Simulation-based Bayesian Model Averaging via Divergences

David Huk

Abstract

It is common in statistics to be faced with multiple choices of models to explain the data generating process (DGP) behind some observed data. The statistician tasked with choosing a model is faced with two mainstream options. Firstly, the model selection approach, where one can use various criteria in order to summarise the fittingness of a given model, selecting the one with the lowest criterion value. Secondly, the model averaging approach, where each candidate model is appraised based on how well it explains the DGP with the final model being constructed as a weighted average of all candidate models using their ability to explain the data as weights. While each approach has its merits, a critical advantage of the latter is the representation of uncertainty between the different candidate models - a concept absent in model selection.

In this short essay, we will focus on Bayesian Model Averaging (BMA) as an answer to model uncertainty. We will review BMA for variable selection in a regression setting with multiple possible explanatory variables. We introduce a simulation-based approach to BMA with the use of Approximate Bayesian Computations (ABC) with divergences. Finally, we assess its effectiveness through a simulation study.

1 Introduction

Consider the following set-up:

Assume a DGP $\mathcal{M}^*$ from which we observe realisations $\mathbf{y} = (y_1, \ldots, y_n)$. Intending to model the DGP, we come up with a set of candidate models $\boldsymbol{\mathcal{M}} = (\mathcal{M}_1, \ldots, \mathcal{M}_k), k \in \mathbb{N}$. Under each model $\mathcal{M}_j$, we assume that y follows a given distribution $p_j(y|\theta_j|\mathcal{M}_j)$. We now ask the question of how exactly does one model the unknown DGP $\mathcal{M}^*$.

BMA gives us the following answer by using rules of conditional probability. We begin by defining the *posterior density* of θ_i under model $\mathcal{M}_j$ and conditional on the observation y as:

$$p_j\left(\theta_j \mid y, \mathcal{M}_j\right) = \frac{p_j\left(\theta_j, y \mid \mathcal{M}_j\right)}{p_j\left(y \mid \mathcal{M}_j\right)}$$

$$= \frac{p_j\left(y \mid \theta_j, \mathcal{M}_j\right) p_j\left(\theta_j \mid \mathcal{M}_j\right)}{\int_{\Theta_j} p_j\left(y \mid \theta_j, \mathcal{M}_j\right) p_j\left(\theta_j \mid \mathcal{M}_j\right) \mathrm{d}\theta_j} \equiv \frac{p_j\left(y \mid \theta_j, \mathcal{M}_j\right) p\left(\theta_j \mid \mathcal{M}_j\right)}{p_j\left(y \mid \mathcal{M}_j\right)} \tag{1}$$

where Θ_j is the parameter space under model j. The last equality defines $p\left(y \mid \mathcal{M}_j\right)$ as the *marginal likelihood of model* j, being the probability of observing the data under model j without conditioning on the parameters. This marginal likelihood is essential in the *posterior probability* of model j given the data defined as

$$P\left(\mathcal{M}_j \mid y\right) = \frac{p_j\left(y \mid \mathcal{M}_j\right) P\left(\mathcal{M}_j\right)}{\sum_{i=1}^{K} p_j\left(y \mid M_i\right) P\left(M_i\right)} \equiv \frac{p_j\left(y \mid \mathcal{M}_j\right) P\left(\mathcal{M}_j\right)}{p(y)} \tag{2}$$

where we implicitly define $p(y)$ as the likelihood of observing y. This is where the model averaging is taking place, as a linear combination of marginal likelihoods and model probabilities, where

each model is weighted by its marginal likelihood. In other words, our beliefs in each model are reinforced by the evidence the data gives in favor of it.
In particular, this model averaging can be used to obtain estimates of relevant quantities across all models. Consider a quantity of interest Δ present and with a common interpretation in all models in $\boldsymbol{\mathcal{M}}$ such as a future observation or even a parameter value. It follows that the marginal posterior distribution across all models is given by

$$p(\Delta \mid y) = \sum_{j=1}^{K} p_j \left(\Delta \mid \boldsymbol{Y}, M_l\right) p_j \left(M_l \mid \boldsymbol{Y}\right) \tag{3}$$

which is an average over all posterior distributions weighted by the posterior model probabilities. In this way, we can obtain combined parameter estimates or predictions from a direct combination of candidate models.
However BMA is not without challenges. Care needs to be taken when specifying the prior distribution over $\boldsymbol{\mathcal{M}}$. Additionally, the computation of marginal likelihoods for each model is non-trivial and often non-availbale in closed form. Finally, in cases where the model space is large, the averaging can become very costly. This is common in variable selection settings for regression or neural network parameter selection where the model space scales exponentially with the dimension of covariates. The first and second challenges have been addressed in GLMs by Raftery [1996] as well as with g-priors in Fernandez et al. [2001], both relying on effective and uninformative priors leading to simplifications of the marginal likelihood computation. The last challenge was address in Madigan et al. [1995] with MC^3, a reversible MCMC-inspired algorithm for exploring large model spaces based on posterior model probabilities.
In this short essay, we will explore the second challenge, namely dealing (or rather *avoid* dealing) with prohibitively expensive or unavailable marginal likelihoods.

2 Likelihood-Free Inference by Simulations

The issue of likelihoods being expensive to evaluate or simply unavailable is not unique to BMA. It is almost an inherent property of Bayesian statistics, where integrating out parameters can become unfeasible due to improper priors and/or missing normalising constants. In fact, even in frequentist statistics, problematic likelihoods are common, be it due to high dimension or complexity of models. In particular, with recent advances in machine learning, it is common to have large Neural Network-based models which are capable of generating samples (GANs) but for which likelihood evaluation is unfeasible. In these situations, one needs to find a workaround in order to not have to rely on likelihood evaluation for performing inference.
One such alternative is specific to *generative models* - models such as GANS from which one can repeatedly draw samples. In fact we can construct a generative model for a parameter simply by defining a prior which one can draw samples from. With models capable of simulating observations, one can perform inference by relying on comparisons between simulated and observed data. In the Bayesian framework, the most popular such approach is Approximate Bayesian Computation (ABC).

ABC for Inference The general idea is as follows. For a given generator model, we wish to infer some parameter of interest θ. We have observed data $\boldsymbol{y}$ with an assumed distribution $p(y|\theta)$ and prior $p(\theta)$, both of which we can sample from. We begin by drawing a sample parameter θ_1 from the prior followed by a sample realisation y' conditional on that parameter. Next, we compare the simulated y' against observed data $\boldsymbol{y}$ ultimately deciding whether or not to retain the parameter θ_1 based on the similarity between simulated and observed data. When repeating this process many times, we end up with a set of parameter values $\{\theta_1, \ldots, \theta_m\}$ which are representative of

our posterior density of θ as we restrict the threshold of acceptance for proposed θ_is. (See eg. Lintusaari et al. [2017] for further details.)
This method gives us a way of performing inference while avoiding any explicit computations; all we are doing is comparing samples of data. We can make use of this for BMA - it has indeed been done already as in Grelaud et al. [2009], Toni et al. [2009] or Marin et al. [2011]. The modification to the method is rather simple and can be summarised in the following steps:

1. Select a model $\mathcal{M}_j$ according to the model prior over $\boldsymbol{\mathcal{M}}$.
2. Draw a parameter θ_j according to the model-specific density $p(\theta^i|\mathcal{M}_j)$.
3. Simulate an observation y' from the model-and-parameter-specific density $p(y|\theta_j, \mathcal{M}_j)$
4. Compare the simulation y' against observed data and retain θ_j if they are "similar enough".

This can be repeated either a large number of times or until a desired number of parameters are accepted. As it may not be obvious from the start how many acceptances one might expect, a good strategy is to run the algorithm for a fixed number of iterations and accept the best 5 or 1% of parameters based on how computationally expensive it is. Once a set of accepted parameters has been collected, we can simply compute the frequencies of accepted parameters coming from different initial models $\mathcal{M}_j$ and use this as approximations to the posterior probabilities of those models. Note that "similar enough" can mean many different ways of comparing data samples.
A common choice for comparing the simulations to observed data are summary statistics, which can lead to information loss. On the other hand, using more data can introduce unwanted variance into the inference. A trade-off ensues between reducing information loss and reducing the variation between simulation-observation pairs Fearnhead and Prangle [2012]. In search for a compromise, the authors in Bernton et al. [2019] introduce a method based on the Wassertein distance in order to compare simulations to observations. Another such attempt is done by Park et al. [2016], where the authors use the maximum mean discrepancy as a dissimilarity measure. Both of these are examples of divergences. But first, let us discuss what a divergence is and how to use it on data.

Statistical Divergences for Sample Comparison Assume we have observed data from distribution $\mathcal{P}^*$ and want to choose a parameter $\boldsymbol{\theta}$ which parameterises a second distribution $\mathcal{P}^{\boldsymbol{\theta}}$ such that the two distributions are as close to each other as possible. A divergence D is then defined as a function of two distributions such that (i) $D(\mathcal{P}^*||\mathcal{P}^{\boldsymbol{\theta}}) \geq 0$ and (ii) $D(\mathcal{P}^*||\mathcal{P}^{\boldsymbol{\theta}}) = 0 \iff \mathcal{P}^* = \mathcal{P}^{\boldsymbol{\theta}}$. Therefore, a divergence can be used to optimise a parameter to recover the best possible model $\mathcal{P}^{\boldsymbol{\theta}}$ for the data generating distribution $\mathcal{P}^*$.
An example divergence is the maximum mean discrepancy (MMD), which for a given choice of kernel k is defined as:

$$\mathrm{MMD}^2\left(\mathcal{P}_x, \mathcal{P}_y\right) = \mathbb{E}_{X_1}\mathbb{E}_{X_2}k\left(X_1, X_2\right) + \mathbb{E}_{Y_1}\mathbb{E}_{Y_2}k\left(Y_1, Y_2\right) - 2\mathbb{E}_{X_1}\mathbb{E}_{Y}k(X_1, Y_1), \tag{4}$$

where X_1, X_2 and Y_1, Y_2 are distributed according to two distributions $\mathcal{P}_x$ and $\mathcal{P}_y$ respectively. Due to the expectation with respect to the distributions, it is straightforward to extend the MMD to an empirical approximation in order to compare samples. As motivated in Bernton et al. [2019] and Park et al. [2016], using these divergences avoids the issues of information loss due to the use of summaries while also being generalisable to high dimensions.

Another possible choice of divergences are scoring rules (SRs). As defined in Gneiting and Raftery [2007], a scoring rule $S(\mathcal{P}, \mathbf{x})$ is a function between a distribution $\mathcal{P}^{\boldsymbol{\theta}}$ and observed data $\mathbf{x}$ as a realisation of a random variable $X \sim \mathcal{P}^*$. Then the *expected scoring rule* is defined as $S(\mathcal{P}, \mathbf{x}) := \mathbb{E}_{Y\sim\mathcal{P}^*}S(\mathcal{P}^{\boldsymbol{\theta}}, Y)$. The SR is termed *proper* if relative to a set of distributions $\mathbf{P}$, if the expected SR is minimised when $\mathcal{P}^* = \mathcal{P}^{\boldsymbol{\theta}}$:

$$S\left(\mathcal{P}^*, \mathcal{P}^*\right) \leq S\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathcal{P}^*\right) \ \forall \mathcal{P}^{\boldsymbol{\theta}}, \mathcal{P}^* \in \mathbf{P}.$$

Furthermore, a SR is termed *strictly proper*, if the minimisation above is unique:

$$S\left(\mathcal{P}^*, \mathcal{P}^*\right) < S\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathcal{P}^*\right) \ \forall \mathcal{P}^{\boldsymbol{\theta}}, \mathcal{P}^* \in \mathbf{P} \text{ s.t. } \mathcal{P}^* \neq \mathcal{P}^{\boldsymbol{\theta}}.$$

By considering the quantity $D_{SR}(\mathcal{P}^*||\mathcal{P}^{\boldsymbol{\theta}}) := S\left(\mathcal{P}^*, \mathcal{P}^{\boldsymbol{\theta}}\right) - S\left(\mathcal{P}^*, \mathcal{P}^*\right)$ for a strictly proper SR, one can see that it defines a divergence. Indeed, (i) is verified by the SR being proper and (ii) is verified by the additional requirement of being strictly proper. This permits the use of strictly proper SRs as divergences to do inference on parameters of a distribution. For instance, one such strictly proper scoring rule is the kernel score, which for an appropriate choice of kernel is identical to the MMD. As for the choice of SR we will use, we introduce the Energy Score as:

$$S_{\mathrm{E}}\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathbf{x}\right) = 2 \cdot \mathbb{E}_{\mathbf{X}' \sim \mathcal{P}^{\boldsymbol{\theta}}} \|\mathbf{X}' - \mathbf{x}\|_2^\beta - \mathbb{E}_{\mathbf{X}'_1, \mathbf{X}'_2 \sim \mathcal{P}^{\boldsymbol{\theta}}} \|\mathbf{X}'_1 - \mathbf{X}'_2\|_2^\beta$$

where $\beta \in (0, 2)$ is a hyperparameter regulating the severity of the divergence for incorrect distributions $\mathcal{P}^{\boldsymbol{\theta}}$. The energy SR is a strictly proper SR for the class of $\mathbf{P}$ such that $\mathbb{E}_{\mathbf{X}' \sim P} \|\mathbf{X}'\|^\beta < \infty$, see Gneiting and Raftery [2007]. It is possible to obtain unbiased estimates of $S_{\mathrm{E}}\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathbf{x}\right)$ by repeated sampling from $\mathcal{P}^{\boldsymbol{\theta}}$, as:

$$\hat{S}_{\mathrm{E}}\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathbf{y}\right) = \frac{2}{m} \sum_{j=1}^{m} \|\mathbf{x}_j - \mathbf{y}\|_2^\beta - \frac{1}{m(m-1)} \sum_{\substack{k=1 \\ k \neq j}}^{m} \|\mathbf{x}_j - \mathbf{x}_k\|_2^\beta$$

where $\mathbf{y}$ are observations and $\mathbf{x}$ are samples. As such, we have a method for inferring parameters of a distribution by comparing the observations to simulated samples. This is equivalent to the following optimisation problem for choosing $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \hat{S}_{\mathrm{E}}\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathbf{x}\right)$ since the first part of D_{SR} is constant in $\boldsymbol{\theta}$.

Proposed Approach Our idea is rather simple and follows from the two previous works using divergences in ABC. What we want to do is perform ABC-based model averaging by using divergences as a discrepancy measure between simulated and observed data. By comparing samples to observed data, we can get a score estimate corresponding to each model $\mathcal{M}_j$. As the solution to the optimisation problem is the θ which minimises the SR, we then pick the set of models associated with the lowest SR estimates. Effectiveley, we replace part (4.) in section 2 by instead evaluating the SR and using that to rank the samples. For the choice of SR, we will use the Energy Score as it allows for easy comparisons of samples.

3 Simulation Study

We consider the common BMA problem of having to decide on which covariates to include in a regression model. We consider 3 possible predictors $(Z_1, Z_2, Z_3) \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_3)$, no interactions and with an $\boldsymbol{\mathcal{M}}$-closed set-up, having the true model $\mathcal{M}^*$ have no intercept and using only two of the predictors as:

$$y|\theta_j, \mathcal{M}_j \sim \mathcal{N}(\begin{pmatrix} z_1 & z_2 & z_3 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ -0.5 \\ 1.3 \end{pmatrix}, 1). \tag{5}$$

We assume a uniform prior among all 8 candidate models and put $\mathcal{N}(0, 1)$ priors on parameters θ_1 and θ_2 while putting a Gamma(1,2) on θ_3 (which is reasonable if we did some exploratory

data analysis beforehand, or have knowledge that there should be a positive trend). We use the Energy SR for assessing the discrepency between observed and simulated data. As done in previous applications of the Energy score (Pacchiardi and Dutta [2021] and Pacchiardi et al. [2021]), we fix $\beta = 1$. We repeat the process discussed in section 2 of drawing a model, cognitional parameters and simulated data 500.000 times. For each draw of simulated data, we sample 10 times from the Gaussian in (5) and compare the simulations to the observed data $\mathbf{y}$. We than rank the model-parameters-simulations triplets by assessing their Energy scores, retaining the lowest 5000.

Model:	(0,0,0)	$(0,0,\theta_3)$	$(0,\theta_2,0)$	$(0,\theta_2,\theta_3)$	$(\theta_1,0,0)$	$(\theta_1,0,\theta_3)$	$(\theta_1,\theta_2,0)$	$(\theta_1,\theta_2,\theta_3)$
Frequency:	0	22.02%	0	61.24%	0	3.56%	0	13.18%

Table 1: Posterior model probabilities after performing ABC via the Energy score.

In Table 1, we show the posterior probabilities of each model. We can see that the method prefers the true model $\mathcal{M}^*$, but still gives some probability to other models, namely those which include θ_3.

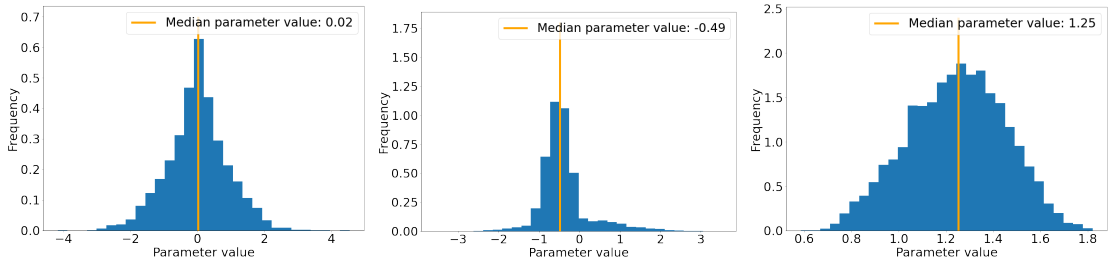

Figure 1: Histograms for parameter values recovered with the ABC via SR approach along with medians. The inferred values coincide

In Figure 1, we show histograms for parameters of retained triplets which are approximations to posteriors with this ABC approach. One can see that we recover the correct values for parameters with θ_3 being slightly under-estimated.

4 Conclusion

In this short essay, we have introduced the idea of using divergences in an ABC framework in order to perform Bayesian model averaging. In particular, we focused on the use of Scoring Rules as divergences and applied the energy score to a simulation study. While the problem considered was simple, we were still able to demonstrate the effectiveness of this approach. We were able to find the correct model and even got posteriors for parameters which were in adherence with the truth.

References

E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Approximate bayesian computation with the wasserstein distance. *arXiv preprint arXiv:1905.03747*, 2019.

P. Fearnhead and D. Prangle. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.

C. Fernandez, E. Ley, and M. F. Steel. Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381–427, 2001.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL `https://doi.org/10.1198/016214506000001437`.

A. Grelaud, J.-M. Marin, C. P. Robert, F. Rodolphe, and J.-F. Taly. Abc likelihood-free methods for model choice in gibbs random fields. 2009.

J. Lintusaari, M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. Fundamentals and recent developments in approximate bayesian computation. *Systematic biology*, 66(1):e66–e82, 2017.

D. Madigan, J. York, and D. Allard. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pages 215–232, 1995.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. Ryder. Approximate bayesian computational methods, 2011.

L. Pacchiardi and R. Dutta. Generalized bayesian likelihood-free inference using scoring rules estimators. *arXiv preprint arXiv:2104.03889*, 2021.

L. Pacchiardi, R. Adewoyin, P. Dueben, and R. Dutta. Probabilistic forecasting with conditional generative networks via scoring rule minimization. *arXiv preprint arXiv:2112.08217*, 2021.

M. Park, W. Jitkrittum, and D. Sejdinovic. K2-abc: Approximate bayesian computation with kernel embeddings. In *Artificial intelligence and statistics*, pages 398–407. PMLR, 2016.

A. E. Raftery. Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–266, 1996.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.