BAYESIAN MODEL AVERAGING IN GLM CONTEXT

Zubia Mansoor
Department of Statistics and Actuarial Science
Simon Fraser University
zubia_mansoor@sfu.ca

Louis Arsenault-Mahjoubi
Department of Statistics and Actuarial Science
Simon Fraser University

Sonny Min
Department of Statistics and Actuarial Science
Simon Fraser University
joosung_min@sfu.ca

August 16, 2020

ABSTRACT

Model uncertainty has been a well-known problem for statisticians and practitioners for decades before the idea of model averaging became popular. In this project, we will demonstrate the shortcomings of standard model selections methods in an introductory example in the linear regression setting on children's cognitive scores data. After illustrating how these methods ignore model uncertainty, we will then present how applying the Bayesian framework in the form of Bayesian model averaging can be used as a technique to remedy the issue. We will see how to interpret the results of our analysis using this new framework, as well as its main benefits and drawbacks. These drawbacks result in difficulties in implementation. To overcome the problems of implementation, we will have to deal with 3 main difficulties: a summation across all models of interest, intractable marginal likelihood integrals, and specifying prior model probabilities. Once we have answered these concerns, with particular emphasis on generalized linear model specific methods, we will apply Bayesian model averaging to a baseball data-set where we fit a generalized linear model. In this second example, we will illustrate the uncertainty in our model choice and compare the predictive performance obtained by using Bayesian model averaging against the best single model selected by standard model selection strategies using different link functions and different implementation methods.

***K*eywords** Bayesian model averaging · Generalized linear models

1 Motivation

Imagine you are a meteorologist who is forecasting the path of a hurricane. In such a sensitive setting, having an accurate prediction and measurement of uncertainty is crucial for early warning. The typical data analysis approach in these statistical problems would involve selecting a 'best' model from a class of models and making inferences based on this single selected model. However, this standard approach suffers from a serious shortcoming as the inferences are dependent on the subset of predictors selected for inclusion in the model. For instance, it is quite possible that two competing models based on different predicted paths of the hurricane will fit the data well and provide sensible inferences. There is, as a result, substantial uncertainty regarding the several potential paths that the hurricane might follow. Thus, proceeding with the 'best' model as the only model that explains the data ignores uncertainty due to model choice and can lead to over-confident inferences and predictions [2][3].

Chatfield [4] and Draper [2] emphasized that mixing models gives better prediction accuracy than using a single model. The basic idea is to use a weighted average of the individual model predictions instead of a single prediction obtained

from the 'best' model. There are a number of lens from which one can view this problem; we present the basics of model averaging through the Bayesian Model Averaging (BMA) framework.

In this section, we will first review the classical method of single model selection using stepwise selection and highlight certain limitations. Following this discussion, we will introduce Bayesian Model Averaging in Section 2 as a Bayesian approach to account for model uncertainty in contrast to the standard methods and subsequently elaborate on the interpretation of the BMA estimates. Section 3 describes its use in linear regression models for the choice of covariates. Next, we briefly go through some benefits and drawbacks of the BMA methodology in Section 4. A detailed description of the implementation of this framework is provided in Section 5. Thereafter, Section 6 shows the application of BMA in the generalized linear models (GLM) setting with different link functions and implementation strategies. Finally, we conclude with some discussions and interesting areas of application.

1.1 Stepwise Selection

A natural strategy for model selection is to choose a single 'best' model and make inferences conditioned on that single selected model. There is a whole host of such techniques and we will focus on one such popular method called stepwise selection [5]. At every step, it evaluates features from the entire collection and determines its suitability to be included in the final model. The main approaches are:

Forward Selection: involves starting with only the intercept in the model. Variables that most improve the model fit are then sequentially added according to a model criterion. This process is repeated until the model can no longer be improved to a statistically significant extent.

Backward Elimination: involves starting with the full model. Variables that have the least impact on the model fit are then sequentially dropped according to the chosen model criterion. This process is repeated until the deletion of variables causes a statistically significant loss of fit.

Bidirectional elimination: a combination of the above two approaches, testing for the inclusion or exclusion of variables at every step.

The above approaches implement stepwise selection using either statistical criterion like the Akaïke information criteria (AIC), the Bayesian information criterion (BIC), or statistical tests.

1.2 Limitations

The standard method of stepwise selection is accompanied by a few limitations. Most notably, they are:

High Variability: Stepwise selection is extremely variable as it depends upon the approach used for model selection. Different selection procedures can lead to different 'best' models being produced each time.

Involves Multiple Testing of Hypotheses: Stepwise selection performed using statistical tests involves sequentially adding and deleting terms based on approximate asymptotic likelihood ratio tests. Due to multiple testing of hypotheses, we end up with an inflated Type-I error and in addition asymptotic approximations tend to break down for small samples. [12]

Ignores Model Uncertainty: Most importantly, by selecting a single model for inferences, stepwise selection ignores model uncertainty and leads to an underestimation of the uncertainty about the quantities of interest.

Up to this point, we have outlined some of the limitations of standard methods and hence acknowledge the need to improve upon these approaches. In the statistical literature, model averaging has been proposed as a way to better reflect the true uncertainty in our estimates. To that extent, we propose the use of *Bayesian Model Averaging* to achieve this goal. While this will be formalized later, we should keep in mind that allowing for the incorporation of model uncertainty into inference forms the crux of this motivation.

2 Bayesian Model Averaging

In this section, we will introduce the BMA framework and delve deeper into making inferences using the concept of model averaging. We will also provide some insights on the interpretation of BMA coefficients which remains to be a contested issue.

2.1 Definition

Bayesian model averaging (BMA) [1] provides a systematic and coherent approach for representing model uncertainty by constructing a probability distribution over all possible models, where each probability provides a measure of how likely the different models are. The fundamental idea is to make inferences based on weighted averages of the quantities of interest where the weights correspond to the model probabilities.

2.2 Model

Let $\mathcal{M} = (M_1, \ldots, M_K)$ be the set of models under consideration.

Using Bayes' theorem, we obtain the posterior probability of each model $p(M_k|data)$. First, we assign a prior probability $p(M_k)$ that M_k is the true model. Then we obtain the marginal likelihood of each model $p(data|M_k)$. By Bayes' rule, we update the posterior probability of each model $p(M_k|data)$ after seeing the data, via the marginal likelihood of model M_k:

$$p(M_k|data) = \frac{p(data|M_k) \times p(M_k)}{\sum_{l=1}^{K} p(data|M_l) \times p(M_l)} \tag{1}$$

where, $p(data|M_k) = \int p(data|\theta_k, M_k)p(\theta_k|M_k)d\theta_k$ is the marginal likelihood of model M_k. Here, θ_k is the vector of parameters for model M_k, $p(\theta_k|M_k)$ is the prior density of θ_k under model M_k and $p(data|\theta_k, M_k)$ is the likelihood.

In this way, the marginal likelihood of each model $p(data|M_k)$ serves to reweight the prior probabilities so that models with higher likelihood receive larger weights while those with lower likelihood receive smaller weights. We renormalize this weighted prior probability by dividing it by the sum $\sum_{l=1}^{K} p(data|M_l) \times p(M_l)$ to get the posterior model probabilities.

Once we have obtained the posterior probability of each model, we can use these probabilities as weights to make inferences and obtain weighted averages of the quantities of interest.

Let Δ be the quantity of interest. Δ can be Y^*, a future observation; β_j, the coefficient of variable X_j; γ_j, the indicator function that variable X_j is included or $p(\beta_j|data)$, the posterior probability of β_j after seeing the data. The posterior distribution of Δ given data is:

$$p(\Delta|data) = \sum_{k=1}^{K} p(\Delta|M_k, data) \times p(M_k|data) \tag{2}$$

Thus, the posterior density, of the quantity of interest, is a weighted average of the densities under each of the individual models where the weights correspond to the posterior model probabilities. Models with high posterior probability receive more weight, while models with low posterior probability are discounted.

Moreover, the posterior mean and variance of Δ are given by:

$$E[\Delta|data] = \sum_{k=1}^{K} \hat{\Delta}_k \times p(M_k|data) \tag{3}$$

$$Var[\Delta|data] = \sum_{k=1}^{K} \{Var[\Delta|M_k, data] + \hat{\Delta}_k^2\} p(M_k|data) - E[\Delta|data]^2 \tag{4}$$

where, $\hat{\Delta}_k = E[\Delta|M_k, data]$. [6] [2]

The posterior mean and variance of Δ use the model-specific expectations and variances weighted by their posterior probabilities.

Since the weights in Equation (1) are probabilities and sum to one, if the 'best' model had posterior probability one, all of the weights would be placed on that single best model. In this case, using BMA would be equivalent to selecting the best model with the highest posterior probability. However, if there are several models that receive substantial weights, they would all account for the uncertainty about the true model.

2.3 Interpretation

In a typical regression problem, we have a response with a set of predictor variables. The coefficient of a predictor variable, in this context, is interpreted as the effect of that predictor after adjustment for the effects of the other predictors included in the specific model under consideration. However, Kass [15] argued that when models are averaged, we lose this simple interpretation. This is because now within each model, the coefficient has a different meaning that depends on the predictors included in that model. An interesting question is raised by Draper [1] of whether BMA really amounts to combining "apples and oranges" and hence is invalid.

In response to Draper, Hoeting et al [1] view BMA as Bayesian inference on an overarching full model that takes into account the complete collection of covariates in the study via assignment of a prior probability of one-half on the inclusion of each covariate.

Raftery [15] summarizes his response to Kass's objection of combining models under BMA. According to Raftery, Equation (2) can be looked at from two different angles. The first being a mixture across different models which is harder to interpret. However, alternatively, we can view it as the posterior distribution from a single model with all the covariates but with a prior distribution that assigns a probability to each coefficient being equal to zero. In doing so, the interpretation of the coefficient of a predictor variable boils down to the effect of that predictor controlling for all the other predictors but allowing for the possibility that they have no effect.

Following this discussion, the usual regression interpretation of the predictor coefficient - "the effect of that predictor after adjustment for the effects of the other predictors" - can be modified for the BMA setting as "the effect of that predictor after adjustment for the possibility of the effects of the other predictors". The underlying difference between BMA and the standard method of selecting a single set of covariates is that *we have adjusted for the possibility of the effects of all the covariates.*

3 BMA in the linear regression setting

In this section, we provide an example to illustrate the shortcomings of the classical model selection methodology. We demonstrate these limitations in the linear regression setting and present how Bayesian model averaging can be used to remedy the issue.

3.1 Data

We have data from a survey of 434 adult American women and their children. [7]

The response variable is *kid_score* which is a continuous response denoting the kid's IQ score. The goal is to model the linear relationship between the four explanatory variables (mother's characteristics) and the response variable.

Table (1) summarizes the response and the 4 predictor variables for 434 observations.

Variable name	Description	Data type
kid_score	Kid's IQ score	Continuous
hs	The mother has a high school degree	Indicator
iq	Mother's IQ score	Continuous
work	The mother worked during the first three years of the kid's life	Indicator
age	Mother's age	Continuous

Table 1: Variable description

3.2 Standard Method

We will first consider the standard method of Stepwise Selection using Backward Elimination according to the Bayesian Information Criterion (BIC). It is defined as:

$$\text{BIC} = -2\ln(\widehat{\text{likelihood}}) + (p+1)\ln(n)$$

Here, n is the number of observations and p is the number of predictors in the model. Note that the model with the smallest BIC is preferable.

We will perform model selection on kid's cognitive score dataset by following the steps detailed below.

- Start with the full model consisting of all the predictor variables: *hs, iq, work, age*.
- Drop one variable at a time and record all BICs.
- Choose the model with the smallest BIC.
- Repeat this process until none of the models yield a decrease in BIC.

Step	Model	Dropped	BIC
Full	kid_score ~ hs + iq +work +age		2541.1
Step 1	kid_score ~ hs + iq +work	[-age]	2535.4
Step 2	kid_score ~ hs + iq	[-work]	2530.6
Step 3	kid_score ~ hs + iq		2530.6
	kid_score ~ hs	[-iq]	2531.7
	kid_score ~ iq	[-hs]	2604

Table 2: Backward Elimination with BIC

Table (2) summarizes the process of backward elimination using BIC on the kid's cognitive dataset. We start with the full model that predicts kid's IQ score from mother's high school status, mother's IQ score, mother's work status, and mother's age. The BIC for the full model is 2541.1. At step 1, we exclude each variable from the full model and record the new BIC. We observe that dropping the variable *age* (relative to dropping any other variable at this step) yields the lowest BIC which is 2535.4. Similarly, at step 2, removing the variable *work* results in the lowest BIC of 2530.6. The current model consists of two explanatory variables namely *hs* and *iq*. Finally, dropping either *hs* or *iq* at step 3 leads to an increase in BIC. This implies that the model observed in step 2 is the 'best model' according to our chosen model criterion. This model predicts kid's IQ score using mother's high school status and mother's IQ score.

The example above shows how BIC can be used to pick a 'best' model. However, there might be several competing models with similar values of BIC. Selecting only the model with the lowest BIC means ignoring the presence of other models that are equally good or can provide useful insights. This methodology, therefore, ignores model uncertainty and hence we leverage the BMA framework in the next subsection to account for that.

3.3 BMA

As stated earlier, we will represent model uncertainty by constructing a probability distribution over all models $(16 = 2^4)$ corresponding to all possible subsets of the four predictor variables.

3.3.1 Fitted Models

We will use the function *bas.lm* in BAS package [17] to obtain the fitted models in R. We use BIC to approximate the marginal likelihood of each model $p(data|M_k)$. We assign a uniform prior $p(M_k); k = 1, \ldots, 16$, to each model in

the set under consideration. Thus, $p(M_k) = \frac{1}{16}$. Table (3) shows us the summary of the top 5 models ordered by their posterior model probabilities.

	P(B != 0 | Y)	model 1	model 2	model 3	model 4	model 5
Intercept	1	1	1	1	1	1
hs	0.611	1	0	0	1	1
iq	1	1	1	1	1	1
work	0.112	0	0	1	1	0
age	0.069	0	0	0	0	1
BF	NA	1	0.562	0.109	0.088	0.061
PostProbs	NA	0.529	0.297	0.058	0.046	0.032
R2	NA	0.214	0.201	0.206	0.216	0.215
dim	NA	3	2	3	4	4
logmarg	NA	-2583.135	-2583.712	-2585.349	-2585.57	-2585.939

Table 3: Summary of top 5 models

The summary table provides information on following statistics for the top 5 models.

Item	Description
P(B!=0 | Y)	Posterior inclusion probability (pip) of each coefficient under data $\boldsymbol{Y}$
0 or 1 in the column	indicator of whether the variable is included in the model
BF	Bayes factor BF[M_k: M_b], where M_b is the model with highest posterior probability
PostProbs	Posterior probability of each model
R2	R-squared in the ordinary least square (OLS) regression
dim	Number of variables (including the intercept) included in the model
logmarg	Log of marginal likelihood of the model

The output suggests that the variable *iq* is included in all the top 5 models while the variable *age* is excluded from the top 4 models. The model with the highest posterior probability includes the intercept, *hs* and *iq*. Based on this posterior probability, we believe that there is a 53% chance that the model with the mother's high school status and the mother's IQ score is the true model. The model with the second-highest posterior probability includes only the intercept and *iq*. There is a non-negligible posterior probability of approximately 0.3 on the model with the mother's IQ score only. The top two models collectively account for 83% of the probability, with the rest 17% being distributed across the remaining 14 models.

3.3.2 Uncertainty Plot

In order to get a more comprehensive overview for model comparison, we look at a visualization of the models that illustrate model uncertainty, beyond the top five models that we considered previously.

In Figure (1), the rows correspond to the predictor variables and the columns represent the 16 possible models. These models are arranged in order of their log posterior odds over the null model (model with only the intercept). Black boxes represent the variables that are excluded whereas the colored box means that the variable is included in the model. The color of each column is proportional to the log of the posterior probabilities. This allows us to view clusters of models that have roughly similar posterior probabilities.

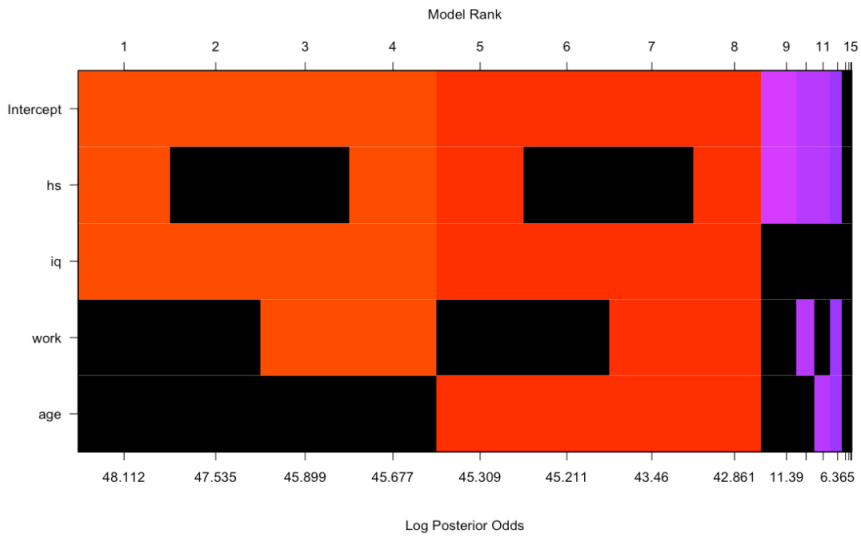

Figure 1: Uncertainty Plot

Again, we see that the model with only *hs* and *iq* has the most mass. Therefore, this group of explanatory variables provides a better description of the response relative to the other subsets of predictors. Each of these predictors is included in exactly 8 models. The variable *iq* appears in the top 8 models with high posterior probabilities and hence is excluded from the lower probability models.

3.3.3 Coefficient Summary under BMA

The summary table (3) and the uncertainty plot in Figure (1) revealed that the top two models are much better supported by the data than the other models. However, among those two models, no one model clearly dominates the other. Hence, there is substantial uncertainty due to model choice. In such a situation, we turn to model averaging using BMA to make predictions and get parameter estimates.

Table (4) displays the predictor variables and the corresponding statistics in the BMA setting. The second column *post mean* displays the BMA coefficients averaged over all the models, including the models where the variable was not selected (implying that the coefficient is zero in this case). These estimates would be used for future predictions. The posterior standard deviation given by *post SD* provides a measure of variability of the coefficient. The last column presents posterior inclusion probabilities (PIP) for each variable. The values in this column are the sum total of the posterior model probabilities for all the models where that variable was included. These values thus provide a measure of how likely the variable will be included in the true model.

Marginal Posterior Summaries of Coefficients:

Using BMA

Based on the top 16 models

	post mean	post SD	post p(B != 0)
Intercept	86.79724	0.87287	1.0000
hs	3.59494	3.35643	0.61064
iq	0.58101	0.06363	1.0000
work	0.36696	1.30939	0.1121
age	0.02089	0.11738	0.06898

Table 4: Coefficient Summary under BMA

We observe that the variable *iq* has a posterior inclusion probability of 1, suggesting that it is very likely that the mother's IQ score should be included in the true model. Thus, virtually all of the posterior model mass rests on models that include *iq*. The variable *hs* also has a high posterior inclusion probability of about 0.61. In contrast, *work* and *age* have relatively small PIPs compared to *iq* and *hs* and hence are less likely to be included in the true model.

3.3.4 Posterior Density Plots

Finally, let's turn to visualizing plausible values for the coefficients, taking into account that there is uncertainty about the best model. Figure (2) displays the plot of the posterior distributions for each of the regression coefficients. Here, the x-axis represents the values that the coefficients take and the y-axis denotes the probability mass. The vertical bar represents the posterior probability that the coefficient is 0. The bell-shaped curve represents the density of plausible values from all the models where the coefficient was non-zero. This is scaled so that the height of the density for non-zero values is the probability that the coefficient is non-zero.

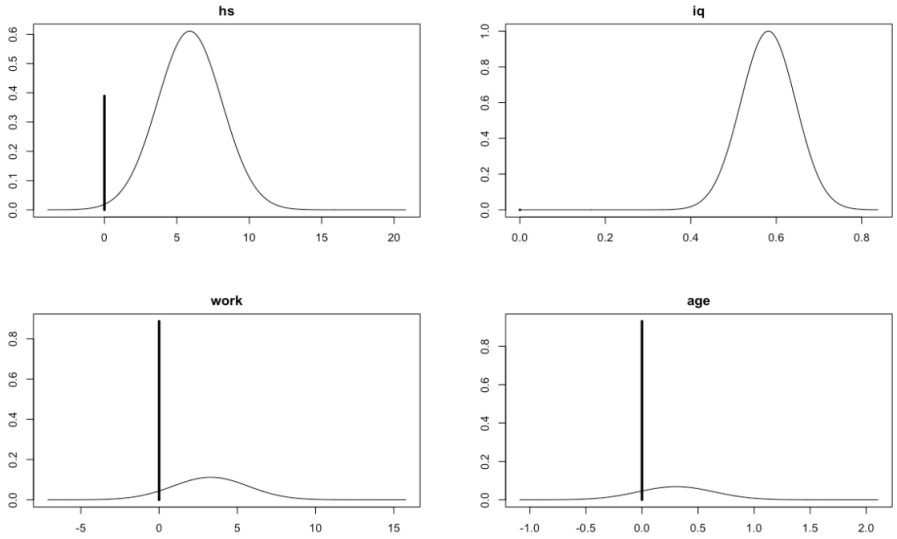

Figure 2: Posterior Density Plots

These posterior density plots draw the same conclusions as the summary table (3). The posterior probability distributions of the variables *age* and *work* have a large point mass at 0 and hence the high vertical bars. Even for the models where it has been forced into, the distribution overlaps 0. Comparatively, the distribution of *hs* has a relatively small mass at 0. For *iq*, the probability that the coefficient is zero is quite small. There is a little tip at 0 for the variable *iq*, indicating that the posterior inclusion probability of *iq* is not exactly 1. The range of plausible values is centered far from 0, also reflecting our beliefs after seeing the data that this variable is very likely to be included in the true model. Thus, since the probability mass for *iq* to be 0 is so small, we are almost certain that the mother's IQ score should be included under Bayesian model averaging.

4 Benefits and Drawbacks of BMA

Now that we have covered the basics of BMA and demonstrated it on a real-world dataset, let us look at some benefits and drawbacks of this framework.

4.1 Benefits

Accounts for model uncertainty: When a single model dominates the posterior model probabilities (PMP), we can use this model as our best guess of the true situation. However, often we remain uncertain about the true model. In this case, we benefit from Bayesian model averaging, where we take the weighted average of all the candidate models, with the

weights provided by the PMPs. This reduces the overconfidence associated with our predictions and estimates by taking into account model uncertainty.

Results in improved predictions: Often in practice, we cannot consistently identify the true model. This induces a prediction error which is more pronounced in methods based on the selection of a single model while it is mitigated by BMA.

Updates its estimates: As the data accumulates, the model weights are continually adjusted and hence the estimates get updated accordingly. On the other hand, standard methods are extremely variable and may lead to different 'best' models with the observation of new data.

Robust to model misspecification: In the context of BMA, we make use of several competing models to make inferences instead of relying on a single model. Thus, the chances that at least one of the models in the full set under consideration is at least approximately correct makes BMA relatively robust to model misspecification.

4.2 Drawbacks

Number of models: The total number of all possible models under consideration in Equation (2) can be large which can render the exhaustive summation infeasible.

Integrals involved: To get the posterior distribution of Δ, the integrals implicit due to Equation (2) can be difficult to compute. In the next section, we will introduce methods to overcome this drawback.

Specification of $p(M_k)$: Choosing the prior model probabilities over the different models can be a challenging problem and has been an interesting area of research. We shed more light on this in Subsection 5.3.

Choosing the class of models: After having overcome the previous limitations, we are still left with selecting the class of models that is used to get the weighted average of the quantities of interest.

In summary, we proposed the use of BMA by highlighting the limitations of standard methods of model selection. Then, we covered the fundamentals of BMA and elaborated on the interpretation of the estimates. Building upon this, we compared BMA with the standard methodology by implementing it on a real-world example. We showed how BMA improved upon classical methods by acknowledging the presence of other competing models. Finally, we ended with some of the benefits and drawbacks associated with this framework. Moving forward, we will introduce methods to tackle some of these drawbacks.

5 Implementation of Bayesian Model Averaging

In this section we will go through various methods used to overcome the main difficulties in implementing Bayesian model averaging (BMA). First, we will see how we can handle the summation involved in averaging over all of our models of interests. This will involve us presenting 2 methods: Occam's Window & Markov Chain Monte Carlo Model Composition (MC^3) and the algorithms to implement them. Secondly, we will give overviews of numerical approximations methods for the marginal likelihood integrals implicit in computing the posterior for our quantity of interest. We will start with a method based on a Bayes factor approximation that is particularly practical in the generalized linear models (GLM) setting before moving on to widely applicable ones that might be less useful for GLMs. Lastly, some guidelines for choosing prior probabilities on the models of interest will be given. One such guideline is simpler to implement, but makes assumptions on the independence of the predictors that are included in the models. The other method relaxes this assumption, but is arguably more difficult to use in practice as it relies on having an expert generate 'imaginary' data.

5.1 Dealing with the summation

The first problem we will address is the summation across the set of all models of interest M:

$$p(\Delta|data) = \sum_{k=1}^{K} p(\Delta|M_k, data) \times p(M_k|data)$$

If we are in a situation where there are p predictors under consideration, this implies that we have $K = 2^p$ models to sum over. Going through all of these models rapidly becomes infeasible. We will look at a way of limiting the scope of the summation with Occam's window and then look at applying Monte Carlo methods to approximate our distribution of interest directly.

5.1.1 Occam's window

The Occam's window method that was introduced in Madigan and Raftery (1994) greatly reduces the number of models under considerations. The aim is to make the summation involved in computing $p(\Delta \mid data)$ manageable by looking only at a subset of M that is selected according to 2 standard principles in scientific practice.

The first of these principles is that models that are not supported by the data relative to our most probable model should be considered discredited. So, we will want to compare all of our models $M_k \subset M$ with our most likely model(s) $M' = \{M_i \subset M : p(M_i \mid data) = max_l\{p(M_l \mid data)\}\}$ and discard the most relatively unlikely ones according to some constant threshold C. Mathematically, this amounts to exclusively considering models in the set

$$A' = \Big\{M_k : \frac{p(M' \mid data)}{p(M_k \mid data)} \leq C\Big\}$$

In their examples, Madigan & Raftery (1994), select $C = 20$ as it is comparable to the 5% p-value cut-off that is often used for hypothesis testing. The choice of C is context dependent. The authors suggest numbers between 10 and 100, while acknowledging that circumstances like evaluating forensic evidence for criminal cases could warrant a value as high as 1000.

The second principle is the eponymous Occam's razor. Stated by William of Occam (circa 1287–1347) as: "Plurality must never be posited without necessity", the principle is now common place in philosophy, science and, in particular, statistics. The heuristic claims that, all other things being equal we should favor explanations with fewer assumptions against more complex ones. In statistics, this is applied when we favor more parsimonious models over their more intricate counterparts. In fact, we see this fact explicitly in the Akaike information criterion and the Bayesian information criterion as both penalize models for every additional parameter they incorporate. In our context, we will remove the models in

$$B = \Big\{M_k : \exists M_l \in A', M_l \subset M_k : \frac{p(M_l \mid D)}{p(M_k \mid data)} > 1\Big\}$$

as they have a simpler sub-model contained in them that is more probable given our data. Combining the 2 principles, we are left to consider only the set $A = A' \setminus B$. Note that although the eliminated models through this procedure should have relatively small posterior probabilities, we will have eliminated most of the models and possibly, considered in their aggregate, what amounts to a large portion of the posterior probability. We now have the problem of finding our A of interest among our full set of models.

Although removing models from M according to the two principles above is seen as desirable in itself, there is no guarantee that using Occam's window reaches the same (or even approximately the same) conclusions as a 'full' BMA. Indeed, in the comment to Hoeting et al. (1999), Draper suggests referring to these as seperate objects (BMA and BMA^*) since the search criteria are only giving the most relatively probable models (with, possibly, a preference for simplicity) they are not necessarily a representative sample of the model space. Hoeting et al. (1999) note that, despite not having theoretical proofs of the validity of BMA over Occam's window as an approximation to the full BMA, in all of their experience the difference between the two was small.

We will therefore introduce an algorithm to find the subset of models of interest A that utilizes the models' posterior probability ratios in what is called Occam's window. The "Down" algorithm, which starts from larger models and goes on removing parameter has the following steps (let O_L & $O_R \in \mathbb{R}$ be two user chosen thresholds with $O_L < O_R$, C be the full set of models and A be the empty set):

- 1) Select model M from C arbitrarily
- 2) $C < -C \setminus M$ & A <- A+M
- 3) Remove a parameter from M to get $M_0 \subset M$

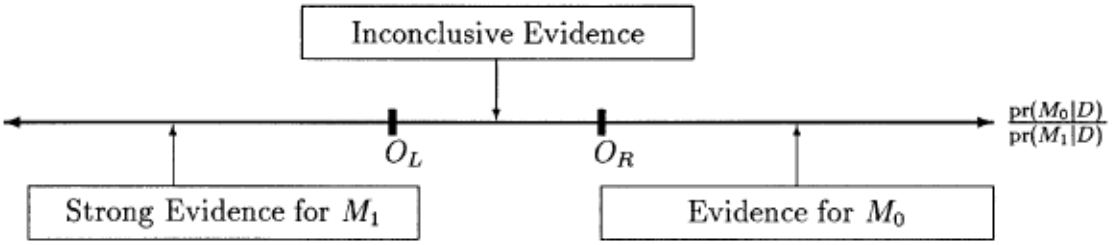

FIG. 1. *Occam's window: interpreting the posterior odds.*

Figure 1 from Hoeting et al. 1999 (note: they do not log the posterior probability ratios like in Madigan & Raftery 1994)

- 4) Compute $B = \frac{P(M_0|data)}{P(M|data)}$
- 5) If $B > O_R$, then $A < -A - M$ & if $M_0 \ni C$, $C < -C + M_0$
- 6) If $O_L \leq B \leq O_R$, then if $M_0 \ni C$, $C < -C + M_0$
- 7) If there are submodels of M left, go to 3
- 8) If $C \neq 0$, go to 1

The authors also present an "Up" algorithm which starts from the null model and sequentially adds terms to the selected model and an in between "Up and Down" algorithm which alternates between removing and adding parameters to selected models. Their experience applying the different methods yielded no significant differences between the three of them. It is for this reason that we present only the "Down" algorithm here.

As for the selection of the thresholds, O_L and O_R, Madigan and Raftery (1994) use $\frac{1}{20}$ & 1 respectively. Subsequent work in the field by Raftery, Madigan, and Volinsky (1996) found improved predictive scores in practice by using $O_L = \frac{1}{20}$ and $O_R = 20$. Using $O_L = \frac{1}{O_R}$ is known as a symmetric Occam's window because the same posterior probability ratio is required from either model for it to be favored over the other. Since this procedure does not give any preference to the simpler model, it discards the Occam's razor principle and only uses the first principle to select models. By reducing O_R, we can choose to enforce the principle of Occam's razor to a greater extent.

An alternative algorithm for finding the set A in a symmetric Occam's window is presented in Volinsky et al. (1997). In this case,

$$A = A' = \left\{ M_k : \frac{p(M' \mid data)}{p(M_k \mid data)} \leq C \right\}$$

Their work is essentially built on algorithms that sought to identify the best models in other contexts.

In the linear regression setting, Furnival and Wilson (1974) introduce the "leaps and bounds" algorithm to identify the most promising subset of our models of interest. The method returns the q best models of each considered size according to residual sum of squares, the maximum likelihood estimates of their parameters, $\hat{\theta}$, the variances of these estimates $var\hat{\theta}$, the model R^2 or any subset of these quantities. Their algorithm utilizes different sweep operators depending on which quantities that we are interested in computing. For model screening, for example, we might only need to compute the residual sum of squares SSR. In the standard linear regression setting,

$$y_{nx1} = X_{nxp}\theta_{px1} + \epsilon_{nx1}$$

for example, we have that if

$$C = \begin{bmatrix} X'X & X'y \\ y'X & y'y \end{bmatrix}$$

then,

$$Sweep(C, X) = \begin{bmatrix} (X'X)^{-} & \hat{\theta} \\ -\hat{\theta} & SSR \end{bmatrix}$$

The algorithm uses matrix operators such as the semi-sweep operator and the Gaussian elimination to rapidly go through which models of a given size have the best fit.

Lawless and Singhal 1978 expand the algorithm to encompass the nonlinear regression setting and allow it to yield the Bayesian information criterion approximations (in the form of approximate likelihood ratio tests). Volinsky et al. use this modified algorithm in the BMA context to find A. This is the algorithm that is employed to solve the summation problem in the BMA R package (Raftery 2020). Volinsky et al. note that for a large enough q, the algorithm will usually provide set desired set of models with some additional ones. In order to further thin the algorithm's output set and better approximate A, the authors suggests keeping only models with posterior model probability above $\frac{1}{C'}$ where C' is another user selected constant and $C' > C$ to avoid losing models that should be in A. They used $C' = C^2$ and found no important discarding of models in A. In fact, in the BMA package, the tunable parameter $OR.fix$ is the choice of exponent in setting $C' = C^{OR.fix}$ and its default value is 2. Of interest is the fact that the implementation of this solution in does not allow for asymmetric Occam's windows.

5.1.2 Monte Carlo Markov Chain Model Composition (MC^3)

MC^3 is a Monte Carlo method presented in Madigan and York (1995) that aims to approximate the distribution of the quantity of interest in BMA $P(\Delta \mid M, data)$ by creating a discrete-time Markov chain with stationary distribution $P(M \mid data)$. The Markov chain $\{M(t)\}$, $t = 1, 2, 3, ..., N$ will take on a model in our set of interest $\mathcal{M}$ at each time step. Given that we are at a specific model M, we must define a set of models from which a move can be proposed, $nbd(M)$. In the setting of graphical models, the authors choose the set of models with one additional link or one less link than the current model. In the GLM context, this suggests using models with one more or one less predictor than the current model as $nbd(M)$. We must then choose a transition density $q(M' \mid M)$, which must be 0 for all models outside $nbd(M)$ and non-zero for all models in $nbd(M)$. Using a symmetric transition density (where $q(M' \mid M) = q(M \mid M')$ for all M, M' in our set of models of interest), we can now apply the Metropolis-Hastings algorithm as follows:

- 1) Choose M_0 arbitrarily.
- 2) for $i \geq 1$, until N
  - Draw $M' \sim q(M' \mid M_{i-1})$
  - With probability

$$min\Big\{1, \frac{p(M' \mid data)}{p(M_{i-1} \mid data)}\Big\}$$

    set $M_i = M'$ otherwise set $M_i = M_{i-1}$

This will yield a sequence $M(1)$,..., $M(N)$ of N models approximately under their posterior distributions $P(M_i \mid data)$. With the help of standard MCMC results we have that if we take $\hat{G} = \frac{1}{N}\sum_{t=1}^{N} g(M(t))$, then we get that $\hat{G} \to E(g(M))\ as\ N \to \infty$. If we set $g(M) = p(\Delta \mid M, data)$, we get convergence to the summation that was needed. Also of interest is the fact that this is the method used in the R package 'BMA' (Raftery et al. 2020).

5.2 Methods for computing the likelihood integrals

In obtaining $P(\Delta \mid data)$, we must perform a weighted sum of our posterior model probabilities $P(M_i \mid data)$, $i = 1, 2, ..., K$ where

$$P(M_i \mid data) = \frac{p(data \mid M_i)p(M_i)}{\sum_{l=1}^{K} p(data \mid M_l)p(M_l)}$$

and,

$$p(data \mid M_k) = \int p(data \mid \theta_k, M_k)p(\theta_k \mid M_k)d\theta_k$$

with θ_l being the vector of parameters in M_k.

First, we will look at a method to compute $P(M_i \mid D)$ directly that is applicable in the GLM context based on Raftery 1996's approximation of the Bayes factor. Then, we will briefly go through other more general means of approximating the integral. One of these will be an application of the Laplace method from Tierney and Kadane 1986 while the other utilizes maximum likelihood estimates of θ_k.

5.2.1 A Bayes factor based approximation for GLMs

The Bayes factor for model M_1 against M_0 is, by definition, a ratio of the marginal likelihoods that we are interested in estimating:

$$BF[M_1 : M_0] = \frac{p(data \mid M_1)}{p(data \mid M_0)}$$

This suggests using approximations for the Bayes factors to estimate them. Raftery (1996) offers us multiple methods of achieving this goal. The most accurate of which is:

$$2log(BF[M_1 : M_0]) \approx D + (E_1 - E_0)$$

Where we are in a Bayesian context with priors on β_k, $k = 1, 2, ..., p$ with $E[\beta_k \mid M_k] = \omega_k$ & $Var[\beta_k \mid M_k] = W_k$. Moreover, $D = 2(l_1(\hat{\beta}_1 - l_0(\hat{\beta}_0))$ is our well-known likelihood ratio test statistic with D following a Chi-squared distribution with the difference of the two models' number of parameters as its degrees of freedom when $M_0 \subset M_1$. We also have l_k being the log-likelihoods $l_k(\hat{\beta}_k) = log(p(data \mid \beta_k, M_k))$. We define

$$E_k = 2\lambda_k(\hat{\beta}_k) + \lambda_k'(\hat{\beta})^T(F_k + G_k)^{-1}(2 - F_k(F_k + G_k)^{-1})\lambda_k'(\hat{\beta}_k) - log|F_k + G_k| + p_k log(2\pi)$$

and F_k is the expected Fisher information matrix, $G_k = W_k^{-1}$, $\lambda_k(\beta_k) = log(p(\beta_k \mid M_k))$ is the log of our prior on β_k. If we have n observations, this approximation strategy has an overall relative error of $O(n^{-\frac{1}{2}})$ and, in particular, when we choose the canonical link, the error improves to $O(n^{-1})$. Raftery (1996) demonstrates this approach in an example using normal priors and provides us with a parametrization that involves the choice of a single quantity. The author also gives methods of choosing this quantity when there is high prior uncertainty. Although the errors associated with the approximation are higher than those with the approximation provided by Tierney and Kadane (1986), they rely quantities that are easily obtainable for generalized linear models. Moreover, this is the method that the R BMA package uses in its bic.glm function (Raftery et al. 2020). Once we have this method for evaluating approximated Bayes factor, we can obtain our posterior probabilities from it in the following way:

Letting $M_0, M_1, ..., M_K$ be our K+1 models under consideration and $O[M_i : M_0] = \frac{p(M_i)}{p(M_0)}$ is the prior odds of model M_i to model M_0.

$$p(M_k \mid data) = \frac{BF[M_k : M_0]O[M_k : M_0]}{\sum_{i=0}^{K} BF[M_i : M_0]O[M_i : M_0]}$$

which is equation (9) from Hoeting et al. (1999). We present a derivation of this fact in section 6, equations (5).

5.2.2 Alternatives for marginal likelihood approximation

In this section, we will obtain the Laplace method based approximation given in Hoeting (1999) of $P(data \mid M_K)$ following the methodology of Tierney and Kadane (1986).If we denote the likelihood $l_k = p(data \mid \theta_k, M_K)$, then we can write

$$p(data \mid M_K) = \int e^{log(l_k)} p(\theta_k \mid M_k) d\theta_k$$

If we now let $L = \frac{log(p(\theta_k|M_k)) + log(l_k)}{n}$ we get,

$$p(data \mid M_K) = \int e^{nL} d\theta_k$$

Then, with $\tilde{\theta}$ being the posterior mode of θ, the Laplace method gives us:

$$\int e^{nL} d\theta_k \approx \int e^{nL\tilde{\theta} - \frac{n(\theta-\tilde{\theta})^2}{2|\psi|}} d\theta_k$$

$$= (2\pi)^{\frac{p_k}{2}} |\psi|^{\frac{1}{2}} e^{nL(\tilde{\theta})}$$

$$= (2\pi)^{\frac{p_k}{2}} |\psi|^{\frac{1}{2}} e^{log(p(\tilde{\theta}_k|M_k)) + log(l_k(\tilde{\theta}))}$$

$$= (2\pi)^{\frac{p_k}{2}} |\psi|^{\frac{1}{2}} l_k(\tilde{\theta}) p(\tilde{\theta_k} \mid M_k)$$

$$= (2\pi)^{\frac{p_k}{2}} |\psi|^{\frac{1}{2}} p(data \mid \tilde{\theta}_k, M_k) p(\tilde{\theta}_k \mid M_k))$$

Where $|\psi|$ is the determinant of minus the inverse Hessian of $log(\frac{p(data|\tilde{\theta}_k, M_k)}{p(\tilde{\theta}_k|M_k)})$. With this we have equation (10) from Hoeting et al. (1999). The error in this approximation is $O(n^{-1})$ which, although it is better than the one in the previous section (if the link-function is not canonical), is harder to obtain in the generalized linear model setting and for large n, the above approximation will still perform satisfactorily in most situations.

Raftery, Madigan and Volinsky (1996) simplify the approximation further noting that for large n, $\tilde{\theta}_k \approx \hat{\theta}_k$ where $\hat{\theta}_k$ are the maximum likelihood estimates of the model parameters and that $log(|\psi|) = -p_k log(n) + O(1)$. This now yields:

$$log(p(data \mid M_K)) \approx log(p(data \mid \hat{\theta}, M_k)) - p_k log(n) + O(1)$$

Another alternative to the Bayes' factor based approximation to compute the marginal likelihood integral is the maximum likelihood based approximation applied by Taplin (1993) to a time series problem. This consists in using:

$$p(\Delta \mid M_k, D) \approx p(\Delta \mid M_k, \hat{\theta}_k, data)$$

Raftery, Madigan and Volinsky (1996) give a brief heuristic argument based on the linear regression setting for why this approximation would perform reasonably well and applied to the BMA context with promising results.

5.3 Setting prior probabilities on models

In the Bayesian model averaging formula to compute our quantity of interest we need $P(M_i)$ for all model M_i. These are probabilities that we must assign to all our models of interest before any data is taken into consideration. Since the number of models could be large, this could be a difficult endeavor without a well-defined and fast procedure. We present two such methods to efficiently generate prior probabilities.

The first works to specify prior for GLMs and is generally applicable when we have a parameter associated with each predictor. The method is suggested in Hoeting et al. (1999). For a model M_i, we specify it's prior probability as follows:

$$p(M_i) = \prod_{j=1}^{p} \pi_j^{\delta_{ij}} (1 - \pi_j)^{(1-\delta_{ij})}$$

With $\pi_j \in [0, 1]$ being the prior probability that $\beta_j \neq 0$ (the parameter associated with the effect of the jth predictor) and δ_{ij} is an indicator that predictor j is in M_i. So, every model with a certain predictor has its prior probability multiplied by the π_j and those without that predictor by $1 - \pi_j$. Note that this simple prior specification strategy assumes that the inclusion of each component in a model is independent from the inclusion or exclusion of others which might not be a realistic assumption.

Selecting $\pi_j = 1/2$ corresponds to the uniform or uninformative prior. Hoeting et al. note that, in their experience, assigning uniform prior probabilities to all models has performed well and yield results that were robust to moderate departures from prior uniformity.

An alternative method to elicit prior distribution from an expert is given in Madigan, Gavrin and Raftery (1995) for the discrete data setting. This method is more involved, but does not assume the independence between component inclusion like the previous method does. The authors create a program to randomly select a variable and its state and then requires the domain expert to fill in the values of the other variables. In completing this exercise, the expert

generated "imaginary data". Then, starting with a uniform prior, and incorporating the imaginary data, we can compute the posterior models probabilities and use these as our updated priors. This methodology provided informative priors that yielded improvements in predictive performance over uniform priors for their particular choice of application. This suggests that although experts can be prone to bias and inconsistencies, this prior information elicitation method can be, at least in some circumstances, useful.

6 Example: Bayesian Model Averaging in GLM context

We provide an example where Bayesian model averaging provides useful insights additional to the traditional techniques in the generalized linear model setting.

6.1 Data: Baseball salary v. Player performance statistics

Table 5 shows a data set with 13 baseball performance score criteria from 336 players.

Variable name	Description	Data type
avg_bat	Batting average	Continuous
OBP	On-base percentage	Continuous
runs	Number of runs	Continuous
hits	Number of hits	Continuous
doubles	Number of doubles	Continuous
triples	Number of triples	Continuous
homeruns	Number of home runs	Continuous
RBI	Number of runs batted in	Continuous
walks	Number of walks	Continuous
s_outs	Number of strike-outs	Continuous
stolen	Number of stolen bases	Continuous
errors	Number of errors	Continuous
FA	Free agency eligibility	Indicator

Table 5: Predictor description

The response variable is Y where $Y_{ij} \sim$ Binomial(π_{ij}), Y_{ij}'s are independent where π_{ij} = probability of a player i in group j earning more than 1 million dollars. Y_{ij} = 1 if the salary of an individual i is greater than 1 million dollars, 0 otherwise. i=1,...,336, j= 1,2 (Free agent eligibility no/yes)

6.2 Standard GLM

We fitted two standard generalized linear models with different link functions: one with logit, and the other with probit.

$$\text{logit: } \log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_{13} x_{13i}$$

$$\text{probit: } \Phi^{-1}(\pi_{ij}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_{13} x_{13i}$$

where ϕ indicates standard normal density with mean zero and variance 1. Table 6 shows the resulting models from each GLM with corresponding standard errors and p-values.

Logit link GLM

	Estimate	SE	z	p-value
(Intercept)	-0.596	1.717	-0.347	0.728
avg_bat	-1.820	17.927	-0.102	0.919
OBP	-12.408	14.592	-0.850	0.395
runs	0.019	0.025	0.771	0.441
hits	0.017	0.016	1.072	0.284
doubles	0.019	0.037	0.509	0.611
triples	0.044	0.098	0.449	0.654
homeruns	0.073	0.055	1.345	0.179
RBI	0.020	0.022	0.938	0.348
walks	0.011	0.023	0.469	0.639
s_outs	-0.031	0.010	-3.149	0.002
stolen	0.022	0.020	1.126	0.260
errors	-0.003	0.032	-0.090	0.928
fa1	2.87	0.40	7.12	0.00

Probit link GLM

	Estimate	SE	z	p-value
(Intercept)	-0.294	0.940	-0.313	0.754
avg_bat	-1.330	9.937	-0.134	0.894
OBP	-6.845	8.154	-0.839	0.401
runs	0.010	0.014	0.715	0.475
hits	0.010	0.009	1.104	0.270
doubles	0.013	0.021	0.629	0.529
triples	0.034	0.055	0.618	0.537
homeruns	0.047	0.031	1.518	0.129
RBI	0.010	0.012	0.808	0.419
walks	0.007	0.013	0.506	0.613
s_outs	-0.018	0.006	-3.290	0.001
stolen	0.012	0.011	1.083	0.279
errors	0.000	0.018	0.015	0.988
fa1	1.62	0.22	7.56	0.00

Table 6: Modelss from standard GLM with logit and probit links

In the standard GLMs, only the FA eligibility and the number of strike outs turned out to be meaningful predictors with p-values lower than the significance level $\alpha = 0.05$. For the rest of this section, we only discuss the results from the logit link case for the standard GLM method since the results from the two links showed negligible differences.

6.3 Variable Selection

The results in table 6 indicate the possible presence of multicollinearity between the predictors inflating the p-values even though some predictors might be important. To extract a reduced model by eliminating possible multi-correlated predictors, we performed variable selection with Bayesian Information Criterion (BIC) in conjunction with backward elimination. Table 7 showed that our model could be reduced to have 5 significant predictors instead of 2. The BIC of the full model was 304.45, and 262.37 for the reduced model.

	Estimate	SE	z	p-value
(Intercept)	-0.969	1.406	-0.689	0.491
OBP	-12.264	4.921	-2.492	0.013
runs	0.058	0.012	4.665	0.000
RBI	0.042	0.012	3.562	0.000
s_outs	-0.025	0.008	-3.214	0.001
fa1	2.88	0.38	7.61	0.00

Table 7: Reduced model from backward elimination by BIC

6.3.1 Words on Bayesian Information Criterion

As briefly mentioned in Section 3.2, the definition of BIC is expressed as:

$$BIC_k = -2\ln(\hat{likelihood_k}) + (p_k + 1)\ln(n)$$

where k= k-th model, p= number of predictors. This can further be rearranged as:

$$BIC_k = n\ln(1 - R_k^2) + (p_k + 1)\ln(n)$$

In most cases, a model with smaller BIC is preferred. The first term of BIC decreases as the model k's R^2 increases, and R^2 tends to rise as we add more predictors into the model. However, the second term of BIC escalates BIC hence penalizing the model complexity. Therefore, BIC allows us to obtain models with relatively high R^2 with low complexity. In the R environment, the 'step()' function with k=log(number of entries) and direction="backward" argument performs backward elimination by BIC where k is the number of degrees of freedom used for the penalty.

6.4 GLM with BMA

6.4.1 Models of interests

As in the standard GLM, we performed BMA GLM for two different link functions: "logit" and "probit":

$$\text{Logit: } log\frac{\pi_{ij}}{1-\pi_{ij}} = \sum_{k=0}^{2^p}[(\beta_{0k} + \beta_{1k}x_{1i} + \beta_{2k}x_{2i} + ... + \beta_{13k}x_{13i}) \times pr(M_k|\text{data})]$$

$$\text{Probit: } \Phi^{-1}(\pi_{ij}) = \sum_{k=0}^{2^p}[(\beta_{0k} + \beta_{1k}x_{1i} + \beta_{2k}x_{2i} + ... + \beta_{13k}x_{13i}) \times pr(M_k|\text{data})]$$

where 2^p indicates the total number of possible combinations of p predictors, and $pr(M_k \mid data)$ is the posterior probability of model(M) k given data. We used uniform prior for the model probabilities under the assumption that we have no prior information on them. In other words, we assumed that $pr(M_k) = \frac{1}{2^p}$ for all k. The equations above indicate that our resulting estimates π_{ij} would be proportional to the weighted sum of the estimates from every model possible, with $pr(M_k \mid data)$ being the weights.

6.4.2 Fitted models

Table 8 shows the top 10 fitted binomial BMA GLMs with logit link ordered by posterior model probabilities (PMP). The combined posterior model probabilities of the top 10 models was 0.6. In the BMA context, each coefficient has a posterior mean, standard deviation, and posterior inclusion probability (PIP) expressed as P(B!=0|D) in percentage, instead of a p-value in the standard GLM. It is also noticeable that each model has a different set of predictors hence different coefficient values. In our model, FA eligibility has posterior inclusion probability 1, suggesting that it is very likely that FA eligibility should be included in the model. Also, the predictors *hits* and *st_outs* have relatively high posterior inclusion probabilities 0.651 and 0.577, respectively. However, *errors* and *doubles* have small PIPs compared to other predictors, indicating that it is not likely that those should be included in the true model.

	P(B!=0|D)	Post Mean	SD	model.1	model.2	model.3	model.4	model.5	model.6	model.7	model.8	model.9	model.10
Intercept	100.00	-3.497	1.947	-5.019	-0.969	-4.343	-4.512	-5.021	-0.993	-4.990	-4.983	-4.688	-4.658
avg_bat	13.00	-1.730	5.422	.	.	.	.	.	-16.120	.	.	.	.
OBP	26.30	-2.864	5.532	.	-12.260	.	.	.	.	.	.	.	.
runs	42.90	0.020	0.026	.	0.058	0.044	.	.	.	.	0.025	.	0.039
hits	65.10	0.020	0.017	0.034	.	.	0.034	0.025	0.044	0.030	0.022	0.024	.
doubles	2.60	0.001	0.008	.	.	.	.	.	.	.	.	.	.
triples	3.60	0.002	0.021	.	.	.	.	.	.	.	.	.	.
homeruns	26.70	0.022	0.042	.	.	.	0.088	.	0.096	0.042	.	.	.
RBI	47.90	0.018	0.021	.	0.042	0.040	.	0.021	.	.	.	0.034	0.025
walks	3.00	0.000	0.003	.	.	.	.	.	.	.	.	.	.
s_outs	57.70	-0.013	0.013	.	-0.025	-0.020	-0.020	.	-0.028	.	.	-0.015	.
stolen	8.60	0.002	0.009	.	.	.	.	.	.	.	.	.	.
errors	1.70	0.000	0.005	.	.	.	.	.	.	.	.	.	.
fa	100.00	2.777	0.371	2.773	2.876	2.729	2.703	2.751	2.685	2.724	2.760	2.766	2.684
nVar				2	5	4	4	3	5	3	3	4	3
BIC				-1,692	-1,692	-1,691	-1,691	-1,691	-1,691	-1,690	-1,689	-1,689	-1,689
PMP				0.119	0.112	0.075	0.060	0.056	0.051	0.048	0.028	0.027	0.024

Table 8: Binomial BMA GLM with logit link

6.4.3 Uncertainty plot

We assessed model uncertainty by first using the uncertainty plots. Table 9 shows the top 20 models ordered by their log posterior odds.

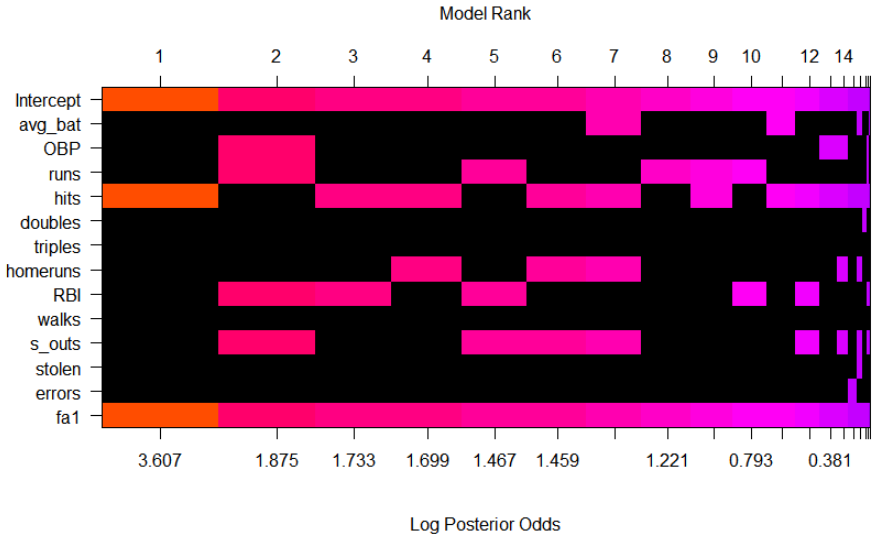

Table 9: Uncertainty plot from BMA GLM

The log posterior odds are computed by:

$$\text{Log Posterior Odds} = \ln(PO[M_m : M_0])$$

where $PO[M_m : M_0]$ indicates the posterior odds between M_m and M_0. The log posterior odds can further be expressed as:

$$\ln(BF[M_m : M_0] \times O[M_m : M_0])$$

where $BF[M_m : M_0]$ is the Bayes factor between M_m and M_0, and $O[M_m : M_0]$ is the prior odds between the two models. We discuss them in details in the following subsection.

The models are evaluated into a number of different clusters, suggesting that there is no evidence that only a small number of the nested models would be enough to explain the data with high probability. Also, we witnessed that the number of hits and the number of runs are rarely included in the same model. This is a strong indication that the two predictors are highly correlated with each other. In practice, runs in baseball were mostly scored as a result of successful hits by batters, so the high correlation seems obvious. Similar pattern was seen for *RBI* and *homeruns*.

6.4.4 Bayes Factors

We first go over the definition of prior odds and posterior odds. Prior odds refers to the ratio of the prior probability of a model to the prior probability of another:

$$\text{Prior odds: } O[M_1 : M_0] = \frac{pr(M_1)}{pr(M_0)}$$

where M_0 is the null model, which has only the intercept term as the predictor. Similarly, the posterior odds indicates the ratio of posterior probabilities of two models:

$$\text{Posterior odds: } PO[M_1 : M_0] = \frac{pr(M_1 \mid data)}{pr(M_0 \mid data)}$$

By Bayes' rule, the equation of the posterior odds can be expressed in terms of Bayes factor and the prior odds:

$$\begin{aligned} PO[M_1 : M_0] &= \frac{(pr(data \mid M_1) \times pr(M_1))/pr(data)}{(pr(data \mid M_0) \times pr(M_0))/pr(data)} \\ &= \frac{(pr(data \mid M_1) \times pr(M_1))}{(pr(data \mid M_0) \times pr(M_0))} \\ &= \frac{pr(data \mid M_1)}{pr(data \mid M_0)} \times \frac{pr(M_1)}{pr(M_0)} \\ &= BF[M_1 : M_0] \times O[M_1 : M_0] \end{aligned}$$

Therefore, the Bayes factor $BF[M_1 : M_0]$ for a model M_1 against another model M_0 given data can be defined as the ratio of the two marginal likelihoods under different models and is equivalent to the ratio of posterior odds and the prior odds of the two models:

$$BF[M_1 : M_0] = \frac{pr(data \mid M_1)}{pr(data \mid M_0)} = \frac{PO[M_1 : M_0]}{O[M_1 : M_0]}$$

In our uncertainty plot, log posterior odds is equivalent to the log of Bayes factors since we assumed a uniform prior for model probabilities, which gives us prior odds 1 for all models.
To interpret the meaning of the Bayes factors, one can use a scale proposed by Harold Jeffreys (1961) as in Table 10. If the Bayes factor is between 1 and 1/3, the evidence against M_0 is not worth bare mention, meaning that there are no significant differences between the two models. If the Bayes factor is between 0.03 and 0.01, the evidence is strong, suggesting that there is significant evidence that M_0 is different from M_1.

$\boldsymbol{BF[M_1 : M_0]}$	**Inverted BF**	**Evidence against M0**
1 to 3	1 to 0.33	No evidence
3 to 20	0.33 to 0.03	Positive
20 to 150	0.03 to 0.01	Strong
> 150	< 0.01	Very Strong

Table 10: Interpreting the Bayes factor using Jeffreys' scale (1961)

where 'inverted BF' refers to $1/BF[M_1 : M_0]$.
However, recently developed software packages commonly use a scale for interpreting Bayes factors proposed by Kass and Raftery (1995) which deals with the natural logarithm of the computed Bayes factors as described in Table 11

$\boldsymbol{2 * log(BF[M_1 : M_0])}$	**Evidence against M0**
0 to 2	No evidence
2 to 6	Positive
6 to 10	Strong
> 10	Very Strong

Table 11: Interpreting the Bayes factor using the log scale proposed by Kass and Raftery (1995)

Reporting in terms of the log scale can be helpful when the likelihoods are very small.

Relationship with the posterior model probability In several software packages, Bayes factors are used in computing posterior model probabilities.

	Post Mean	P(B!=0|D)	GLM coefs	p-values
Intercept	-3.497	100.00	-0.596	0.728
fa	2.777	100.00	2.87	0.00
hits	0.020	65.10	0.017	0.284
s_outs	-0.013	57.70	-0.031	0.002
RBI	0.018	47.90	0.020	0.348
runs	0.020	42.90	0.019	0.441
homeruns	0.022	26.70	0.073	0.179
OBP	-2.864	26.30	-12.408	0.395
avg_bat	-1.730	13.00	-1.820	0.919
stolen	0.002	8.60	0.022	0.260
triples	0.002	3.60	0.044	0.654
walks	0.000	3.00	0.011	0.639
doubles	0.001	2.60	0.019	0.611
errors	0.000	1.70	-0.003	0.928

Table 12: (Left) Posterior means and posterior inclusion probabilities from BMA GLM, (Right) Coefficients and p-values from standard full GLM

Suppose we have models $M_m,\ m = 0, 1, ..., 2^p$. The posterior probability of each model is given by:

$$\begin{aligned} pr(M_m|D) &= \frac{\text{marginal likelihood of } M_m \times pr(M_m)}{\sum_{j=0}^{2^p} \text{marginal likelihood of } M_j \times pr(M_j)} \\ &= \frac{pr(\text{D}|M_m)pr(M_m)}{\sum_{j=0}^{2^p} pr(D|M_j)pr(M_j)} \\ &= \frac{pr(D|M_m) \times pr(M_m)/(pr(D|M_b) \times pr(M_b))}{\sum_{j=1}^{2^p}[pr(D|M_j) \times pr(M_j)/(pr(D|M_b) \times pr(M_b))]} \\ &= \frac{BF[M_m : M_b] \times O[M_m : M_b]}{\sum_{j=1}^{2^p} BF[M_j : M_b] \times O[M_j : M_b]} \end{aligned} \tag{5}$$

Therefore, we sum up this section with the conclusion that Bayes factors play vital role in determining the model uncertainty as well as computing posterior model probabilities, which are the critical features of Bayesian model averaging.

6.5 Comparing BMA GLM to standard GLM

In this section, we compare the results from our BMA GLM to the standard GLM.

6.5.1 Coefficients: BMA GLM and standard full GLM

Table 12 shows the posterior means of the coefficients drawn from fitted BMA GLM with logit link with their posterior inclusion probabilities, and the coefficients and the corresponding p-values from the fitted standard full GLM with logit link. They are ordered by the posterior inclusion probabilities. It is noticeable that, altough the direction of the effects is identical, the magnitudes differ. Also, although each p-value can tell us whether the predictor's effect on the mean probability is significant or not, interpretation on the importance of each predictor in explaining the data with the posterior inclusion probability is likely to be more intuitive and simple to comprehend for most people than with the p-values. For example, *hits* has posterior inclusion probability 0.651, which can be interpreted as the probability that *hits* should be included in the true model is 65.10 percent. However, the predictor's p-value is 0.284, which suggests the effect of the predictor is not statistically significant under our significance level $\alpha = 0.05$, but the value itself cannot be interpreted in the same way as the PIP value and is arguably harder to comprehend for most non-professionals.
Some researchers may argue that averaging over different models is an example of an 'adding apples and oranges' problem, and hence trying to make interpretations on the BMA coefficients, and their inclusion probabilities seem inappropriate. However, referring to Hoeting et al(1999), BMA can be viewed as standard Bayesian inference for just

	Post Mean	P(B!=0|D)	Step coefs	p-values
Intercept	-3.497	100.00	-0.969	0.491
fa	2.777	100.00	2.88	0.00
hits	0.020	65.10		
s_outs	-0.013	57.70	-0.025	0.001
RBI	0.018	47.90	0.042	0.000
runs	0.020	42.90	0.058	0.000
homeruns	0.022	26.70		
OBP	-2.864	26.30	-12.264	0.013
avg_bat	-1.730	13.00		
stolen	0.002	8.60		
triples	0.002	3.60		
walks	0.000	3.00		
doubles	0.001	2.60		
errors	0.000	1.70		

Table 13: (Left) Posterior means and posterior inclusion probabilities from BMA GLM, (Right) Coefficients and p-values from standard reduced GLM

one model, the full model in which all variables are included. The twist is that the prior allows for the possibility that some of the coefficients might be equal to zero(or, essentially, equivalently close to zero). Once we recast the way we think of BMA in this way, in terms of just one model, the "apples and oranges" problem disappears.

6.5.2 Coefficients: BMA GLM and standard reduced GLM

Table 13 shows the comparison of the coefficients between the same BMA GLM and the standard reduced GLM. Similar conclusions can be made about the posterior inclusion probabilities and the p-values, as we discussed in the full model case. However, it was interesting to observe that the reduced model failed to capture some of the predictors with relatively higher inclusion probabilities. Notably, we suspected *hits* would likely be included, but it was not. We speculate that it was a result of the backward elimination. In backward elimination, once removal of a predictor reduces BIC, it never gets included back in the model. Consequently, we tested fitting another reduced model with the stepwise selection instead of backward elimination, but the resulting model was identical. Further investigation on the reason could be conducted in future research.

6.6 Predictive performance

Before presenting the results, we briefly discuss methods for assessing the success of various modeling strategies. A primary purpose of statistical analysis is to make forecasts (Dawid, 1984). Similarly, Bernado and Smith (1994, page 238) argue that when comparing different strategies, all other things being equal, we should select a model that consistently assigns higher probabilities to the events that truly occur. This means that measuring how well a model predicts observation is one way to scale the efficacy of the models.

6.6.1 Partial performance score (PPS)

One measure of predictive performance uses the logarithmic scoring rule of Good(1952). The predictive log score measures the predictive performance of an individual model(M) using the sum of the logarithms of the observed ordinates of the predictive density for each observation in the test set (Hoeting et al. 1999). That is, our predictive performance of BMA can be measured with:

$$-\sum_{d\in D^{test}} \ln[\sum_{M\in A} pr(d|M^{train}, D^{train})pr(M^{train}|D^{train})]$$

and we call the resulting score *partial predictive score* (PPS)
We introduce a process of measuring the predictive performance:

1. Randomly split the dataset into training data(D^{train}) and test data(D^{test})
2. Train model M^{train} using D^{train}
3. Estimate the responses on the subjects in D^{test} with M^{train}
4. Compute PPS = $-\sum_{d\in D^{test}} \ln[\sum_{M\in A} pr(d|M^{train}, D^{train})pr(M^{train}|D^{train})]$

Closer the estimations in 3. to the true response in the test data, larger the likelihood resulting in a small PPS. Therefore, models with smaller PPS are considered to be more efficient.

6.6.2 Results

To assess how BMA GLM performs compared to other GLMs, we fitted four different types of models:

- Standard full GLM: 13 predictors, BIC = 304.45
- Standard reduced GLM: Backward elimination using BIC, 5 predictors, BIC = 262.37
- BMA GLM with Occam's window: $O_R = 20, O_L = \frac{1}{20}$(Raftery, Madigan, and Volinsky, 1996), averaging over $2^p = 53$ models
- BMA GLM without Occam's window: Averaging over $2^p = 8,192$ models

All procedures were done equally for both binomial *logit* link and binomial *probit* link to assess the effect of the link functions.
In practice, BMA GLM without Occam's window did not average over the entire 8,192 models. The models after top 170 models with higher probabilities had posterior model probabilities almost 0 (most of them smaller than 10^{-172}), hence their contributions to estimations were negligible. The software package automatically removed them after fitting the models to reduce the computational intensity for estimations.

model	PPS	model	PPS
Full glm(logit)	63.279	Full glm(probit)	63.533
Stepwise glm(logit)	59.410	Stepwise glm(probit)	59.582
BMA OW(logit)	58.236	BMA OW(probit)	58.678
BMA No-OW(logit)	58.039	BMA No-OW(probit)	58.321

Table 14: PPS of different modelling strategies

Table 14 show the resulting partial predictive scores for the competing modeling strategies. Standard full GLM of both link functions had the highest PPS, and BMA GLMs without Occam's window had the lowest PPS. The difference in PPS can be viewed as an increase in predictive performance per event by a factor of $exp(\frac{\text{Difference in PPS}}{\text{Number of entries in } D^{test}})$ (Hoeting et al. 1999). For instance, the difference in PPS between standard full GLM(logit) and BMA GLM(logit) with no Occam's window is 5.24. This can be comprehended as BMA GLM(logit) with no Occam's window performed exp(5.24/168) = 1.032 or by about 3.2% more effectively than the standard full GLM(logit). With the same logic, we conclude that the BMA GLM(logit) performed up to 0.8% more effectively than the standard reduced GLM(logit). This improvement may seem very small, but there were cases where BMA strategy performed up to 10% better than the traditional strategies(Hoeting et al. 1999), which can be considered significant. We suggest finding out deterministically how much BMA strategy can outperform the standard methods could further be investigated in the future.

7 Conclusion

In summary, we have looked at a linear regression setting with real data on child cognitive scores and applied the standard model selection tools to pick a single best model for inference on it. In addition to having other problems,

we then highlighted how selecting a single model to use for that inference ignored model uncertainty. That is to say, once a model is selected, in the standard procedures, we discard all other models and all the information that they might provide. To resolve the issue, we introduce Bayesian model averaging and display its results on the same linear regression example to illustrate its ability to take into consideration inputs from more than one model. These results were then given a new interpretation within the framework of Bayesian model averaging that is different from the standard one. In the next section, we've presented the solutions that are used for implementing BMA with a particular emphasis on those useful in the generalized linear model context and on the ones used in the R 'BMA' package. Since it is this package that we used in the final section where we go through another real data set example. This was a situation where generalized linear models were our models of choice. We compared different link functions, different choices of tunable parameters in implementation of BMA to single best models chosen by standard model selection methods. For the standard methods, we illustrated model uncertainty throughout both our examples. We saw how, in the particular cases that were presented, using BMA provided improvements in predictive performance over a single model.

8 Extension and other areas of application

In this section, we will briefly look at other areas where Bayesian model averaging is being applied. Hinne 2015 highlights meta-analysis and network analysis as two such fields.

Meta-analysis is where we seek to pool the results of different studies measuring (or attempting to measure) the same effects. The agglomeration of studies aims at generating more robust conclusions then any of the individual studies sampled. In meta-analysis, we often encounter the dilemma of whether we should set the in-between study variance to 0 (meaning that they are all measures of the same underlying effect size) or if we have some between study variations. Bayesian model averaging allows us to side-step this dilemma by never having to chose one particular model, but instead to average over them.

In network analysis, we will attempt to model the interactions between large groups of entities represented as nodes. The number of possible models is determined by the total number of different link combinations and is typically large. The large number of models of interests makes it so that we rarely have a single one that has the majority of the posterior probability. Instead, there is an substantial amount of model uncertainty that remains and many models with most links in common will have similar posterior probabilities. Then, Bayesian model averaging is an interesting option to avoid having to select a single model.

9 Contributions

Zubia: Section 1 (Motivation), Section 2 (Bayesian Model Averaging), Section 3 (BMA in the linear regression setting), Section 4 (Benefits and Drawbacks of BMA).

Louis: Abstract, Section 5 (Implementation), Section 7 (Conclusion), Section 8 (Extension and other areas of application), References.

Sonny: Section 6 (Example: Bayesian Model Averaging in GLM context).

References

[1] Hoeting, Jennifer A., et al. Bayesian model averaging: a tutorial. In *Statistical Science*, pages 382–401, 1999.

[2] Draper, D. "Assessment and Propagation of Model Uncertainty" (Disc: P71-97). *Journal of the Royal Statistical Society, Series B, Methodological, 57, 45–70* (1995).

[3] Hodges, J. S. "Uncertainty, Policy Analysis and Statistics" (C/R: P276-291). *Statistical Science, 2, 259–275* (1987).

[4] Chatfield, C. "Model uncertainty, data mining and statistical inference." *Journal of the Royal Statistical Society A 158, 419–466* (1995).

[5] Efroymson, M. A. "Multiple Regression Analysis." *In Ralston, A. and Wilf, HS, editors, Mathematical Methods for Digital Computers. Wiley.* (1960).

[6] Raftery, A.E. " Bayesian model selection in structural equation models." *In Testing Stuctural Equation Models (K.A. Bollen and J.S. Long, eds.), Beverly Hills: Sage, pp. 163-180* (1993).

[7] Gelman, A., Hill, J. "Data analysis using regression and multilevel/hierarchical model." *Cambridge: Cambridge University Press.* (2007).

[8] Furnival, George M., and Robert W. Wilson. Regressions by leaps and bounds. *Technometrics* 16.4 (1974): 499-511.

[9] Max Hinne et al. A Conceptual Introduction to Bayesian Model Averaging (Preprint), 2014.

[10] Lawless, J. F., and Kishore Singhal. Efficient screening of nonnormal regression models. *Biometrics* (1978): 318-327.

[11] Madigan, David, Jonathan Gavrin, and Adrian E. Raftery. Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Communications in Statistics-Theory and Methods* 24.9 (1995): 2271-2292.

[12] Madigan, David and Raftery, Adrian E. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89.428. pages 1535-1546, 1994.

[13] Madigan, David and York, Jeremy. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique* (1995): 215-232.

[14] Raftery, Adrian E. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* 83.2 (1996): 251-266.

[15] Raftery, Adrian E., David Madigan, and Chris T. Volinsky. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian statistics* 5 (1996): 323-349.

[16] Raftery et al. Package 'BMA'. *R document*, 2020.

[17] Clyde et al. Package 'BAS'. *R document*, 2020.

[18] Taplin, Ross H. Robust likelihood calculation for time series. *Journal of the Royal Statistical Society: Series B (Methodological)* 55.4 (1993): 829-836.

[19] Tierney, Luke, and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association* 81.393 (1986): 82-86.

[20] Volinsky, Chris T., et al. Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46.4 (1997): 433-448.