

Using stacking to average Bayesian predictive distributions

Yuling Yao^{*}, Aki Vehtari[†], Daniel Simpson[‡] and Andrew Gelman[§]

Abstract. Bayesian model averaging is flawed in the $\mathcal{M}$ -open setting in which the true data-generating process is not one of the candidate models being fit. We take the idea of *stacking* from the point estimation literature and generalize to the combination of predictive distributions. We extend the utility function to any proper scoring rule and use Pareto smoothed importance sampling to efficiently compute the required leave-one-out posterior distributions. We compare stacking of predictive distributions to several alternatives: stacking of means, Bayesian model averaging (BMA), Pseudo-BMA using AIC-type weighting, and a variant of Pseudo-BMA that is stabilized using the Bayesian bootstrap. Based on simulations and real-data applications, we recommend stacking of predictive distributions, with bootstrapped-Pseudo-BMA as an approximate alternative when computation cost is an issue.

Keywords: Bayesian model averaging, model combination, proper scoring rule, predictive distribution, stacking, Stan.

1 Introduction

A general challenge in statistics is prediction in the presence of multiple candidate models or learning algorithms $\mathcal{M} = (M_1, \dots, M_K)$. Choosing one model that can give optimal performance for future data can be unstable and wasteful of information (see, e.g., Piironen and Vehtari, 2017). An alternative is model averaging, which tries to find an optimal model combination in the space spanned by all individual models. In Bayesian context, the natural target for prediction is to find a predictive distribution that is close to the true data generating distribution (Gneiting and Raftery, 2007; Vehtari and Ojanen, 2012).

Ideally, we would avoid the Bayesian model combination problem by extending the model to include the separate models M_k as special cases (Gelman, 2004). In practice, constructing such an expansion requires a lot of conceptual and computational effort. Hence, in this paper we focus on simpler tools that work with existing inferences from models that have been fitted separately.

To appear in *Bayesian Analysis*.

^{*}Department of Statistics, Columbia University, New York, NY, yy2619@columbia.edu

[†]Helsinki Institute of Information Technology, Department of Computer Science, Aalto University, Finland, Aki.Vehtari@aalto.fi

[‡]Department of Statistical Sciences, University of Toronto, Canada, dp.simpson@gmail.com

[§]Department of Statistics and Department of Political Science, Columbia University, New York, NY, gelman@stat.columbia.edu

This paper is organized as follows. In Section 2, we give a brief review of some existing model averaging methods. Then we propose our stacking method in Section 3. In Section 4, we compare stacking, Bayesian model averaging, and several other alternatives through a Gaussian mixture model, a series of linear regression simulations, two real data examples, and an application in variational inference. We conclude with Section 5 where we give general recommendations.

2 Existing approaches

In Bayesian model comparison, the relationship between the true data generator and the model list $\mathcal{M} = (M_1, \dots, M_K)$ can be classified into three categories: $\mathcal{M}$ -closed, $\mathcal{M}$ -complete and $\mathcal{M}$ -open. We adopt the following definition from [Bernardo and Smith \(1994\)](#) (see also [Key et al. \(1999\)](#), and [Clyde and Iversen \(2013\)](#)):

- $\mathcal{M}$ -closed means the true data generating model is one of $M_k \in \mathcal{M}$, although it is unknown to researchers.
- $\mathcal{M}$ -complete refers to the situation where the true model exists and is out of model list $\mathcal{M}$. But we still wish to use the models in $\mathcal{M}$ because of tractability of computations or communication of results, compared with the actual belief model. Thus, one simply finds the member in $\mathcal{M}$ that maximizes the expected utility (with respect to the true model).
- $\mathcal{M}$ -open refers to the situation in which we know the true model M_t is not in $\mathcal{M}$, but we cannot specify the explicit form $p(\tilde{y}|y)$ because it is too difficult conceptually or computationally, we lack time to do so, or do not have the expertise, etc.

Bayesian model averaging If all candidate models are generative, the Bayesian solution is to simply average the separate models, weighing each by its marginal posterior probability. This is called *Bayesian model averaging* (BMA) and is optimal if the method is evaluated based on its frequency properties evaluated over the joint prior distribution of the models and their internal parameters ([Madigan et al., 1996](#); [Hoeting et al., 1999](#)). If $y = (y_1, \dots, y_n)$ represents the observed data, then the posterior distribution for any quantity of interest Δ is $p(\Delta|y) = \sum_{k=1}^K p(\Delta|M_k, y)p(M_k|y)$. In this expression, each model is weighted by its posterior probability,

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

and this expression depends crucially on the marginal likelihood under each model, $p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k$.

BMA is appropriate for the $\mathcal{M}$ -closed case. In $\mathcal{M}$ -open and $\mathcal{M}$ -complete cases, BMA will asymptotically select the one single model on the list that is closest in Kullback-Leibler (KL) divergence.

A further problem with BMA is that the marginal likelihood is sensitive to the specific prior $p(\theta_k|M_k)$ in each model. For example, consider a problem where a parameter has

been assigned a normal prior distribution with center 0 and scale 10, and where its estimate is likely to be in the range $(-1, 1)$. The chosen prior is then essentially flat, as would also be the case if the scale were increased to 100 or 1000. But such a change would divide the posterior probability of the model by roughly a factor of 10 or 100.

Stacking *Stacking* (Wolpert, 1992; Breiman, 1996; LeBlanc and Tibshirani, 1996) is a direct approach for averaging point estimates from multiple models. In supervised learning, where the data are $((x_i, y_i), i = 1, \dots, n)$ and each model M_k has a parametric form $\hat{y}_k = f_k(x|\theta_k)$, stacking is done in two steps (Ting and Witten, 1999). First, each model is fitted separately and the leave-one-out (LOO) predictor $\hat{f}_k^{(-i)}(x_i) = E[y_i | \hat{\theta}_k, y_{-i}, M_k]$ is obtained for each model k and each data point i . In the second step, a weight for each model is obtained by minimizing the LOO mean squared error

$$\hat{w} = \arg \min_w \sum_{i=1}^n \left(y_i - \sum_k w_k \hat{f}_k^{(-i)}(x_i) \right)^2. \quad (2.1)$$

Breiman (1996) claims that either a positive constraint $w_k \geq 0, k = 1, \dots, K$, or a simplex constraint: $w_k \geq 0, \sum_{k=1}^K w_k = 1$ guarantees a solution. Better predictions may be attainable using regularization (Merz and Pazzani, 1999; Yang and Dunson, 2014). Finally, the point prediction for a new data point with feature vector $\tilde{x}$ is

$$\hat{\tilde{y}} = \sum_{k=1}^K \hat{w}_k f_k(\tilde{x} | \hat{\theta}_{k, y_{1:n}}).$$

It is not surprising that stacking typically outperforms BMA when the criterion is mean squared predictive error (Clarke, 2003), because BMA is not optimized to this task. Wong and Clarke (2004) emphasize that the BMA weights reflect the fit to the data rather than evaluating the prediction accuracy. On the other hand, stacking is not widely used in Bayesian model combination because the classical stacking only works with point estimates, not the entire posterior distribution (Hoeting et al., 1999).

Clyde and Iversen (2013) give a Bayesian interpretation for stacking by considering model combination as a decision problem when the true model M_t is not in the model list. If the decision is of the form $a(y, w) = \sum_{k=1}^K w_k \hat{y}_k$, then the expected utility under quadratic loss is,

$$E_{\tilde{y}}[u(\tilde{y}, a(y, w)) | y] = - \int ||\tilde{y} - \sum_{k=1}^K w_k \hat{y}_k||^2 p(\tilde{y} | y, M_t) d\tilde{y},$$

where $\hat{\tilde{y}}_k$ is the predictor of new data $\tilde{y}$ in model k . Hence, the stacking weights are the solution to the LOO estimator

$$\hat{w} = \arg \max_w \frac{1}{n} \sum_{i=1}^n u(y_i, a(y_{-i}, w)),$$

where $a(y_{-i}, w) = \sum_{k=1}^K w_k E[y_i | y_{-i}, M_k]$.

Le and Clarke (2017) prove the stacking solution is asymptotically the Bayes solution. With some mild conditions on distributions, the following asymptotic relation holds:

$$\int l(\tilde{y}, a(y, w)) p(\tilde{y} | y) d\tilde{y} - \frac{1}{n} \sum_{i=1}^n l(y_i, a(y_{-i}, w)) \xrightarrow{L_2} 0,$$

where l is the squared loss, $l(\tilde{y}, a) = (\tilde{y} - a)^2$. They also prove that when the action is a predictive distribution $a(y_{-i}, w) = \sum_{k=1}^K w_k p(y_i | y_{-i}, M_k)$, the asymptotic relation still holds for negative logarithm scoring rules.

However, most early literature limited stacking to averaging *point* predictions, rather than *predictive distributions*. In this paper, we extend stacking from minimizing the squared error to maximizing scoring rules, hence make stacking applicable to combining a set of Bayesian posterior predictive distributions. We argue this is the appropriate version of Bayesian model averaging in the $\mathcal{M}$ -open situation.

Akaike weights and pseudo Bayesian model averaging Leave-one-out cross-validation is related to various information criteria (see, e.g. Vehtari and Ojanen, 2012). In case of maximum likelihood estimates, leave-one-out cross-validation is asymptotically equal to Akaike’s information criterion (AIC, Stone, 1977). In a statistical model with the number of parameters to be k and the maximized likelihood to be $\hat{L}$, $AIC = -2 \log \hat{L} + 2k$. Akaike (1978) proposed to use $\exp(-\frac{1}{2}AIC)$ for model weighting (see also Burnham and Anderson, 2002; Wagenmakers and Farrell, 2004). More recently we have seen also Watanabe-Akaike information criterion (WAIC, Watanabe, 2010) and leave-one-out cross-validation estimates used to compute model weights following the idea of AIC weights.

In a Bayesian setting Geisser and Eddy (1979; see also, Gelfand 1996) proposed pseudo Bayes factors where marginal likelihoods $p(y | M_k)$ are replaced with a product of Bayesian leave-one-out cross-validation predictive densities $\prod_{i=1}^n p(y_i | y_{-i}, M_k)$. Following the naming by Geisser and Eddy, we call AIC-type weighting which uses Bayesian cross-validation predictive densities as *pseudo Bayesian model averaging* (Pseudo-BMA).

Exact leave-one-out cross-validation can be computationally costly. For example, in the econometric literature, Geweke and Amisano (2011, 2012) suggest averaging prediction models by maximizing predictive log score, while they only consider time series due to the computational challenges of exact LOO for general data structures. In the present paper we demonstrate that Pareto smoothed importance sampling leave-one-out cross-validation (PSIS-LOO) (Vehtari et al., 2017a,b) is a practically efficient way to calculate the needed leave-one-out predictive densities $p(y_i | y_{-i}, M_k)$.

In this paper we show that the uncertainty in the future data distribution should be taken into account when computing Pseudo-BMA weights. We will propose an AIC-type weighting using the Bayesian bootstrap and the expected log predictive density (elpd), which we call Pseudo-BMA+ weighting. We show that although Pseudo-BMA+ weighting gives better results than regular BMA or Pseudo-BMA weighting (in $\mathcal{M}$ -open settings),

it is still inferior to the log score stacking. Due to its simplicity we use Pseudo-BMA+ weighting as an initial guess for optimization procedure in the log score stacking.

Other model weighting approaches Besides BMA, stacking, and AIC-type weighting, some other methods have been introduced to combine Bayesian models. [Gutiérrez-Peña and Walker \(2005\)](#) consider using a nonparametric prior in the decision problem stated above. Essentially they are fitting a mixture model with a Dirichlet process, yielding a posterior expected utility of

$$U_n(w_k, \theta_k) = \sum_{i=1}^n \sum_{k=1}^K w_k f_k(y_i | \theta_k).$$

They then solve for the optimal weights $\hat{w}_k = \arg \max_{w_k, \theta_k} U_n(w_k, \theta_k)$.

[Li and Dunson \(2016\)](#) propose model averaging using weights based on divergences from a reference model in $\mathcal{M}$ -complete settings. If the true data generating density function is known to be f^* , then an AIC-type weight can be defined as

$$w_k = \frac{\exp(-n\text{KL}(f^*, f_k))}{\sum_{k=1}^K \exp(-n\text{KL}(f^*, f_k))}. \quad (2.2)$$

The true model can be approximated with a reference model M_0 with density $f_0(\cdot | \theta_0)$ using nonparametric methods like Gaussian process or Dirichlet process, and $\text{KL}(f^*, f_k)$ can be estimated by its posterior mean,

$$\widetilde{\text{KL}}_1(f_0, f_k) = \int \int \text{KL}(f_0(\cdot | \theta_0), f_k(\cdot | \theta_k)) p(\theta_k | y, M_k) p(\theta_0 | y, M_0) d\theta_k d\theta_0,$$

or by the Kullback-Leibler divergence for posterior predictive distributions,

$$\widetilde{\text{KL}}_2(f_0, f_k) = \text{KL}\left(\int f_0(\cdot | \theta_0) p(\theta_0 | y, M_0) d\theta_0, \int f_k(\cdot | \theta_k) p(\theta_k | y, M_k) d\theta_k\right).$$

Here, $\widetilde{\text{KL}}_1$ corresponds to Gibbs utility, which can be criticized for not using the posterior predictive distributions ([Vehtari and Ojanen, 2012](#)). Although asymptotically the two utilities are identical, and $\widetilde{\text{KL}}_1$ is often computationally simpler than $\widetilde{\text{KL}}_2$.

Let $p(\tilde{y} | y, M_k) = \int f_k(\tilde{y} | \theta_k) p(\theta_k | y, M_k) d\theta_k$, $k = 0, \dots, K$, then

$$\widetilde{\text{KL}}_2(f_0, f_k) = - \int \log p(\tilde{y} | y, M_k) p(\tilde{y} | y, M_0) d\tilde{y} + \int \log p(\tilde{y} | y, M_0) p(\tilde{y} | y, M_0) d\tilde{y}.$$

As the entropy of the reference model $\int \log p(\tilde{y} | y, M_0) p(\tilde{y} | y, M_0) d\tilde{y}$ is constant, the corresponding terms cancel out in the weight (2.2), leaving

$$w_k = \frac{\exp(n \int \log p(\tilde{y} | y, M_k) p(\tilde{y} | y, M_0) d\tilde{y})}{\sum_{k=1}^K \exp(n \int \log p(\tilde{y} | y, M_k) p(\tilde{y} | y, M_0) d\tilde{y})}.$$

It is proportional to the exponential expected log predictive density, where the expectation is taken with respect to the reference model M_0 . Comparing with definition 3.6 in Section 3.4, this method could be called Reference-Pseudo-BMA.

3 Theory and methods

We label the classical stacking procedure (2.1) as *stacking of means* because it combines models by minimizing the mean squared error of the point estimate. In general, we can use a proper scoring rule (or equivalently the underlying divergence) to compare distributions. After choosing that, stacking can be extended to combining the whole distributions.

3.1 Stacking using proper scoring rules

Adapting the notation of Gneiting and Raftery (2007), we label Y as the random variable on the sample space $(\Omega, \mathcal{A})$ that can take values on $(-\infty, \infty)$. $\mathcal{P}$ is a convex class of probability measure on Ω . Any member of $\mathcal{P}$ is called a probabilistic forecast. A *scoring rule* is a function $S : \mathcal{P} \times \Omega \rightarrow \mathbb{R} = [\infty, \infty]$ such that $S(P, \cdot)$ is $\mathcal{P}$ -quasi-integrable for all $P \in \mathcal{P}$. In the continuous case, every distribution $P \in \mathcal{P}$ is identified with its density function p .

For two probabilistic forecasts P and Q , we write $S(P, Q) = \int S(P, \omega) dQ(\omega)$. A scoring rule S is called *proper* if $S(Q, Q) \geq S(P, Q)$ and *strictly proper* if equality holds only when $P = Q$ almost surely. A proper scoring rule defines the divergence $d : \mathcal{P} \times \mathcal{P} \rightarrow (0, \infty)$ as $d(P, Q) = S(Q, Q) - S(P, Q)$. For continuous variables, some popularly used scoring rules include:

1. *Quadratic score*: $QS(p, y) = 2p(y) - \|p\|_2^2$ with the divergence $d(p, q) = \|p - q\|_2^2$.
2. *Logarithmic score*: $\text{LogS}(p, y) = \log(p(y))$ with $d(p, q) = \text{KL}(q, p)$. The logarithmic score is the only proper local score assuming regularity conditions.
3. *Continuous-ranked probability score*: $\text{CRPS}(F, y) = -\int_{\mathbb{R}} (F(y') - 1(y' \geq y))^2 dy'$ with $d(F, G) = \int_{\mathbb{R}} (F(y) - G(y))^2 dy$, where F and G are the corresponding distribution functions.
4. *Energy score*: $\text{ES}(P, y) = \frac{1}{2} \mathbb{E}_P \|Y - Y'\|_2^\beta - \mathbb{E}_P \|Y - y\|_2^\beta$, where Y and Y' are two independent random variables from distribution P . When $\beta = 2$, this becomes $\text{ES}(P, y) = -\|\mathbb{E}_P(Y) - y\|^2$. The energy score is strictly proper when $\beta \in (0, 2)$ but not when $\beta = 2$.
5. *Scoring rules depending on first and second moments*: Examples include $S(P, y) = -\log \det(\Sigma_P) - (y - \mu_P)^T \Sigma_P^{-1} (y - \mu_P)$, where μ_P and Σ_P are the mean vector and covariance matrix of distribution P .

The ultimate goal of stacking a set of K predictive distributions built from the models $\mathcal{M} = (M_1, \dots, M_K)$ is to find the distribution in the convex hull $\mathcal{C} = \{\sum_{k=1}^K w_k p(\cdot | M_k) : \sum_k w_k = 1, w_k \geq 0\}$ that is optimal according to some given criterion. In this paper, we propose the use of proper scoring functions to define the optimality criterion.

If we define $\mathcal{S}_1^K = \{w \in [0, 1]^K : \sum_{k=1}^K w_k = 1\}$, then we can write the stacking problem as

$$\min_{w \in \mathcal{S}_1^K} d\left(\sum_{k=1}^K w_k p(\cdot | y, M_k), p_t(\cdot | y)\right) \text{ or } \max_{w \in \mathcal{S}_1^K} S\left(\sum_{k=1}^K w_k p(\cdot | y, M_k), p_t(\cdot | y)\right), \quad (3.1)$$

where $p(\tilde{y}|y, M_k)$ is the predictive density of new data $\tilde{y}$ in model M_k that has been trained on observed data y and $p_t(\tilde{y}|y)$ refers to the true distribution.

An empirical approximation to (3.1) can be constructed by replacing the full predictive distribution $p(\tilde{y}|y, M_k)$ evaluated at a new datapoint $\tilde{y}$ with the corresponding LOO predictive distribution $\hat{p}_{k,-i}(y_i) = \int p(y_i|\theta_k, M_k)p(\theta_k|y_{-i}, M_k)d\theta_k$. The corresponding stacking weights are the solution to the optimization problem

$$\max_{w \in \mathcal{S}_1^K} \frac{1}{n} \sum_{i=1}^n S\left(\sum_{k=1}^K w_k \hat{p}_{k,-i}, y_i\right). \quad (3.2)$$

The stacked estimate of the predictive density is

$$\hat{p}(\tilde{y}|y) = \sum_{k=1}^K \hat{w}_k p(\tilde{y}|y, M_k). \quad (3.3)$$

When using logarithmic score (corresponding to Kullback-Leibler divergence), we call this *stacking of predictive distributions*:

$$\max_{w \in \mathcal{S}_1^K} \frac{1}{n} \sum_{i=1}^n \log \sum_{k=1}^K w_k p(y_i|y_{-i}, M_k).$$

The choice of scoring rules can depend on the underlying application. Stacking of means (2.1) corresponds to the energy score with $\beta = 2$. The reasons why we prefer stacking of predictive distributions (corresponding to the logarithmic score) to stacking of means are: (i) the energy score with $\beta = 2$ is not a strictly proper scoring rule and can give rise to identification problems, and (ii) without further smoothness assumptions, every proper local scoring rule is equivalent to the logarithmic score (Gneiting and Raftery, 2007).

3.2 Asymptotic behavior of stacking

The stacking estimate (3.1) finds the optimal predictive distribution within the convex set $\mathcal{C}$, that is the closest to the data generating process with respect to the chosen scoring rule. This is different from Bayesian model averaging, which asymptotically with probability 1 will select a single model: the one that is closest in KL divergence to the true data generating process.

Solving for the stacking weights in (3.2) is an M-estimation problem. Under some mild conditions (Le and Clarke, 2017; Clyde and Iversen, 2013; Key et al., 1999), for either the logarithmic scoring rule or the energy score (negative squared error) and a given set of weights $w_1 \dots w_K$, as sample size $n \rightarrow \infty$, the following asymptotic limit holds:

$$\frac{1}{n} \sum_{i=1}^n S\left(\sum_{k=1}^K w_k \hat{p}_{k,-i}, y_i\right) - \mathbb{E}_{\tilde{y}|y} S\left(\sum_{k=1}^K w_k p(\tilde{y}|y, M_k), \tilde{y}\right) \xrightarrow{L_2} 0.$$

Thus the leave-one-out-score is a consistent estimator of the posterior score. In this sense, stacking gives optimal combination weights asymptotically.

In terms of [Vehtari and Ojanen \(2012, Section 3.3\)](#), the proposed stacking of predictive distributions is the M_* -optimal projection of the information in the actual belief model M_* to $\hat{w}$, where explicit specification of M_* is avoided by re-using data as a proxy for the predictive distribution of the actual belief model and the weights w_k are the free parameters.

3.3 Pareto smoothed importance sampling

One challenge in calculating the stacking weights proposed in (3.2) is that we need the leave-one-out (LOO) predictive density,

$$p(y_i|y_{-i}, M_k) = \int p(y_i|\theta_k, M_k)p(\theta_k|y_{-i}, M_k)d\theta_k.$$

Exact LOO requires refitting each model n times. To avoid this onerous computation, we use the following approximate method. For the k -th model, we fit to all the data, obtaining S simulation draws $\theta_k^s (s = 1, \dots, S)$ from the full posterior $p(\theta_k|y, M_k)$ and calculate

$$r_{i,k}^s = \frac{1}{p(y_i|\theta_k^s, M_k)} \propto \frac{p(\theta_k^s|y_{-i}, M_k)}{p(\theta_k^s|y, M_k)}. \quad (3.4)$$

The ratio $r_{i,k}^s$ has a density function and can be unstable, due to a potentially long right tail. This problem can be resolved using Pareto smoothed importance sampling (PSIS, [Vehtari et al., 2017a](#)). For each fixed model k and data y_i , we fit the generalized Pareto distribution to a set of largest importance ratios $r_{i,k}^s$, and calculate the expected values of the order statistics of the fitted generalized Pareto distribution. These value are used to obtain the smoothed importance weight $w_{i,k}^s$, which is used to replace $r_{i,k}^s$. For details of PSIS, see [Vehtari et al. \(2017a\)](#). PSIS-LOO importance sampling ([Vehtari et al., 2017b](#)) computes the LOO predictive density as

$$\begin{aligned} p(y_i|y_{-i}, M_k) &= \int p(y_i|\theta_k, M_k) \frac{p(\theta_k|y_{-i}, M_k)}{p(\theta_k|y, M_k)} p(\theta_k|y, M_k) d\theta_k \\ &\approx \frac{\sum_{s=1}^S w_{i,k}^s p(y_i|\theta_k^s, M_k)}{\sum_{s=1}^S w_{i,k}^s}. \end{aligned} \quad (3.5)$$

The reliability of the PSIS approximation can be determined by the estimated shape parameter $\hat{k}$ in the generalized Pareto distribution. For the left-out data points where $\hat{k} > 0.7$, [Vehtari et al. \(2017b\)](#) suggest replacing the PSIS approximation of those problematic cases by the exact LOO or k -fold cross-validation.

One potential drawback of LOO is the large variance when the sample size is small. We see in simulations that when the ratio of relative sample size to the effective number of parameters is small, the weighting can be unstable. How to adjust this small sample behavior is left for the future research.

3.4 Pseudo-BMA

In this paper, we also consider an AIC-type weighting using leave-one-out cross-validation. As mentioned in Section 2, these weights estimate the same quantities as [Li and Dunson \(2016\)](#) that use the divergence from the reference model based inference.

To maintain comparability with the given dataset and to get easier interpretation of the differences in scale of effective number of parameters, we define the *expected log pointwise predictive density* (elpd) for a new dataset $\tilde{y}$ as a measure of predictive accuracy of a given model for the n data points taken one at a time ([Gelman et al., 2014](#); [Vehtari et al., 2017b](#)). In model M_k , $\text{elpd}^k = \sum_{i=1}^n \int p_t(\tilde{y}_i) \log p(\tilde{y}_i|y, M_k) d\tilde{y}_i$, where $p_t(\tilde{y}_i)$ denotes the true distribution of future data $\tilde{y}_i$.

Given observed data y and model k , we use LOO to estimate the elpd as

$$\widehat{\text{elpd}}_{\text{loo}}^k = \sum_{i=1}^n \log \hat{p}(y_i|y_{-i}, M_k) = \sum_{i=1}^n \log \left(\frac{\sum_{s=1}^S w_{i,k}^s p(y_i|\theta_k^s, M_k)}{\sum_{s=1}^S w_{i,k}^s} \right).$$

The Pseudo-BMA weighting for model k is defined as

$$w_k = \frac{\exp(\widehat{\text{elpd}}_{\text{loo}}^k)}{\sum_{k=1}^K \exp(\widehat{\text{elpd}}_{\text{loo}}^k)}. \quad (3.6)$$

However, this estimation doesn't take into account the uncertainty resulting from having a finite number of proxy samples from the future data distribution. Taking into account the uncertainty would regularize the weights making them go further away from 0 and 1.

The computed estimate $\widehat{\text{elpd}}_{\text{loo}}^k$ is defined as the sum of n independent components so it is trivial to compute their standard errors by computing the standard deviation of the n pointwise values ([Vehtari and Lampinen, 2002](#)). As in equation 3.5, define

$$\widehat{\text{elpd}}_{\text{loo},i}^k = \log \hat{p}(y_i|y_{-i}, M_k),$$

and then we can calculate

$$\text{se}(\widehat{\text{elpd}}_{\text{loo},i}^k) = \sqrt{\sum_{i=1}^n (\widehat{\text{elpd}}_{\text{loo},i}^k - \widehat{\text{elpd}}_{\text{loo}}^k/n)^2}.$$

A simple modification of weights is to use the log-normal approximation:

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

Finally, the Bayesian bootstrap (BB) can be used to compute uncertainties related to LOO estimation ([Vehtari and Lampinen, 2002](#)). The Bayesian bootstrap ([Rubin,](#)

1981) gives simple non-parametric approximation to the distribution. Having samples of $z_1, \dots, z_n$ from a random variable Z , it is assumed that posterior probabilities for all observed z_i have the distribution $\text{Dirichlet}(1, \dots, 1)$ and values of Z that are not observed have zero posterior probabilities. Thus, each BB replication generates a set of posterior probabilities $\alpha_{1:n}$ for all observed $z_{1:n}$,

$$\alpha_{1:n} \sim \text{Dirichlet}(\overbrace{1, \dots, 1}^n), \quad P(Z = z_i | \alpha) = \alpha_i.$$

This leads to one BB replication of any statistic $\phi(Z)$ that is of interest:

$$\hat{\phi}(Z | \alpha) = \sum_{i=1}^n \alpha_i \phi(z_i).$$

The distribution over all replicated $\hat{\phi}(Z | \alpha)$ (i.e., generated by repeated sampling of α) produces an estimation for $\phi(Z)$.

As the distribution of $\widehat{\text{elpd}}_{\text{loo},i}^k$ is often highly skewed, BB is likely to work better than the Gaussian approximation. In our model weighting, we can define

$$z_i^k = \widehat{\text{elpd}}_{\text{loo},i}^k, \quad i = 1, \dots, n.$$

We sample vectors $(\alpha_{1,b}, \dots, \alpha_{n,b})_{b=1, \dots, B}$ from the Dirichlet $(\overbrace{1, \dots, 1}^n)$ distribution, and compute the weighted means,

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

Then a Bayesian bootstrap sample of w_k with size B is,

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

and the final adjusted weight of model k is,

$$w_k = \frac{1}{B} \sum_{b=1}^B w_{k,b}, \tag{3.7}$$

which we call Pseudo-BMA+ weight.

4 Simulation examples

4.1 Gaussian mixture model

This simple example helps us understand how BMA and stacking behave differently. It also illustrates the importance of the choice of scoring rules when combining distributions.

Suppose the observed data $y = (y_i, i = 1, \dots, n)$ come independently from a normal distribution $N(3.4, 1)$, not known to the data analyst, and there are 8 candidate models, $N(\mu_k, 1)$ with $\mu_k = k$ for $1 \leq k \leq 8$. This is an $\mathcal{M}$ -open problem in that none of the candidates is the true model, and we have set the parameters so that the models are somewhat separate but not completely distinct in their predictive distributions.

For BMA with a uniform prior $\Pr(M_k) = \frac{1}{8}, k = 1, \dots, 8$, we can write the posterior distribution explicitly:

$$\hat{w}_k^{\text{BMA}} = P(M_k|y) = \frac{\exp(-\frac{1}{2} \sum_{i=1}^n (y_i - \mu_k)^2)}{\sum_{k'} \exp(-\frac{1}{2} \sum_{i=1}^n (y_i - \mu_{k'})^2)},$$

from which we see that $\hat{w}_3^{\text{BMA}} \xrightarrow{P} 1$ and $\hat{w}_k^{\text{BMA}} \xrightarrow{P} 0$ for $k \neq 3$ as sample size $n \rightarrow \infty$. Furthermore, for any given n ,

$$\begin{aligned} E_{y \sim N(\mu, 1)}[\hat{w}_k^{\text{BMA}}] &\propto E_y \left(\exp(-\frac{1}{2} \sum_{i=1}^n (y_i - \mu_k)^2) \right) \\ &\propto \left(\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2} ((y - \mu_k)^2 + (y - \mu)^2)\right) dy \right)^n \\ &\propto \exp\left(-\frac{n(\mu_k - \mu)^2}{4}\right), \end{aligned}$$

This example is simple in that there is no parameter to estimate within each of the models: $p(\tilde{y}|y, M_k) = p(\tilde{y}|M_k)$. Hence, in this case the weights from Pseudo-BMA and Pseudo-BMA+ are the same as the BMA weights.

For stacking of means, we need to solve

$$\hat{w} = \arg \min_w \sum_{i=1}^n (y_i - \sum_{k=1}^8 w_k k)^2, \quad s.t. \sum_{k=1}^8 w_k = 1, \quad w_k \geq 0.$$

This is nonidentifiable because the solution contains any vector $\hat{w}$ satisfying

$$\sum_{k=1}^8 \hat{w}_k = 1, \quad \hat{w}_k \geq 0, \quad \sum_{k=1}^8 \hat{w}_k k = \frac{1}{n} \sum_{i=1}^n y_i.$$

For point prediction, the stacked prediction is always $\sum_{k=1}^8 \hat{w}_k k = \frac{1}{n} \sum_{i=1}^n y_i$, but it can lead to different predictive distributions $\sum_{k=1}^8 \hat{w}_k N(k, 1)$. To get one reasonable result, we transform the least squares optimization to the following normal model and assign a uniform prior to w :

$$y_i \sim N\left(\sum_{k=1}^8 w_k k, \sigma^2\right), \quad p(w_1, \dots, w_8, \sigma) = 1.$$

Then we could use the posterior means of w as model weights.

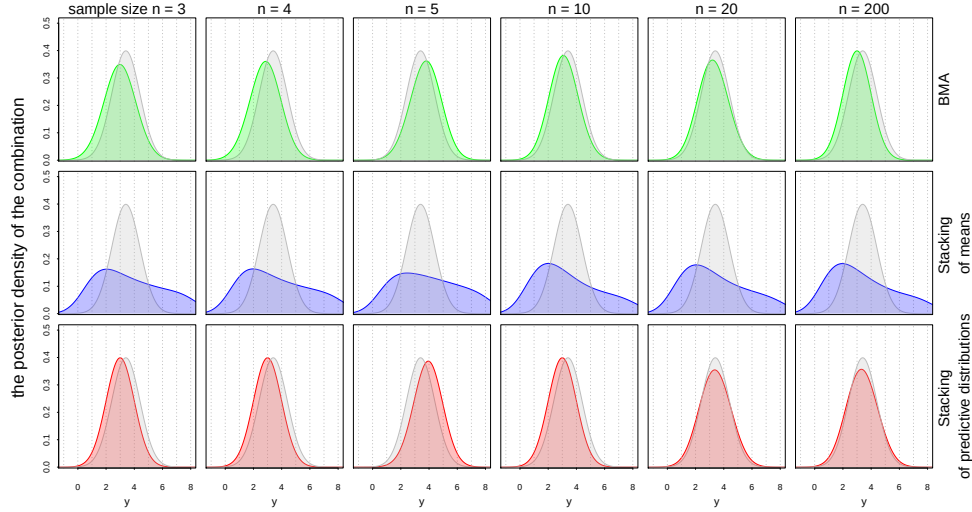


Figure 1: For the Gaussian mixture example, the predictive distribution $p(\tilde{y}|y)$ of BMA (green curve), stacking of means (blue) and stacking of predictive distributions (red). In each graph, the gray distribution represents the true model $N(3.4, 1)$. Stacking of means matches the first moment but can ignore the distribution. For this $\mathcal{M}$ -open problem, stacking of predictive distributions outperforms BMA as sample size increases.

For stacking of predictive distributions, we need to solve

$$\max_w \sum_{i=1}^n \log \left(\sum_{k=1}^8 w_k \exp \left(-\frac{(y_k - k)^2}{2} \right) \right), \quad s.t. \sum_{k=1}^8 w_k = 1, \quad w_k \geq 0$$

In fact, this example is a density estimation problem. [Smyth and Wolpert \(1998\)](#) first applied stacking to non-parametric density estimation, which they called *stacked density estimation*. It can be viewed as a special case of our stacking method.

We compare the posterior predictive distribution $\hat{p}(\tilde{y}|y) = \sum_k \hat{w}_k p(\tilde{y}|y, M_k)$ for these three methods of model averaging. Figure 1 shows the predictive distributions in one simulation when the sample size n varies from 3 to 200. Stacking of means (the middle row of graphs) gives an unappealing predictive distribution, even if its point estimate is reasonable. The broad and oddly spaced distribution here arises from nonidentification of w , and it demonstrates the general point that stacking of means does not even try to match the shape of the predictive distribution. The top and bottom row of graphs show how BMA picks up the single model that is closest in KL divergence, while stacking picks a combination; the benefits of stacking becomes clear for large n .

In this trivial non-parametric case, stacking of predictive distributions is almost the same as fitting a mixture model, except for the absence of the prior. The true model $N(3.4, 1)$ is actually a convolution of single models rather than a mixture, hence no approach can recover the true one from the model list. From Figure 2 we can compare

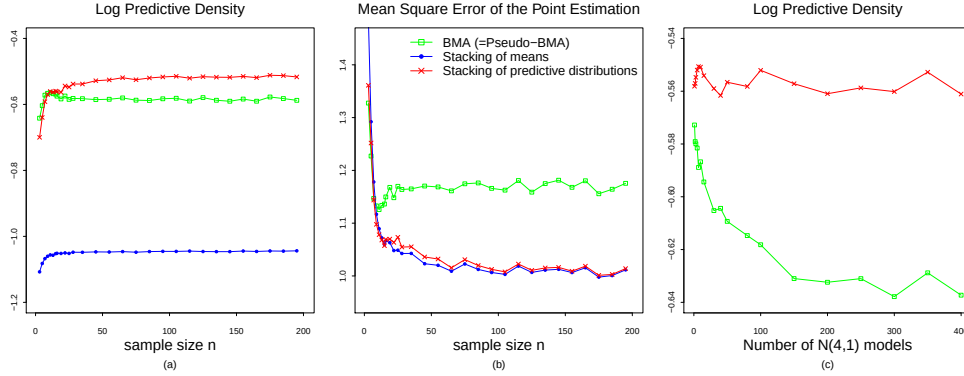


Figure 2: (a) The left panel shows the expected log predictive density of the combined distribution under BMA, stacking of means and stacking of predictive distributions. Stacking of predictive distributions performs best for moderate and large sample sizes. (b) The middle panel shows the mean squared error treating the posterior mean of $\hat{y}$ as a point estimation. Stacking of predictive distributions gives almost the same optimal mean squared error as stacking of means, both of which perform better than BMA. (c) The right panel shows the expected log predictive density of stacking and BMA when adding some more $N(4, 1)$ models to the model list, where sample size is fixed to be 15. All average log scores and errors are calculated through 500 repeated simulation and 200 test data generating from the true distribution.

the mean squared error and the mean logarithmic score of these three combination methods. The log scores and errors are calculated through 500 repeated simulations and 200 test data. The left panel shows the logarithmic score (or equivalent, expected log predictive density) of the predictive distribution. Stacking of predictive distributions always gives a larger score except for extremely small n . In the middle panel, it shows the mean squared error by considering the posterior mean of predictive distribution to be a point estimate, even if it is not our focus. In this case, it is not surprising to see that stacking of predictive distributions gives almost the same optimal mean squared error as the stacking of means, both of which are better than the BMA. Two distributions close in KL divergence are close in each moment, while the reverse does not necessarily hold. This illustrates the necessity of matching the *distributions*, rather than matching the *moments*.

Stacking depends only on the space expanded by all candidate models, while BMA or Pseudo-BMA weighting may be misled by such model expansion. If we add another $N(4, 1)$ as the 9th model in the model list above, stacking will not change at all in theory, even though it becomes non-strictly-convex and has infinite same-height mode. For BMA, it is equivalent to putting double prior mass on the original 4th model, which doubles the final weights for it. The right panel of Figure 2 shows such phenomenon: we fix sample size n to be 15 and add more and more $N(4, 1)$ models. As a result, BMA (or Pseudo-BMA weighting) puts larger weight on $N(4, 1)$ and behaves worse, while the stacking is essentially unchanged. It illustrates another benefit of stacking compared to BMA or Pseudo-BMA weights. If the performance of a combination method decays as

the list of candidate models is expanded, this may indicate disastrous performance if there are many similar weak models on the candidate list. We are not saying BMA can never work in this case. In fact some other methods are proposed to make BMA overcome such drawbacks. For example, [George \(2010\)](#) establishes dilution priors to compensate for model space redundancy for linear models, putting smaller weights on those models that are close to each other. [Fokoue and Clarke \(2011\)](#) introduce prequential model list selection to obtain an optimal model space. But we propose stacking as a more straightforward solution.

4.2 Linear subset regressions

The previous section demonstrates a simple example of combining several different nonparametric models. Now we turn to the parametric case. This example comes from [Breiman \(1996\)](#) who compares stacking to model selection. Suppose the true model is

$$Y = \beta_1 X_1 + \cdots + \beta_J X_J + \epsilon,$$

where $\epsilon \sim N(0, 1)$. All the covariates X_j are independently from $N(5, 1)$. The number of predictors J is 15. The coefficient β is generated by

$$\beta_j = \gamma \left((1_{|j-4|<h}(h - |j-4|)^2 + (1_{|j-8|<h})(h - |j-8|)^2 + (1_{|j-12|<h})(h - |j-12|)^2) \right),$$

where γ is determined by fixing the signal-to-noise ratio such that

$$\frac{\text{Var}(\sum_j \beta_j X_j)}{1 + \text{Var}(\sum_j \beta_j X_j)} = \frac{4}{5}.$$

The value h determines the number of nonzero coefficients in the true model. For $h = 1$, there are 3 “strong” coefficients. For $h = 5$, there are 15 “weak” coefficients. In the following simulation, we fix $h = 5$. We consider the following two cases:

1. $\mathcal{M}$ -open: Each subset contains only one single variable. Hence, the k -th model is a univariate linear regression with the k -th variable X_k . We have $K = J = 15$ different models in total. One advantage of stacking and Pseudo-BMA weighting is that they are not sensitive to prior, hence even a flat prior will work, while BMA can be sensitive to the prior. For each single model $M_k : Y \sim N(\beta_k X_k, \sigma^2)$, we set prior $\beta_k \sim N(0, 10)$, $\sigma \sim \text{Gamma}(0.1, 0.1)$.
2. $\mathcal{M}$ -closed: Let model k be the linear regression with subset $(X_1, \dots, X_k)$. Then there are still $K = 15$ different models. Similarly, in model $M_k : Y \sim N(\sum_{j=1}^k \beta_j X_j, \sigma^2)$, we set prior $\beta_j \sim N(0, 10)$, $j = 1, \dots, k$, $\sigma \sim \text{Gamma}(0.1, 0.1)$.

In both cases, we have seven methods for combining predictive densities: (1) stacking of predictive distributions, (2) stacking of means, (3) Pseudo-BMA, (4) Pseudo-BMA+, (5) best model selection by mean LOO value, (6) best model selection by marginal likelihood, and (7) BMA. We generate a test dataset $(\tilde{x}_i, \tilde{y}_i)$, $i = 1, \dots, 200$ from the underlying true distribution to calculate the out of sample logarithm scores for the combined distribution under each method and repeat the simulation 100 times to compute the expected predictive accuracy of each method.

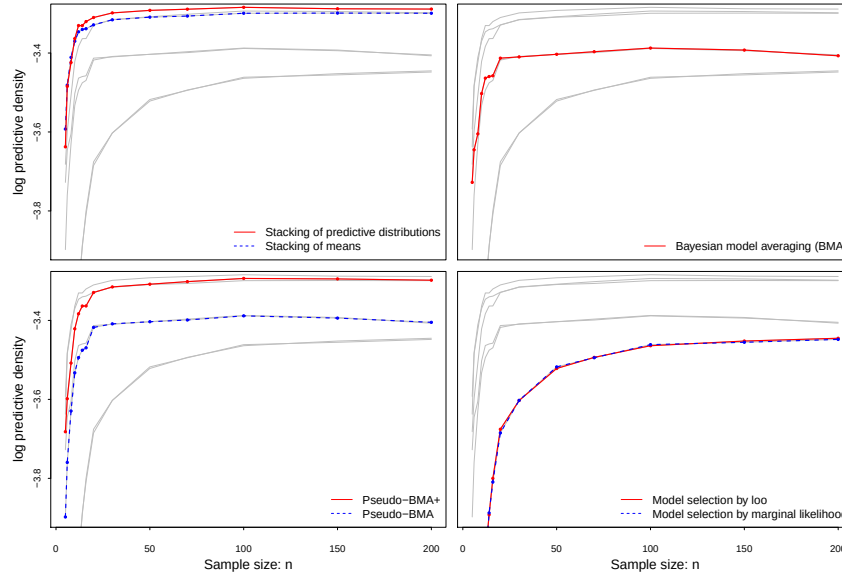


Figure 3: Mean log predictive densities of 7 combination methods in the linear regression example: the k -th model is a univariate regression with the k -th variable ($1 \leq k \leq 15$). We evaluate the log predictive densities using 100 repeated experiments and 200 test data.

Figure 3 shows the expected out-of-sample log predictive densities for the seven methods, for a set of experiments with sample size n ranging from 5 to 200. Stacking outperforms all other methods even for small n . Stacking of predictive distributions is asymptotically better than any other combination method. Pseudo-BMA+ weighting dominates naive Pseudo-BMA weighting. Finally, BMA performs similarly to Pseudo-BMA weighting, always better than any kind of model selection, but that advantage vanishes in the limit since BMA picks up one model. In this $\mathcal{M}$ -open setting, model selection can never be optimal.

The results change when we move to the second case, in which the k -th model contains variables $X_1, \dots, X_k$ so that we are comparing models of differing dimensionality. The problem is $\mathcal{M}$ -closed because the largest subset contains all the variables, and we have simulated data from this model. Figure 4 shows the mean log predictive densities of the seven combination methods in this case. For a large sample size n , almost all methods recover the true model (putting weight 1 on the full model), except BMA and model selection based on marginal likelihood. The poor performance of BMA comes from the parameter priors: recall that the optimality of BMA arises when averaging over the priors and not necessarily conditional on any particular chosen set of parameter values. There is no general rule to obtain a “correct” prior that accounts for the complexity for BMA in an arbitrary model space. Model selection by LOO can recover the true model, while selection by marginal likelihood cannot due to the same prior problems. Once again, BMA eventually becomes the same as model selection by marginal likelihood, which is much worse than any other methods asymptotically.

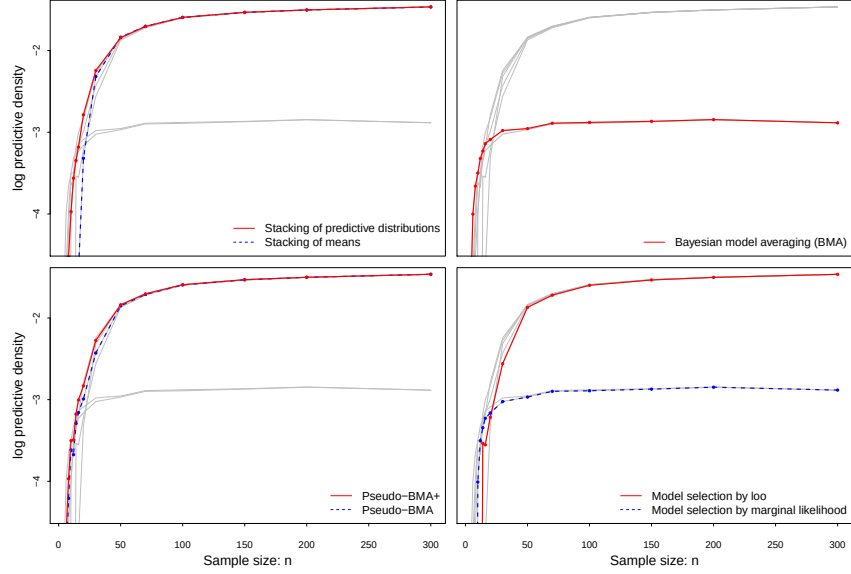


Figure 4: Mean log predictive densities of 7 combination methods in the linear regression example: the k -th model is the regression with the first k variables ($1 \leq k \leq 15$). We evaluate the log predictive densities using 100 repeated experiments and 200 test data.

In this example, stacking is unstable for extremely small n . In fact, our computations for stacking of predictive distributions and Pseudo-BMA depend on the PSIS approximation to $\log p(y_i|y_{-i})$. If this approximation is crude, then the second step optimization cannot be accurate. It is known that the parameter $\hat{k}$ in the generalized Pareto distribution can be used to diagnose the accuracy of PSIS approximation. When $\hat{k} > 0.7$ for a datapoint, we cannot trust the PSIS-LOO estimate and so we re-run the full inference scheme on the dataset with that particular point left out (see [Vehtari et al., 2017b](#)).

Figure 5 shows the comparison of the mean elpd estimated by LOO and the mean elpd calculated using 200 independent test data for each model and each sample size in the simulation described above. The area of each dot in Figure 5 represents the relative complexity of the model as measured by the effective number of parameters divided by sample size. We evaluate the effective number of parameters using LOO ([Vehtari et al., 2017b](#)). The sample size n varies from 30 to 200 and variable size is fixed to be 20. Clearly, the relationship is far from the line $y = x$ for extremely small sample size, and the relative bias ratio ($\text{elpd}_{\text{loo}}/\text{elpd}_{\text{test}}$) depends on model complexity. Empirically, we have found the approximation to be poor when the sample size is less than 5 times the number of parameters. Further diagnostics for PSIS are described by [Vehtari et al. \(2017a\)](#).

As a result, in the small sample case, LOO can lead to relatively large variance, which makes the stacking of predictive distributions and Pseudo-BMA/ Pseudo-BMA+

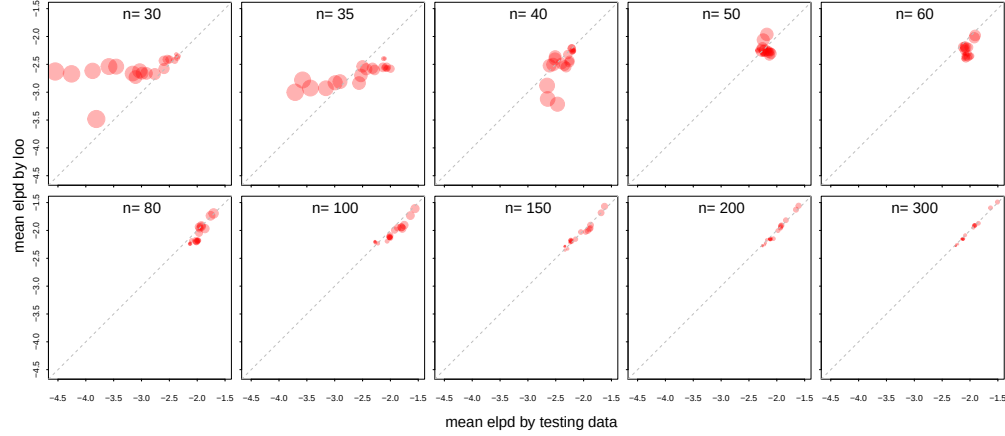


Figure 5: Comparison of the mean elpd estimated by LOO and the mean elpd calculated from test data, for each model and each sample size in the simulation described above. The area of each dot represents the relative complexity of the model as measured by the effective number of parameter divided by sample size.

unstable, with performance improving quickly as n grows.

4.3 Comparison with mixture models

Stacking is inherently a two-step procedure. In contrast, when fitting a mixture model, one estimates the model weights and the status within parameters in the same step. In a mixture model, given a model list $\mathcal{M} = (M_1, \dots, M_K)$, each component in the mixture occurs with probability w_k . Marginalizing out the discrete assignments yields the joint likelihood

$$p(y|w_{1:K}, \theta_{1:K}) = \sum_{k=1}^K w_k p(y|\theta_k, M_k).$$

The mixture model seems to be the most straightforward continuous model expansion. Nevertheless, there are several reasons why we may prefer stacking to fitting a mixture model. Firstly, MCMC methods for mixture models are difficult to implement and generally quite expensive. Secondly, if the sample size is small or several components in the mixture could do the same thing, the mixture model can face non-identification or instability problem unless a strong prior is added.

Figure 6 shows a comparison of mixture models and other model averaging methods in a numerical experiment, in which the true model is

$$Y \sim N(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2, 1), \quad \beta_k \text{ is generated from } N(0, 1),$$

and there are 3 candidate models, each containing one covariate:

$$M_k : Y \sim N(\beta_k X_k, \sigma_k^2), \text{ with a prior } \beta_k \sim N(0, 1), \quad k = 1, 2, 3.$$

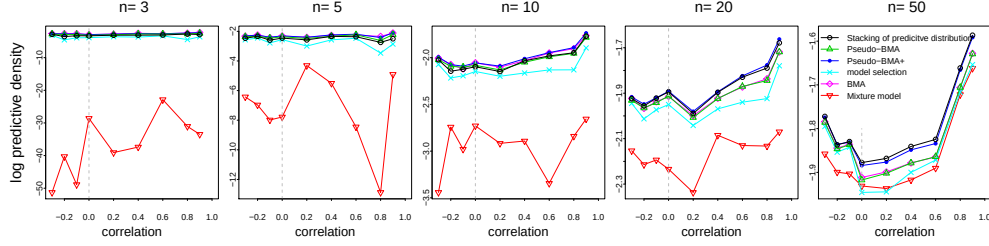


Figure 6: Log predictive densities of the combined distribution obtained by stacking of predictive distributions, BMA, Pseudo-BMA, Pseudo-BMA+, model selection by marginal likelihood, and mixture models. In each case, we evaluate the predictive density by 100 testing data and 100 repeated simulations. The correlation of variables ranges from -0.3 to 0.9 , and sample size ranges from 3 to 50. Stacking of predictive distributions and Pseudo-BMA+ outperform mixture models in all cases.

In the simulation, we generate the design matrix by $\text{Var}(X_i) = 1$ and $\text{Cor}(X_i, X_j) = \rho$. ρ determines how correlated these models are and it ranges from -0.3 to 0.9 .

Figure 6 shows that both the performance of mixture models and single model selection are worse than any other model averaging methods we suggest, even though the mixture model takes much longer time to run (about 30 more times) than stacking or Pseudo-BMA+. When the sample size is small, the mixture model is too complex to fit. On the other hand, stacking of predictive distributions and Pseudo-BMA+ outperform all other methods with a moderate sample size.

Clarke (2003) argues that the effect of (point estimation) stacking only depends on the space spanned by the model list, hence he suggests putting those “independent” models on the list. Figure 6 shows high correlations do not hurt stacking and Pseudo-BMA+ in this example.

4.4 Variational inference with different initial values

In Bayesian inference, the posterior density of parameters $\theta = (\theta_1, \dots, \theta_m)$ given observed data $y = (y_1 \dots y_n)$ can be difficult to compute. Variational inference can be used to give a fast approximation for $p(\theta|y)$ (for a recent review, see Blei et al., 2017). Among a family of distributions $\mathcal{Q}$, we try to find one $q \in \mathcal{Q}$ such that the Kullback-Leibler divergence to the true posterior distribution is minimized:

$$q^* = \arg_{q \in \mathcal{Q}} \min \text{KL}(q(\theta), p(\theta|y)) = \arg_{q \in \mathcal{Q}} \min (\mathbb{E}_q \log q(\theta) - \mathbb{E}_q \log p(\theta, y)). \quad (4.1)$$

One widely used variational family is mean-field family where parameters are assumed to be mutually independent $\mathcal{Q} = \{q(\theta) : q(\theta_1, \dots, \theta_m) = \prod_{j=1}^m q_j(\theta_j)\}$. Some recent progress is made to run variational inference algorithm in a black-box way. For example, Kucukelbir et al. (2017) implement *Automatic Differentiation Variational Inference* in Stan (Stan Development Team, 2017). Assuming all parameters θ are continuous and model likelihood is differentiable, it transforms θ into real coordinate space $\mathbb{R}^m$ through

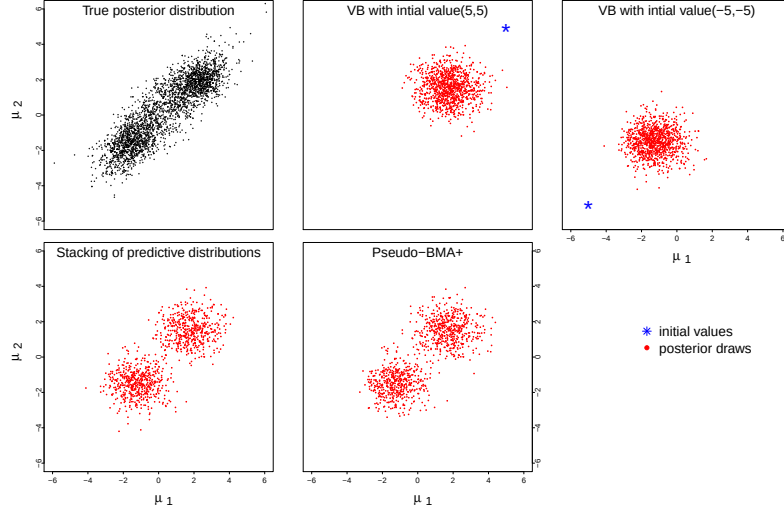


Figure 7: (1) A multi-modal posterior distribution of (μ_1, μ_2) . (2–3) Posterior draws from variational inference with different initial values. (4–5) Averaged posterior distribution using stacking of predictive distributions and Pseudo-BMA+ weighting.

$\zeta = T(\theta)$ and uses normal approximation $p(\zeta|\mu, \sigma^2) = \prod_{j=1}^m N(\zeta_j|\mu_j, \sigma_j^2)$. Plugging this into (4.1) leads to an optimization problem over (μ, σ^2) , which can be solved by stochastic gradient descent. Under some mild condition, it eventually converges to a local optimum q^* . However, q^* may depend on initialization since such optimization problem is in general non-convex, particularly when the true posterior density $p(\theta|y)$ is multi-modal.

Stacking of predictive distributions and Pseudo-BMA+ weighting can be used to average several sets of posterior draws coming from different approximation distributions. To do this, we repeat the variational inference K times. At time k , we start from a random initial point and use stochastic gradient descent to solve the optimization problem (4.1), ending up with an approximation distribution q_k^* . Then we draw S samples $(\theta_k^{(1)}, \dots, \theta_k^{(S)})$ from $q_k^*(\theta)$ and calculate the importance ratio $r_{i,k}^s$ defined in (3.4) as $r_{i,k}^s = 1/p(y_i|\theta_k^{(s)})$. After this, the remaining steps follow as before. We obtain stacking or Pseudo-BMA+ weights w_k and average all approximation distributions as $\sum_{k=1}^K w_k q_k^*$.

Figure 7 gives a simple example that the averaging strategy helps adjust the optimization uncertainty of initial values. Suppose the data is two-dimensional $y = (y^{(1)}, y^{(2)})$ and the parameter is $(\mu_1, \mu_2) \in \mathbb{R}^2$. The likelihood $p(y|\mu_1, \mu_2)$ is given by

$$y^{(1)} \sim \text{Cauchy}(\mu_1, 1), \quad y^{(2)} \sim \text{Cauchy}(\mu_2, 1).$$

A $N(0, 1)$ prior is assigned to $\mu_1 - \mu_2$. We generate two observations $(y_1^{(1)} = 3, y_1^{(2)} = 2)$ and $(y_2^{(1)} = -2, y_2^{(2)} = -2)$. The first panel shows the true posterior distribution of

	Full model		BMA		Stacking of predictive distributions		Pseudo-BMA+ weighting	
	Mean Candidate	Voter- specific	Mean Candidate	Voter- specific	Mean Candidate	Voter- specific	Mean Candidate	Voter- specific
prox. adv.	-3.05 (1.32)	-2.01 (1.06)	-0.22 (0.95)	0.75 (0.68)	0.00 (0.00)	0.00 (0.00)	-0.02 (0.08)	0.04 (0.24)
direct. adv.	7.95 (2.85)	4.18 (1.36)	3.58 (2.02)	2.36 (0.84)	2.56 (2.32)	1.93 (1.16)	1.60 (4.91)	1.78 (1.22)
incumb. adv.	1.06 (1.20)	1.14 (1.19)	1.61 (1.24)	1.30 (1.24)	0.48 (1.70)	0.34 (0.89)	0.66 (1.13)	0.54 (1.03)
quality adv.	3.12 (1.24)	2.38 (1.22)	2.96 (1.25)	2.74 (1.22)	2.20 (1.71)	2.30 (1.52)	2.05 (2.86)	1.89 (1.61)
spend adv.	0.27 (0.04)	0.27 (0.04)	0.32 (0.04)	0.31 (0.04)	0.31 (0.07)	0.31 (0.03)	0.31 (0.04)	0.30 (0.04)
partisan adv.	0.06 (0.05)	0.06 (0.05)	0.08 (0.06)	0.07 (0.06)	0.01 (0.04)	0.00 (0.00)	0.03 (0.05)	0.03 (0.05)
constant	53.3 (1.2)	52.0 (0.8)	51.4 (1.0)	51.6 (0.8)	51.9 (1.1)	51.6 (0.7)	51.5 (1.2)	51.4 (0.8)

Figure 8: Regression coefficients and standard errors in the voting example, from the full model (columns 1–2), the averaged subset regression model using BMA (columns 3–4), stacking of predictive distributions (columns 5–6) and Pseudo-BMA+ (columns 7–8). *Democratic proximity advantage* and *Democratic directional advantage* are two highly correlated variables. *Mean candidate* and *Voter-specific* are two datasets that provide different measurements on candidates’ ideological placement.

$\mu = (\mu_1, \mu_2)$, which is bimodal. We run mean-field normal variational inference in Stan, with two initial values to be $(\mu_1, \mu_2) = (5, 5)$ and $(-5, -5)$ separately. This produces two distinct approximation distributions as shown in panel 2 and 3. We then draw 1000 samples each from these two distributions and use stacking or Pseudo-BMA+ to combine them. The lower 2 panels show the averaged posterior distributions. Though neither can recover the true distribution, the averaged version is closer to it.

4.5 Proximity and directional models of voting

Adams et al. (2004) use US Senate voting data from 1988 to 1992 to study voters’ preference for the candidates who propose policies that are similar to their political beliefs. They introduce two similar variables that indicate the distance between voters and candidates. *Proximity voting comparison* represents the i -th voter’s comparison between the candidates’ ideological positions:

$$U_i(D) - U_i(R) = (x_R - x_i)^2 - (x_D - x_i)^2,$$

where x_i represents the i -th voter’s preferred ideological position, and x_D and x_R represent the ideological positions of the Democratic and Republican candidates, respectively. In contrast, the i -th voter’s *directional comparison* is defined by

$$U_i(D) - U_i(R) = (x_D - X_N)(x_i - X_N) - (x_R - X_N)(x_i - X_N),$$

where X_N is the neutral point of the ideology scale.

Finally, all these comparison is aggregated in the party level, leading to two party-level variable *Democratic proximity advantage* and *Democratic directional advantage*. The sample size is $n = 94$.

For both of these two variables, there are two ways to measure candidates’ ideological positions x_D and x_R , which lead to two different datasets. In the *Mean candidate* dataset, they are calculated by taking the average of all respondents’ answers in the relevant

state and year. In the *Voter-specific* dataset, they are calculate by using respondents' own placements of the two candidates. In both datasets, there are 4 other party-level variables.

The two variables *Democratic proximity advantage* and *Democratic directional advantage* are highly correlated. [Montgomery and Nyhan \(2010\)](#) point out that Bayesian model averaging is an approach to helping arbitrate between competing predictors in a linear regression model. They average over all 2^6 linear subset models excluding those containing both variables *Democratic proximity advantage* and *Democratic directional advantage*, (i.e., 48 models in total). Each subset regression is with the form

$$M_\gamma : y | (X, \beta_0, \beta_\gamma) \sim N(\beta_0 + X_\gamma \beta_\gamma, \sigma^2).$$

Accounting for the different complexity, they used the hyper- g prior ([Liang et al., 2008](#)). Let ϕ to be the inverse of the variance $\phi = \frac{1}{\sigma^2}$. The hyper- g prior with a hyper-parameter α is,

$$\begin{aligned} p(\phi) &\propto \frac{1}{\phi}, \\ \beta | (g, \phi, X) &\sim N\left(0, \frac{g}{\phi} (X^T X)^{-1}\right), \\ p(g|\alpha) &= \frac{\alpha - 2}{2} (1 + g)^{-\alpha/2}, \quad g > 0. \end{aligned}$$

The first two columns of Figure 8 show the linear regression coefficients as estimated using least squares. The remaining columns show the posterior mean and standard error of the regression coefficients using BMA, stacking of predictive distributions, and Pseudo-BMA+ respectively. Under all three averaging strategies, the coefficient of *proximity advantage* is no longer statistically significantly negative, and the coefficient of *directional advantage* is shrunk. As fit to these data, stacking puts near-zero weights on all subset models containing *proximity advantage*, whereas Pseudo-BMA+ weighting always gives some weight to each model. In this example, averaging subset models by stacking or Pseudo-BMA+ weighting gives a way to deal with competing variables, which should be more reliable than BMA according to our previous argument.

4.6 Predicting well-switching behavior in Bangladesh

Many wells in Bangladesh and other South Asian countries are contaminated with natural arsenic. People whose wells have arsenic levels that exceed a certain threshold are encouraged to switch to nearby safe wells (for background details, see [Gelman and Hill \(2006, Chapter 5.4\)](#)). We are analyzing a dataset including 3020 respondents to find factors predictive of the well switching. The outcome variable is

$$y_i = \begin{cases} 1, & \text{if household } i \text{ switched to a safe well.} \\ 0, & \text{if household } i \text{ continued using its own well.} \end{cases}$$

And we consider following input variables:

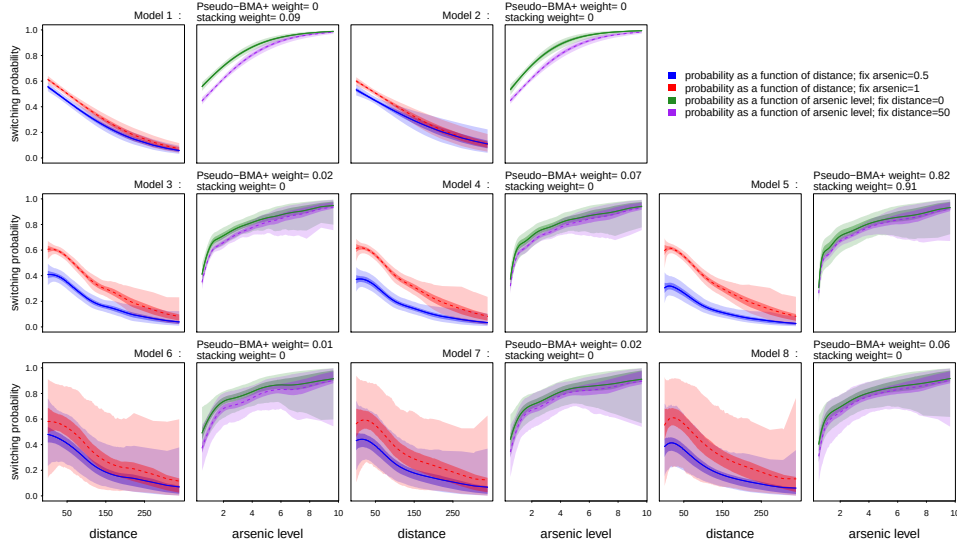


Figure 9: The posterior mean, 50% and 95% confidence interval of the well switching probability in Models 1–8. For each model, the switching probability is shown as a function of (a) the distance to the nearest safe well or (b) the arsenic level of the existing well. In each subplot, other input variables are held constant. The model weights by stacking of predictive distributions and Pseudo-BMA+ are printed above each panel.

- **dist**: the distance (in meters) to the closest known safe well,
- **arsenic**: the arsenic level (in 100 micrograms per liter) of the respondent’s well,
- **assoc**: whether a member of the household is active in any community association,
- **educ**: the education level of the head of the household.

We start with what we call Model 1, a simple logistic regression with all variables above as well as a constant term,

$$y \sim \text{Bernoulli}(\theta),$$

$$\theta = \text{logit}^{-1}(\beta_0 + \beta_1 \text{dist} + \beta_2 \text{arsenic} + \beta_3 \text{assoc} + \beta_4 \text{educ}).$$

Model 2 contains the interaction between distances and arsenic levels,

$$\theta = \text{logit}^{-1}(\beta_0 + \beta_1 \text{dist} + \beta_2 \text{arsenic} + \beta_3 \text{assoc} + \beta_4 \text{educ} + \beta_5 \text{dist} \times \text{arsenic}).$$

Furthermore, we can use spline to capture the nonlinear relational between the logit switching probability and the distance or the arsenic level. Model 3 contains the B-splines for the distance and the arsenic level with polynomial degree 2,

$$\theta = \text{logit}^{-1}(\beta_0 + \beta_1 \text{dist} + \beta_2 \text{arsenic} + \beta_3 \text{assoc} + \beta_4 \text{educ} + \alpha_{dis} B_{dis} + \alpha_{ars} B_{ars}),$$

where B_{dis} is the B-spline basis of distance with the form $(B_{dis,1}(\text{dist}), \dots, B_{dis,q}(\text{dist}))$ and $\alpha_{dis}, \alpha_{ars}$ are vectors. We also fix the number of spline knots to be 10. Model 4 and 5 are the similar models with 3-degree and 5-degree B-splines, respectively.

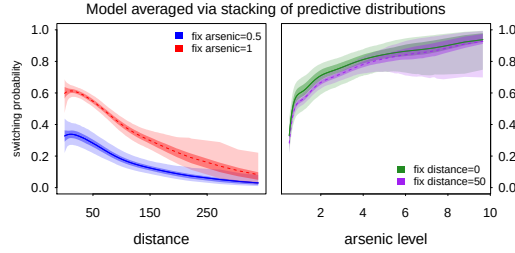


Figure 10: The posterior mean, 50% and 95% confidence interval of the well switching probability in the combined model via stacking of predictive distributions. Pseudo-BMA+ weighting gives a similar result for the combination.

Next, we can add a bivariate spline to capture nonlinear interactions,

$$\theta = \text{logit}^{-1}(\beta_0 + \beta_1 \text{dist} + \beta_2 \text{arsenic} + \beta_3 \text{assoc} + \beta_4 \text{educ} + \beta_5 \text{dist} \times \text{arsenic} + \alpha B_{\text{dis,ars}}),$$

where $B_{\text{dis,ars}}$ is the bivariate spline basis with the degree to be 2×2 , 3×3 and 5×5 in Model 6, 7 and 8 respectively.

Figure 9 shows the inference results in all 8 models, which are summarized by the posterior mean, 50% confidence interval and 95% confidence interval of the probability of switching from an unsafe well as a function of the distance or the arsenic level. Any other variables `assoc` and `educ` are fixed at their means. It is not obvious from these results which one is the best model. Spline models give a more flexible shape, but also introduce more variance for posterior estimation.

Finally, we run stacking of predictive distributions and Pseudo-BMA+ to combine these 8 models. The calculated model weights are printed above each panel in Figure 9. For both combination methods, Model 5 (univariate splines with degree 5) accounts for the majority share. Model 8 is the most complicated one, but both stacking and Pseudo-BMA+ avoid overfitting by assigning it a negligible weight.

Figure 10 shows the posterior mean, 50% confidence interval, and 95% confidence interval of the switching probability in the stacking-combined model. Pseudo-BMA+ weighting gives a similar combination result for this example. At first glance, the combination looks quite similar to Model 5, while it may not seem necessary to put an extra 0.09 weight on Model 1 in stacking combination since Model 1 is completely contained in Model 5 if setting $\alpha_{\text{dis}} = \alpha_{\text{ars}} = 0$. However, Model 5 is not perfect since it predicts that the posterior mean of switching probability will decrease as a function of the distance to the nearest safe well, for small distances. In fact, without further control, it is not surprising to find boundary fluctuation as a main drawback for higher order splines. This decreasing trend around the left boundary is flatter in the combined distribution since the combination contains part of straightforward logistic regression (in stacking weights) or lower order splines (in Pseudo-BMA+ weights). In this example the sample size $n = 3020$ is large, hence we have reasons to believe stacking of predictive distributions gives the optimal combination.

5 Discussion

5.1 Sparse structure and high dimensions

Yang and Dunson (2014) propose to combine multiple point forecasts, $f = \sum_{k=1}^K w_k f_k$, through using a Dirichlet aggregation prior, $w \sim \text{Dirichlet}(\frac{\alpha}{K^\gamma}, \dots, \frac{\alpha}{K^\gamma})$, and the adaptive regression. Their goal is to impose the sparsity structure (certain models can receive zero weights). They show their combination algorithm can achieve the minimax squared risk among all convex combinations,

$$\sup_{f_1, \dots, f_K \in F_0} \inf_{\hat{f}} \sup_{f_\lambda^* \in F_\Gamma} E \|\hat{f} - f_\lambda^*\|^2,$$

where $F_0 = (f : \|f\|_\infty \leq 1)$.

The stacking method can also adapt to sparsity through stronger regularizations. When the dimension of model space is high, we can use a hierarchical prior on w in estimation (3.2) to pull toward sparsity if that is desired.

5.2 Constraints and regularity

In point estimation stacking, the simplex constraint is the most widely used regularization so as to overcome potential problems with multicollinearity. Clarke (2003) suggests relaxing the constraint to make it more flexible.

When combining distributions, there is no need to worry about multicollinearity except in degenerate cases. But in order to guarantee a meaningful posterior predictive density, the simplex constraint becomes natural, which is satisfied automatically in BMA and Pseudo-BMA weighting. As mentioned in the previous section, stronger priors can be added.

Another assumption is that the separate posterior distributions are combined linearly. There could be gains from going beyond convex linear combinations. For instance, in the subset regression example when each individual model is a univariate regression, the true model distribution is a convolution instead of a mixture of each possible models distribution. Both of them lead to the additive model in the point estimation, so stacking of the means is always valid, while stacking of predictive distributions is not possible to recover the true model in the convolution case.

Our explanation is that when the model list is large, the convex span should be large enough to approximate the true model. And this is the reason why we prefer adding stronger priors to make the estimation of weights stable in high dimensions.

5.3 General recommendations

The methods discussed in this paper are all based on the idea of fitting models separately and then combining the estimated predictive distributions. This approach is limited in that it does not pool information between the different model fits: as such, it is only

ideal when the K different models being fit have nothing in common. But in that case we would prefer to fit a larger super-model that includes the separate models as special cases, perhaps using an informative prior distribution to ensure stability in inferences.

That said, in practice it is common for different sorts of models to be set up without any easy way to combine them, and in such cases it is necessary from a Bayesian perspective to somehow aggregate their predictive distributions. The often-recommended approach of Bayesian model averaging can fail catastrophically in that the required Bayes factors can depend entirely on arbitrary specifications of noninformative prior distributions. Stacking is a more promising general method in that it is directly focused on performance of the combined predictive distribution. Based on our theory, simulations, and examples, we recommend stacking (of predictive distributions) for the task of combining separately-fit Bayesian posterior predictive distributions. As an alternative, Pseudo-BMA+ is computationally cheaper and can serve as an initial guess for stacking. The computations can be done in R and Stan, and the optimization required to compute the weights connects directly to the predictive task.

Appendix A. Implementation in Stan and R

For n observations and S posterior draws in each model, the $S \times n$ matrix of log likelihood values, $(\log p(y_i|\theta_s))_{1 \leq s \leq S, 1 \leq i \leq n}$, can be computed from the generated quantities block in a Stan program. The leave-one-out likelihood is obtained through the approach of [Vehtari et al. \(2017b\)](#). For the example in Section 4.2, the k -th model is a linear regression with the k -th covariate. We save the following Stan code in the file `regression.stan`:

```
data {
  int n; // number of observations
  int p; // number of variables
  vector[n] y;
  matrix[n, p] X;
}
parameters {
  vector[p] beta;
  real<lower=0> sigma;
}
transformed parameters {
  vector[n] theta;
  theta = X * beta;
}
model {
  y ~ normal(theta, sigma);
  beta ~ normal(0, 10);
  sigma ~ gamma(0.1, 0.1);
}
generated quantities { //to generate log likelihood matrices,
  vector[n] log_lik;
  for (i in 1:n)
    log_lik[i] = normal_lpdf(y[i] | theta[i], sigma);
}
```

In R we can simulate the log likelihood matrices from all K models and save them as a list:

```
library("rstan")
log_lik_list <- list()
for (k in 1:K){
  # Fit the k-th model with Stan
  fit <- stan("regression.stan", data=list(y=y, X=X[,k], n=length(y), p=1))
  log_lik_list[[k]] <- extract(fit)[["log_lik"]]
}
```

The function `model_weights` in the `loo` package¹ in R can give model combination

¹ The R package `loo` implements leave-one-out cross-validation using Pareto Smoothed Importance Sampling. It can be downloaded from <https://github.com/stan-dev/loo/tree/yuling-stacking>.

weights according to stacking of predictive distributions, Pseudo-BMA and Pseudo-BMA+ weighting. We can choose whether to use the Bayesian bootstrap to make further regularization for Pseudo-BMA, if computation time is not a concern.

```
model_weights_1 <- model_weights(log_lik_list, method="stacking")
model_weights_2 <- model_weights(log_lik_list, method="pseudobma", BB=TRUE)
```

in one simulation with six models to combine, the output gives us the computed weights under each approach:

The stacking weights are:

```
[1,] "Model 1" "Model 2" "Model 3" "Model 4" "Model 5" "Model 6"
[2,] "0.25"    "0.06"    "0.09"    "0.25"    "0.35"    "0.00"
```

The Pseudo-BMA+ weights using the Bayesian bootstrap are:

```
[1,] "Model 1" "Model 2" "Model 3" "Model 4" "Model 5" "Model 6"
[2,] "0.28"    "0.05"    "0.08"    "0.30"    "0.28"    "0.00"
```

For reasons discussed in the paper, we generally recommend stacking for combining separate Bayesian predictive distributions.

References

- Adams, J., Bishin, B. G., and Dow, J. K. (2004). “Representation in Congressional Campaigns: Evidence for Discounting/Directional Voting in U.S. Senate Elections.” *Journal of Politics*, 66(2): 348–373. [20](#)
- Akaike, H. (1978). “On the likelihood of a time series model.” *The Statistician*, 217–235. [4](#)
- Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons. [2](#)
- Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). “Variational inference: A review for statisticians.” *Journal of the American Statistical Association*, 112(518): 859–877. [18](#)
- Breiman, L. (1996). “Stacked regressions.” *Machine Learning*, 24(1): 49–64. [3](#), [14](#)
- Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer, 2nd edition. [4](#)
- Clarke, B. (2003). “Comparing Bayes model averaging and stacking when model approximation error cannot be ignored.” *Journal of Machine Learning Research*, 4: 683–712. [3](#), [18](#), [24](#)
- Clyde, M. and Iversen, E. S. (2013). “Bayesian model averaging in the M-open framework.” In Damien, P., Dellaportas, P., Polson, N. G., and Stephens, D. A. (eds.), *Bayesian Theory and Applications*, 483–498. Oxford University Press. [2](#), [3](#), [7](#)

- Fokoue, E. and Clarke, B. (2011). “Bias-variance trade-off for prequential model list selection.” *Statistical Papers*, 52(4): 813–833. [14](#)
- Geisser, S. and Eddy, W. F. (1979). “A Predictive Approach to Model Selection.” *Journal of the American Statistical Association*, 74(365): 153–160. [4](#)
- Gelfand, A. E. (1996). “Model determination using sampling-based methods.” In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.), *Markov Chain Monte Carlo in Practice*, 145–162. Chapman & Hall. [4](#)
- Gelman, A. (2004). “Parameterization and Bayesian modeling.” *Journal of the American Statistical Association*, 99(466): 537–545. [1](#)
- Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multi-level/Hierarchical Models*. Cambridge University Press. [21](#)
- Gelman, A., Hwang, J., and Vehtari, A. (2014). “Understanding predictive information criteria for Bayesian models.” *Statistics and Computing*, 24(6): 997–1016. [9](#)
- George, E. I. (2010). “Dilution priors: Compensating for model space redundancy.” In *Borrowing Strength: Theory Powering Applications A Festschrift for Lawrence D. Brown*, 158–165. Institute of Mathematical Statistics. [14](#)
- Geweke, J. and Amisano, G. (2011). “Optimal prediction pools.” *Journal of Econometrics*, 164(1): 130–141. [4](#)
- (2012). “Prediction with misspecified models.” *American Economic Review*, 102(3): 482–486. [4](#)
- Gneiting, T. and Raftery, A. E. (2007). “Strictly proper scoring rules, prediction, and estimation.” *Journal of the American Statistical Association*, 102(477): 359–378. [1](#), [6](#), [7](#)
- Gutiérrez-Peña, E. and Walker, S. G. (2005). “Statistical decision problems and Bayesian nonparametric methods.” *International Statistical Review*, 73(3): 309–330. [5](#)
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). “Bayesian model averaging: A tutorial.” *Statistical Science*, 14(4): 382–401. [2](#), [3](#)
- Key, J. T., Pericchi, L. R., and Smith, A. F. M. (1999). “Bayesian model choice: What and why.” *Bayesian Statistics*, 6: 343–370. [2](#), [7](#)
- Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). “Automatic differentiation variational inference.” *Journal of Machine Learning Research*, 18(1): 430–474. [18](#)
- Le, T. and Clarke, B. (2017). “A Bayes interpretation of stacking for M-complete and M-open settings.” *Bayesian Analysis*, 12(3): 807–829. [4](#), [7](#)
- LeBlanc, M. and Tibshirani, R. (1996). “Combining estimates in regression and classification.” *Journal of the American Statistical Association*, 91(436): 1641–1650. [3](#)

- Li, M. and Dunson, D. B. (2016). “A framework for probabilistic inferences from imperfect models.” *ArXiv e-prints:1611.01241*. 5, 9
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). “Mixtures of g priors for Bayesian variable selection.” *Journal of the American Statistical Association*, 103(481): 410–423. 21
- Madigan, D., Raftery, A. E., Volinsky, C., and Hoeting, J. (1996). “Bayesian model averaging.” In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models*, 77–83. 2
- Merz, C. J. and Pazzani, M. J. (1999). “A principal components approach to combining regression estimates.” *Machine Learning*, 36(1-2): 9–32. 3
- Montgomery, J. M. and Nyhan, B. (2010). “Bayesian model averaging: Theoretical developments and practical applications.” *Political Analysis*, 18(2): 245–270. 21
- Piironen, J. and Vehtari, A. (2017). “Comparison of Bayesian predictive methods for model selection.” *Statistics and Computing*, 27(3): 711–735. 1
- Rubin, D. B. (1981). “The Bayesian bootstrap.” *Annals of Statistics*, 9(1): 130–134. 9
- Smyth, P. and Wolpert, D. (1998). “Stacked density estimation.” In *Advances in Neural Information Processing Systems*, 668–674. 12
- Stan Development Team (2017). *Stan modeling language: User’s guide and reference manual*. Version 2.16.0, <http://mc-stan.org/>. 18
- Stone, M. (1977). “An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike’s Criterion.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 44–47. 4
- Ting, K. M. and Witten, I. H. (1999). “Issues in stacked generalization.” *Journal of Artificial Intelligence Research*, 10: 271–289. 3
- Vehtari, A., Gelman, A., and Gabry, J. (2017a). “Pareto smoothed importance sampling.” *ArXiv e-print:1507.02646*. 4, 8, 16
- (2017b). “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.” *Statistics and Computing*, 27(5): 1413–1432. 4, 8, 9, 16, 26
- Vehtari, A. and Lampinen, J. (2002). “Bayesian model assessment and comparison using cross-validation predictive densities.” *Neural Computation*, 14(10): 2439–2468. 9
- Vehtari, A. and Ojanen, J. (2012). “A survey of Bayesian predictive methods for model assessment, selection and comparison.” *Statistics Surveys*, 6: 142–228. 1, 4, 5, 8
- Wagenmakers, E.-J. and Farrell, S. (2004). “AIC model selection using Akaike weights.” *Psychonomic bulletin & review*, 11(1): 192–196. 4
- 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. 4
- Wolpert, D. H. (1992). “Stacked generalization.” *Neural Networks*, 5(2): 241–259. 3

Wong, H. and Clarke, B. (2004). “Improvement over Bayes prediction in small samples in the presence of model uncertainty.” *Canadian Journal of Statistics*, 32(3): 269–283.

[3](#)

Yang, Y. and Dunson, D. B. (2014). “Minimax Optimal Bayesian Aggregation.” *ArXiv e-prints:1403.1345*. [3](#), [24](#)

Acknowledgments

We thank the U.S. National Science Foundation, Institute for Education Sciences, Office of Naval Research, and Defense Advanced Research Projects Administration for partial support of this work. We also thank the Editor, Associate Editor, and two anonymous referees for their valuable comments that helped strengthen this manuscript.