Optimizing Prediction for Policy Analysis Using Bayesian Model Averaging: Applications to Large-Scale Educational Assessments

David Kaplan
Department of Educational Psychology

APPAM 2018, Washington, DC
The key feature of Bayesian inference is the quantification of uncertainty about statistical parameters and models.

This uncertainty is encoded in the so-called prior distribution on the parameters and models.

Our degree of uncertainty about parameters and models can be more or less informative – based on scientific judgement (or a lack thereof).
Within the Bayesian framework, parameters are not the only unknown elements.

The Bayesian framework recognizes that models themselves possess uncertainty insofar as a particular model is typically chosen based on prior knowledge of the problem at hand, and the variables that have been used in previously specified models.

This form of uncertainty often goes unnoticed.

One popular approach to addressing the problem of model uncertainty lies in the method of *Bayesian model averaging* (BMA).
Early work by Leamer (1978) laid the foundation for Bayesian model averaging.

Fundamental theoretical work on Bayesian model averaging was conducted by Madigan and his colleagues (Madigan & Raftery, 1994; Kass & Raftery, 1995; Raftery, 1997; Hoeting et al. 1999).

Bayesian model averaging has been applied to a wide variety content social science domains, with major developments in weather forecasting: (e.g. Sloughter, Gneiting, & Raftery, 2013.)

Bayesian model averaging has been implemented in the R software programs “BMA" and “BMS".
Method of BMA

- Let $\Upsilon$, represent a quantity of interest that we wish to predict.

- Next, consider a set of competing models $M_k$, $k = 1, 2, \ldots, K$ that are not necessarily nested.

- The posterior distribution of $\Upsilon$ given data $y$ can be written as

$$p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|y).$$

where $p(M_k|y)$ is the posterior probability of model $M_k$ written as

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}, \quad l \neq k.$$
The interesting feature of equation (2) is that \( p(M_k | y) \) will likely be different for different models.

The term \( p(y | M_k) \) can be expressed as an integrated likelihood

\[
p(y | M_k) = \int p(y | \theta_k, M_k) p(\theta_k | M_k) d\theta_k, \tag{3}
\]

where \( p(\theta_k | M_k) \) is the prior distribution of \( \theta_k \) under model \( M_k \) (Raftery et al., 1997).

Thus, BMA provides an approach for combining models specified by researchers.

The advantage of BMA has been discussed in Madigan and Raftery (1994), who showed that BMA provides better predictive performance than that of a single model based on a proper scoring rules.
**Method of BMA**

- BMA is difficult to implement.
  1. The number of terms in \( p(\gamma | y) = \sum_{k=1}^{K} p(\gamma | M_k) p(M_k | y) \) can be quite large and the corresponding integrals are hard to compute.
  2. Eliciting \( p(M_k) \) may not be straightforward. The uniform prior \( 1/M \) is often used.
  3. Choosing the class of models to average over is also challenging.

- The problem of reducing the overall number of models that one could incorporate in the summation has led to a solution based on the notion of *Occam’s window* (Madigan and Raftery, 1994).
In the “BMA" program a *unit information prior* (Kass & Raftery, 1995; Raftery, 1998) is placed on model parameters.

- A weakly informative (data-based) prior that is diffused over the region of the likelihood where parameter values are considered mostly plausible, but not overly spread out.

- Can be considered the prior of an individual with unbiased but weak prior information (Hoff, 2009)

- Priors on the model space are equivalent for all models: $\frac{1}{M}$.

- BMA addresses the uncertainty of the model and its parameters.

- Informative priors would be hard to elicit. Other default non-informative parameter priors need to be examined.
A key characteristic of statistics is to develop accurate predictive models (Dawid, 1984).

All other things being equal, a given model is to be preferred over other competing models if it provides better predictions of what actually occurred.

We need to decide on rules for gauging predictive accuracy – often termed **scoring rules**.

Scoring rules provide a measure of the accuracy of probabilistic forecasts.

A forecast can be said to be “well-calibrated" if the assigned probabilities of the outcome match the actual proportion of times that the outcome occurred.
Two examples of scoring rules

- The K-L divergence between the two distributions can be written as

\[
I(f, g) = \int f(x) \log \left( \frac{f(x)}{g(x|\theta)} \right) dx
\]  

(4)

where \( I(f, g) \) is the "information lost when \( g \) is used to approximate \( f \).

- The K-L statistic is related to the logarithmic scoring rule (Dawid & Musio, 2015).
For studies that examine the prediction of a dichotomous outcome we will use the Brier score defined as

$$Brier = \frac{1}{T} \sum_{t=1}^{T} (f_t - o_t)^2,$$

Brier scores are so-called *proper scoring rules*.

In both examples, the theory of BMA states that using model averaged coefficients will attain better frequentist prediction performance than any other model based on the log-score rule (Good, 1952) (and other proper scoring rules).
An Important Assumption

- It is assumed that the true model, say, $M_T$ is one of the models in the set of models $M_k$, $k = 1, 2, \ldots, K$.

- This assumption is referred to as the $M$–closed framework.

- The $M$–closed framework can be contrasted with the $M$–completed framework and the $M$–open framework.

- In the $M$–closed framework, it makes sense to assign prior probabilities that $M_T$ is in the space of models. This is the framework for this paper.

- The application of the indifference prior is the conventional default.
The initial model for Bayesian model averaging can be written for the $i^{th}$ student ($i = 1, 2, \ldots, N$) as

$$\text{READING}_i = \beta_0 + \beta_1 (\text{GENDER}_i) + \beta_2 (\text{NATIVE}_i) + \beta_3 (\text{SLANG}_i) + \beta_4 (\text{ESCS}_i) + \beta_5 (\text{JOYREAD}_i) + \beta_6 (\text{DIVREAD}_i) + \beta_7 (\text{MEMOR}_i) + \beta_8 (\text{ELAB}_i) + \beta_9 (\text{CSTRAT}_i) + \epsilon_i,$$

The sample comes from approximately PISA (2009)-eligible students in the United States ($N \sim 5000$). Reading proficiency PV (READING), GENDER (male=0, female=1), immigrant status (NATIVE), language that the students use (SLANG: coded 1 if the test language is the same as language at home, 0 otherwise), economic, social and cultural status of the students (ESCS), enjoyment of reading (JOYREAD) and diversity in reading (DIVREAD), memorization strategies (MEMOR), elaboration strategies (ELAB), and control strategies (CSTRAT).
Table 1: Bayesian model averaging results for full multiple regression model

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Post Prob (β ≠ 0)</th>
<th>Avg coef</th>
<th>SD</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTERCEPT</td>
<td>1.00</td>
<td>493.63</td>
<td>2.11</td>
<td>494.86</td>
<td>491.67</td>
<td>492.77</td>
<td>496.19</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.42</td>
<td>2.72</td>
<td>3.54</td>
<td>.</td>
<td>6.46</td>
<td>6.84</td>
<td>.</td>
</tr>
<tr>
<td>NATIVE</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>SLANG</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>ESCS</td>
<td>1.00</td>
<td>30.19</td>
<td>1.24</td>
<td>30.10</td>
<td>30.36</td>
<td>30.18</td>
<td>29.90</td>
</tr>
<tr>
<td>JOYREAD</td>
<td>1.00</td>
<td>29.40</td>
<td>1.40</td>
<td>29.97</td>
<td>28.93</td>
<td>27.31</td>
<td>28.35</td>
</tr>
<tr>
<td>DIVREAD</td>
<td>0.92</td>
<td>-4.01</td>
<td>1.68</td>
<td>-4.44</td>
<td>-4.28</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>MEMOR</td>
<td>1.00</td>
<td>-18.61</td>
<td>1.31</td>
<td>-18.47</td>
<td>-18.76</td>
<td>-18.99</td>
<td>-18.70</td>
</tr>
<tr>
<td>CSTRAT</td>
<td>1.00</td>
<td>27.53</td>
<td>1.46</td>
<td>27.62</td>
<td>27.43</td>
<td>27.27</td>
<td>27.45</td>
</tr>
<tr>
<td>$R^2$</td>
<td></td>
<td></td>
<td></td>
<td>0.340</td>
<td>0.341</td>
<td>0.339</td>
<td>0.338</td>
</tr>
<tr>
<td>PMP</td>
<td>0.54</td>
<td>0.37</td>
<td>0.05</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Extensions of BMA

1. BMA for Structural Equation Modeling

2. BMA for Propensity Score Analysis
   - Kaplan & Chen (2014), *MBR*
We argue that so-called “Bayesianly” proper approaches to multiple imputation, although correctly accounting for uncertainty in imputation model parameters, ignores the uncertainty in the imputation model itself.

We address imputation model uncertainty by implementing Bayesian model averaging as part of the imputation process.

We implement Bayesian model averaging for multiple imputation under the chained equations approach to Bayesian normal theory-based multiple imputation.

Our results show the extent of imputation model uncertainty and a modest benefit of conducting multiple imputation using BMA based on the Kullback-Liebler divergence measure.
We have developed a two-step approach to BMA applied to growth curve modeling.

1. Estimate a growth curve model with time-invariant predictors and obtain growth estimates (intercept, slope) for each country.

2. Regress growth outcomes on predictors using Bayesian model-averaged coefficients.

Application to country-level gender differences in math using data from TIMSS.

Preliminary results show better prediction and forecasting using BMA compared to using conventional GCM with time-invariant predictors.
Future Directions

1. Comparisons to frequentist approaches to model averaging (Yavuz & Kaplan, in progress)
2. Handling complex sampling designs
3. Examining the consequences of the $M$-closed framework assumption.
4. Assessing the impact of the default priors.
Our published research papers, code, and data can be found at

http://bise.wceruw.org/index.html
Portions of the research reported in this paper were supported by the Institute of Education Sciences, U.S. Department of Education (2011–2014), through Grant R305D110001 to the University of Wisconsin - Madison. The opinions expressed are those of the authors and do not necessarily represent views of the Institute or the U.S. Department of Education.

Additional support for the work by Kaplan and Stăncel-Piatak was made possible by the International Association for the Evaluation of Educational Achievement (IEA).

I would like to thank my following co-authors

- Cassie Chen (ETS)
- Danielle Lee (National Council of State Boards of Nursing)
- Agnes Stăncel-Piatak (IEA, Hamburg)
- Sinan Yavus (PhD student, UW-Madison)
Thank You