On the Utility of Bayesian Model Averaging for Optimizing Prediction: Two Case Studies

David Kaplan

Department of Educational Psychology

THE UNIVERSITY of WISCONSIN MADISON

Bayesian Methods for Social Policy Research and Evaluation, Washington, DC 2017

- This work was conducted in collaboration with Danielle Lee.

Introduction

- Of critical importance to education policy is the monitoring of trends in education over time.
- The United Nations Sustainable Development Goals identified Goal 4 as focusing on quality education for all.
- Many of the stated targets under Goal 4 focus on reducing the gender gap in quality education.
- Goal 4.6 focuses on achieving literacy and numeracy for men and women.
- Developing optimal predictive models allows researchers and policy makers to assess cross-country progress and forecasts toward that goal.

- Our interest lies in developing LSAs for optimal prediction and forecasting in education policy contexts.
- As with political forecasting and weather forecasting, we need to account for model uncertainty.

- Our interest lies in developing LSAs for optimal prediction and forecasting in education policy contexts.

- As with political forecasting and weather forecasting, we need to account for model uncertainty.

> *"Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are."(Hoeting, et al 1999, pg. 382)*

- Our interest lies in developing LSAs for optimal prediction and forecasting in education policy contexts.
- As with political forecasting and weather forecasting, we need to account for model uncertainty.

> *"Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are."(Hoeting, et al 1999, pg. 382)*

- *Bayesian model averaging* (BMA) is a framework for modeling and estimation that addresses the problem of model uncertainty
- Madigan and Raftery (1994) show that BMA provides better predictive performance than that of any single model based on a log-score rule.

History of BMA Research

- 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; Raftery, 1997; Hoeting et al. 1999; Clyde,1999).
- Draper (1995) has discussed how model uncertainty can arise even in the context of experimental designs.
- Kass and Raftery (1995) provide a review of Bayesian model averaging and the costs of ignoring model uncertainty.
- A more recent review of the general problem of model uncertainty can be found in Clyde & George (2004) .
- Bayesian model averaging has been implemented in the R software programs “BMA” and “BMS”.

- Bayesian model averaging has been applied to a wide variety content domains.
  - Economics (Fernandez, Ley, & Steele, 2001)
  - Bioinformatics of gene express (Yeung, Bumgarner, Raftery, 2005)
  - Weather forecasting (Sloughter, Gneiting, & Raftery, 2013)
- Within education BMA has been applied causal inference within propensity score analysis (Kaplan & Chen, 2014).
- I will discuss an extension of Bayesian model averaging to structural equation modeling with applications to education can be found in Kaplan & Lee (2015).

Bayesian model averaging

- Let Υ be a predicted observation and M_k, $k = 1, 2, \ldots, K$ a set of competing models that are not necessarily nested.

- The posterior distribution of Υ given data y can be written as

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

meaning that the predicted value given the data is a weighted sum of the predicted value under each model multiplied by a measure of the quality of each model.

- The posterior model probability is obtained as

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}, \qquad l \neq k. \tag{2}$$

and will be different for different models. Finally,

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

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

- BMA provides an approach for combining models specified by researchers.
- The number of terms in $p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|y)$ can be quite large and the corresponding integrals are hard to compute.
    - Solutions to reducing the size of the model space are based on search algorithms such as *Occam's window* and MC^3 (Madigan and Raftery, 1994).
- It is common to use the constant prior $1/M$ across the model space and a non-informative prior for model parameters.

Scoring Rules

- 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 must decide on rules for gauging predictive accuracy – referred to as *scoring rules* (Gneiting & Raftery, 2007).
- Scoring rules provide a measure of the accuracy of probabilistic forecasts, and 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.

- We will evaluate predictive performance using the 90% predictive coverage criterion and the log of the percent predictive coverage for continuous outcomes, referred to as the *log-score*.

- For the prediction of a dichotomous outcome we can use the Brier score (Brier, 1950) defined as

$$Brier = \frac{1}{T}\sum_{t=1}^{T}(f_t - o_t)^2, \tag{4}$$

- The log-score and the Brier score are so-called *proper scoring rules* insofar as the score is maximized (or minimized in the case of the Brier score) when the reported forecast probability is the same as true probability.

Case Study 1: BMA in Regression Analysis

- The sample comes from approximately PISA (2009)-eligible students in the United States ($N \sim 5000$).
- Background variables: gender (gender), immigrant status (native), language (1 = test language is the same as language at home, 0 otherwise) and economic, social, and cultural status of the student (escs).
- Student reading attitudes: enjoyment of reading reading (joyread), diversity in reading (divread), memorization strategies (memor), elaboration strategies (elab), control strategies (cstrat).
- The outcome was the first plausible value of the PISA 2009 reading assessment.

- To gauge predictive performance we follow Hoeting, et al., (1999).

  1. Randomly divide the data set into "model averaging" data and "prediction" data.
  2. Fit a single frequentist model, Bayesian model, and BMA to the model averaging data.
  3. Predict the final dependent variable for the prediction data with the results from the model averaging data.
  4. Compare their predictive performance based on a 90% predictive coverage interval and the log-score.

- The initial model for Bayesian model averaging can be written for the i^{th} student $(i = 1, 2, \ldots, N)$ as

$$\begin{aligned} READING_i = {} & \beta_0 + \beta_1(GENDER_i) + \beta_2(NATIVE_i) + \\ & \beta_3(SLANG_i) + \beta_4(ESCS_i) + \beta_5(JOYREAD_i) + \\ & \beta_6(DIVREAD_i) + \beta_7(MEMOR_i) + \\ & \beta_8(ELAB_i) + \beta_9(CSTRAT_i) + \epsilon_i, \end{aligned} \tag{5}$$

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).

Bayesian Model Averaging

Table 1: Bayesian model averaging results for full multiple regression model

Predictor	Post Prob	Avg coef	SD	Model 1	Model 2	Model 3	Model 4
Full Model							
INTERCEPT	1.00	493.63	2.11	494.86	491.67	492.77	496.19
GENDER	0.42	2.72	3.54	.	6.46	6.84	.
NATIVE	0.00	0.00	0.00	.	.	.	.
SLANG	0.00	0.00	0.00	.	.	.	.
ESCS	1.00	30.19	1.24	30.10	30.36	30.18	29.90
JOYREAD	1.00	29.40	1.40	29.97	28.93	27.31	28.35
DIVREAD	0.92	-4.01	1.68	-4.44	-4.28	.	.
MEMOR	1.00	-18.61	1.31	-18.47	-18.76	-18.99	-18.70
ELAB	1.00	-15.24	1.26	-15.37	-14.95	-15.43	-15.90
CSTRAT	1.00	27.53	1.46	27.62	27.43	27.27	27.45
R^2				0.340	0.341	0.339	0.338
BIC				-1993.72	-1992.98	-1988.72	-1988.51
PMP				0.54	0.37	0.05	0.04

Table 2: Comparison of the predictive performance

Regression model	Percent of predictive coverage	Log score of predictive coverage
Bayesian model averaging	95.69%	-0.04
Best model from Bayesian model averaging	90.50%	-0.10
Bayesian model	90.50%	-0.10
Frequentist model	90.45%	-0.10

Case Study 2: BMA for SEM (Kaplan & Lee, 2015, *SEM*)

- The general steps of our BMA-SEM method are as follows
  1. Specify an initial model of interest recognizing that this may not be the model that generated the data.
  2. Reduce the model space using the up and down algorithm by treating a path diagram as a DAG.
  3. Obtain the weighted average of structural parameters over each model, weighted by the posterior model probabilities.
  4. Compare predictive performance of the BMA-SEM to the initially specified SEM by computing the reduced form of the models and calculating the log-score or the predictive coverage.

- A detailed simulation study by Kaplan & Lee (2015) revealed
  1. No difference among methods when the true model is known. Scoring rules give same results.
  2. No influence of width of Occam's window on predictive coverage.
- Our method performs no worse than when the true model is known.
- Our simulation study establishes the groundwork for applying the method to real data.

Case Study 2

- For the case study, we use data from PISA 2009 to estimate a model relating reading proficiency to a set of background and reading strategy variables.
- The sample was collected from PISA-eligible students in the United States, and the sample size was 5,053.
- The sample was split into a model averaging set (N = 2,526) and a predictive testing set (N = 2,527).

Case Study 2

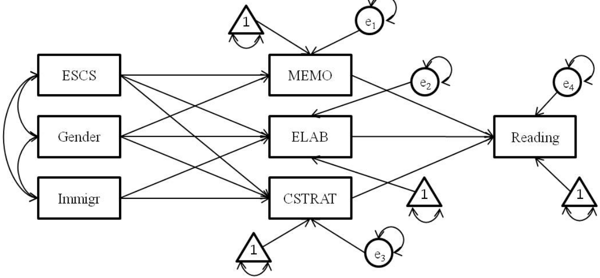

Figure 1: Initial path model.

Table 3: Selected models by BMA-SEM with the C = 100 for the PISA data

Parameter[a]	M_1	M_2	M_3
MEMO∼ESCS	●	●	
ELAB∼ESCS	●	●	●
CSTRAT∼ESCS	●	●	●
Reading∼ESCS	●	●	●
MEMO∼Gender	●	●	●
ELAB∼Gender	●	●	●
CSTRAT∼Gender	●	●	●
Reading∼Gender	●	●	●
MEMO∼Immigr	●	●	●
ELAB∼Immigr			
CSTRAT∼Immigr		●	
Reading∼Immigr			
ELAB∼MEMO	●	●	●
CSTRAT∼MEMO	●	●	●
Reading∼MEMO	●	●	●
CSTRAT∼ELAB	●	●	●
Reading∼ELAB	●	●	●
Reading∼CSTRAT	●	●	●
BIC	39461.68	39464.74	39465.15
PMP	0.72	0.15	0.13

Note. [a] ∼ refers to regression of left-hand variable onto right-hand variable; PMP = posterior model probability.

Table 4: Comparison of the result of Bayesian model averaging to the result of the Bayesian SEM.

	Bayesian model averaging			BSEM			
Predictor	mean$(\beta\|y)$	SD$(\beta\|y)$	P$(\beta \neq 0\|y)\%$	EAP	SD	95% PPI	
MEMOR∼ESCS	0.09	0.03	100.00	0.07	0.02	0.02	0.12
MEMOR∼GENDER	0.18	0.04	100.00	0.18	0.04	0.09	0.27
MEMOR∼NATIVE	-0.20	0.06	100.00	—			
ELAB∼ESCS	0.16	0.03	100.00	0.16	0.03	0.11	0.21
ELAB∼GENDER	0.00	0.00	0.00	-0.05	0.04	-0.14	0.04
ELAB∼NATIVE	-0.20	0.06	100.00	-0.20	0.06	-0.32	-0.09
CSTRAT∼ESCS	0.29	0.02	100.00	0.29	0.02	0.25	0.34
CSTRAT∼GENDER	0.25	0.04	100.00	0.25	0.04	0.18	0.33
CSTRAT∼NATIVE	-0.20	0.06	100.00	-0.20	0.05	-0.30	-0.09
JOYREAD∼ESCS	0.13	0.02	100.00	—			
JOYREAD∼GENDER	0.64	0.04	100.00	—			
JOYREAD∼NATIVE	0.00	0.00	0.00	—			
JOYREAD∼MEMOR	-0.11	0.02	100.00	-0.11	0.02	-0.16	-0.06
JOYREAD∼ELAB	0.08	0.02	100.00	0.04	0.02	-0.01	0.08
JOYREAD∼CSTRAT	0.28	0.02	100.00	0.36	0.02	0.31	0.41
READING∼JOYRREAD	0.34	0.02	100.00	0.34	0.02	0.31	0.38
MEMOR∼1	0.02	0.06	100.00	-0.14	0.03	-0.20	-0.08
ELAB∼1	0.04	0.05	100.00	0.06	0.06	-0.05	0.18
CSTRAT∼1	-0.07	0.06	100.00	-0.08	0.05	-0.17	0.02
JOYREAD∼1	-0.36	0.03	100.00	-0.02	0.02	-0.06	0.02
READING∼1	5.00	0.02	100.00	5.00	0.02	4.96	5.03
MEMOR∼∼MEMOR	1.19	0.03	100.00	1.20	0.02	1.14	1.27
ELAB∼∼ELAB	1.23	0.03	100.00	1.23	0.02	1.17	1.30
CSTRAT∼∼CSTRAT	1.18	0.03	100.00	0.98	0.02	0.95	1.01
JOYREAD∼∼JOYREAD	0.89	0.03	100.00	0.98	0.02	0.95	1.01
READING∼∼READING	0.76	0.02	100.00	0.98	0.02	0.95	1.01

Note. N = 2,490; ∼ refers to regression of left-hand variable onto right-hand variable; ∼1 refers to intercept; ∼∼ refers to variance;

EAP, expected a posteriori; SD, posterior standard deviation; PPI, posterior probability interval.

Table 5: Ninety percent coverage and log-score for PISA example

Method	90% Coverage	Log-score
BMA-SEM (4)	0.90	-0.11
BMA-SEM (20)	0.90	-0.11
BMA-SEM (100)	0.90	-0.11
FSEM	0.88	-0.13
BSEM	0.88	-0.13

Conclusions

- The question of using a model for some purpose beyond theory building suggests assessing the accuracy of a model's predictions.
- Thus, we are less concerned about the fit of a model and more concerned about finding a model that will predict well.
- Building optimally predictive models requires good model "calibration", i.e. predictions aligning with real world outcomes.

Conclusions (cont'd)

- Model selection and averaging is a very large topic in statistics.
- This problem has been addressed in both frequentist and Bayesian contexts.
    - AIC, BIC, DIC.
- There are also methods of frequentist model averaging based on Akaike weights.
- Newer methods based on the "lasso" (from statistics and machine learning) are also used to develop predictive models.
- Our current work is looking at direct comparisons along a common metric of scoring, with applications to large-scale educational assessments.

Conclusions (cont'd)

- Handling complex sampling designs
  - Essential for large-scale surveys
  - Not trivial
- Cross-System growth regressions
  - ILSAs are longitudinal at the country level.
  - Possible for ILSAs depending on the outcome of interest.

Conclusions (cont'd)

- Within the Bayesian world, BMA is known to yield a model that will perform better than any given sub-model on the criteria of predictive accuracy.
- The idea is that although not all models are equally good (as measured by their PMPs), all models do contain some important information.
- By combining models, while accounting for model uncertainty, we obtain a model with optimal predictive performance.

Thank you