly Gaussian with $\mathbf{b_{-i}}$; the covariance between $\ell_i(\theta_i)$ and $b_j$ is $z_j L_i(\theta_i)$ where

$$L_i(\theta_i) = \int_{\Theta_i} \Sigma_i(\theta_i, \theta_i')\, \pi_i(\theta_i')\, \mathrm{d}\theta_i'\,, \tag{15}$$

a result that follows from our assumption that all GP priors are mutually independent and the fact that GPs are closed under integration.

Lastly, we note that observing $\mathbf{z}$ is equivalent to observing that $\mathbf{b_{-i}} = \mathbf{0}$; in essence, this follows because an observation of $\mathbf{z}$ collapses the joint Gaussian distribution between all model evidences down to a hyperplane, where each point with support under the conditional belief satisfies the invariant that $b_j = \bar{z}_j a_j + z_j \sum_{j' \in [k]\,, j' \neq j} a_{j'} = 0$ $\forall\ j \in [k]$. Thus, we can conclude that $\ell_i(\theta_i) \mid \mathbf{z}$ and $\ell_i(\theta_i) \mid \mathbf{b_{-i}} = \mathbf{0}$ have the same distribution. Using the fact that Gaussians are closed under conditioning, it follows that $\ell_i(\theta_i) \mid \mathbf{z}$ is a Gaussian random variable and its entropy is $H\big(\ell_i(\theta_i) \mid \mathbf{z}\big) = \frac{1}{2}\log 2\pi e\, \Sigma_{i|\mathbf{z}}(\theta_i, \theta_i)$ where

$$\Sigma_{i|\mathbf{z}}(\theta_i, \theta_i) = \Sigma_i(\theta_i, \theta_i) - \begin{bmatrix} z_1 L_i(\theta_i) \\ \vdots \\ z_{i-1} L_i(\theta_i) \\ z_{i+1} L_i(\theta_i) \\ \vdots \\ z_k L_i(\theta_i) \end{bmatrix}^T \begin{bmatrix} s_{1,1} & \ldots & s_{1,k} \\ \vdots & \ddots & \vdots \\ s_{i-1,1} & \ldots & s_{i-1,k} \\ s_{i+1,1} & \ldots & s_{i+1,k} \\ \vdots & \ddots & \vdots \\ s_{k1} & \ldots & s_{kk} \end{bmatrix}^{-1} \begin{bmatrix} z_1 L_i(\theta_i) \\ \vdots \\ z_{i-1} L_i(\theta_i) \\ z_{i+1} L_i(\theta_i) \\ \vdots \\ z_k L_i(\theta_i) \end{bmatrix}. \tag{16}$$

The mutual information between $\ell_i(\theta_i)$ and $\mathbf{z}$ can now be framed as an expectation over $\mathbf{z}$:

$$I(\ell_i(\theta_i); \mathbf{z}) = \tfrac{1}{2}\log \Sigma_i(\theta_i, \theta_i) - \tfrac{1}{2}\int \log\big(\Sigma_{i|\mathbf{z}}(\theta_i, \theta_i)\big)\, p(\mathbf{z})\, \mathrm{d}\mathbf{z}. \tag{17}$$

To design our next observation, we choose to evaluate the likelihood at the point maximizing the expected information gain about $\mathbf{z}$, where we maximize over the parameter spaces of *all models.* Formally, we choose to observe the likelihood $\ell_{i^*}(\theta^*_{i^*})$ where

$$\theta^*_i = \underset{\theta_i \in \Theta_i}{\arg\max}\, I\big(\ell_i(\theta_i); \mathbf{z}\big)\,, \tag{18}$$

$$i^* = \underset{i \in \{1,\ldots,k\}}{\arg\max}\, \theta^*_i\,. \tag{19}$$

Unfortunately, the integral in (17) is intractable. Luckily, it can be accurately and efficiently estimated using standard numerical techniques. Specifically, we can generates samples of $\mathbf{z}$ by drawing from the posterior distribution over $\mathbf{a} = [a_1, \ldots, a_k]$ using (9). Also note that this estimation is only performed as a means of selecting likelihood observation locations; the effect of inaccuracies in this estimate on the estimated model evidences will be negligible.

As a final note, observe that it is imperative to omit an element from the vector $\mathbf{b_{-i}}$ as specifying $k-1$ of the $k$ model posteriors fully determines the last one: the covariance matrix of the full random vector $\mathbf{b} = [b_1, \ldots, b_k]$ is singular, making the square matrix in (16) non-invertible. The choice to omit $b_i$ is largely arbitrary, although it does simplify the notation slightly.

## 6. Experiments

We perform experiments on synthetic and real-world data in which we compare our proposed method against round-robin BQ, where likelihood evaluations are evenly distributed between all model parameter spaces, and two Monte Carlo based benchmarks: bridge sampling (Meng & Wong, 1996) and reversible jump MCMC (Green, 1995). Our implementation of bridge sampling follows the one described by Gronau et al. (2017): specifically, we use the optimal bridge function defined by Meng & Wong (1996) and a Gaussian proposal distribution with moments fit to samples from the true posterior distribution (as suggested by Overstall & Forster (2010)). Our choice of diffeomorphism for reversible jump MCMC varies by experimental setting and is described in the relevant sections below. For all BQ methods, constant-mean GP priors with Matérn covariance functions ($\nu = 3/2$) were placed on the log of the model likelihoods and all GP hyperparameters were fit in accordance with the framework defined by Chai & Garnett (2019). Our implementation of round-robin BQ uses uncertainty sampling to select locations to observe log-likelihoods, as proposed by Gunter et al. (2014).

### 6.1. Synthetic Experiments

For our synthetic experiments, we consider a model selection task between two zero-mean GP models: one chosen to have a squared exponential covariance and one chosen to have a Matérn covariance with $\nu = 5/2$. The observed dataset $\mathcal{D}$ consists of $5d$ observations from a $d$-dimensional, zero-mean GP with a squared exponential covariance. Each model is parameterized by the $d$ length-scales of their respective covariance functions (for the sake of simplicity, all other GP hyperparameters were set to be the same as the true, data-generating GP's). In this setting, prior knowledge of the experimental setting suggests that the two likelihood functions are similar. Thus an appropriate choice of diffeomorphism is the identity function and the corresponding Jacobian is always 1. The intractable integrals associated with our proposed method are estimated using 10 000 simple Monte Carlo (SMC) samples, i.e., samples drawn from the probability distribution being integrated against.

We allot a budget of $50d$ total likelihood evaluations and initialize each BQ based method with $5d$ randomly sampled likelihood observations from both model parameter spaces. We run experiments with $d$ ranging from 1 to 4 and for each value of $d$, we consider 100 different, randomly sampled observed datasets. All methods are evaluated on the fractional error of their $z_1$ estimates with ground truth values being determined by exhaustive SMC.

As Figure 2 shows, our proposed method outperforms all benchmarks compared against. Furthermore, the difference in performance between our proposed method and both round-robin BQ and bridge sampling is significant at the 1% significance level across all dimensions according to a one-sided paired $t$-test; the difference between our proposed method and reversible jump MCMC is significant at the 1% significance level for $d$ =1, 2, and 3.

### 6.2. Real-World Experiments

Our real-world application is a model selection problem from the field of astrophysics. Given spectrographic observations of quasar emissions, astrophysicts are interested in inferring the existence of damped Lyman-$\alpha$ absorbers (DLAs) between the quasar and earth. DLAs are large gaseous clouds containing neutral hydrogen at high densities. The distribution of DLAs throughout the universe is of interest as it informs galaxy formation models. Their location and size can be inferred from quasar spectra as they cause distinctive dips in the observed flux at well-defined wavelengths. Garnett et al. (2017) developed a model that specifies the likelihood that a quasar emission spectrum contains arbitrarily many DLAs. Their likelihood model for $n$ DLAs is parameterized by $2n$ parameters, two for each putative DLA: one that corresponds to its size and one that corresponds to its distance from earth. Garnett et al. (2017) also specified a data-driven prior over these parameters; computing the model evidence for any number of DLAs requires integrating the likelihood against this prior, an intractable integral.

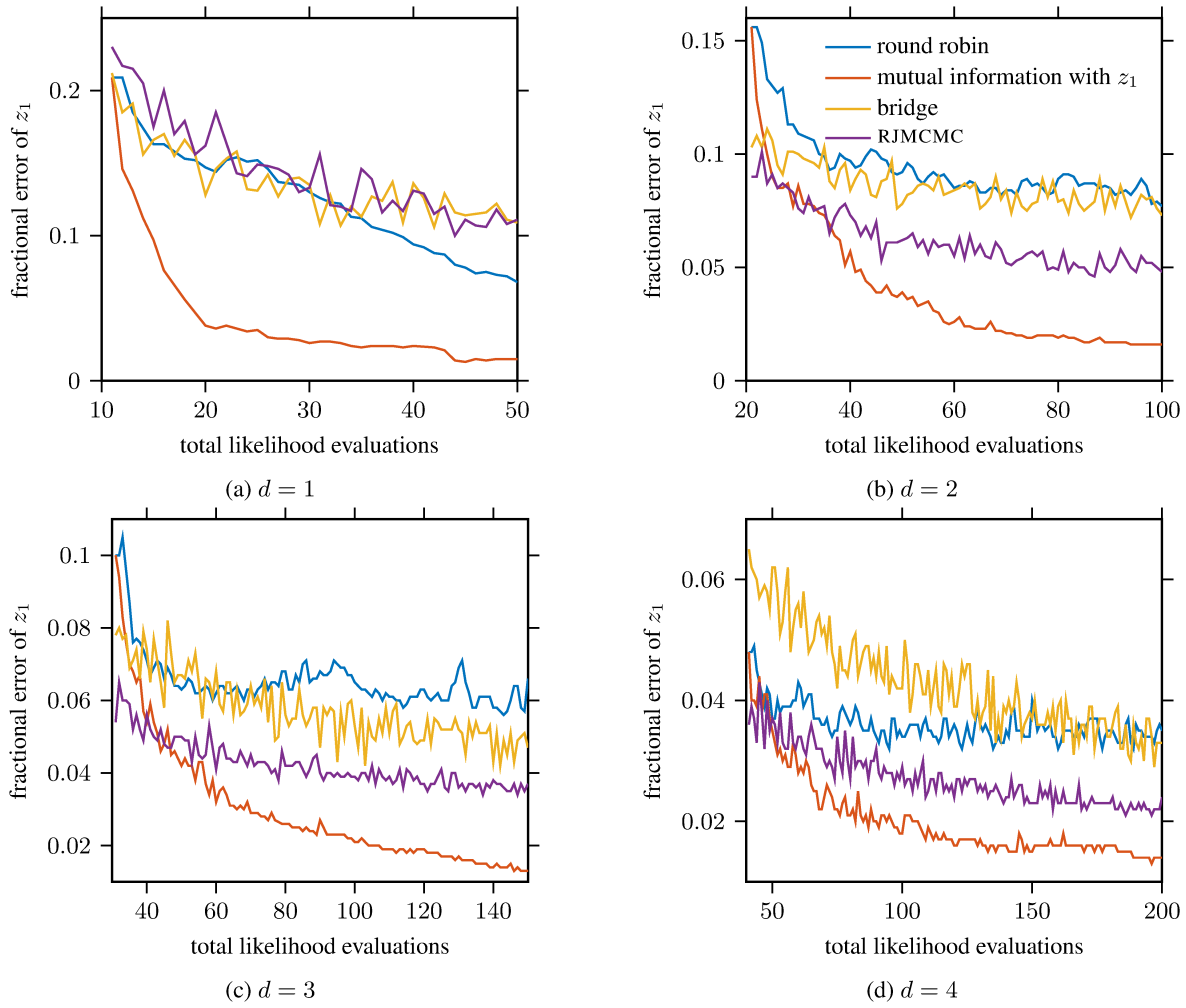



*Figure 2.* Fractional errors of $z_1$ for all tested methods as a function of the total number of model likelihood observations

For both of the experiments below, the diffeomorphism associated with our implementation of reversible jump MCMC is again the identity function and the corresponding Jacobian factor is 1. The intractable integrals associated with our proposed method are estimated using quasi-Monte Carlo (Caflisch, 1998). We evaluated all methods on the absolute error of their estimates of the log Bayes factor: $\log B_{ij} = \log z_i - \log z_j$ (Jeffreys, 1961; Kass & Raftery, 1995), another potential quantity of interest in model selection tasks. We consider the absolute error instead of the fractional error as the target quantity is a log value. We make use of Bartolucci et al. (2006)'s work to translate the output Markov chain into a log odds estimate.

#### 6.2.1. Two Models

In our first experiment on this dataset, we consider two candidate models for each quasar emission spectrum: the first corresponding to a single DLA and the second corresponding to two DLAs. In this experimental setting, we use the data-driven prior of Garnett et al. (2017) as the proposal distribution for the two additional parameters when transitioning from the single DLA model to the two DLA model. We select 20 spectra from phase III of the Sloan Digital Sky Survey (SDSS–III) (Eisenstein et al., 2011) where the model posteriors for the above models are known *a priori* to be between 0.4 and 0.6; this choice makes the model selection task difficult in some sense. We allot a budget of 150 total likelihood evaluations and initialize each BQ based method with 25 randomly sampled likelihood observations from both model parameter spaces. We repeat the experiment 5 times for each spectra, using a different initialization for each trial.

As Figure 3 shows, our proposed method outperforms all benchmarks compared against. The difference in performance between our proposed method and both Monte Carlo based benchmarks is significant at the 1% significance level according to a one-sided paired $t$-test.

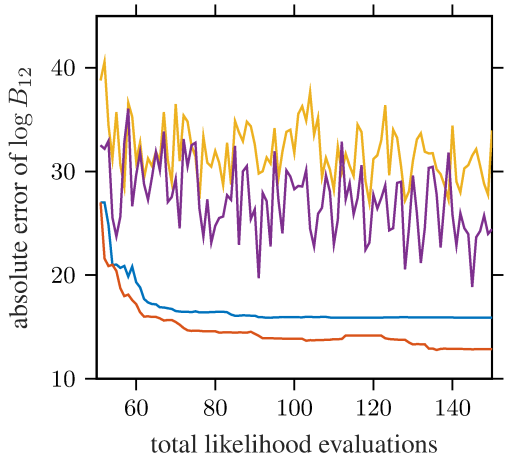



*Figure 3.* Absolute errors of log odds for all tested methods as a function of the total number of model likelihood observations. This figure shares a legend with Figure 2

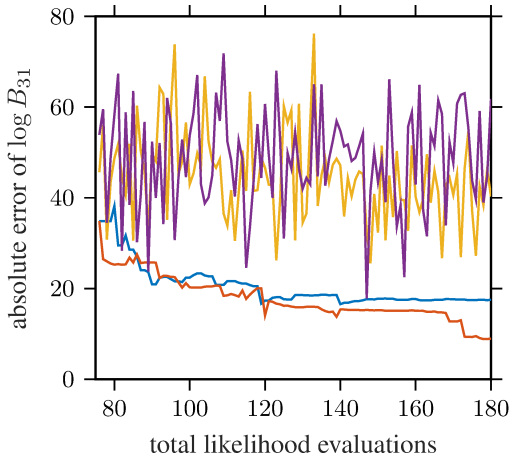
#### 6.2.2. Three Models

To demonstrate the ability of our method to generalize to greater than two models, we ran an additional experiment on this dataset where we consider three candidate models for a quasar emission spectrum known to contain three DLAs: in addition to the two models from the previous section, we consider a third model that corresponds to the presence of three DLAs. We again use the data-driven prior of Garnett et al. (2017) as the proposal distribution for the two additional parameters when transitioning from the two DLA model to the three DLA model. We allot a budget of 180 total likelihood evaluations and initialize each BQ based method with 25 randomly sampled likelihood observations from each model parameter spaces. We repeat the experiment 10 times, using a different initialization for each trial.

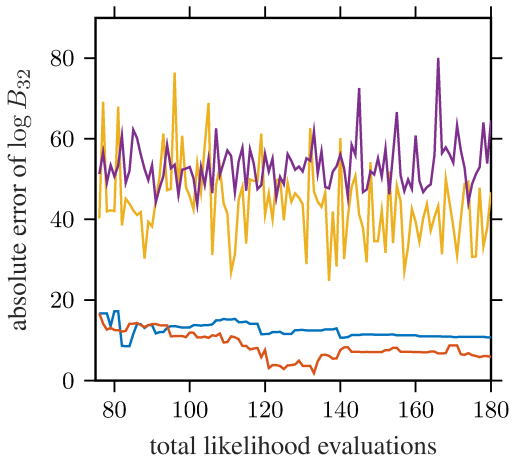
As Figure 4 shows, our proposed method once again outperforms all benchmarks compared against. The difference in performance between our proposed method and both Monte Carlo based benchmarks is significant at the 5% significance level according to a one-sided paired $t$-test.

### 6.3. Model Choice Experiments

In situations where one is performing model choice as opposed to model averaging, the quantity of interest is not the model posterior probabilities but rather the model with the highest posterior probability, a related but different object. We considered an alternative acquisition function that targets this quantity, which we briefly present here. Given two candidate models $\mathcal{M}_1$ and $\mathcal{M}_2$, the goal in model choice is to determine the value of the indicator random variable $[z_1 > z_2]$, where we have adopted the Iverson bracket notation. Therefore, instead of considering the mutual information between $\ell_i(\theta_i)$ and $\mathbf{z}$, one could conceivably consider the mutual information between $\ell_i(\theta_i)$ and $[z_1 > z_2]$ directly. Formally, this quantity can be expressed as




(a) $B_{31}$




(b) $B_{32}$

*Figure 4.* Absolute errors of log odds against $\mathcal{M}_3$ for all tested methods as a function of the total number of model likelihood observations. This figure shares a legend with Figure 2

$$I\big([z_1 > z_2]; \ell_i(\theta_i)\big) = H\big([z_1 > z_2]\big) - H\big([z_1 > z_2] \mid \ell_i(\theta_i)\big)\,. \tag{20}$$

Much like our proposed method, this alternative acquisition function searches over both models' parameter spaces for the next evaluation location:

$$\theta_i^* = \underset{\theta_i \in \Theta_i}{\arg\max}\, I\big([z_1 > z_2]; \ell_i(\theta_i)\big)\,, \tag{21}$$

for $i \in \{1, 2\}$. Because the quantity $H\big([z_1 > z_2]\big)$ does not involve either $\ell_1(\theta_1)$ or $\ell_2(\theta_2)$, it can be safely ignored when searching for this maximum. Unfortunately, the second term

in the mutual information (20) is intractable, much like the integral in (17). However, writing

$$H\big([z_1 > z_2] \mid \ell_i(\theta_i)\big) = \int H\Big(\Pr\big([z_1 > z_2] \mid \ell_i(\theta_i)\big)\Big)\, p\big(\ell_i(\theta_i)\big)\, \mathrm{d}\ell_i(\theta_i)\,, \quad (22)$$

we may recognize the expression as a one-dimensional integral that can also be estimated numerically.

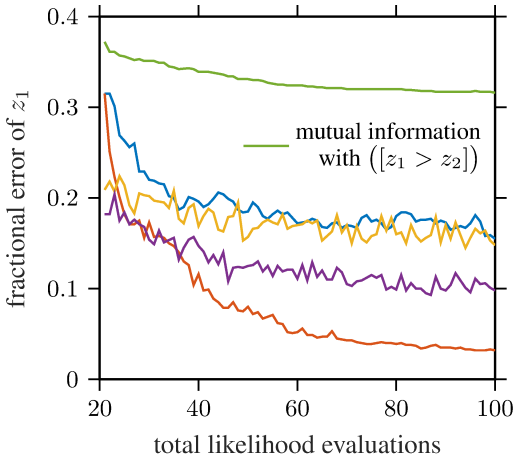



*Figure 5.* Fractional errors of $z_1$ for all tested methods as well as an alternative method considered specifically for the task of model choice. Please refer to Figure 2 for the omitted legend entries.

Despite the arguably more rational choice to target $[z_1 > z_2]$ instead of $\mathbf{z}$, this approach did not perform well in our experiments. Figure 5 shows the performance of this method in the 2-dimensional synthetic experimental setting described above; this method's relative performance continues to drop off as the number of dimensions increases. We also considered a different performance metric: the fraction of trials where the model with the higher ground truth posterior probability is correctly identified. It is reasonable to expect an acquisition function that targets $[z_1 > z_2]$ to outperform our method that targets $\mathbf{z}$ when considering this metric. The results for the 2-dimensional synthetic experimental setting are shown in Figure 6.

We attribute the poor performance of this alternative acquisition function to the fact that the implied alternative objective of this acquisition function, the entropy of $[z_1 > z_2]$, is less reliable than the entropy of $\mathbf{z}$ and thus, this method has a tendency to become overly confident too quickly. If at any point one of the models achieves a much higher posterior model probability, then this method samples the model likelihoods at functionally uninformative locations until the budget of evaluations has been expended. This is because from the perspective of this alternative method, the objective has already been optimized: $\Pr(z_1 > z_2)$ will either be very close to zero or very close to one with very little uncertainty.

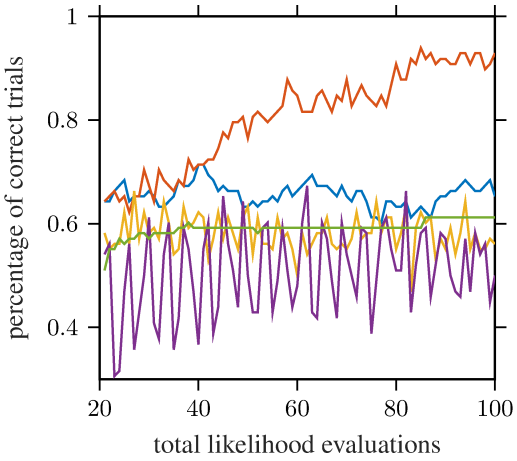



*Figure 6.* Fraction of trials where the "correct" model or model with the higher ground truth posterior probability would have been selected for all tested methods as well as an alternative method considered specifically for the task of model choice. Please refer to Figure 2 for the omitted legend entries.

In future work, we hope to improve the performance of this alternative method, potentially by targeting the random variable $z_1 - z_2$ instead of $[z_1 > z_2]$ as we hypothesize that incorporating the magnitude of the difference will encourage continued exploration of the model parameter spaces.

## 7. Conclusion

In this paper, we have presented a novel, BQ based method for automated model selection. Our proposed method makes use of a novel acquisition function that targets the entropy of the posterior model probabilities, quantities specifically relevant to the task of model selection. This allows our method to actively sample locations across multiple model parameter spaces simultaneously. Our experiments conducted on real-world and synthetic data show that our proposed method can outperform both previously published BQ approaches to model selection as well as Monte Carlo based model selection techniques in terms of achieving accurate posterior model probability estimates.

One exciting line of inquiry that we hope to study in future work concerns the design of trans-dimensional covariance functions. In certain settings, our assumption that the model evidences are independent does not accurately reflect our *a priori* knowledge e.g. in model selection tasks with nested model parameters, we should expect model evidences to be at least slightly correlated. These types of relationships could be captured by a multi-task GP (Cressie, 1993; Yu et al., 2005) where the covariance between model likelihoods is learned alongside individual model likelihoods simultaneously.

## Acknowledgements

HC and RG were supported by the National Science Foundation under award numbers IIA–1355406 and IIS–1845434. JT was supported by EPSRC and MRC CDT in Next Generational Statistical Science under award number EP/L016710/1. We also extend thanks to the BigBayes Group at Oxford University for access to computational resources.

## References

Bach, F. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

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

Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.

Bernardo, J. M. and Smith, A. F. *Bayesian Theory*. John Wiley & Sons, 1994.

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., and Sejdinovic, D. Probabilistic Integration: A Role in Statistical Computation? *arXiv preprint arXiv:1512.00933v6 [stat.ML]*, 2015.

Caflisch, R. E. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, 7:1–49, 1998.

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

Chai, H. and Garnett, R. Improving Quadrature for Constrained Integrands. In *Proceedings of The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Cressie, N. A. *Statistics for Spatial Data*. John Wiley & Sons, 1993.

Diaconis, P. Bayesian Numerical Analysis. *Statistical Decision Theory and Related Topics*, 4(1):163–175, 1988.

Eisenstein, D. J., Weinberg, D. H., Agol, E., Aihara, H., Allende Prieto, C., Anderson, S. F., Arns, J. A., Aubourg, É., Bailey, S., Balbinot, E., and et al. SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems. *The Astronomical Journal*, 142:72, September 2011.

Garnett, R., Ho, S., Bird, S., and Schneider, J. Detecting Damped Lyman-$\alpha$ Absorbers with Gaussian Processes. *Monthly Notices of the Royal Astronomical Society*, 472 (2):1850–1865, 2017.

Godsill, S. J. On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics*, 10(2):230–248, 2001.

Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81:80–97, 2017.

Gunter, T., Osborne, M. A., Garnett, R., Hennig, P., and Roberts, S. J. Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature. *Advances in Neural Information Processing Systems (NIPS) 27*, 2014.

Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57 (1):97–109, 1970.

Jeffreys, H. *Theory of Probability*. Oxford University Press, 3rd edition, 1961.

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. Convergence guarantees for kernel-based quadrature rules in misspecified settings. In *Advances in Neural Information Processing Systems (NIPS) 29*, 2016.

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *arXiv preprint arXiv:1709.00147*, 2017.

Karvonen, T., Oates, C. J., and Särkkä, S. A Bayes–Sard Cubature Method. *arXiv preprint arXiv:1804.03016*, 2018.

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

Larkin, F. M. Gaussian measure in Hilbert space and applications in numerical analysis. *Rocky Mountain Journal of Mathematics*, 2(3):379–422, 1972.

Meng, X. and Wong, W. H. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6(4):831–860, 1996.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Minka, T. P. Deriving quadrature rules from Gaussian processes. Technical report, Statistics Department, Carnegie Mellon University, 2000.

Neal, R. M. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

O'Hagan, A. Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29:245–260, 1991.

Osborne, M. A., Garnett, R., Ghahramani, Z., Duvenaud, D., Roberts, S. J., and Rasmussen, C. E. Active learning of model evidence using Bayesian quadrature. *Advances in Neural Information Processing Systems (NIPS) 25*, 2012.

Overstall, A. M. and Forster, J. J. Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics and Data Analysis*, 54(12): 3269–3288, 2010.

Rasmussen, C. E. and Ghahramani, Z. Bayesian Monte Carlo. *Advances in Neural Information Processing Systems (NIPS) 16*, 2003.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Skilling, J. Nested sampling. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 735: 395–405, 2004.

Snelson, E., Ghahramani, Z., and Rasmussen, C. E. Warped Gaussian processes. *Advances in Neural Information Processing Systems (NIPS) 17*, 2004.

Vehtari, A. *Bayesian model assessment and selection using expected utilities*. PhD thesis, Helsinki University of Technology; Teknillinen korkeakoulu, 2001.

Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

Yu, K., Tresp, V., and Schwaighofer, A. Learning Gaussian processes from multiple tasks. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 2005.