Bayesian Stacking in Multilevel Models

.

.

.

.

Abstract

The issue of model uncertainty has been gaining interest in education and the social sciences community over the years, and the dominant methods for handling model uncertainty are based on Bayesian inference, and particularly, Bayesian model averaging. However, Bayesian model averaging assumes that the true data-generating model is within the candidate model space. Unlike Bayesian model averaging, the method of Bayesian stacking can account for model uncertainty without assuming that a true model exists. An issue with Bayesian stacking, however, is that it is an optimization technique that uses predictor-independent model weights and is, therefore, not fully Bayesian. Bayesian hierarchical stacking, proposed by Yao, Pirš, Vehtari, and Gelman (2021), further incorporates uncertainty by applying a hyperprior to the stacking weights. Considering the importance of multilevel models commonly applied in educational settings, this paper investigates the predictive performance of original Bayesian stacking and Bayesian hierarchical stacking along with two other readily available weighting methods, pseudo-BMA and pseudo-BMA bootstrap (PBMA and PBMA+) that are based on Bayesian model averaging weighting, via a simulation study and a real data example from PISA 2018. Predictive performance is measured by the Kullback-Leibler divergence scores. Although the differences in predictive performance among these four weighting methods in Bayesian stacking are small, we still find that Bayesian hierarchical stacking performs as well as PBMA and PBMA+ or even better in the $\mathcal{M}$-complete and $\mathcal{M}$-open setting.

Introduction

Issues of Model Uncertainty

Model selection and model uncertainty have been a general challenge in statistical inference for decades. The problem has been summarized by, among others, Hoeting, Madigan, Raftery, and Volinsky (1999) who wrote

> "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."(pg. 382)

Similar sentiments have been expressed earlier by Leamer (1978) and Draper et al. (1987). In education studies, it is common to encounter data with a hierarchical structure, for example, students nested within schools. To better capture the nesting effects. However, the majority of studies select a single best model using different modeling methods such as the ordinary least squares approach (OLS), maximum likelihood estimation, or even Bayesian methods. The issue with these approaches lies in the assumption that the selected model is the one that actually generated the data thus ignoring the typical practice of searching for a best-fitting model.

$\mathcal{M}$-frameworks

When considering the problem of model selection from a Bayesian perspective, there are three types of relationships between the true data generating model (DGM) and substantive models that need to be considered (see Bernardo & Smith, 2000): $\mathcal{M}$-closed, $\mathcal{M}$-complete, and $\mathcal{M}$-open.

- In the $\mathcal{M}$-closed setting, the true DGM is within the candidate model list. That is, the true DGM M_t is one of the $M_k \in \mathcal{M}$, where $k = 1, 2, \ldots, K$ denotes the number of models.
- In the $\mathcal{M}$-complete setting, the true DGM exists but is not within the model list. Rather, a list of models is considered, with each model serving as a reasonable proxy to some true DGM.
- In the $\mathcal{M}$-open setting, not only is the true DGM M_t not in $\mathcal{M}$ but also the explicit form $p(\tilde{y}|y) = p\left(\tilde{y}|M_t, y\right)$ cannot be specified.

Conventional modeling approaches such as LASSO might be able to obtain satisfactory predictive power within the $\mathcal{M}$-closed setting (e.g. Muthukrishnan & Rohini, 2016). However, in the $\mathcal{M}$-complete and $\mathcal{M}$-open setting, we cannot assume the true DGM is within the candidate model space. In a multilevel setting, for instance, students nested within different schools will lead to a larger degree of uncertainty compared to the fixed effects models. Therefore, instead of relying on a single best model, averaging the information across different models could potentially be a superior option. There are numerous methods to evaluate multiple candidate models as well as handling model uncertainty from a Bayesian perspective. For instance, Kass and Raftery (1995) proposed selecting the maximum a posterior (MAP) model; Geisser and Eddy (1979) suggested choosing a model based on AIC-type weighting. Through the joint efforts of many researchers (Leamer, 1978; Madigan & Raftery, 1994; Raftery, Madigan, & Hoeting, 1997; Hoeting et al., 1999; Clyde, 1999, 2003; Draper, 1995), *Bayesian model averaging* (BMA) was introduced to handle model uncertainty and obtain optimal predictive performance as measured by scoring rules such as the Kullback-Leibler divergence (KLD) (Kullback & Leibler, 1951; Kullback, 1959, 1987) or the log predictive score (Good, 1952; Bernardo & Smith, 2000).

Problems with Bayesian Model Averaging

Generally speaking, BMA averages coefficients across a large space of models weighted by each model's marginal posterior probability. That is, for a candidate model list $\mathcal{M} = (M_1, \ldots, M_K)$, the posterior probability of the quantity of interest θ (e.g. a predicted value, denoted as $\tilde{y}$) can be expressed as:

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

where $y_1, y_2, ..., y_n$ are the observed data. Each model is weighted by the posterior model probability:

$$p\left(M_k|y\right) = \frac{p\left(y|M_k\right) p\left(M_k\right)}{\sum_{k=1}^{K} p\left(y|M_k\right) p\left(M_k\right)}, \tag{2}$$

where the weights $\mathbf{w}$ are a constrained simplex: $\mathcal{S}_K = \left\{\mathbf{w} : \sum_{k=1}^{K} w_k = 1; w_k \in [0,1], \forall k\right\}$, where $\mathcal{S}_K$ is the set of weights where all the weights sum to one and are all between zero and one. The marginal likelihood of y given each model is

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

where θ_k are the parameters associated with model k.

A perusal of the extant literature shows applications of Bayesian model averaging primarily to economics (e.g. Fernández, Ley, & Steel, 2001), political science (e.g. Montgomery & Nyhan, 2010), bioinformatics of gene express (e.g. Yeung, Bumgarner, & Raftery, 2005), weather forecasting (e.g. Raftery, Gneiting, Balabdaoui, & Polakowski, 2005; Sloughter, Gneiting, & Raftery, 2013). Recent work by Kaplan and his colleagues (Kaplan & Lee, 2015; Kaplan & Yavuz, 2019; Kaplan & Chen, 2014; Kaplan & Lee, 2018; Kaplan & Huang, 2021; Kaplan, 2021), have discussed and extended Bayesian model averaging primarily to problems in large-scale educational assessments, such as the

Program for International Student Assessment (PISA) (OECD, 2002). However, that work did not address the multilevel nature of these assessments wherein students are sampled within schools.[1]

Although Bayesian model averaging has shown good out-of-sample predictive performance (Hoeting et al., 1999), the general problem of using BMA lies in its major assumptions: BMA assumes $\mathcal{M}$-closed setting. More specifically, in $\mathcal{M}$-open and $\mathcal{M}$-complete setting where the true DGM is not within the candidate model set, BMA still combines the information obtained from each candidate model for prediction. Another issue with BMA is its sensitivity to the choices of priors on the model parameters. Different $p\left(\theta_k|M_k\right)$ can lead to quite different results. Therefore, the accurate predictive performance of BMA requires the correct specification of the prior information.

Our Contributions

This paper examines the predictive performance of multilevel models under the $\mathcal{M}$-complete and $\mathcal{M}$-open framework using Bayesian stacking (BS). The focus of our paper is on original stacking and a newly developed approach based on Bayesian hierarchical stacking (BHS) weights proposed by Yao et al. (2021), to be described below. We also compare these approaches to two other readily available methods for producing stacking weights. Our contribution to the literature is three-fold. First, BHS is very new and to our knowledge has not been systematically compared to other forms of stacking. Second, this paper conducts the comparison of these different weighting schemes in Bayesian stacking in the three $\mathcal{M}$-frameworks: one empirical study to represent the $\mathcal{M}$-open setting and two simulation studies to construct the $\mathcal{M}$-closed and $\mathcal{M}$-complete setting. Third, stacking, in general, is not well-known in the educational literature, where it is common to implement multilevel models to study within and between school predictors of academic and

[1] We recognize that multilevel models can be considered Bayesian hierarchical models (see e.g. Gelman et al., 2014). For ease of discussion, in this paper we make a distinction between the generic idea of Bayesian hierarchical models, and the specific types of multilevel models found in educational applications (see e.g. Raudenbush & Bryk, 2002)

non-academic outcomes (Raudenbush & Bryk, 2002). Thus, questions such as improving the predictive performance of multilevel models should be of general interest to education and social science researchers.

The remainder of this paper is organized as follows. In the next section, we provide an overview of the stacking weights that will be used in this study. They include the original stacking weights proposed by Yao, Vehtari, Simpson, and Gelman (2018a) and the newly developed hierarchical stacking weights proposed by Yao et al. (2021) which constitute the focus of this paper. In addition, we include two other types of weights that have been proposed for stacking, including so-called *pseudo-BMA* (PBMA) and *pseudo-BMA bootstrapping* (PBMA+) weights. This is followed by an empirical study using United States data from the 2018 cycle of the Program on International Student Assessment (PISA) (OECD, 2018). This is then followed by the details of our simulation design investigating the predictive performance of these four types of stacking weights in the context of multilevel models. The paper closes with conclusions and directions for further research.

Types of Bayesian Stacking Weights

Stacking, as one of the ensemble methods, was first introduced in machine learning studies by Wolpert (1992). Later, Clyde and Iversen (2013) adapted it to the Bayesian framework. In this section, we briefly review the relevant background of different weighting methods in Bayesian stacking: original stacking weights, Bayesian hierarchical stacking, Pseudo-BMA, and Pseudo-BMA bootstrap.

Original Stacking Weights

Following a recent review by Kaplan (2021), stacking involves weighting the predictive distribution obtained from multiple candidate models based on different scoring rules such as KLD to reach the optimal prediction. In our example later, the outcome of interest will be students' scores on the reading literacy assessment from PISA data in 2018.

First, we enumerate all the candidate models which can be denoted as $f_k(x)$ with different covariates x.

$$y = f_k(x) + \epsilon \tag{4}$$

As such, the optimal predictions will be the weighted combinations of these candidate models. In particular, $\hat{f}_k$ is used to estimate f_k and $\tilde{y}$ is the predictive distribution based on y

$$\tilde{y} = \sum_{k=1}^{K} \hat{w}_k \hat{f}_k(x). \tag{5}$$

To minimize the loss function between the weighted combination of predictive distribution and the actual outcome distribution, $\hat{w}_K$ is then computed based on an optimization function

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} w_k \hat{f}_{k,-i}(x_i) \right)^2 \tag{6}$$

where $\hat{f}_{k,-i}(x_i)$ is an estimate of f_k based on $n-1$ observations, leaving the i^{th} observation out. In Bayesian stacking, leave-one-out cross validation (LOO-CV) is used to compute $\hat{f}_{k,-i}(x_i)$. Similar to q-fold CV which holds one fold out for validation data set, in LOO-CV, each observation serves as the validation set with the remaining $n-1$ observations serving as the training set. To be precise, the expected log point-wise predictive density (ELPD) can be derived as

$$\text{ELPD} = \sum_{i=1}^{n} \int p_t(\tilde{y}_i) \log\, p(\tilde{y}_i|y) d\tilde{y}_i,$$

where $p_t(\tilde{y}_i)$ represents true DGM process for the predicted values $\tilde{y}_i$. By leaving i_{th} data point out one at a time, the Bayesian LOO estimates will be

$$\text{ELPD}_{loo} = \sum_{i=1}^{n} \log p(y_i|y_{-i}), \quad \text{where} \quad p(y_i|y_{-i}) = \int p(y_i|\theta) p(\theta|y_{-i}) d\theta$$

LOO-CV can be implemented by the R software program loo (Vehtari, Gabry, Yao, & Gelman, 2019).[2]

Hierarchical Stacking Weights

To provide the flexibility of using predictor-dependent weights, within the Bayesian framework, Yao et al. (2021) proposed the method of *Bayesian hierarchical stacking*. More specifically, for our example using PISA data, the weights assigned to student i in school j for the model to predict the math scores should be different from the weights assigned to student i' in school j' for the prediction of math scores. Accounting for differences in the weights for each observation can explain their unique characteristics and thus lead to a more precise prediction for each observation. The main difference between original BS and BHS lies in the computation of stacking weights. Instead of using a specific weight simplex in BMA and BS, Yao et al. (2021) suggested using a weight function $\mathbf{w}(x) = (w_1(x), \ldots, w_k(x))$. Replacing the weight simplex which is used in BS with a weight function can turn a simplex space into a function space, and allow for more flexibility to capture the variability inherent in the data. In the first stage, Yao et al. (2021) fitted the individual model with a holdout dataset $\mathcal{D}'$, which has an identical distribution to that of the observation dataset $D = \{y_i, x_i\}_{i=1}^n$, then computed

$$p\left(\tilde{y}_i|x_i, M_k, \mathcal{D}'\right) = \int p\left(\tilde{y}_i|x_i, M_k, \theta_k\right) p\left(\theta_k|M_k, \mathcal{D}'\right) d\theta_k \tag{7}$$

In the second stage, x_i and y_i are plugged in to obtain the pointwise full likelihood:

$$p\left(\tilde{y}_i|w, \mathcal{D}', x_i\right) = \sum_{k=1}^{K} w_k\left(x_i\right) p\left(\tilde{y}_i|x_i, M_k, \mathcal{D}'\right) \tag{8}$$

---

[2] The *widely applicable information criterion* (WAIC) has also been advocated for model selection. Although the WAIC and LOO-CV are asymptotically equivalent (Watanabe, 2010), the implementation of LOO-CV in the loo package is more robust in finite samples with weak priors or influential observations (Vehtari, Gelman, & Gabry, 2017)

By integrating out the holdout data set $\mathcal{D}'$, the likelihood can be obtained through an approximation by the leave-one-out predictive density:

$$p\left(\tilde{y}_i|\mathbf{w}, x_i\right) := \mathbb{E}_{\mathcal{D}'}\left[p\left(\tilde{y}_i|\mathbf{w}, x_i, \mathcal{D}'\right)\right] \approx \sum_{k=1}^{K} w_k\left(x_i\right) p_{k,-i} \tag{9a}$$

After summing over $\tilde{y}_i$, the goal for BHS is to maximize the likelihood function:

$$\log(p(\mathcal{D}|\mathbf{w})) \approx \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} w_k\left(x_i\right) p_{k,-i}\right) \tag{10}$$

Yao et al. (2021) brings the concept of pooling into the stacking framework for the weight functions: namely, complete pooling, no pooling, and partial pooling methods. The first approach is *complete pooling* which is the same as the original stacking approach: using a weight simplex and each predictor has the same weight $w_k(x) = w_k$. Another approach is *no-pooling stacking* which makes separate optimization of the objective function in Equation (10) for each x_i independently. The last method is *partial pooling* stacking, which requires specifying an appropriate hierarchical prior $p^{\text{prior}}(\cdot)$ so that the posterior distribution of the stacking weights can be obtained by

$$\log p(\mathbf{w}(\cdot)|\mathcal{D}) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} w_k\left(x_i\right) p_{k,-i}\right) + \log p^{\text{prior}}\left(\mathbf{w}\right) + \text{ constant}, \quad \mathbf{w}(\cdot) \in \mathcal{S}_K \tag{11}$$

For discrete predictors with J categories, $x = 1, \ldots, J$, a softmax transformation is taken to convert the simplex matrix space S_K^J to an unconstrained space $\mathbb{R}^{J(K-1)}$,

$$w_{jk} = \frac{\exp\left(\alpha_{jk}\right)}{\sum_{k=1}^{K} \exp\left(\alpha_{jk}\right)}, 1 \leq k \leq K-1, 1 \leq j \leq J; \quad \alpha_{jK} = 0, 1 \leq j \leq J, \tag{12}$$

where $\alpha_{jk} \in \mathbb{R}$ is the log odds ratio of model k with reference to model M_K for category j. Yao et al. (2021) suggested using a normal hierarchical prior on α_{jk} since it can pool the unconstrained weights towards the mean $(\mu_1, \ldots, \mu_{K-1})$. And the unconstrained model

weights are conditional on priors of μ and σ:

$$\alpha_{jk}|\mu_k, \sigma_k \sim \mathcal{N}\left(\mu_k, \sigma_k\right), k = 1, \ldots, K-1, j = 1, \ldots, J$$

From the Bayesian perspective, to account for the uncertainty of all the parameters with the joint posterior distribution $p(\alpha, \mu, \sigma|\mathcal{D})$, a hyperprior to μ and σ can be assigned. For instance,

$$\mu_k \sim \mathcal{N}\left(\mu_0, \tau_\mu\right), \quad \sigma_k \sim \mathcal{N}^+\left(0, \tau_\sigma\right), \quad k = 1, \ldots, K-1,$$

where $\mathcal{N}^+$ is the half-normal distribution. In this sense, no pooling stacking describes the situation when the σ_k approaches infinity and every model has its own weight whereas the complete pooling method means the σ_k approaches zero, where every model has the same weights. The joint posterior density using partial pooling methods will be:

$$\begin{aligned} \log p(\alpha, \mu, \sigma|\mathcal{D}) = & \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} w_k\left(x_i\right) p_{k,-i}\right) \\ & + \sum_{k=1}^{K-1}\sum_{j=1}^{J} \log p^{\text{prior}}\ \left(\alpha_{jk}|\mu_k, \sigma_k\right) \sum_{k=1}^{K-1} \log p^{\text{\ hyper prior}}\ \left(\mu_k, \sigma_k\right) \end{aligned} \tag{13}$$

For continuous and hybrid predictors, the weights can be modeled additively to allow for more structure:

$$\begin{aligned} w_{1:K}(x) &= \text{softmax}\left(w^*_{1:K}(x)\right), \\ \text{where} \quad w^*_k(x) &= \mu_k + \sum_{m=1}^{M} \alpha_{mk} f_m(x), k \leq K-1, w^*_K(x) = 0, \end{aligned} \tag{14}$$

and where $\{f_m : \mathcal{X} \to \mathbb{R}\}$ are m distinct features, $w_k^*(x)$ is the combination of the prior mean μ_k, and the additive functions $\alpha_{mk} f_m(x)$.[3] The final joint posterior density will be:

$$\log p(\alpha, \mu, \sigma|\mathcal{D}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} w_k \left(x_i\right) p_{k,-i} \right) + \sum_{k=1}^{K-1} \sum_{j=1}^{J} \log p^{\text{prior}} \left(\alpha_{jk}|\mu_k, \sigma_k\right) \sum_{k=1}^{K-1} \log p^{\text{hyper prior}} \left(\mu_k, \sigma_k\right) \tag{15}$$

Yao et al. (2021) also discussed different choices of priors and recommended as a general rule using weakly informative priors such as using a half-normal prior on the model scale parameters rather than half-Cauchy or inverse-gamma priors because the latter two will lead to larger dispersion. However, researchers can choose different priors based on the purpose of their research and the structure of the data.

Pseudo-BMA Stacking Weights

In addition to BS and BHS, we will also examine the performance of stacking weights based on so-called *pseudo-BMA* (PBMA) proposed by Geisser and Eddy (1979, see also; Gelfand, 1996; Yao et al., 2018a). The basic idea behind PBMA is as follows. First, as discussed in Yao et al. (2021), LOO-CV has connections to other types of weights that can be used for stacking. For example, in the case of maximum likelihood estimation LOO-CV weights are asymptotically equivalent to Akaike information criterion (AIC) weights (Akaike, 1973) that are used in frequentist model averaging applications (Yao et al., 2018a, see also; Burnham & Anderson, 2002; Fletcher, 2018). As a method of model selection, earlier work by Geisser and Eddy (1979, see also; Gelfand, 1996) criticized the underpinnings of Bayes factors and suggested substituting the marginal likelihood of the k^{th} model, $p(y|M_k)$, used in the calculation of Bayes factors with Bayesian leave-one-out cross-validation predictive densities, defined as $\prod_{i=1}^{n} p(y_i|y_{-i}, M_k)$. Yao et al. (2018a) refer to AIC weighting using LOO-CV predictive densities as PBMA. To elaborate, the PBMA

[3] The softmax function converts a vector of real numbers into a vector of probabilities (Stan Development Team, 2021).

weighting is

$$w_k = \frac{\exp\left(\widehat{\text{elpd}}^k_{\text{loo}}\right)}{\sum_{k=1}^{K}\exp\left(\widehat{\text{elpd}}^k_{\text{loo}}\right)}$$

Considering the uncertainty that arises from the finite samples of future data, Vehtari and Lampinen (2002) computed and used the standard error for the point-wise ELPD to modify the weight function by lognormal approximation:

$$w_k = \frac{\exp\left(\widehat{\text{elpd}}^k_{\text{loo}} - \frac{1}{2}\,\text{se}\left(\widehat{\text{elpd}}^k_{\text{loo}}\right)\right)}{\sum_{k=1}^{K}\exp\left(\widehat{\text{elpd}}^k_{\text{loo}} - \frac{1}{2}\,\text{se}\left(\widehat{\text{elpd}}^k_{\text{loo}}\right)\right)}.$$

Pseudo-BMA+ Stacking Weights

The difficulty with PBMA weights is that they do not take into account uncertainty in the LOO estimation of the weights. To address this Yao, Vehtari, Simpson, and Gelman (2018b) proposed an approach that combines the Bayesian bootstrap (see Rubin, 1981) with the ELPD defined earlier. They refer to this approach as *pseudo-BMA+* (PBMA+). Following Yao et al. (2018b), the essential idea behind PBMA+ is that the posterior distribution of the realizations $z_i, (i = 1, \ldots, n)$, of a random variable Z has a Dirichlet(1,. . . ,1) distribution. That is, in Bayesian stacking, define for each model k,

$$z_i^k = \widehat{\text{ELPD}}^k_{loo,i}$$

Taking B bootstrap samples $(\pi_{1,b}, \ldots, \pi_{n,b})$, $b = 1, \ldots, B$ from $\overbrace{\text{Dirichlet}(1,\ldots,1)}^{n}$ allows us to calculate the weighted means as

$$\bar{z}_b^k = \sum_{i=1}^{n} \pi_{i,b} z_i^k$$

From here, a Bayesian bootstrap sample of the stacking weight for model k based on bootstrap samples of size B can be obtained as

$$w_{k,b} = \frac{\exp(n\bar{z}_b^k)}{\sum_{k=1}^{K} \exp(n\bar{z}_b^k)}, \qquad b = 1, \ldots, B$$

leading to the final PBMA+ weight for model k

$$w_k = \frac{1}{B} \sum_{b=1}^{B} w_{k,b}$$

Of importance to this paper, Yao et al. (2018b) showed that PBMA+ performs better than BMA and PBMA in $\mathcal{M}$-open settings, but not as well as stacking in terms of the log predictive densities. This paper adds to the existing literature by comparing original stacking and hierarchical stacking weights to PBMA and PBMA+ weights in the context of multilevel models applied to large-scale assessments.

Empirical Study

This section examines the predictive performance of BS, BHS, PBMA, and PBMA+ using data from PISA 2018 under $\mathcal{M}$-open setting. This is because we are neither able to know if the true DGM is within the candidate model sets nor are we assuming that each model in the ensemble is a "proxy" model for the true DGM. PISA is a triennial international survey that aims to evaluate education systems across the world (79 countries) and measures 15-year-olds' ability to use their cognitive outcomes such as reading, mathematics, and science knowledge and skills to meet real-life challenges. There are 4838 participants randomly selected for this study and we specify four models to be our candidate models using nineteen covariates (see Table 1 for details) and the first plausible value of the reading assessment as the dependent variable:

Our candidate models for the ensemble are as follows:

Table 1
PISA 2018 predictors of reading scores

Variable Name	Variable Label
FEMALE	Sex (1=Female)
ESCS	Index of economic, social and cultural status
METASUM	Meta-cognition: summarising
PERFEED	Perceived feedback
HOMEPOS	Home possessions
ADAPTIVE	Adaptive instruction
TEACHINT	Perceived teacher's interest
ICTRES	ICT resources
JOYREAD	Joy/Like reading
COMPETE	Competitiveness
WORKMAST	Work mastery
GFOFAIL	General fear of failure
SWBP	Subjective well-being: Positive affect
MASTGOAL	Mastery goal orientation
BELONG	Subjective well-being: Sense of belonging to school
SCREADCOMP	Perception of reading competence
SCREADDIFF	Perception of reading difficulty
PISADIFF	Perception of difficulty of the PISA test
PV1READ	First plausible value reading score

- Model 1 includes the demographic measures (FEMALE, ESCS, HOMEPOS, ICTRES) and a random intercept and a random slope for ICTRES nested within schools.
- Model 2 investigates the effects of attitudes and behaviors on reading scores (JOYREAD, PISADIFF, SCREADCOMP, SCREADDIFF) with a random intercept for the school effects.
- Model 3 consists of predictors about academic mindset and students' general well-being (METASUM, GFOFAIL, MASTGOAL, SWBP, WORKMAST, ADAPTIVITY, COMPETE) and a random intercept accounting for school effects.
- Model 4 examines the effects of students' attitudes toward the school on reading scores (PERFEED, TEACHINT, BELONG) with a random intercept and a random slope of TEACHINT nested in schools.

Two sample sizes are examined: (a) a small sample of 400 students in total with 20 students in each school ($n_j = 20$), and (b) the full PISA 2018 sample size of 4838 with approximately 176 participating schools. The results for model weights are summarized in the upper panel of Table 2. Model 2 is preferred over the other models. This is the same with all the stacking methods for both the small sample size and the full sample size. However, we can see that PBMA and PBMA+ tend to put the majority of weight on a single model. By contrast, BS and BHS have a more balanced weighting scheme on all the models without relying on one model.

The lower panel of Table 2 summarizes the results of the Kullback-Leibler Divergence (KLD) score (also referred to as *relative entropy* (Kullback & Leibler, 1951; Kullback, 1959, 1987) for sample sizes of 200 and 4838. Here we consider two distributions, $p(y)$ and $g(y|\theta)$, where $p(y)$ denotes the distribution of observed reading literacy scores, and $g(y|\theta)$ denotes the prediction of these reading scores based on a model. The KLD between these two distributions can be written as

$$\text{KLD}(f, g) = \int p(y) \log \left(\frac{p(y)}{g(y|\theta)} \right) dy \tag{16}$$

where $\text{KLD}(f, g)$ is the information lost when $g(y|\theta)$ is used to approximate $p(y)$. For example, the actual reading outcome scores might be compared to the predicted outcome using Bayesian model averaging along with different choices of model and parameter priors. The model with the lowest KLD measure is deemed best in the sense that the information lost when approximating the actual reading outcome distribution with the distribution predicted on the basis of the model is the lowest.

Generally speaking, there is not much difference between all the KLDs in both small and large samples. However, we can find that the KLD obtained from BHS is the lowest in the large sample and the second lowest in the small sample. In $\mathcal{M}$-open setting, we will never know what the explicit form of the true DGM is, and thus we question the predictive

performance obtained from PBMA since it put 100% weights on model 2. Even though there is not a large difference in terms of KLDs across these four Bayesian stacking approaches, BHS obtains the best predictive performance.

Table 2
Model Weights and Predictive Performance Comparisons for PISA 2018 Example

	n = 200				n = 4838			
Model weights	BS	BHS	PBMA	PBMA+	BS	BHS	PBMA	PBMA+
Model1	0.010	0.076	0.000	0.000	0.199	0.011	0.000	0.029
Model2	0.615	0.541	1.000	0.943	0.644	0.588	1.000	0.884
Model3	0.375	0.318	0.000	0.057	0.157	0.391	0.000	0.087
Model4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Predictive scores								
KLD	0.032	0.018	0.017	0.030	0.076	0.040	0.042	0.050

Simulation Study

In this section, we implement simulation studies to examine the predictive performance for all Bayesian stacking methods under $\mathcal{M}$-closed and $\mathcal{M}$-complete setting.

Data Generation

Since model 2 yields the highest weights in the empirical study, we generate data based on model 2, which includes four covariates and one random intercept for the school. For the sake of simplicity, we use x_{1ij}, x_{2ij}, x_{3ij}, and x_{4ij} to denote the covariates which are normally distributed with the same mean and variance as the corresponding attitude and behaviors related variables in the PISA data. To distinguish $\mathcal{M}$-closed and $\mathcal{M}$-complete setting, we add a scalar for the Gaussian noise ϵ_{ij}. More specifically,

$$y_{ij} = \beta_{00} + \beta_{01}x_{1ij} + \beta_{02}x_{2ij} + \beta_{03}x_{3ij} + \beta_{04}x_{4ij} + U_{0j} + \sigma\epsilon_{ij} \tag{17}$$

, where y_{ij} is the response variable for student i in school j, x_{pij} denotes person i who is in school j, and has a value on variable p. The parameter β_{00} is overall intercept, β_{0p} denote the regression coefficients for x_{pij}. U_{0j} denotes the random intercept for school effects. We

set σ to be zero for the $\mathcal{M}$-closed setting, which indicates the data is generated exactly according to the true DGM without any noise. In $\mathcal{M}$-complete setting, we set σ to be five, which indicates that we know the explicit form of the true DGM, but the generated data does not totally depend on the true DGM. In this way, we can include the true DGM in the ensemble without error for both situations to examine the predictive performance in $\mathcal{M}$-closed and $\mathcal{M}$-complete setting.

PISA 2018 data has an approximately average number of 30 students nested in 150 schools, which is a 1:5 ratio for within-group sample size with respect to between-group sample size. To mimic the real data setting, we generate the data with a small sample of 500 where the number of schools n_j is set to be 50 and the number of students in each school n_i is 10. For the large sample, we set n_i to be 30 and n_j to be 150. In addition, Hedges and Hedberg (2007) has demonstrated that the intra-class correlation (ICC) in educational data generally falls in the range between 0.1 to 0.25. Therefore, we set the ICC to be 0.1, 0.2, and 0.3 when we generate the data.

After generating the data, we fit the simulated data using the rstanarm packages (Goodrich, Gabry, Ali, & Brilleman, 2022) in the statistical software environment R (R Core Team, 2022) version 4.2.1. Given that we are interested in how the school effects bring in randomness in the prediction using all four Bayesian stacking methods, all of our candidate models include a random intercept for school. Therefore, there are 15 candidate models (denoted as M1 to M15) in total with different choices of the combination of four covariates (i.e., $\binom{4}{1}, \binom{4}{2}, \binom{4}{3}, \binom{4}{4}$). Table 3 summarizes the covariates in each candidate model. We used *stan_lmer* function in rstanarm to fit all the candidate models, and extracted the weighted densities for each model using *loo* function in R package loo (Vehtari et al., 2022). For BHS, we used rstan (Stan Development Team, 2020) to specify the corresponding priors, hyper-priors, predictor-dependent weights, and log-likelihoods. All software codes for the simulation study are available in a GitHub repository (https://github.com/huskyh233/BayesStacking-Multilevel).

Table 3
Summary of covariates in each candidate model

Models	Covariates
Model 1	x_{1ij}
Model 2	x_{2ij}
Model 3	x_{3ij}
Model 4	x_{4ij}
Model 5	$x_{1ij} + x_{2ij}$
Model 6	$x_{1ij} + x_{3ij}$
Model 7	$x_{1ij} + x_{4ij}$
Model 8	$x_{2ij} + x_{3ij}$
Model 9	$x_{2ij} + x_{4ij}$
Model 10	$x_{3ij} + x_{4ij}$
Model 11	$x_{1ij} + x_{2ij} + x_{3ij}$
Model 12	$x_{1ij} + x_{2ij} + x_{4ij}$
Model 13	$x_{1ij} + x_{3ij} + x_{4ij}$
Model 14	$x_{2ij} + x_{3ij} + x_{4ij}$
Model 15	$x_{1ij} + x_{2ij} + x_{3ij} + x_{4ij}$

Convergence

To begin, we examined the two MCMC convergence criteria of all the models. Conventional measures for convergence are the effective sample sizes (ESS) and the potential scale reduction factor ($\hat{R}s$). The effective sample size is a measure of the number of independent MCMCs, which is proportional to the number of iterations of the MCMC algorithm and the autocorrelation present in the samples. The closer the ESS is to the number of samples taken from the posterior distribution (accounting for warm-up samples and thinning), the better convergence the model has achieved.

When implementing an MCMC algorithm, one of the most important diagnostics is the *potential scale reduction factor* (see e.g., Gelman & Rubin, 1992a; Gelman, 1996; Gelman & Rubin, 1992c), often denoted as *Rhat* or $\hat{R}$. This diagnostic is based on an analysis of variance and is intended to assess convergence among several parallel chains with varying starting values and is measured by the between-chain variance and the underestimate is measured by the within-chain variance (Gelman, 1996). The idea is that if the ratio of these two sources of variance is equal to one, then this is evidence that the

chains have converged. If the $\hat{R} > 1.01$ this may be a cause for concern.

A problem with $\hat{R}$ originally noted by Gelman et al. (2014) and further discussed in Vehtari, Gelman, Simpson, Carpenter, and Bürkner (2021) is that it sometimes does not detect non-stationarity, in the sense of the average or variability in the chains changing over the iteration history. A relatively new version of the potential scale reduction factor is available in Stan. This version is referred to as the *Split* $\hat{R}$ and is obtained by splitting the chain in two and then calculating the *Split* $\hat{R}$ on twice as many chains. So, if one uses four chains with 5000 iterations per chain, the *Split* $\hat{R}$ is based on eight chains with 2500 iterations per chain.An $\hat{R}$ less than 1.01 indicates convergence (Gelman & Rubin, 1992b).

With the 100 replications used in this study, the relative prediction bias for all the fitted models was less than 10%, which indicates that 100 replications are adequate. With 10,000 iterations, all the models managed to converge with the effective samples at around 2,000, and all $\hat{R}s$ were less than 1.01.

Model Weights

In this section, we investigate the difference in model weights obtained by different weighting methods in Bayesian stacking. Since BHS uses predictor-dependent weights for each covariate, we take an average of the weights for each predictor and compare it with the other BS methods.

Figure 1 shows the model weights for both the small sample and the large sample across different ICC in the $\mathcal{M}$-closed setting ($\sigma = 0$). As expected, the true DGM yields the highest weights within the $\mathcal{M}$-closed setting. More specifically, PBMA and PBMA+ assign 100 % weight to model 15 while BHS and BS assign particularly small amounts of weight to other models. This becomes more obvious in the large sample (N = 4500), where BHS and BS appear to put more weight in model 12. There is not much difference in model weights when ICC changes from 0.1 to 0.2 and 0.3. The variation in the model weights appears to be mainly due to the sample size for all the Bayesian stacking weighting methods.

Similarly, Figure 2 shows the model weights in the $\mathcal{M}$-complete setting ($\sigma = 5$). Generally speaking, the weights are more "spread out" in the $\mathcal{M}$-complete setting. For instance, when the sample size is 500, model 12 stands out in terms of model weights, besides model 15. To wit, model 15 does not attain almost 100% weights as it does in the $\mathcal{M}$-closed setting. However, in the large sample, model 15 still outperforms others in terms of model weights. This is true for PBMA, and PBMA+ in both small and large samples. Therefore, we anticipate that as the sample size approaches infinity, the weights assigned to model 15 using these two methods will approach one. As for BHS, both model 12 and model 15 obtain noticeable weights when N = 500 and N = 4500. Unlike BHS, PBMA, and PBMA+, there is no single model that obtains distinguishable weights using BS. That is, all of the fifteen models obtain non-zero weights and there is no prominent model.

Predictive Performance

After examining the difference in model weights, in this section, we compare the predictive performance for different weighting methods in Bayesian stacking. We computed the average KLD as well as the corresponding standard deviation to examine the predictive performance. In addition, we also want to investigate the computation time for running different methods in a high-speed computing system (Center for High Throughput Computing, 2006). Table 4 and 5 summarize these results under $\mathcal{M}$-closed setting and $\mathcal{M}$-complete setting respectively.

From Table 4, we can see that the average KLD obtained from different weighting methods is almost identical. That is, They are all close to zero with a small standard deviation. In line with our expectations, in $\mathcal{M}$-closed setting, there is no big difference in predictive performance using different weighting methods in Bayesian stacking. This situation is constant in terms of different sample sizes and ICC. Therefore, when the true DGM is in the candidate model sets, using predictor-dependent weighting does not appear to play a significant role in increasing predictive performance. However, in terms of computational efficiency, we can clearly find a high gap between BS and the other three

methods, which is more obvious in the sample size. For instance, when ICC is 0.1, for the sample of 500 participants, BS yields the longest computation time, 11.495 seconds. The implementation time of BHS is the second longest, while PBMA and PBMA+ are the shortest. However, in the sample of 5000, it takes 1324.899 seconds for BS to finish the computation. Though there is no big difference among the other three methods, the implementation of BHS takes the shortest time while PBMA and PBMA+ take the middle place. Therefore, for the sake of computational efficiency as well as the prediction capacity, BHS, PBMA, and PBMA+ might be a better choice than BS for multilevel models in $\mathcal{M}$-closed setting.

Table 5 summarises the results when the scaler of the error term is five, and thus the $\mathcal{M}$-complete setting. Generally speaking, there is more variability in terms of the average KLDs and their corresponding standard deviation, with shorter computation time, compared to the $\mathcal{M}$-closed setting. When ICC = 0.1, there is not much difference in average KLD obtained by all four weighting methods in both small and large samples. However, we can still find that the average KLD in the BS cell is the highest, BHS yields the second highest while PBMA and PBMA+ are the smallest. The gap in average KLD between BS and the other three methods is larger in the small sample. The standard deviation of KLD is also larger in the small sample, compared to the ones in the large sample. As for the computation efficiency, it takes a longer time for BS to finish the implementation process than other methods. This is consistent when the ICC and the sample sizes vary. The results in the condition of ICC = 0.2 are similar to the ones when ICC = 0.1. When ICC becomes 0.3, in the small sample when n = 500, BHS obtains the lowest average KLD compared to other methods. That is, the average KLD in the BHS cell is 0.199, while the other three methods are almost identically equal to 0.226. In addition, the standard deviation of KLD is also the smallest in the BHS cell. By contrast, both the mean and standard deviation for KLD are almost identical in the large sample. Given the results above, we can find that the predictive performance using different weighting

methods in Bayesian stacking is fairly similar. However, when the intraclass correlation correlation (0.1 to 0.3), BHS appears to stand out in terms of predictive performance than others in the small sample.

Table 4
Kullback–Leibler divergence across different sample sizes with different ICC in $\mathcal{M}$-closed setting

ICC	Sample Sizes (n_i*n_j)	Methods	KLD Mean	KLD SD	Time
0.1	500 (10*50)	BS	0.000	0.000	11.495
		BHS	0.001	0.004	7.600
		PBMA	0.000	0.000	3.73
		PBMA+	0.000	0.000	3.81
	4500 (30*150)	BS	0.001	0.003	1324.899
		BHS	0.000	0.000	11.540
		PBMA	0.000	0.000	12.958
		PBMA+	0.000	0.000	13.450
0.2	500 (10*50)	BS	0.000	0.000	10.520
		BHS	0.000	0.003	7.331
		PBMA	0.000	0.000	3.581
		PBMA+	0.000	0.000	3.660
	4500 (30*150)	BS	0.002	0.004	1659.788
		BHS	0.000	0.000	12.521
		PBMA	0.000	0.000	13.160
		PBMA+	0.000	0.000	13.650
0.3	500 (10*50)	BS	0.000	0.000	12.646
		BHS	0.000	0.004	7.428
		PBMA	0.000	0.000	3.483
		PBMA+	0.000	0.000	3.561
	4500 (30*150)	BS	0.002	0.007	1489.477
		BHS	0.000	0.000	11.197
		PBMA	0.000	0.000	12.622
		PBMA+	0.000	0.000	13.122

Conclusion and Discussion

This paper examined different weighting schemes in Bayesian stacking for large-scale assessment data with a multilevel structure. Four methods were investigated including original BS, PBMA, PBMA+, and BHS. Unlike the BS method which is not fully Bayesian, BHS not only incorporates predictor-dependent weighting but also adds priors and hyperpriors to the weights, which allows for more flexibility in predicting the different

Table 5
Kullback–Leibler divergence across different sample sizes with different ICC in $\mathcal{M}$-complete setting

ICC	Sample Sizes (n_i*n_j)	Methods	KLD Mean	KLD SD	Time
0.1	500 (10*50)	BS	0.190	0.272	28.891
		BHS	0.180	0.379	6.662
		PBMA	0.172	0.302	3.498
		PBMA+	0.170	0.292	3.572
	4500 (30*150)	BS	0.095	0.043	267.785
		BHS	0.091	0.036	12.543
		PBMA	0.091	0.035	13.883
		PBMA+	0.091	0.036	14.376
0.2	500 (10*50)	BS	0.210	0.281	25.978
		BHS	0.210	0.473	6.779
		PBMA	0.196	0.332	3.556
		PBMA+	0.196	0.337	3.629
	4500 (30*150)	BS	0.102	0.050	196.303
		BHS	0.100	0.039	10.533
		PBMA	0.100	0.038	12.105
		PBMA+	0.100	0.039	12.560
0.3	500 (10*50)	BS	0.226	0.326	26.877
		BHS	0.199	0.259	6.951
		PBMA	0.226	0.380	3.538
		PBMA+	0.227	0.380	3.613
	4500 (30*150)	BS	0.111	0.042	192.495
		BHS	0.111	0.042	10.891
		PBMA	0.112	0.042	12.487
		PBMA+	0.112	0.042	12.970

data structures. In this study, we focused on data with a multilevel structure with different between-group variability in the three $\mathcal{M}$-frameworks. Broadly speaking, though the prediction discrepancy among the different weighing schemes in Bayesian stacking does not appear to be very obvious, we can still find that their predictive performance varies depending on 1. which $\mathcal{M}$-framework they are in (e.g., the scalar of the error term, σ); 2. the between-group variability indicated by the ICC; 3. and the sample size. To be precise, in the $\mathcal{M}$-open setting, BHS and PBMA yield the best predictive performance. Nevertheless, the predictive performance of PBMA is questionable given the fact that it assigns 100 % weight to model 2, which is arguable in $\mathcal{M}$-open setting. Our simulation studies mimic the $\mathcal{M}$-closed and $\mathcal{M}$-complete setting. In the $\mathcal{M}$-closed setting, all of the

methods obtain approximately zero KLD and assign more or less 100 % weight on the true DGM. However, in the $\mathcal{M}$-complete setting, BHS stands out in the situation where the between-group variability is large (ICC = 0.3). Therefore, we anticipate that as we increase the ICC as well as σ for the Gaussian noise, BHS will have a larger deviation in the prediction with other methods. For instance, adding more layers in the multilevel structure such as a three-level hierarchy, or more complex group relationships such as crossed random effects, could increase the between-group variability. Another option would be varying σ in the simulation study to examine if BHS still obtains the best predictive performance when there are different amounts of noise in the $\mathcal{M}$-complete setting. Considering the computation efficiency, BHS, PBMA, and PBMA+ might be better choices for start-ups, compared to BS.

There are several limitations in this study, and thus there are several directions for future research. First, we applied the same hyper-priors which Yao et al. (2021) used in their paper. Therefore, it would be useful to explore how different priors and hyper-priors affect predictive performance using BHS. Second, in this study, we used parametric regression methods to model the weights in BHS. In the future, it may be interesting to use non-parametric methods to compute the weights for different predictors, which might lead to more flexibility without having any assumptions. A third limitation derives from specific issues associated with the construction of large-scale assessments. Specifically, although the present paper investigated the performance of various stacking weights for multilevel models, the full utility of stacking for large-scale assessments will require incorporating plausible values and sampling weights. Finally, the data structure we investigated in this study has fixed group memberships, which is not always the case in large-scale assessments. Thus, it will be interesting to explore stacking methods with more complicated data structures of relevance to education research, such as multiple membership models.

To conclude, a number of approaches now exist to address the problem of model uncertainty with applications to the social and behavioral sciences, and although Bayesian

model averaging remains a very popular and often adequate procedure for addressing model uncertainty, it rests on the $\mathcal{M} - closed$ assumption, which analysts may feel uncomfortable holding. Bayesian stacking methods relax this assumption, and a variety of choices for stacking weights are available and easily implemented through open-source software. For this study, we examined a variety of stacking methods for multilevel models applied to large-scale educational assessments and found Bayesian hierarchical stacking to be, by far, a promising approach for calculating stacking weights with respect to predictive performance and recommend its routine implementation for stacking statistical models.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory.* Budapest: Akademiai Kiado.

Bernardo, J. M., & Smith, A. F. M. (2000). *Bayesian theory.* New York: Wiley.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (Second ed.). New York: Springer.

Center for High Throughput Computing. (2006). *Center for high throughput computing.* Center for High Throughput Computing. Retrieved from `https://chtc.cs.wisc.edu/` doi: 10.21231/GNT1-HW21

Clyde, M. A. (1999). Bayesian model averaging and model search strategies (with discussion). In J. M. Bernardo, A. P. Dawid, J. O. Berger, & A. F. M. Smith (Eds.), *Bayesian statistics, 6* (pp. 157–185). Oxford: Oxford University Press.

Clyde, M. A. (2003). Model averaging. In *Subjective and objective Bayesian statistics* (pp. 320–335). Hoboken, N. J.: Wiley-Interscience.

Clyde, M. A., & Iversen, E. S. (2013). Bayesian model averaging in the M-open framework. In *Bayesian theory and applications* (pp. 483–498). Oxford: Oxford University Press.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B)*, *57*, 55–98.

Draper, D., Hodges, J. S., Leamer, E. E., Morris, C. N., , & Rubin, D. B. (1987). *A research agenda for assessment and propagation of model uncertainty* (Tech. Rep.). Santa Monica, CA: Rand Corporation. Retrieved from `https://www.rand.org/pubs/notes/N2683.html` (N-2683-RC,)

Fernández, C., Ley, E., & Steel, M. F. J. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, *16*, 563–576.

Fletcher, D. (2018). *Model averaging.* Berlin: Springer.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of*

the American Statistical Association, *74*(365), 153–160.

Gelfand, A. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain monte carlo in practice* (pp. 145–161). Boca Raton: Chapman & Hall.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 131–143). New York: Chapman & Hall.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehatari, A., & Rubin, D. B. (2014). *Bayesian data analysis, 3rd edition*. London: Chapman and Hall.

Gelman, A., & Rubin, D. B. (1992a). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Gelman, A., & Rubin, D. B. (1992b). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.

Gelman, A., & Rubin, D. B. (1992c). A single series from the Gibbs sampler provides a false sense of security. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 625–631). Oxford: Oxford University Press.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *14*, 107–114.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2022). *rstanarm: Bayesian applied regression modeling via Stan.* Retrieved from `https://mc-stan.org/rstanarm/` (R package version 2.21.3)

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87.

Hoeting, J. A., Madigan, D., Raftery, A., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417.

Kaplan, D. (2021). On the Quantification of Model Uncertainty: A Bayesian Perspective.

Psychometrika, *86*(1), 215–238. Retrieved from https://doi.org/10.1007/s11336-021-09754-5

Kaplan, D., & Chen, J. (2014). Bayesian model averaging for propensity score analysis. *Multivariate Behavioral Research*, *49*, 505–517.

Kaplan, D., & Huang, M. (2021). Bayesian probabilistic forecasting with large-scale educational trend data: A case study using NAEP. *Large-scale Assessments in Education*, *9*(1), 1–31.

Kaplan, D., & Lee, C. (2015). Bayesian model averaging over directed acyclic graphs with implications for the predictive performance of structural equation models. *Structural Equation Modeling*. doi: 10.1080/10705511.2015.1092088

Kaplan, D., & Lee, C. (2018). Optimizing prediction using Bayesian model averaging: Examples using large-scale educational assessments. *Evaluation Review*. doi: 10.1177/0193841X18761421

Kaplan, D., & Yavuz, S. (2019). An approach to addressing multiple imputation model uncertainty using Bayesian model averaging. *Multivariate Behavioral Research*. Retrieved from https://doi.org/10.1080/00273171.2019.1657790 (PMID: 31538505) doi: 10.1080/00273171.2019.1657790

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Kullback, S. (1959). *Information theory and statistics*. New York: John Wiley and Sons.

Kullback, S. (1987). The Kullback-Leibler distance. *The American Statistician*, *41*, 340–341.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. New York: Wiley.

Madigan, D., & Raftery, A. (1994). Model selection and accounting for model uncertainly in graphical models using Occam's window. *Journal of the American Statistical*

Association, *89*, 1535–1546.

Montgomery, J. M., & Nyhan, B. (2010). Bayesian model averaging: Theoretical developments and practical applications. *Political Analysis*, *18*, 245–270.

Muthukrishnan, R., & Rohini, R. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. In *2016 ieee international conference on advances in computer applications (icaca)* (pp. 18–20).

OECD. (2002). *PISA 2000 technical report*. Paris: Organization for Economic Cooperation and Development.

OECD. (2018). *PISA 2018 Technical Report*. Paris: OECD. Retrieved from `https://www.oecd.org/pisa/data/pisa2018technicalreport/`

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raftery, A., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*, 1155-1174.

Raftery, A., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*, 179–191.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Second ed.). Thousands Oaks, CA: Sage Publications.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, *9*, 130–134.

Sloughter, J. M., Gneiting, T., & Raftery, A. (2013). Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Monthly Weather Review*, *141*, 2107–2119.

Stan Development Team. (2020). *RStan: the R interface to Stan.* Retrieved from `http://mc-stan.org/` (R package version 2.21.1)

Stan Development Team. (2021). *Stan modelling language users guide and reference manual v. 2.27.* Stan Development Team Sydney, Australia.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2022). *loo: Efficient leave-one-out cross-validation and waic for bayesian models.*

Retrieved from https://mc-stan.org/loo/ (R package version 2.5.1)

Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2019). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.* Retrieved from https://CRAN.R-project.org/package=loo (R package version 2.1.0)

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. doi: 10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis.*. Retrieved from https://doi.org/10.1214/20-BA1221

Vehtari, A., & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural computation*, *14*(10), 2439–2468.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571-3594.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*, 241 - 259.

Yao, Y., Pirš, G., Vehtari, A., & Gelman, A. (2021). Bayesian hierarchical stacking: Some models are (somewhere) useful. *Bayesian Analysis*, *1*(1), 1–29.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018a). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*(3), 917–1007.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018b). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*, 917–1007. Retrieved from https://doi.org/10.1214/17-BA1091 doi: 10.1214/17-BA1091

Yeung, K. Y., Bumgarner, R. E., & Raftery, A. (2005). Bayesian model averaging: development of an improved multi-class, gene selection, and classification tool for microarray data. *Bioinformatics*, *21*, 2394–2402.

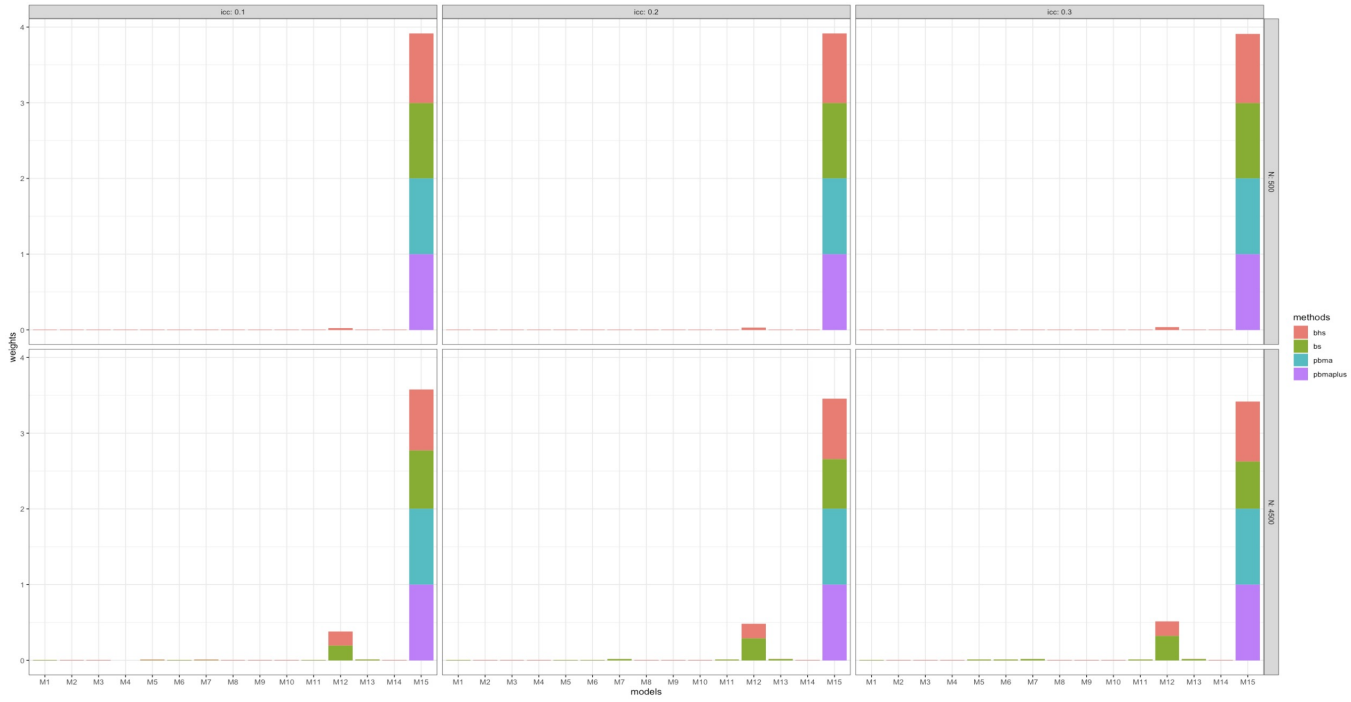

Figure 1. Model weights across different sample sizes with different ICC in $\mathcal{M}$-closed setting

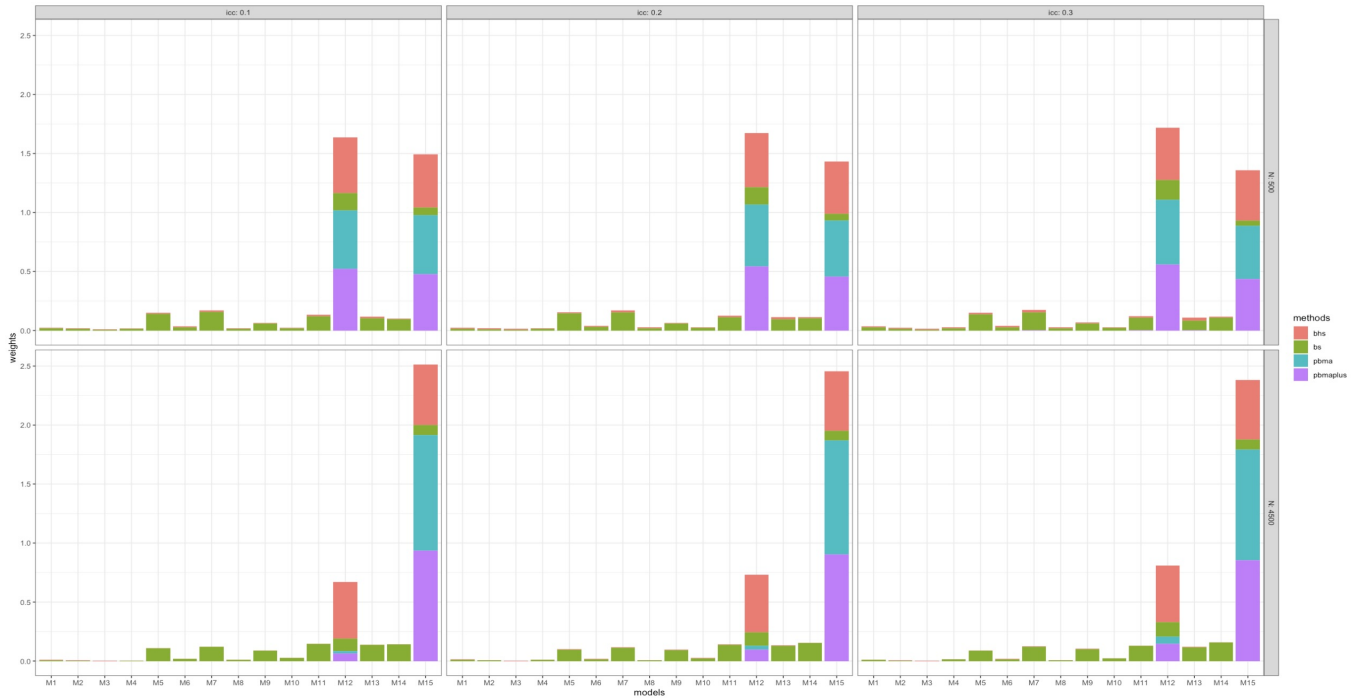

Figure 2. Model weights across different sample sizes with different ICC in $\mathcal{M}$-complete setting