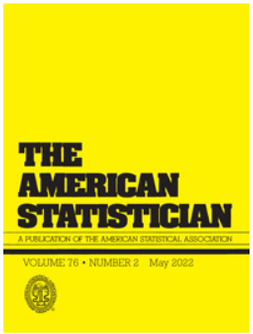

The American Statistician

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Black Box Variational Bayesian Model Averaging

Vojtech Kejzlar, Shrijita Bhattacharya, Mookyong Son & Tapabrata Maiti

To cite this article: Vojtech Kejzlar, Shrijita Bhattacharya, Mookyong Son & Tapabrata Maiti (2022): Black Box Variational Bayesian Model Averaging, The American Statistician, DOI: 10.1080/00031305.2022.2058611

To link to this article: https://doi.org/10.1080/00031305.2022.2058611

THE AMERICAN STATISTICIAN
2022, VOL. 00, NO. 0, 1–12: Statistical Computing and Graphics
https://doi.org/10.1080/00031305.2022.2058611

Black Box Variational Bayesian Model Averaging

Vojtech Kejzlar[a], Shrijita Bhattacharya[b], Mookyong Son[b], and Tapabrata Maiti[b]

[a]Department of Mathematics and Statistics, Skidmore College, Saratoga Springs, NY; [b]Department of Statistics and Probability, Michigan State University, East Lansing, MI

ABSTRACT

For many decades now, Bayesian Model Averaging (BMA) has been a popular framework to systematically account for model uncertainty that arises in situations when multiple competing models are available to describe the same or similar physical process. The implementation of this framework, however, comes with a multitude of practical challenges including posterior approximation via Markov chain Monte Carlo and numerical integration. We present a Variational Bayesian Inference approach to BMA as a viable alternative to the standard solutions which avoids many of the aforementioned pitfalls. The proposed method is "black box" in the sense that it can be readily applied to many models with little to no model-specific derivation. We illustrate the utility of our variational approach on a suite of examples and discuss all the necessary implementation details. Fully documented `Python` code with all the examples is provided as well.

ARTICLE HISTORY
Received June 2021
Accepted March 2022

KEYWORDS
Bayesian inference; Markov chain Monte Carlo; Model evidence; Model selection; Model uncertainty; Variational Bayes

1. Introduction

The existence of several competing models to solve the same or similar problem is a common scenario across scientific applications. One typically encounters a slew of candidate models during standard regression analysis with multiple predictors. Another widely familiar example is a numerical weather prediction with multitudes of forecasting models available. The routine practice in this situation is to select a single model and then make inference based on this model which ignores a major component of uncertainty—model uncertainty (Leamer 1978). Bayesian model averaging (BMA) is the natural Bayesian framework to systematically account for uncertainty due to several competing models.

For any quantity of interest $\boldsymbol{\Delta}$, such as a future observation or an effect size, the BMA posterior density $p(\boldsymbol{\Delta}|\boldsymbol{d})$ corresponds to the mixture of posterior densities of the individual models $p(\boldsymbol{\Delta}|\boldsymbol{d}, M)$ weighted by their posterior model probabilities $p(M|\boldsymbol{d})$:

$$p(\boldsymbol{\Delta}|\boldsymbol{d}) = \sum_{M\in\mathcal{M}} p(\boldsymbol{\Delta}|\boldsymbol{d}, M)p(M|\boldsymbol{d}), \tag{1}$$

where $\mathcal{M}$ denotes the space of all models, $\boldsymbol{d} = (d_1, \ldots, d_n)$ are given datapoints, and $d_i = (x_i, y_i)$ for $i = 1, \ldots, n$ are input-observation pairs. Note, if the space of models is $\mathcal{M} = \{M_1, \ldots, M_K\}$, then the formula in (1) can be equivalenty written as $p(\boldsymbol{\Delta}|\boldsymbol{d}) = \sum_{k=1}^{K} p(\boldsymbol{\Delta}|\boldsymbol{d}, M_k)p(M_k|\boldsymbol{d})$ which is a more commonly used notation for BMA. We stick to the notation in (1) to facilitate the developments in Section 2.2.

The posterior probability of a model M is given by a simple application of the Bayes' theorem:

$$p(M|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|M)p(M)}{\sum_{M'\in\mathcal{M}} p(\boldsymbol{d}|M')p(M')}. \tag{2}$$

Due to the mixture form of the density (1), determining these probabilities is the key to successful implementation of the BMA framework. To do so, one first needs to assign a suitable prior probability $p(M)$ that M is the *true model* (assuming there is one such, among the models considered). Hoeting et al. (1999) notes that,

> *When there is little prior information about the relative plausibility of the models considered, the assumption that all models are equally likely a priory is a reasonable "neutral" choice.*

One can, nevertheless, choose informative prior distributions when prior information about the likelihood of each model is available. Eliciting an informative prior is a nontrivial task, but Madigan, Gavrin, and Raftery (1995) provide some guidance in the context of graphical models that can be applied in other settings as well.

The second component of model's posterior probability is the model's marginal likelihood, also known as model evidence,

$$p(\boldsymbol{d}|M) = \int p(\boldsymbol{d}|\boldsymbol{\theta}_M, M)p(\boldsymbol{\theta}_M|M)\mathrm{d}\boldsymbol{\theta}_M, \tag{3}$$

where $\boldsymbol{\theta}_M$ is the set of model-specific parameters ($\boldsymbol{\theta}_M = (\boldsymbol{\beta}, \sigma)$ in regression problems), $p(\boldsymbol{\theta}_M|M)$ is their prior distribution,

CONTACT Vojtech Kejzlar vkejzlar@skidmore.edu Department of Mathematics and Statistics, Skidmore College, Saratoga Springs, NY.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS.

and $p(\boldsymbol{d}|\boldsymbol{\theta}_M, M)$ is the model's data likelihood. The evaluation of model evidence is one of the main reasons why BMA becomes computationally challenging in practice, because a closed form solution is available only in special scenarios for the exponential family of distributions with conjugate priors, and thus the integral (3) requires approximation. Some problem-specific algorithms have been developed for direct sampling from BMA posterior density (1) such as the Markov chain Monte Carlo (MCMC) model composition for linear regression models (MC^3) (Raftery, Madigan, and Hoeting 1997).

A vast body of literature was produced over the past 30 years on the topic of model evidence approximation, with the simplest approach being the Monte Carlo (MC) integration. The advantage of MC integration is in the method's ease of implementation, however, one typically needs to generate a large number of samples from prior distribution to achieve reasonable convergence. A popular improvement to the simple MC integration is the harmonic mean estimator which makes the use of samples from posterior distribution of model parameters and therefore converges more quickly. On the other hand, it can be unstable, and it tends to overestimate the evidence (see Raftery et al. 2007; Lenk 2009). A large class of statistically efficient estimators is based on importance sampling that relies on draws from an importance density which approximates the joint posterior density of model parameters. However, a poor choice of importance density may lead to a huge loss of efficiency. See Neal (2001), Friel and Pettitt (2008), and Pajor (2017) for some examples of estimators with importance sampling. Another classical method is the Laplace approximation. This corresponds to a second order Taylor expansion of the log-likelihood around its maximum, which makes the likelihood normal. Laplace method is efficient for well-behaved likelihoods. We refer the reader to Kass and Raftery (1995), Ardia et al. (2012) and Friel and Wyse (2012) for a complete survey of popular approximation methods for the model evidence. Additionally, the more recently proposed Nested Sampling algorithm by Skilling (2006) and expanded by Feroz, Hobson, and Bridges (2009) provides another alternative to the aforementioned approaches.

Here we want to point out that the definition of BMA relies on the assumption that the true model which represents the physical reality is within the models being considered (i.e., $\mathcal{M}$-closed setting). BMA can lead to misleading results when the true model is not included (i.e., $\mathcal{M}$-open setting). For instance, a scenario with two models—one mediocre and one "perfect" almost everywhere with a large deviation from the truth at a single point of the input space—will typically result in the selection of the mediocre model. Similarly, BMA can also lead to a suboptimal performance under model misspecification (Clarke 2003; Masegosa 2020). Using BMA in the $\mathcal{M}$-open setting additionally creates a logical tension between interpreting $p(M|\boldsymbol{d})$ and $p(M)$ as probabilities of M being the true model and knowing that the true model is not in $\mathcal{M}$. One can perhaps reconcile this tension by considering $p(M|\boldsymbol{d})$ and $p(M)$ as the probabilities of M being a useful description of physical reality. In what follows, we will assume that the reader is comfortable with assigning a prior over $\mathcal{M}$, even in the $\mathcal{M}$-open setting. We refer to Bernardo and Smith (1994) for a detailed discussion about the conceptual differences between the $\mathcal{M}$-closed and $\mathcal{M}$-open settings and to Fragoso, Bertoli, and Louzada (2018) for a recent survey of BMA methodology. A decision-theoretic approach to account for model uncertainty in $\mathcal{M}$-open setting is presented in Clyde and Iversen (2013). Recently, Phillips et al. (2021) proposed a model-mixing approach for the case when the list of models considered does not contain the true model. Both of these methods address the inadequacy of BMA in $\mathcal{M}$-open setting by not considering models as an extension of the parameter space.

Despite its conceptual and computational challenges, BMA has a long history of use in both natural sciences and humanities because of a superior predictive performance that is theoretically guaranteed (Bernardo and Smith 1994). Geweke (1999) introduced BMA in economics and later in other fields such as political and social sciences. See the recent review on the use of BMA in Economics by Steel (2020). BMA has also been applied to the medical sciences (Balasubramanian et al. 2014; Schorning et al. 2016), ecology and evolution (Silvestro et al. 2014; Hooten and Hobbs 2015), genetics (Wei, Visweswaran, and Cooper 2011; Wen 2015), machine learning (Clyde, Ghosh, and Littman 2011; Hernández et al. 2018; Mukhopadhyay and Dunson 2020), and lately in nuclear physics (Neufcourt et al. 2019; Neufcourt et al. 2020a; Kejzlar et al. 2020).

In this article, we present a Variational Bayesian Inference (VBI) approach to BMA. VBI is a useful alternative to the sampling-based approximation via MCMC that approximates a target density through optimization. Statisticians and computer scientists (starting with Peterson and Anderson 1987; Jordan et al. 1999) have been widely using variational techniques because they tend to be faster and easier to scale to massive datasets. Our method is based on the variational inference algorithm with reparameterization gradients developed by Titsias and Lázaro-Gredilla (2014) and Kucukelbir et al. (2017) which can be applied to many models with minimum additional derivations. The proposed approach, which we shall call the black box variational BMA (VBMA), is a one step procedure that simultaneously approximates model evidences and posterior distributions of individual models while enjoying all the advantages (and disadvantages) of VBI. Here we note that this is not the first time a VBI is used in the context of BMA. For instance, Latouche and Robin (2016) developed a variational Bayesian approach specifically for averaging of graphon functions, and Jaureguiberry, Vincent, and Richard (2014) use VBI and BMA for audio source separation. However, the VBMA is a general algorithm that can be applied directly to a wide class of models including Bayesian neural networks, generalized linear models, and Gaussian process models.

1.1. Outline of this Article

In Section 2, we provide a brief overview of VBI and derive our proposed VBMA algorithm. Then, in Section 3, we present a collection of examples that include standard linear regression, logistic regression, and Bayesian model selection. To fully showcase the computational benefits of VBMA, we consider Gaussian process models for residuals of separation energies of atomic nuclei. We compare VBMA with direct sampling BMA via MC^3 and with MCMC posterior approximation and evidence computed using MC integration. A fully documented `Python` code with our algorithm and examples is available at

https://github.com/kejzlarv/BBVBMA. Finally, in Section 4, we discuss the pros and cons of VBMA and provide a list of sensible machine learning applications for the proposed methodology.

2. BMA via Variational Bayesian Inference

2.1. Variational Bayesian Inference

VBI strives to approximate a target posterior distribution through optimization. One first considers a family of distributions $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$, indexed by a variational parameter $\boldsymbol{\lambda}$, over the space of model parameters and subsequently finds a member of this family q^* closest to the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{d})$. The simplest variational family is the mean-field family which assumes independence of all the components in $\boldsymbol{\theta}$ but many other families of variational distributions exist; see Wainwright and Jordan (2008), Hoffman and Blei (2015), Ranganath, Tran, and Blei (2016), Tran, Blei, and Airoldi (2015), Tran, Ranganath, and Blei (2017), Rezende and Mohamed (2015), Kingma et al. (2016), Kucukelbir et al. (2017), Fortunato, Blundell, and Vinyals (2017), Papamakarios, Pavlakou, and Murray (2017), Papamakarios et al. (2021), Kobyzev, Prince, and Brubaker (2021), and Weilbach et al. (2020). The recent work of Ambrogioni et al. (2021) and the references therein provide a detailed discussion of these classes of variational families and their associated implementation challenges. The approximate distribution q^* is chosen to minimize the Kullback–Leibler (KL) divergence of $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ from $p(\boldsymbol{\theta}|\boldsymbol{d})$:

$$q^* = \underset{q(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\operatorname{argmin}} KL(q(\boldsymbol{\theta}|\boldsymbol{\lambda})||p(\boldsymbol{\theta}|\boldsymbol{d})). \tag{4}$$

Finding q^* is done in practice by maximizing an equivalent objective function (Jordan et al. 1999), the *evidence lower bound* (ELBO):

$$\mathcal{L}(q) = \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\lambda})}\Big[\log p(\boldsymbol{d},\boldsymbol{\theta}) - \log q(\boldsymbol{\theta}|\boldsymbol{\lambda})\Big]. \tag{5}$$

The ELBO is the sum between the negative KL divergence of the variational distribution from the true posterior distribution and the log of the marginal data distribution $p(\boldsymbol{d})$. The term $\log p(\boldsymbol{d})$ is constant with respect to $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$. It is also a lower bound on $\log p(\boldsymbol{d})$ for any choice of $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$. ELBO can be optimized via standard coordinate- or gradient-ascent methods. However, these techniques are inefficient for large datasets, and so it has become common practice to use the stochastic gradient ascent (SGA) algorithm. SGA updates $\boldsymbol{\lambda}$ at the tth iteration according to

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t + \rho_t \tilde{l}(\boldsymbol{\lambda}_t), \tag{6}$$

where $\tilde{l}(\boldsymbol{\lambda})$ is a realization of the random variable $\tilde{\mathcal{L}}(\boldsymbol{\lambda})$ which is an unbiased estimate of the gradient $\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$.

Let us now assume that $\log p(\boldsymbol{d},\boldsymbol{\theta})$ and $\log q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ are differentiable functions with respect to $\boldsymbol{\theta}$, and that the random variable $\boldsymbol{\theta}$ can be reparameterized using a differentiable transformation $t(\boldsymbol{z},\boldsymbol{\lambda})$ of an auxiliary variable $\boldsymbol{z}$ so that $\boldsymbol{z} \sim \psi(\boldsymbol{z})$ and $\boldsymbol{\theta} = t(\boldsymbol{z},\boldsymbol{\lambda})$ imply $\boldsymbol{\theta} \sim q(\boldsymbol{\theta}|\boldsymbol{\lambda})$. It is assumed that $\psi(\boldsymbol{z})$ exists in a standard form so that any parameter mean vector is set to zero and scale parameters are set to one. For example, if we consider a real valued θ with normal variation family $q(\theta|\mu,\sigma^2)$, then $t(z,(\mu,\sigma)) = z\sigma + \mu$ and $z \sim \text{Normal}(0,1)$. Note that the variational parameters are part of the transformation and not the auxiliary distribution. The gradient of the ELBO can be then expressed as the following expectation with respect to the auxiliary distribution $\psi(\boldsymbol{z})$ (Titsias and Lázaro-Gredilla 2014; Kucukelbir et al. 2017):

$$\nabla_{\boldsymbol{\lambda}}\mathcal{L}(q) = \mathbb{E}_{\psi(z)}\Big[\nabla_{\boldsymbol{\theta}}(\log p(\boldsymbol{d},\boldsymbol{\theta}) - \log q(\boldsymbol{\theta}|\boldsymbol{\lambda})) \times \nabla_{\boldsymbol{\lambda}} t(\boldsymbol{z},\boldsymbol{\lambda})\Big]. \tag{7}$$

The expectation (7) does not have a closed form in general, nevertheless, one can use S samples from $\psi(\boldsymbol{z})$ to construct its unbiased MC estimate for the SGA (6)

$$\tilde{l}(\boldsymbol{\lambda}) = \frac{1}{S}\sum_{s=1}^{S}\Big[\nabla_{\boldsymbol{\theta}}(\log p(\boldsymbol{d}, t(\boldsymbol{z}[s],\boldsymbol{\lambda})) - \log q(t(\boldsymbol{z}[s],\boldsymbol{\lambda})|\boldsymbol{\lambda})) \times \nabla_{\boldsymbol{\lambda}} t(\boldsymbol{z}[s],\boldsymbol{\lambda})\Big], \tag{8}$$

where $\boldsymbol{z}[s] \sim \psi(\boldsymbol{z})$. Since the differentiability assumptions and the reparameterization trick allows the use of autodifferentiation to take gradients, the method is *black box* in nature (Kucukelbir et al. 2017). The disadvantage of reparameterization gradient is that it requires differentiable models, that is, models with no discrete variables. One can use the so called score gradient (Ranganath, Gerrish, and Blei 2014) for models with discrete variables which is also black box in nature, however, the variance of the gradient estimates can be large and lead to unreliable results (Ruiz, Titsias, and Blei 2016). The estimate (8) can be conveniently used in the SGA algorithm which converges to a local maximum of $\mathcal{L}(\boldsymbol{\lambda})$ (global for $\mathcal{L}(\boldsymbol{\lambda})$ concave (Bottou, Le Cun, and Bengio 1997)) when the learning rate ρ_t follows the Robbins–Monro conditions (Robbins and Monro 1951)

$$\sum_{t=1}^{\infty}\rho_t = \infty, \qquad \sum_{t=1}^{\infty}\rho_t^2 < \infty. \tag{9}$$

Choosing an optimal learning rate ρ_t can be challenging in practice. Ideally, one would want the rate to be small in situations where MC estimates of the ELBO gradient are erratic (large variance) and large when the MC estimates are relatively stable (small variance). The elements of variational parameter λ can also differ in scale, and the selected learning rate should accommodate these varying, potentially small, scales. The ever increasing abundance of stochastic optimization in machine learning applications spawned development of numerous algorithms for element-wise adaptive scale learning rates. We use the Adam algorithm (Kingma and Ba 2014) which is a popular and easy-to-implement adaptive rate algorithm. However, there are many other frequently used algorithms such as the AdaGrad (Duchi, Hazan, and Singer 2011), the ADADELTA (Zeiler 2012), or the RMSprop (Tieleman and Hinton 2012). The step size associated with Adam is kept constant throughout the article. However, as pointed out in Shazeer and Stern (2018), one may achieve a better performance by making use of linear ramp-up followed by some form of decay (Vaswani et al. 2017). In the supplementary material, we provide the results based on RMSprop for comparison. We did not observe significant differences between RMSprop and Adam for the examples in Section 3.

Below, we extend the standard VBI that approximates a distribution of model parameters to a scenario where a distribution over the model space needs to be also approximated.

2.2. Black Box Variational BMA

For any quantity of interest $\boldsymbol{\Delta}$, such as a future observation or an effect size, the BMA posterior density $p(\boldsymbol{\Delta}|\boldsymbol{d})$ corresponds to the mixture of posterior densities of the individual models $p(\boldsymbol{\Delta}|\boldsymbol{d}, M)$ weighted by their posterior model probabilities $p(M|\boldsymbol{d})$ as in Equations (1)–(3).

In order to facilitate variational inference, we reformulate the problem of BMA as follows

$$p(\boldsymbol{\Delta}|\boldsymbol{d}) = \int p(\boldsymbol{\Delta}|\boldsymbol{d}, M, \boldsymbol{\theta}_M) p(M, \boldsymbol{\theta}_M|\boldsymbol{d}) \mathrm{d}\mu(M, \boldsymbol{\theta}_M) \tag{10}$$

where μ is the product measure of counting and Lebesgue. Note, the expression (10) indeed summarizes the Equations (1)–(3) in one step. In practice, the most difficult quantity to compute is $p(M, \boldsymbol{\theta}_M|\boldsymbol{d})$. We shall now consider the joint posterior distribution of the model M and its corresponding parameter $\boldsymbol{\theta}_M$ as our parameter of interest. Note, the above notation allows for the dependence of $\boldsymbol{\theta}$ on the model M. This is needed as the indexing parameter of each model could differ in dimension, distribution, etc. As explained in Section 2, there has been a plethora of literature in using variational inference to obtain the posterior distribution $\boldsymbol{\theta}$ for a given model. In this section, we adapt the variational inference to obtain the joint distribution of the model and the parameter together in one stroke.

We next assume a variational approximation to the posterior distribution $p(M, \boldsymbol{\theta}_M|\boldsymbol{d})$ of the form $q(M, \boldsymbol{\theta}_M|\boldsymbol{\lambda}_M) = q(M)q(\boldsymbol{\theta}_M|M, \boldsymbol{\lambda}_M)$, where $q(M)$ is the variational model weight of model M and $q(\boldsymbol{\theta}_M|\boldsymbol{\lambda}_M)$ is the variational distribution of $\boldsymbol{\theta}_M$ under model M and is indexed by its corresponding variational parameter $\boldsymbol{\lambda}_M$. Note that we treat $(M, \boldsymbol{\theta}_M)$ as a random variable which takes on the values $(m, \boldsymbol{\theta}_m)$ for varying values of $m \in \mathcal{M} = \{M_1, \ldots, M_K\}$. The density of this random variable is given by $p(m, \boldsymbol{\theta}_m|\boldsymbol{d})$ under the true posterior, and by $q(m)q(\boldsymbol{\theta}_m|\boldsymbol{\lambda}_m)$ under the variational posterior. Here $q(m)$ for $m \in \mathcal{M}$ can be any categorical distribution satisfying $\sum_{m\in\mathcal{M}} q(m) = 1$. For each $m \in \mathcal{M}$, $q(\boldsymbol{\theta}_m|\boldsymbol{\lambda}_m)$ can be any parameteric distribution indexed by the parameters $\boldsymbol{\lambda}_m$. Possible choices of $q(\boldsymbol{\theta}_m|\boldsymbol{\lambda}_m)$ include but are not restricted to mean field variational family of the form $q(\boldsymbol{\theta}_m|\boldsymbol{\lambda}_m) = \prod_i q(\theta^i_m|\lambda^i_m)$. We additionally assume that $\boldsymbol{\theta}_M$ can be reparameterized using a differentiable transformation $t(\boldsymbol{z}_M, \boldsymbol{\lambda}_M)$ of an auxiliary variable $\boldsymbol{z}_M$ so that $\boldsymbol{z}_M \sim \psi(\boldsymbol{z}_M)$ and $\boldsymbol{\theta}_M = t(\boldsymbol{z}_M, \boldsymbol{\lambda}_M)$. We avoid the inherent dependence of $t(\cdot)$ on M to simplify the notation.

Thus, the optimal variational distribution q^* is given by

$$q^* = \underset{q}{\operatorname{argmin}}\, KL(q(M, \boldsymbol{\theta}_M|\boldsymbol{\lambda}_M)||p(M, \boldsymbol{\theta}_M|\boldsymbol{d})). \tag{11}$$

Again, in the KL expression above, we assume that M is a random variable whose values are the individual models in the model space $\mathcal{M}$. As explained in Section 2, finding q^* is obtained in practice by maximizing an equivalent objective function (Jordan et al. 1999), the ELBO

$$\mathcal{L}(q) = \mathbb{E}_{q(M,\boldsymbol{\theta}_M|\boldsymbol{\lambda}_M)}\Big[\log p(\boldsymbol{d}, M, \boldsymbol{\theta}_M) - \log q(M, \boldsymbol{\theta}_M|\boldsymbol{\lambda}_M)\Big], \tag{12}$$

this time, subject to constraint $\sum_{M\in\mathcal{M}} q(M) = 1$. Since $q(\boldsymbol{\theta}_M|\boldsymbol{\lambda}_M)$ is a parameteric family indexed by the parameters $\boldsymbol{\lambda}_M$, it is indeed a valid density function. However, since the categorical distribution $q(M)$ has freely varying parameters, the constraint $\sum_{M\in\mathcal{M}} q(M) = 1$ is imposed. To accommodate the constraint, using Lagrange multipliers, we optimize

$$\begin{aligned}\mathcal{L}(q) = \mathbb{E}_{q(M,\boldsymbol{\theta}_M|\boldsymbol{\lambda}_M)}[&\log p(\boldsymbol{d}, M, \boldsymbol{\theta}_M) \\ &- \log q(M, \boldsymbol{\theta}_M|\boldsymbol{\lambda}_M)] - \varrho\left(\sum_{M\in\mathcal{M}} q(M) - 1\right).\end{aligned}$$

The ELBO can be simplified further

$$\begin{aligned}&\mathbb{E}_{q(M,\boldsymbol{\theta}_M|\boldsymbol{\lambda}_M)}[\log p(\boldsymbol{d}, M, \boldsymbol{\theta}_M) - \log q(M, \boldsymbol{\theta}_M|\boldsymbol{\lambda}_M)] \\ &\quad - \varrho\left(\sum_{M\in\mathcal{M}} q(M) - 1\right) \\ &= \mathbb{E}_{q(M,\boldsymbol{\theta}_M|\boldsymbol{\lambda}_M)}[\log p(\boldsymbol{d}|M, \boldsymbol{\theta}_M) + \log p(\boldsymbol{\theta}_M|M) + \log p(M) \\ &\quad - \log q(\boldsymbol{\theta}_M|M, \boldsymbol{\lambda}_M) - \log q(M)] - \varrho\left(\sum_{M\in\mathcal{M}} q(M) - 1\right) \\ &= \sum_{M\in\mathcal{M}} q(M)\mathbb{E}_{q(\boldsymbol{\theta}_M|M,\boldsymbol{\lambda}_M)}[\log p(\boldsymbol{d}|M, \boldsymbol{\theta}_M) + \log p(\boldsymbol{\theta}_M|M) \\ &\quad + \log p(M) - \log q(\boldsymbol{\theta}_M|M, \boldsymbol{\lambda}_M) - \log q(M)] \\ &\quad - \varrho\left(\sum_{M\in\mathcal{M}} q(M) - 1\right).\end{aligned}$$

Since the parameters $q(M)$ for $M \in \mathcal{M}$ do not depend on the variational parameters $\boldsymbol{\lambda}_M$, therefore, $\nabla_{\lambda_M}\varrho(\sum_{M\in\mathcal{M}} q(M) - 1) = 0$. Thus, one obtains $\nabla_{\lambda_M}\mathcal{L}(q) = q(M)\mathcal{G}_M$ where

$$\begin{aligned}\mathcal{G}_M = \mathbb{E}_{\psi(\boldsymbol{z}_M)}[&\nabla_{\boldsymbol{\theta}_M}(\log p(\boldsymbol{d}|M, \boldsymbol{\theta}_M) + \log p(\boldsymbol{\theta}_M|M) \\ &- \log q(\boldsymbol{\theta}_M|M, \boldsymbol{\lambda}_M)) \times \nabla_{\boldsymbol{\lambda}_M} t(\boldsymbol{z}_M, \boldsymbol{\lambda}_M)].\end{aligned}$$

To estimate the quantity $\mathcal{G}_M$, we can generate multiple samples from the distribution $\psi(\boldsymbol{z}_M)$ and then use the MC estimate

$$\begin{aligned}\widehat{\mathcal{G}}_M = \frac{1}{S}\sum_{s=1}^{S}[&\nabla_{\boldsymbol{\theta}_M}(\log p(\boldsymbol{d}|M, t(\boldsymbol{z}_M[s], \boldsymbol{\lambda}_M)) \\ &+ \log p(t(\boldsymbol{z}_M[s], \boldsymbol{\lambda}_M)|M) - \log q(t(\boldsymbol{z}_M[s], \boldsymbol{\lambda}_M)|M, \boldsymbol{\lambda}_M)) \\ &\times \nabla_{\boldsymbol{\lambda}_M} t(\boldsymbol{z}_M[s], \boldsymbol{\lambda}_M)].\end{aligned}$$

To derive the update of $q(M)$, note that

$$\begin{aligned}\nabla_{q(M)}\mathcal{L}(q) = &\underbrace{\mathbb{E}_{q(\boldsymbol{\theta}_M|M,\boldsymbol{\lambda}_M)}[(\log p(\boldsymbol{d}|M, \boldsymbol{\theta}_M) + \log p(\boldsymbol{\theta}_M|M) \\ &\quad - \log q(\boldsymbol{\theta}_M|M, \boldsymbol{\lambda}_M))]}_{\mathcal{L}_M} \\ &+ \log p(M) - \log q(M) - 1 - \varrho,\end{aligned}$$

where $\mathcal{L}_M$ is nothing but the ELBO under a fixed model M. Equating the above derivative to 0, we get a closed form expression for $q(M)$ as

$$q(M) = \exp(\mathcal{L}_M + \log p(M) - 1 - \varrho) \propto \exp(\mathcal{L}_M + \log p(M)).$$

It only remains to generate the quantity $\mathcal{L}_M$, which we get again by multiple samples from the distribution $\psi(\boldsymbol{z}_M)$ and then use the MC estimate

$$\begin{aligned}\widehat{\mathcal{L}}_M = \frac{1}{S}\sum_{s=1}^{S}[&\log p(\boldsymbol{d}|M, t(\boldsymbol{z}_M[s], \boldsymbol{\lambda}_M)) + \log p(t(\boldsymbol{z}_M[s], \boldsymbol{\lambda}_M)|M) \\ &- \log q(t(\boldsymbol{z}_M[s], \boldsymbol{\lambda}_M)|M, \boldsymbol{\lambda}_M)].\end{aligned}$$

This allows us to get Algorithm 1 for VBMA.

Algorithm 1 Black Box Variational BMA

Start with an initial choice of $(\boldsymbol{\lambda}_M, q(M))_{M\in\mathcal{M}}$ and a learning rate ρ.

repeat

By generating $\boldsymbol{z}_M[1],\ldots,\boldsymbol{z}_M[S]$ from $\psi(\boldsymbol{z}_M)$, calculate

$$\begin{aligned}\widehat{\mathcal{G}}_M = \frac{1}{S}\sum_{s=1}^{S}[&\nabla_{\boldsymbol{\theta}_M}(\log p(\boldsymbol{d}|M, t(\boldsymbol{z}_M[s],\boldsymbol{\lambda}_M))\\ &+\log p(t(\boldsymbol{z}_M[s],\boldsymbol{\lambda}_M)|M)\\ &-\log q(t(\boldsymbol{z}_M[s],\boldsymbol{\lambda}_M)|M,\boldsymbol{\lambda}_M))\\ &\times\nabla_{\boldsymbol{\lambda}_M} t(\boldsymbol{z}_M[s],\boldsymbol{\lambda}_M)]\end{aligned}$$

Update $\boldsymbol{\lambda}_M$ as

$$\boldsymbol{\lambda}_M = \boldsymbol{\lambda}_M + \rho q(M)\widehat{\mathcal{G}}_M$$

Using the already generated $\boldsymbol{z}_M[1],\ldots,\boldsymbol{z}_M[S]$, calculate

$$\begin{aligned}\widehat{\mathcal{L}}_M = \frac{1}{S}\sum_{s=1}^{S}[&\log p(\boldsymbol{d}|M, t(\boldsymbol{z}_M[s],\boldsymbol{\lambda}_M))\\ &+\log p(t(\boldsymbol{z}_M[s],\boldsymbol{\lambda}_M)|M)\\ &-\log q(t(\boldsymbol{z}_M[s],\boldsymbol{\lambda}_M)|M,\boldsymbol{\lambda}_M)]\end{aligned}$$

Update $q(M)$ as

$$\widetilde{q}(M) = \exp(\widehat{\mathcal{L}}_M + \log p(M))$$

and $q(M) = \widetilde{q}(M)/\sum_{M\in\mathcal{M}}\widetilde{q}(M)$.

until Convergence of $\widehat{\mathcal{L}}(q)$ where

$$\widehat{\mathcal{L}}(q) = \sum_{M\in\mathcal{M}} q(M)\widehat{\mathcal{L}}_M$$

2.2.1. Implementation Details and Variational Families

The general form of VBMA algorithm allows the user to select the variational family that is most appropriate for the problem at hand. As we noted in Section 2, there is a vast pool of candidate families that vary by their expressiveness and ability to capture complex structure of unknown parameters many of which can be used in reparameterization gradients. In the subsequent applications, we shall consider mean-field variational families with normal distributions for real valued variables and log-normal distributions for positive variables. Despite its simplicity, the mean-field variational family can approximate a wide class of posteriors and is good enough to achieve consistency for the variational posterior for a wide class of models (Wang and Blei 2018; Zhang and Gao 2020; Bhattacharya and Maiti 2021). Moreover, all the strictly positive variational parameters λ will be transformed as

$$\tilde{\lambda} = \log(e^{\lambda} - 1) \tag{13}$$

to avoid constrained optimization. See Appendix A for the details on the reparameterization of normal and log-normal mean-field families.

Besides the choice of suitable variational family for VBMA, another practical consideration needs to be made regarding the updates of variational parameter given by $\boldsymbol{\lambda}_M = \boldsymbol{\lambda}_M + \rho q(M)\widehat{\mathcal{G}}_M$. Since each step directly depends on the variational approximation of the posterior model probabilities $q(M)$, the updates can be computationally unstable unless the ELBO of each individual model is close to convergence. We therefore recommend setting $q(M) := 1/K$, where K is the number of models considered, until the variational approximation of the posterior model probabilities stabilizes. Additionally, we recommend to compute the final values of $\widetilde{q}(M)$, and $q(M)$, respectively, as an average of the last several hundred iterations of the Algorithm (1) for a greater reliability of the estimates.

3. Examples

Below, we provide a suite of illustrative real data examples to demonstrate how VBMA serves as a viable alternative to approximate the BMA posterior distribution. First, we analyze the U.S. crime data under the standard linear regression model. Second, we consider a logistic regression model for a heart disease dataset. We also show that VBMA provides a convenient solution to Bayesian model selection with Bayes factors. To fully showcase the computational benefits of VBMA, we study Gaussian process models for the residuals of separation energies of atomic nuclei where the standard MCMC-based implementation is challenging in practice. Each of the examples looks at a situation with several competing models without any prior knowledge of which is better; thus, we set the prior model weights to be uniform over the model space. All the reparameterization gradients in the following applications were obtained using the autodifferentiation engine in `Python` package `PyTorch` (Paszke et al. 2019).

3.1. Bayesian Linear Regression

In this example, we compare VBMA with the MCMC algorithm MC^3 using the aggregated crime data on 47 U.S. states of Vandaele (1978) which has been considered by Raftery, Madigan, and Hoeting (1997) to illustrate the efficiency of BMA in regression scenario with a multitude of candidate models. For simplicity, we concentrate only on a minimal subset of 3 out of 15 predictors of the crime rate and following Raftery, Madigan, and Hoeting (1997), we log transformed all the continuous variables (predictors were also centered).

Given the response variable y, we consider models of the form

$$y = \beta_0 + \sum_{j=1}^{p}\beta_j x_j + \epsilon, \tag{14}$$

where $x_1,\ldots,x_p$ is a subset of a set of candidate predictors $x_1,\ldots,x_k$. In this specific example, we consider three predictors: x_1 corresponding to the percentage of males age 14–24, x_2 corresponds to the probability of imprisonment, and x_3 contains the mean years of schooling in the state. We assign ϵ a normal distribution with mean zero and precision ϕ. The ϵ's are assumed to be independent for distinct cases. For the parameters in each model (14), we use Zellner's g-prior (Zellner 1986; Raftery, Madigan, and Hoeting 1997)

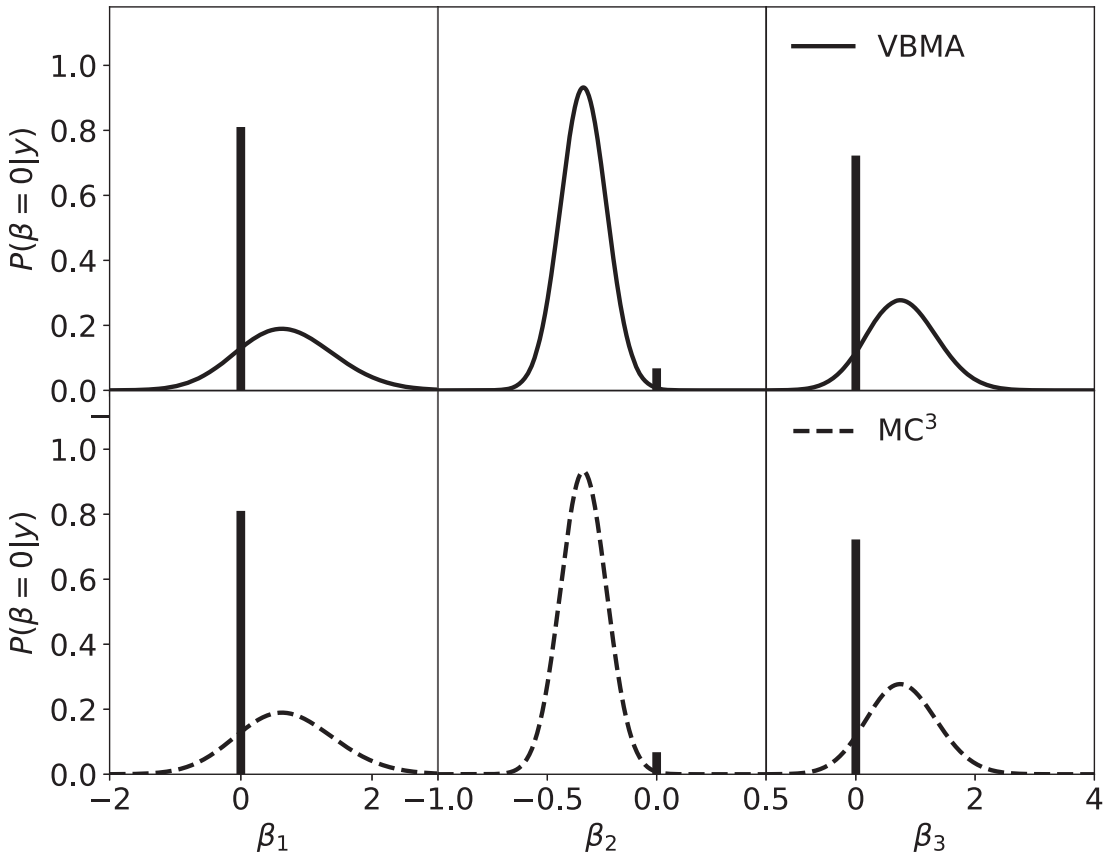

Figure 1. Posterior distributions for predictor slopes based on VBMA (first row) and MC^3 (second row). Namely, β_1 corresponds to the percentage of males 14–24, β_2 to the probability of imprisonment, and β_3 to the mean years of schooling. The density is scaled so that the maximum of the density is equal to $\mathbb{P}(\beta \neq 0|\boldsymbol{d})$. The spike corresponds to $\mathbb{P}(\beta = 0|\boldsymbol{d})$.

$$\begin{aligned} \phi &\propto 1/\phi, \\ \beta_0 &\propto 1, \\ \beta_1, \ldots \beta_p &\propto N(0, g(\boldsymbol{X}'\boldsymbol{X})^{-1}/\phi), \end{aligned}$$

where $g = n$ and $\boldsymbol{X}$ is the design matrix. Zellner's g-prior is one of the most popular conjugate Normal-Gamma prior distributions for linear models that is convenient and provides Bayesian computation with marginal likelihoods that can be evaluated analytically.

Table 1. The top four linear regression models of the crime data according to their posterior model probabilities.

	Inclusion				$p(M\|d)$	
Model	Intercept	x_1	x_2	x_3	MC^3	VBMA
0	*		*		0.58	0.57
1	*		*	*	0.17	0.15
2	*	*	*		0.11	0.11
3	*	*	*	*	0.07	0.05

NOTE: The star indicates the inclusion of predictor in the model and the model ID is provided for easier referencing. Comparison between the VBMA and the MC^3 based averaging is shown.

3.1.1. Results

Table 1 shows the estimates of model posterior probabilities obtained with VBMA and through the MCMC algorithm MC^3 for the top four models. The VBMA results are based on a pretraining sequence of 500 iterations with the model probabilities set to 1/8 and 200 iterations of updating according to Algorithm 1. Ten MC samples from the variational distributions were used to estimate the ELBO gradient. The displayed probabilities were determined as the average over the last 100 iterations of the algorithm to ensure stability of the estimates. The MC^3 results were computed with `R` package `BAS` (Clyde, Ghosh, and Littman 2011). Clearly, the VBMA based values closely match the MC^3 with small deviations for the models with lower posterior probabilities. However, this difference does not dramatically impact the data analysis.

Besides the model posterior probabilities, one can asses the fidelity of VBMA using the posterior distributions of regression coefficients based on the model average. Figure 1 shows the posterior distributions for the coefficients of the percentage of males 14–24, the probability of imprisonment, and the mean years of schooling based on the model averaging results. The figure additionally displays $\mathbb{P}(\beta = 0|\boldsymbol{d})$ obtained by first summing the posterior model probabilities across the models for each predictor and then subtracting the value from one. We can see that the posterior distributions based on VBMA coincide with those obtained with MC^3. On the other hand, and somewhat expectantly, the computational overhead needed to compute the reparameterization gradients is unnecessarily large for the simple case of linear regression. The pretraining

sequence took 4–10 sec per model and the averaging of all eight models took approximately 27 sec on a mid-range laptop. Contrary to that, the MC^3 estimates were instantaneous for all the practical purposes. We shall start seeing the computational efficiency of VBMA in the subsequent applications.

Predictive Performance. Similar to Raftery, Madigan, and Hoeting (1997), we asses the predictive ability of VBMA by randomly splitting the U.S. crime data into a training and a testing dataset. A 50–50 split was chosen here due to a relatively small size of the dataset. We subsequently re-run the VBMA (and MC^3) using the training dataset. The predictive performance was measured through coverage of Bayesian predictive intervals (equal-tails) with the credibility level ranging from 10% to 90% with 10% increments. A $(1-\alpha)\times 100\%$ prediction interval is a posterior credible interval within which a (predicted) observation falls with probability $(1-\alpha)$. An equal-tail interval is chosen so that the posterior probability of being below the interval is as likely as being above it (Gelman et al. 2013). Figure 2 shows the predictive coverage of the two methods plotted against each other with the diagonal dashed line indicating a perfect agreement between the methods. We can see that the coverages for the procedures match in general with small discrepancies at lower quantiles. Additionally, we compare the model averaging predictions with those obtained by the best models according to the adjusted R^2 and Mallows' C_p under both VBMA and MC^3. Adjusted R^2 and Mallows' C_p are commonly used model selection and evaluation criteria in a regression setting. Adjusted R^2 measures quality of the model in terms of total variability explained, whereas Mallows' C_p estimates the size of the bias that is introduced into the predicted responses by having a model that is missing one or more important predictors (James et al. 2013). Both of these model selection strategies lead to M_3 as the best model. However, M_3 generally under-performed the model averaging and underestimated the declared coverage. We can see that by the general shift of the respected curve in comparison to the averaging results.

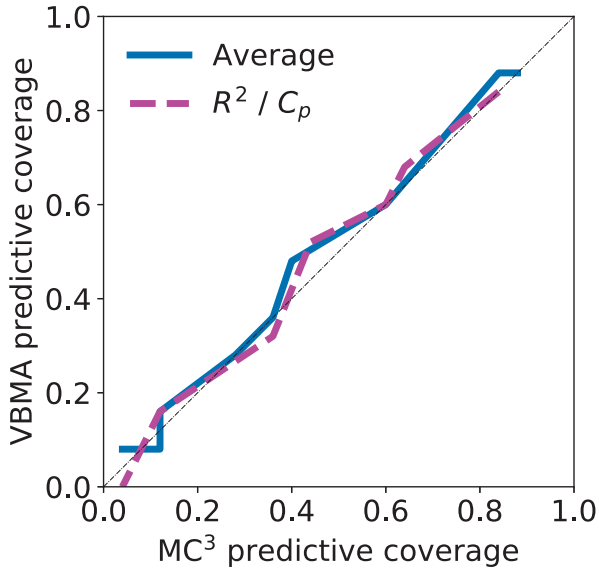

Figure 2. Comparison between coverages of Bayesian predictive interval (equal-tails) on a testing set of 22 observations. The horizontal axis corresponds to the predictive coverage of MC^3 based intervals and the vertical axis to the coverage of VBMA. The diagonal line corresponds to the perfect agreement between the two methods.

3.2. Bayesian Logistic Regression

Unlike the standard linear regression, generalized linear models such as logistic regression exemplify the slew of challenges that one can encounter when implementing BMA. First, the evaluation of the evidence integral does not have an analytic form and the integration can be high-dimensional. Additionally, direct sampling from the BMA posterior through MC^3 algorithm is not available. One therefore needs to approximate the evidence integral and consequently approximate the BMA posterior with MC samples from the mixture of the posteriors of each of the individual models.

Here we illustrate the utility of VBMA on the analysis of heart disease data (Dua and Graff 2017) to asses the factors that contribute to the risk of heart attack. The models used are logistic regression models with logit link function of the form

$$\log\left(\frac{\mathbb{P}(y=1)}{\mathbb{P}(y=0)}\right)=\beta_0+\sum_{j=1}^{p}\beta_j x_j, \tag{15}$$

where $y = 1$ corresponds to subjects with higher chance of heart attack, and $y = 0$ to those with a smaller chance of heart attack. In this example, we shall consider five predictors, namely x_1 is the serum cholestoral in g/dl, x_2 is their resting blood pressure on admission to the hospital, x_3 is the biological sex, x_4 is the age, and x_5 is the maximum hear rate achieved during examination. This gives the total of 32 candidate models. All the continuous variables were again log transformed and centered. For the parameters in each model (15), we use independent normal prior distributions.

3.2.1. Results

Table 2 shows the estimates of model posterior probabilities obtained with VBMA as compared to those computed using MC integration for the top eight models. The VBMA results are based on a pretraining sequence of 500 iterations with the model probabilities set to 1/32 and 100 iterations of updating according to Algorithm 1. Ten MC samples from the variational distributions were used to estimate the ELBO gradient. The MC-based posterior model probabilities are based on 7.5×10^5 samples. This large number of samples was necessary in order to achieve reasonable convergence. We again observe a close match of the VBMA model posteriors with the MC model posteriors.

Figure 3 shows the posterior distribution of regression coefficients based on the model average. The MCMC results were obtained with No-U-Turn sampler (Homan and Gelman 2014) implemented in `Python` package for Bayesian statistical modeling `PyMC3` (Salvatier, Wiecki, and Fonnesbeck 2016). Analogically to the linear regression example, VBMA algorithm with reparameterization gradients captures the posterior distributions well including the parameter uncertainties. When it comes to the computation times, we start seeing the benefits of VBMA for nonconjugate models. The pretraining sequence took 3–4 sec per model and the averaging of all 32 models took approximately 20 sec on a mid-range laptop. On the other hand,

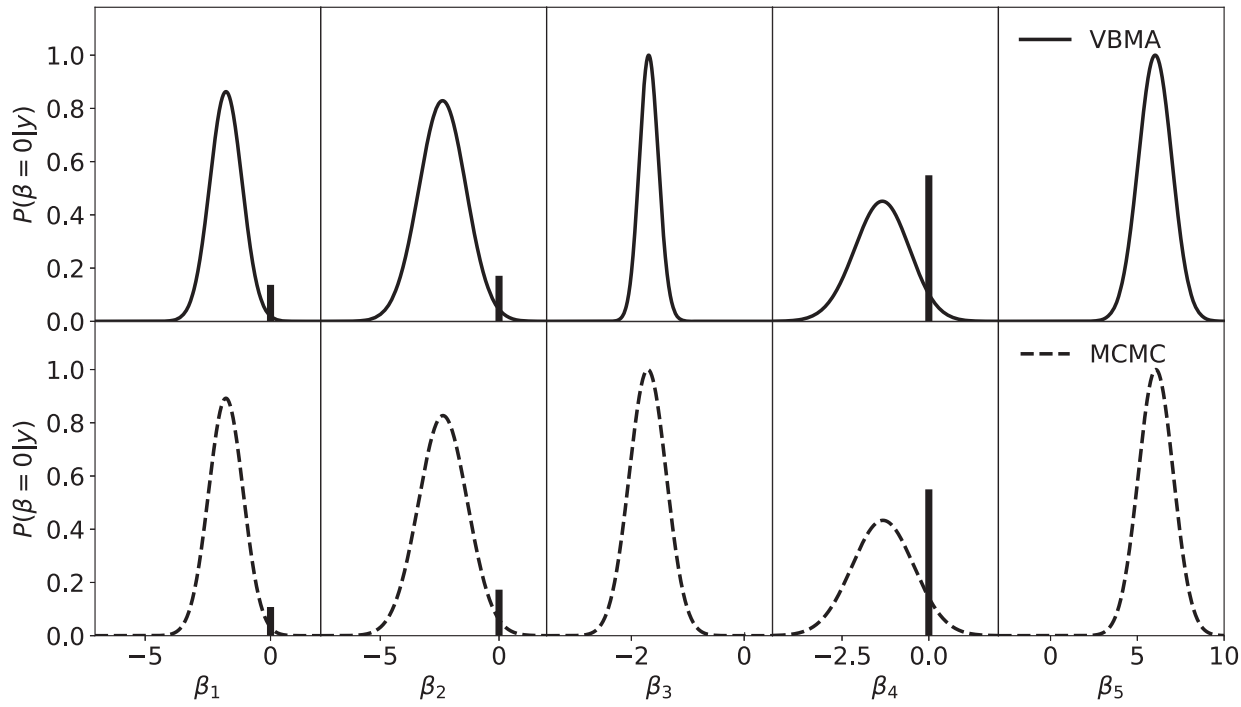

Figure 3. Posterior distributions for predictor slopes based on VBMA (first row) and MCMC (second row). Namely, β_1 corresponds to the serum cholestoral in g/dl, β_2 to the resting blood pressure on admission to the hospital, β_3 to the biological sex, β_4 to the age, and β_5 is the maximum hear rate achieved during examination. The density is scaled so that the maximum of the density is equal to $\mathbb{P}(\beta \neq 0|\boldsymbol{d})$. The spike corresponds to $\mathbb{P}(\beta = 0|\boldsymbol{d})$.

Table 2. The top eight logistic regression models of the heart rate data according to their posterior model probabilities.

	Inclusion						$p(M\|d)$	
Model	Intercept	x_1	x_2	x_3	x_4	x_5	MC	VBMA
1	*	*	*	*		*	0.45	0.43
2	*	*	*	*	*	*	0.28	0.28
3	*	*		*	*	*	0.09	0.09
4	*	*		*		*	0.06	0.06
5	*		*	*	*	*	0.05	0.06
6	*		*	*		*	0.04	0.05
7	*			*	*	*	0.01	0.02
8	*			*		*	< 0.01	< 0.01

NOTE: The star indicates the inclusion of predictor in the model and the model ID is provided for easier referencing. Comparison between the VBMA and the MC based averaging is shown.

the MC estimates of evidence integrals required 20 min per model, and 20–50 sec was needed to obtain 3×10^4 samples via No-U-Turn sampler.

3.3. Bayesian Model Selection

Here, we demonstrate that the VBMA algorithm can be conveniently applied in the generalization of Bayesian hypotheses testing, that is, model selection with Bayes factors. Instead of averaging, suppose that we wish to compare the two Bayesian models,

$$M_0 : \boldsymbol{d} \sim p(\boldsymbol{d}|\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \sim p(\boldsymbol{\theta}_0), \qquad M_1 : \boldsymbol{d} \sim p(\boldsymbol{d}|\boldsymbol{\theta}_1), \boldsymbol{\theta}_1 \sim p(\boldsymbol{\theta}_1),$$

where the definition of the parameter $\boldsymbol{\theta}$ may differ between models. Then, the Bayes factor B_{01} in support of model M_0 is given by

$$B_{01} = \frac{p(\boldsymbol{d}|M_0)}{p(\boldsymbol{d}|M_1)} = \frac{p(M_0|\boldsymbol{d})p(M_1)}{p(M_1|\boldsymbol{d})p(M_0)}, \tag{16}$$

where $p(M_i|\boldsymbol{d})$ for $i \in \{0, 1\}$ is the model's posterior probability defined in Equation (2). The quantity B_{01} is the ratio of the posterior odds of model M_0 to its prior odds and represents the information about the evidence provided by the data in favor of model M_0 as opposed to M_1 (Kass and Raftery 1995). It should be clear from the definition of (16) that the Bayesian model selection suffers from exactly the same computational challenges as BMA. To this extent, VBMA directly approximates posterior probabilities of individual models and Bayes factors can be conveniently computed as a byproduct of the algorithm without the need of approximating the model evidence (3).

3.3.1. Linear and Logistic Regression Examples

To illustrate the Bayesian model selection via VBMA, we consider the following hypotheses for both linear and logistic regression examples above and compare the VBMA based results with their MC counterparts:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0. \tag{17}$$

For the linear regression case, this corresponds to comparing models M_1 and M_3. For the logistic regression example, we need to compare models M_5 and M_2. Table 3 presents the respective Bayes factor approximations. For both linear and logistic regression examples, VBMA approximations qualitatively agree with the MC based approximations. The results show that the U.S crime data favor the linear regression model with $\beta_1 = 0$,

Table 3. Bayes factors obtained via VBMA approximation and MC methods.

Example	Bayes factor	
	MC	VBMA
Linear regression	2.43	3.00
Logistic regression	0.18	0.21

NOTE: Models M_1 and M_3 are considered for the linear regression example and models M_5 and M_2 for the logistic regression example. Bayes factor larger than 1 indicates selection of the model with $\beta_1 = 0$ and vice versa.

and the heart disease data favor logistic regression model with $\beta_1 \neq 0$.

3.4. Nuclear Mass Predictions

As an illustration of VBMA in a scenario where application of standard MCMC-based inference is challenging in practice, we study the separation energies of atomic nuclei which were the subject of various recent machine learning applications (Gaussian process modeling) in the field of nuclear physics (Neufcourt et al. 2018, 2019, 2020b). Namely, our focus is the two-neutron separation energy (S_{2n}) which is a fundamental property of atomic nucleus and is defined as the energy required to remove two neutrons from the nucleus. The S_{2n} values can be obtained through a nuclear mass difference. The knowledge of separation energies determines the limits of nuclear existence and predictions of these quantities can help guide the experimental research at future rare isotope facilities.

In this example, we shall consider 6 state-of-the-art nuclear mass models based on the nuclear density functional theory (DFT) (Nazarewicz 2016): the Skyrme energy density functionals SkM* (Bartel et al. 1982), SkP (Dobaczewski, Flocard, and Treiner 1984), SLy4 (Chabanat et al. 1995), SV-min (Klüpfel et al. 2009), UNEDF0 Kortelainen et al. (2010), and UNEDF1 (Kortelainen et al. 2012). These are global nuclear mass models, because they are capable of reliably describing the whole nuclear chart. Our analysis closely follows that of Neufcourt et al. (2020b), where we consider the statistical model for the differences $y_i = S_{2n}^{\text{exp}}(\boldsymbol{x}_i) - S_{2n}^{\text{th}}(\boldsymbol{x}_i)$ between the observed experimental data and the predictions given by the theoretical models of the form

$$y_i = f(\boldsymbol{x}_i) + \sigma\epsilon_i, \tag{18}$$

where $\boldsymbol{x}_i = (Z_i, N_i)$ corresponds to the proton number Z_i and the neutron number N_i of a nucleus. The function $f(\cdot)$ represents the systematic discrepancy between the underlying physical process and the theoretical mass model. The quantity $\sigma\epsilon_i$ is the scaled experimental error which is assumed to be iid normal with mean zero. For the systematic discrepancy, Neufcourt et al. (2020b) take a Gaussian process (GP) on the two dimensional space $x = (Z, N)$:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(\beta, k(\boldsymbol{x}, \boldsymbol{x}')), \tag{19}$$

where β is the constant mean and k is the squared exponential covariance function characterized by the scale η and characteristic correlation ranges ν_Z and ν_N:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \eta^2 e^{-\frac{(Z-Z')^2}{2\nu_Z^2} - \frac{(N-N')^2}{2\nu_N^2}}. \tag{20}$$

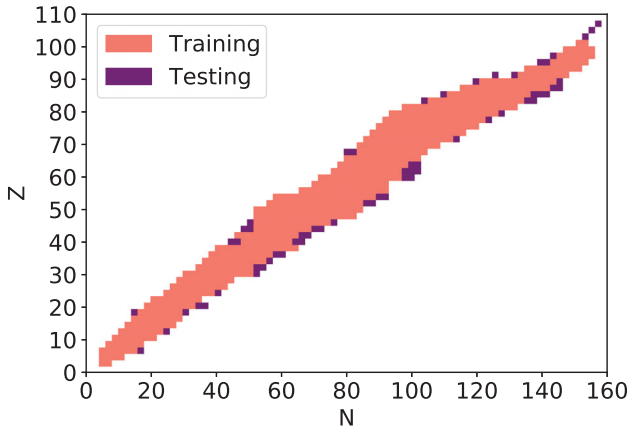

Figure 4. The nuclear chart of even–even and odd–even nuclei divided into the training and testing datasets for the GP modeling of the residuals of two-neutron separation energies S_{2n}. Z corresponds to the proton number and N is the neutron number.

The GP with covariance (20) is a sensible nonparameteric model for the systematic discrepancy as it is expected to be relatively smooth and stationary (Neufcourt et al. 2020a). Since neither of the six Skyrme energy functionals a-priory stands out on the full nuclear domain, using BMA for averaging or model selection is a logical approach here that will allow for predictions with realistically quantified uncertainties.

As the experimental observations, we take the most recent measured values of two-neutron separation energies from the AME2003 dataset (Audi, Wapstra, and Thibault 2003) as training data ($n = 1029$) and keep all additional data tabulated in AME2016 (Wang et al. 2017) for a testing dataset ($n = 120$). The domains of these datasets are depicted in Figure 4. Note that we use both even-even (meaning both Z and N are even) and odd–even nuclei jointly for the training to fully account for the correlations between systematic discrepancies unlike Neufcourt et al. (2020b) who fitted independent GPs on the two domains separately to make the computations manageable. Using the proposed VBMA approach, we are able to do computations in matter of minutes which would take dozens of hours using the standard MCMC approximation.

3.5. Results

The performance of VBMA in averaging of the GP enhanced nuclear mass models was compared with BMA based on the posterior approximation by No-U-Turn sampler and the MC estimates of evidence integrals. Similarly to the previous examples, the VBMA results are based on a pretraining sequence of 300 iterations with the model probabilities set to 1/6. This pretraining sequence lead to the selection of UNEDF1 ($p(M|\boldsymbol{d}) = 1$) with stable ELBOs and so no further training was needed. Ten MC samples from the variational distributions were used to estimate the reparameterization gradient gradient. The resulting *root-mean-square error* (RMSE) on the testing dataset of 120 nuclei was 0.406 MeV as compared to the RMSE of 0.419 MeV given by the MCMC approximation. These RMSE values are consistent with those obtained by Neufcourt et al. (2020b). The MCMC results are based on 2×10^4 posterior samples (10^4

burn-in) and the MC posterior model probabilities are based on 2×10^5 samples which also lead to the selection of UNEDF1.

The MCMC implementation proved to be significantly more time consuming. It required 15–20 hr per model to generate the posterior samples and about 6 hr for the MC integration. On the other hand, the variational approach required between 20 and 25 min per model which clearly demonstrates the utility of VBMA in more complex modeling scenarios. The fidelity of Bayesian predictive intervals was equivalent between the VBMA and MCMC similarly to the linear regression example. We refer the reader to the supplementary materials for additional results containing the study of predictive coverage.

4. Discussion

We presented a VBI approach to BMA that avoids some of the practical challenges that burdens the standard MCMC based approaches to approximate the BMA posterior, especially numerical evaluation of the evidence integral and long sampling times of MCMC sampler. The fidelity of the method was demonstrated on a series of pedagogical examples including the averaging of linear and logistic regression models and Bayesian model selection via Bayes factors. To fully showcase the computational benefits of VBMA, we applied our methodology to nuclear mass models with GP model for systematic discrepancies. The observed speed-up in the case of GP modeling was at least 50-fold compared to the standard MCMC approaches.

The proposed procedure is "black box" in the sense that it can be readily applied to wide range of models with minimal additional derivations needed. For instance, VBMA can be conveniently applied to nonconjugate models including generalized linear models, Bayesian neural networks, and Deep latent Gaussian models (Blei, Kucukelbir, and McAuliffe 2017).

Additionally, VBMA is a general VBI algorithm and the presented implementation with the Adam learning rate and the mean-field variational family is just one of many implementations. One can consider any adaptive learning rate and other variational family available in the literature. VBMA can also be simply modified for greater scalability in the scenarios with complex machine learning models that need to be fitted to massive datasets. First, one can subsample from the data to construct computationally cheap noisy estimates of ELBO gradients. Second, the nature of VBMA allows for immediate parallelization across the models. To achieve a faster convergence of the algorithm, VBMA can be augmented with Rao-Blackwellizaiton (Casella and Robert 1996), control variates (Ross 2006), and importance sampling (Ruiz, Titsias, and Blei 2016) to even further reduce the variance of noisy gradient estimators.

Of course, the use of VBI comes at a cost and one cannot avoid the general pitfalls of variational methods. Using mean-field families can lead to posterior distributions with underestimated uncertainties in cases of highly correlated parameters. One can improve the fidelity of posteriors by using more complex variational family that does not assume independence of unknown parameters (Blei, Kucukelbir, and McAuliffe 2017; Wang and Blei 2018). The choice of adaptive learning rate can be sometimes challenging in practice, and one may observe significant differences among the adaptive learning rates, and a careful sensitivity analysis must be performed. For the models considered in this work, we do not observe significant differences between the RMSprop and the Adam (see the supplementary materials). The reparameterization gradient used in Algorithm 1 works only for differentiable models with no discrete variables. One can use the score gradient (Ranganath, Gerrish, and Blei 2014) for models with discrete variables which is also black box in nature, however, the variance of the gradient estimates is larger than that of reparameterization gradients. Finally, the computational overhead of VBMA for simple models (such as linear regression) can be too high to achieve any meaningful advantage, however, the use of VBMA in complex models can lead to significant computational gains.

Appendix A. Parameterization of Variational Families

A.1. Normal Variational Family

Let us consider a real valued parameter θ with normal variation family $q(\theta|\mu, \sigma^2)$ parameterized by the mean μ and variance σ^2. Under the transformation (13), we get the following expressions for the log likelihood of the variational distribution

$$\log q(\theta|\mu, \lambda_\sigma) = -\frac{1}{2}\log[\log(e^{\lambda_\sigma}+1)] - \frac{1}{2}\log 2\pi - \frac{1}{2}\frac{(\theta-\mu)^2}{\log(e^{\lambda_\sigma}+1)}. \tag{21}$$

The reparameterization gradient is then obtained with $z \sim$ Normal $(0, 1)$ so that

$$\theta = t(z, (\mu, \lambda_\sigma)) = z \times \sqrt{\log(e^{\lambda_\sigma}+1)} + \mu.$$

A.2. Log-Normal Variational Family

For a positive-valued parameters, we shall consider a log-normal variational family $q(\theta|m, \tau^2)$ parameterized by the mean m and variance τ^2:

$$\log q(\theta|m, \lambda_\tau) = -\log\theta - \frac{1}{2}\log[\log(e^{\lambda_\tau}+1)] - \frac{1}{2}\log 2\pi - \frac{1}{2}\frac{(\log\theta-\mu)^2}{\log(e^{\lambda_\tau}+1)}$$

The reparameterization gradient is then obtained with $z \sim$ Normal$(0, 1)$ so that

$$\theta = t(z, (\mu, \lambda_\tau)) = e^{z\times\sqrt{\log(e^{\lambda_\tau}+1)}+\mu}.$$

Supplementary Materials

The supplementary material contains some additional numerical results for the VBMA of linear regression models, logistic regression models, and nuclear mass models. The results were obtained using the RMSprop adaptive learning rate as compared to the Adam learning rate results presented in the main article.

Acknowledgments

The authors thank the reviewers, the Associate Editor, and the Editor for their helpful comments and ideas. This work was supported in part through computational resources and services provided by the Institute for Cyber-Enabled Research at Michigan State University.

Funding

The research is partially supported by the National Science Foundation funding DMS-1952856, DMS-2124605, DMS-1924724, and OAC-2004601.

References

Ambrogioni, L., Lin, K., Fertig, E., Vikram, S., Hinne, M., Moore, D., and van Gerven, M. (2021), "Automatic Structured Variational Inference," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of Proceedings of Machine Learning Research, eds. A. Banerjee, and K. Fukumizu, pp. 676–684. PMLR. [3]

Ardia, D., Baştürk, N., Hoogerheide, L., and van Dijk, H. K. (2012), "A Comparative Study of Monte Carlo Methods for Efficient Evaluation of Marginal Likelihood," *Computational Statistics and Data Analysis*, 56, 3398–3414. 1st issue of the Annals of Computational and Financial Econometrics Sixth Special Issue on Computational Econometrics. [2]

Audi, G., Wapstra, A., and Thibault, C. (2003), "The AME2003 Atomic Mass Evaluation: (ii). Tables, Graphs and References," *Nuclear Physics A*, 729, 337–676. [9]

Balasubramanian, J. B., Visweswaran, S., Cooper, G. F., and Gopalakrishnan, V. (2014), "Selective Model Averaging with Bayesian Rule Learning for Predictive Biomedicine," *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2014, 17–22. [2]

Bartel, J., Quentin, P., Brack, M., Guet, C., and Håkansson, H.-B. (1982), "Towards a Better Parametrisation of Skyrme-like Effective Forces: A Critical Study of the SkM Force," *Nuclear Physics A*, 386, 79–100. [9]

Bernardo, J. M., and Smith, A. F. M. (1994), *Reference Analysis*, Chapter Inference. Wiley. [2]

Bhattacharya, S., and Maiti, T. (2021), "Statistical Foundation of Variational Bayes Neural Networks," *Neural Networks*, 137, 151–173. [5]

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112, 859–877. [10]

Bottou, L., Le Cun, Y., and Bengio, Y. (1997), "Global Training of Document Processing Systems Using Graph Transformer Networks," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pp. 489–493. IEEE. [3]

Casella, G., and Robert, C. P. (1996), "Rao-Blackwellisation of Sampling Schemes," *Biometrika*, 83, 81–94. [10]

Chabanat, E., Bonche, P., Haensel, P., Meyer, J., and Schaeffer, R. (1995), "New Skyrme Effective Forces for Supernovae and Neutron Rich Nuclei," *Physica Scripta*, 1995, 231–233. [9]

Clarke, B. (2003), "Comparing Bayes Model Averaging and Stacking When Model Approximation Error Cannot be Ignored," *The Journal of Machine Learning Research*, 4, 683–712. [2]

Clyde, M., and Iversen, E. (2013), *Bayesian Model Averaging in the M-Open Framework*, pp. 484–498. [2]

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011), "Bayesian Adaptive Sampling for Variable Selection and Model Averaging," *Journal of Computational and Graphical Statistics*, 20, 80–101. [2,6]

Dobaczewski, J., Flocard, H., and Treiner, J. (1984), "Hartree-Fock-Bogolyubov Description of Nuclei Near the Neutron-Drip Line," *Nuclear Physics A*, 422, 103–139. [9]

Dua, D., and Graff, C. (2017), "UCI Machine Learning Repository." [7]

Duchi, J., Hazan, E., and Singer, Y. (2011), "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," *The Journal of Machine Learning Research*, 12, 2121–2159. [3]

Feroz, F., Hobson, M. P., and Bridges, M. (2009), "Multinest: An Efficient and Robust Bayesian Inference Tool for Cosmology and Particle Physics," *Monthly Notices of the Royal Astronomical Society*, 398, 1601–1614. [2]

Fortunato, M., Blundell, C., and Vinyals, O. (2017), "Bayesian Recurrent Neural Networks," arXiv preprint arXiv: 1704.02798. [3]

Fragoso, T. M., Bertoli, W., and Louzada, F. (2018), "Bayesian Model Averaging: A Systematic Review and conceptual Classification," *International Statistical Review*, 86, 1–28. [2]

Friel, N., and Pettitt, A. N. (2008), "Marginal Likelihood Estimation via Power Posteriors," *Journal of the Royal Statistical Society*, Series B, 70, 589–607. [2]

Friel, N., and Wyse, J. (2012), "Estimating the Evidence—A Review," *Statistica Neerlandica*, 66, 288–308. [2]

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013), *Bayesian Data Analysis* (3rd ed.), Boca Raton, FL: CRC Press. [7]

Geweke, J. (1999), "Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication," *Econometric Reviews*, 18, 1–73. [2]

Hernández, B., Raftery, A. E., Pennington, S. R., and Parnell, A. C. (2018), "Bayesian Additive Regression Trees Using Bayesian Model Averaging," *Statistics and Computing*, 28, 869–890. [2]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–401. [1]

Hoffman, M., and Blei, D. (2015), "Stochastic Structured Variational Inference," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* (Vol. 38), San Diego, CA, pp. 361–369. PMLR. [3]

Homan, M. D., and Gelman, A. (2014), "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, 15, 1351–1381. [7]

Hooten, M. B., and Hobbs, N. T. (2015), "A Guide to Bayesian Model Selection for Ecologists," *Ecological Monographs*, 85, 3–28. [2]

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning: with Applications in R*, New York, NY: Springer New York. [7]

Jaureguiberry, X., Vincent, E., and Richard, G. (2014), "Variational Bayesian Model Averaging for Audio Source Separation," in *2014 IEEE Workshop on Statistical Signal Processing (SSP)*, pp. 33–36. [2]

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233. [2,3,4]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [2,8]

Kejzlar, V., Neufcourt, L., Nazarewicz, W., and Reinhard, P.-G. (2020), "Statistical Aspects of Nuclear Mass Models," *Journal of Physics G: Nuclear and Particle Physics*, 47, 094001. [2]

Kingma, D., and Ba, J. (2014), "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations*. [3]

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016), "Improved Variational Inference with Inverse Autoregressive Flow," in *Advances in Neural Information Processing Systems* (Vol. 29), eds. D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc. [3]

Klüpfel, P., Reinhard, P.-G., Bürvenich, T. J., and Maruhn, J. A. (2009), "Variations on a Theme by Skyrme: A Systematic Study of Adjustments of Model Parameters," *Physical Review C*, 79, 034310. [9]

Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2021), "Normalizing Flows: An Introduction and Review of Current Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43, 3964–3979. [3]

Kortelainen, M., Lesinski, T., Moré, J. J., Nazarewicz, W., Sarich, J., Schunck, N., Stoitsov, M. V., and Wild, S. M. (2010), "Nuclear Energy Density Optimization," *Physical Review C*, 82, 024313. [9]

Kortelainen, M., McDonnell, J., Nazarewicz, W., Reinhard, P.-G., Sarich, J., Schunck, N., Stoitsov, M. V., and Wild, S. M. (2012), "Nuclear Energy Density Optimization: Large Deformations," *Physical Review C*, 85, 024304. [9]

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017), "Automatic Differentiation Variational Inference," *Journal of Machine Learning Research*, 18, 1–45. [2,3]

Latouche, P., and Robin, S. S. (2016), "Variational Bayes Model Averaging for Graphon Functions and Motif Frequencies Inference in W-graph Models," *Statistics and Computing*, 26, 1173–1185. [2]

Leamer, E. E. (1978), *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, New York: Wiley. [1]

Lenk, P. (2009), "Simulation Pseudo-bias Correction to the Harmonic Mean Estimator of Integrated Likelihoods," *Journal of Computational and Graphical Statistics*, 18, 941–960. [2]

Madigan, D., Gavrin, J., and Raftery, A. E. (1995), "Eliciting Prior Information to Enhance the Predictive Performance of Bayesian Graphical Models," *Communications in Statistics – Theory and Methods*, 24, 2271–2292. [1]

Masegosa, A. (2020), "Learning Under Model Misspecification: Applications to Variational and Ensemble Methods," in *Advances in Neural Information Processing Systems* (Vol. 33), eds. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, pp. 5479–5491. Curran Associates, Inc. [2]

Mukhopadhyay, M., and Dunson, D. B. (2020), "Targeted Random Projection for Prediction from High-dimensional Features," *Journal of the American Statistical Association*, 115, 1998–2010. [2]

Nazarewicz, W. (2016), "Challenges in Nuclear Structure Theory," *Journal of Physics G: Nuclear and Particle Physics*, 43, 044002. [9]

Neal, R. (2001), "Annealed Importance Sampling," *Statistics and Computing*, 11, 125–139. [2]

Neufcourt, L., Cao, Y., Giuliani, S., Nazarewicz, W., Olsen, E., and Tarasov, O. B. (2020a), "Beyond the Proton Drip Line: Bayesian Analysis of Proton-Emitting Nuclei," *Physical Review C*, 101, 014319. [2,9]

——— (2020b), "Quantified Limits of the Nuclear Landscape," *Physical Review C*, 101, 044307. [9]

Neufcourt, L., Cao, Y., Nazarewicz, W., Olsen, E., and Viens, F. (2019), "Neutron Drip Line in the Ca Region from Bayesian Model Averaging," *Physical Review Letters*, 122, 062502. [2,9]

Neufcourt, L., Cao, Y., Nazarewicz, W., and Viens, F. (2018), "Bayesian Approach to Model-Based Extrapolation of Nuclear Observables," *Physical Review C*, 98, 034318. [9]

Pajor, A. (2017), "Estimating the Marginal Likelihood Using the Arithmetic Mean Identity," *Bayesian Analysis*, 12, 261–287. [2]

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021), "Normalizing Flows for Probabilistic Modeling and Inference," *Journal of Machine Learning Research*, 22, 1–64. [3]

Papamakarios, G., Pavlakou, T., and Murray, I. (2017), "Masked Autoregressive Flow for Density Estimation," in *Advances in Neural Information Processing Systems* (Vol. 30), eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc. [3]

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019), "Pytorch: An Imperative Style, High-performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc. [5]

Peterson, C., and Anderson, J. R. (1987), "A Mean Field Theory Learning Algorithm for Neural Networks," *Complex Systems*, 1, 995–1019. [2]

Phillips, D. R., Furnstahl, R. J., Heinz, U., Maiti, T., Nazarewicz, W., Nunes, F. M., Plumlee, M., Pratola, M. T., Pratt, S., Viens, F. G., and Wild, S. M. (2021), "Get on the Band Wagon: A Bayesian Framework for Quantifying Model Uncertainties in Nuclear Dynamics," *Journal of Physics. G, Nuclear and Particle Physics*, 48, 1–39. [2]

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191. [2,5,7]

Raftery, A. E., Newton, M. A., Satagopa, J. M., and Krivitsk, P. N. (2007), "Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity," *Bayesian Statistics*, 8, 1–45. [2]

Ranganath, R., Gerrish, S., and Blei, D. (2014), "Black Box Variational Inference," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of Proceedings of Machine Learning Research, pp. 814–822. PMLR. [3,10]

Ranganath, R., Tran, D., and Blei, D. M. (2016), "Hierarchical Variational Models," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, Volume 48, ICML'16, pp. 2568–2577. JMLR. [3]

Rezende, D., and Mohamed, S. (2015), "Variational Inference with Normalizing Flows," in *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of Proceedings of Machine Learning Research, Lille, France, eds. F. Bach and D. Blei, pp. 1530–1538. PMLR. [3]

Robbins, H., and Monro, S. (1951), "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, 22, 400–407. [3]

Ross, S. M. (2006), *Simulation* (4th ed.), Orlando, FL: Academic Press, Inc. [10]

Ruiz, F. J. R., Titsias, M. K., and Blei, D. M. (2016), "Overdispersed Black-Box Variational Inference," in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, Arlington, VA, pp. 647–656. AUAI Press. [3,10]

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016), "Probabilistic Programming in Python Using pymc3," *PeerJ Computer Science*, 2, e55. [7]

Schorning, K., Bornkamp, B., Bretz, F., and Dette, H. (2016), "Model Selection Versus Model Averaging in Dose Finding Studies," *Statistics in Medicine*, 35, 4021–4040. [2]

Shazeer, N., and Stern, M. (2018), "Adafactor: Adaptive Learning Rates with Sublinear Memory Cost," in *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of Proceedings of Machine Learning Research, eds. J. Dy and A. Krause, pp. 4596–4604. PMLR. [3]

Silvestro, D., Schnitzler, J., Liow, L. H., Antonelli, A., and Salamin, N. (2014), "Bayesian Estimation of Speciation and Extinction from Incomplete Fossil Occurrence Data," *Systematic Biology*, 63, 349–367. [2]

Skilling, J. (2006), "Nested Sampling for General Bayesian Computation," *Bayesian Analysis*,1, 833–860. [2]

Steel, M. F. J. (2020), "Model Averaging and its use in Economics," *Journal of Economic Literature*, 58, 644–719. [2]

Tieleman, T., and Hinton, G. (2012), "Lecture 6.5—RmsProp: Divide the Gradient by a Running Average of its Recent Magnitude," COURSERA: Neural Networks for Machine Learning. [3]

Titsias, M., and Lázaro-Gredilla, M. (2014), "Doubly Stochastic Variational Bayes for Non-conjugate Inference," in *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, Bejing, China, eds. E. P. Xing and T. Jebara, pp. 1971–1979. PMLR. [2,3]

Tran, D., Blei, D. M., and Airoldi, E. M. (2015). "Copula Variational Inference," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NeurIPS'15, pp. 3564–3572, Cambridge, MA: MIT Press. [3]

Tran, D., Ranganath, R., and Blei, D. M. (2017), "Hierarchical Implicit Models and Likelihood-Free Variational Inference," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NeurIPS'17, pp. 5529–5539. [3]

Vandaele, W. (1978), "Participation in Illegitimate Activities-Ehrlich Revisited (from Deterrence and Incapacitation-Estimating the Effects of Criminal Sanctions on Crime Rates, pp. 270–335 (alfred blumstein et al, ed.-see ncj-44669). [5]

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. U., and Polosukhin, I. (2017), "Attention is All You Need," in *Advances in Neural Information Processing Systems*, volume 30, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc. [3]

Wainwright, M. J., and Jordan, M. I. (2008), "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends® in Machine Learning*, 1, 1–305. [3]

Wang, M., Audi, G., Kondev, F. G., Huang, W. J., Naimi, S., and Xu, X. (2017), "The AME2016 Atomic Mass Evaluation (II). Tables, Graphs and References," *Chinese Physics C*, 41, 030003. [9]

Wang, Y., and Blei, D. M. (2018), "Frequentist Consistency of Variational Bayes," *Journal of the American Statistical Association*, 114, 1147–1161. [5,10]

Wei, W., Visweswaran, S., and Cooper, G. F. (2011), "The Application of Naive Bayes Model Averaging to Predict Alzheimer's Disease from Genome-Wide Data," *Journal of the American Medical Informatics Association*, 18, 370–375. [2]

Weilbach, C., Beronov, B., Wood, F., and Harvey, W. (2020), "Structured Conditional Continuous Normalizing Flows for Efficient Amortized Inference in Graphical Models," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, eds. S. Chiappa and R. Calandra, pp. 4441–4451. PMLR. [3]

Wen, X. (2015), "Bayesian Model Comparison in Genetic Association Analysis: Linear Mixed Modeling and SNP Set Testing," *Biostatistics*, 16, 701–712. [2]

Zeiler, M. D. (2012), "Adadelta: An Adaptive Learning Rate Method," ArXiv 1212.5701. [3]

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Brune de Finetti*, eds. P. Goel and A. Zellner, pp. 233–243, New York: Elsevier. [5]

Zhang, F., and Gao, C. (2020), "Convergence Rates of Variational Posterior Distributions," *The Annals of Statistics*, 48, 2180–2207. [5]