Bayesian Analysis (2018) **13**, Number 3, pp. 917–1007

Using Stacking to Average Bayesian Predictive Distributions (with Discussion)

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, 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 $\mathcal{M} = (M_1, \ldots, M_K)$. 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.

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

[*]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

 https://doi.org/10.1214/17-BA1091

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. We provide the R and Stan code in the Suplement material (Yao et al., 2018).

2 Existing approaches

In Bayesian model comparison, the relationship between the true data generator and the model list $\mathcal{M} = (M_1, \ldots, M_K)$ 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 $y = (y_1, \ldots, y_n)$ 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 $((x_i, y_i), i = 1, \ldots, n)$ 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 $\hat{f}_k^{(-i)}(x_i) = \mathrm{E}[y_i|\hat{\theta}_{k,y_{-i}}, M_k]$ 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

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

Breiman (1996) claims that either a positive constraint $w_k \geq 0, k = 1, \ldots K$, 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,

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

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

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 $\hat{L}$, $\text{AIC} = -2\log\hat{L} + 2k$. Akaike (1978) proposed to use $\exp(-\frac{1}{2}\text{AIC})$ 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

$$w_k = \frac{\exp\big(-n\mathrm{KL}(f^*, f_k)\big)}{\sum_{k=1}^{K} \exp\big(-n\mathrm{KL}(f^*, f_k)\big)}. \tag{2}$$

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 $\mathrm{KL}(f^*, f_k)$ can be estimated by its posterior mean,

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

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

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

Here, $\widetilde{\mathrm{KL}}_1$ 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 $\widetilde{\mathrm{KL}}_1$ is often computationally simpler than $\widetilde{\mathrm{KL}}_2$.

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

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

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), leaving

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

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 8 in Section 3.4, this method could be called Reference-Pseudo-BMA.

3 Theory and methods

We label the classical stacking procedure (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 $S : \mathcal{P} \times \Omega \to \bar{\mathbb{R}} = [\infty, \infty]$ 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 $d : \mathcal{P} \times \mathcal{P} \to (0, \infty)$ as $d(P, Q) = S(Q, Q) - S(P, Q)$. For continuous variables, some popularly used scoring rules include:

1. *Quadratic score:* $\mathrm{QS}(p, y) = 2p(y) - ||p||_2^2$ with the divergence $d(p, q) = ||p - q||_2^2$.
2. *Logarithmic score:* $\mathrm{LogS}(p, y) = \log(p(y))$ with $d(p, q) = \mathrm{KL}(q, p)$. The logarithmic score is the only proper local score assuming regularity conditions.
3. *Continuous-ranked probability score:* $\mathrm{CRPS}(F, y) = -\int_{\mathbb{R}} (F(y') - 1(y' \geq y))^2 dy'$ 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:* $\mathrm{ES}(P, y) = \frac{1}{2}\mathrm{E}_P ||Y - Y'||_2^\beta - \mathrm{E}_P ||Y - y||_2^\beta$, where Y and Y' are two independent random variables from distribution P. When $\beta = 2$, this becomes $\mathrm{ES}(P, y) = -||\mathrm{E}_P(Y) - y||^2$. 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 $\mathcal{M} = (M_1, \ldots, M_K)$ is to find the distribution in the convex hull $\mathcal{C} = \{\sum_{k=1}^K w_k \times p(\cdot|M_k) : \sum_k w_k = 1, w_k \geq 0\}$ 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

$$\min_{w \in \mathcal{S}_1^K} d\Big(\sum_{k=1}^K w_k p(\cdot|y, M_k), p_t(\cdot|y)\Big) \text{ or } \max_{w \in \mathcal{S}_1^K} S\Big(\sum_{k=1}^K w_k p(\cdot|y, M_k), p_t(\cdot|y)\Big), \tag{3}$$

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) 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

$$\max_{w \in \mathcal{S}_1^K} \frac{1}{n} \sum_{i=1}^{n} S\Big(\sum_{k=1}^{K} w_k \hat{p}_{k,-i}, y_i\Big). \tag{4}$$

The stacked estimate of the predictive density is

$$\hat{p}(\tilde{y}|y) = \sum_{k=1}^{K} \hat{w}_k p(\tilde{y}|y, M_k). \tag{5}$$

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 (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) 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 (4) 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 $n \to \infty$, the following asymptotic limit holds:

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

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 (4) 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

$$r_{i,k}^s = \frac{1}{p(y_i|\theta_k^s, M_k)} \propto \frac{p(\theta_k^s|y_{-i}, M_k)}{p(\theta_k^s|y, M_k)}. \tag{6}$$

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

$$\begin{aligned} p(y_i|y_{-i}, M_k) &= \int p(y_i|\theta_k, M_k)\frac{p(\theta_k|y_{-i}, M_k)}{p(\theta_k|y, M_k)}p(\theta_k|y, M_k)d\theta_k \\ &\approx \frac{\sum_{s=1}^S w_{i,k}^s p(y_i|\theta_k^s, M_k)}{\sum_{s=1}^S w_{i,k}^s}. \end{aligned} \tag{7}$$

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

$$w_k = \frac{\exp(\widehat{\text{elpd}}_{\text{loo}}^k)}{\sum_{k=1}^K \exp(\widehat{\text{elpd}}_{\text{loo}}^k)}. \tag{8}$$

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 (7), 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 $z_1, \ldots, z_n$ from a random variable Z, it is assumed that posterior probabilities for all observed z_i have the distribution Dirichlet$(1, \ldots, 1)$ 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 $z_{1:n}$,

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

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

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

We sample vectors $(\alpha_{1,b}, \ldots, \alpha_{n,b})_{b=1,\ldots,B}$ from the Dirichlet $(\overbrace{1, \ldots, 1}^{n})$ 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,

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

and the final adjusted weight of model k is,

$$w_k = \frac{1}{B} \sum_{b=1}^{B} w_{k,b}, \tag{9}$$

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 $y = (y_i, i = 1, \ldots, n)$ come independently from a normal distribution N$(3.4, 1)$, not known to the data analyst, and there are 8 candidate

models, $\mathrm{N}(\mu_k, 1)$ with $\mu_k = k$ for $1 \le k \le 8$. 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 $\Pr(M_k) = \frac{1}{8}, k = 1, \ldots, 8$, we can write the posterior distribution explicitly:

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

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

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

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

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

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

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

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 $\sum_{k=1}^8 \hat{w}_k \mathrm{N}(k, 1)$. To get one reasonable result, we transform the least squares optimization to the following normal model and assign a uniform prior to w:

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

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

For stacking of predictive distributions, we need to solve

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

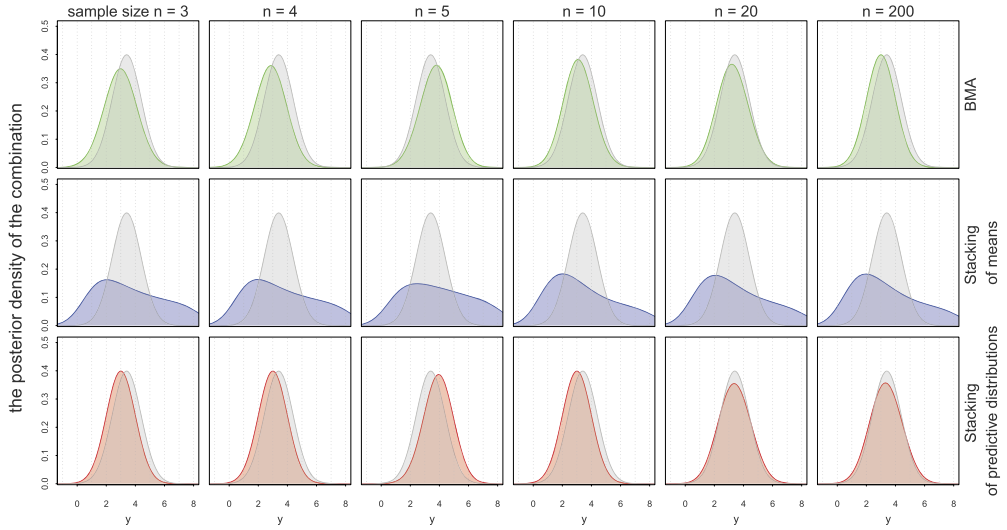

Figure 1: For the Gaussian mixture example, the predictive distribution $p(\tilde{y}|y)$ of BMA (green curve), stacking of means (blue) and stacking of predictive distributions (red). In each graph, the gray distribution represents the true model $\mathrm{N}(3.4, 1)$. Stacking of means matches the first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

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 $\mathrm{N}(3.4, 1)$ 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 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

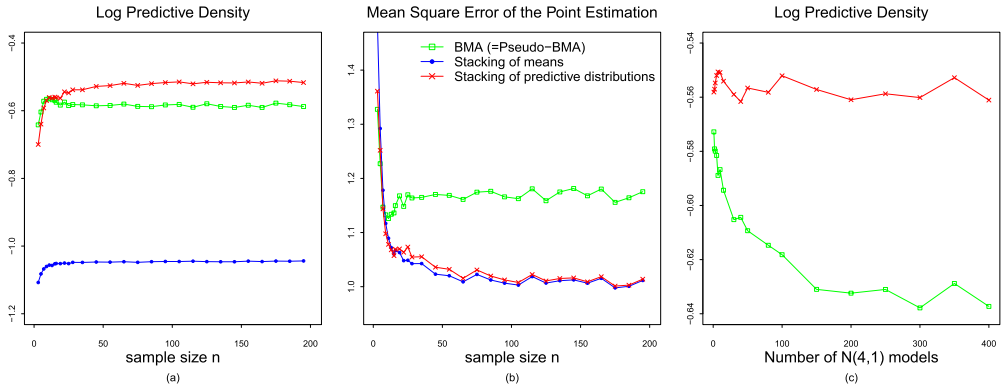

Figure 2: (a) The left panel shows the expected log predictive density of the combined distribution under BMA, stacking of means and stacking of predictive distributions. Stacking of predictive distributions performs best for moderate and large sample sizes. (b) The middle panel shows the mean squared 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 $\mathrm{N}(4,1)$ 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.

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 by misled by such model expansion. If we add another $\mathrm{N}(4,1)$ 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 $\mathrm{N}(4,1)$ models. As a result, BMA (or Pseudo-BMA weighting) puts larger weight on $\mathrm{N}(4,1)$ 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 non-parametric 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 $\epsilon \sim \mathrm{N}(0, 1)$. All the covariates X_j are independently from $\mathrm{N}(5, 1)$. 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

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

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 $M_k : Y \sim \mathrm{N}(\beta_k X_k, \sigma^2)$, we set prior $\beta_k \sim N(0, 10)$, $\sigma \sim \mathrm{Gamma}(0.1, 0.1)$.
2. $\mathcal{M}$-closed: Let model k be the linear regression with subset $(X_1, \ldots, X_k)$. Then there are still $K = 15$ different models. Similarly, in model $M_k : Y \sim \mathrm{N}(\sum_{j=1}^{k} \beta_j X_j, \sigma^2)$, we set prior $\beta_j \sim N(0, 10), j = 1, \ldots, k$, $\sigma \sim \mathrm{Gamma}(0.1, 0.1)$.

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 $(\tilde{x}_i, \tilde{y}_i)$, $i = 1, \ldots, 200$ 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.

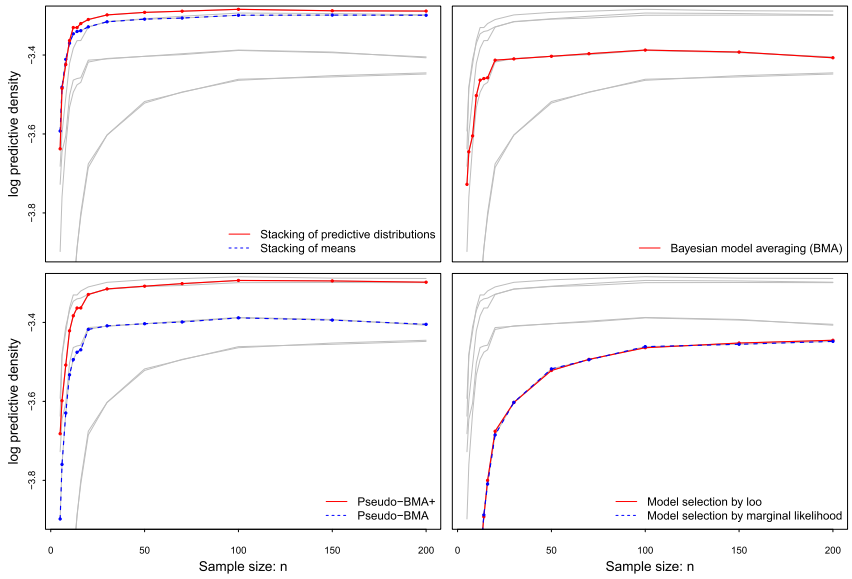

Figure 3: Mean log predictive densities of 7 combination methods in the linear regression example: the k-th model is a univariate regression with the k-th variable ($1 \leq k \leq 15$). We evaluate the log predictive densities using 100 repeated experiments and 200 test data.

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 $X_1, \ldots, X_k$ 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

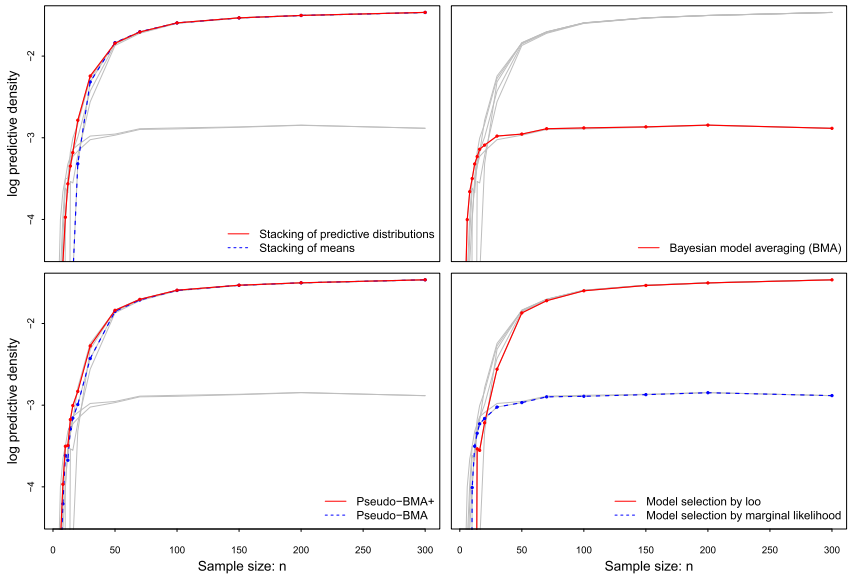

Figure 4: Mean log predictive densities of 7 combination methods in the linear regression example: the k-th model is the regression with the first k variables ($1 \leq k \leq 15$). We evaluate the log predictive densities using 100 repeated experiments and 200 test data.

again, BMA eventually becomes the same as model selection by marginal likelihood, which is much worse than any other methods asymptotically.

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).

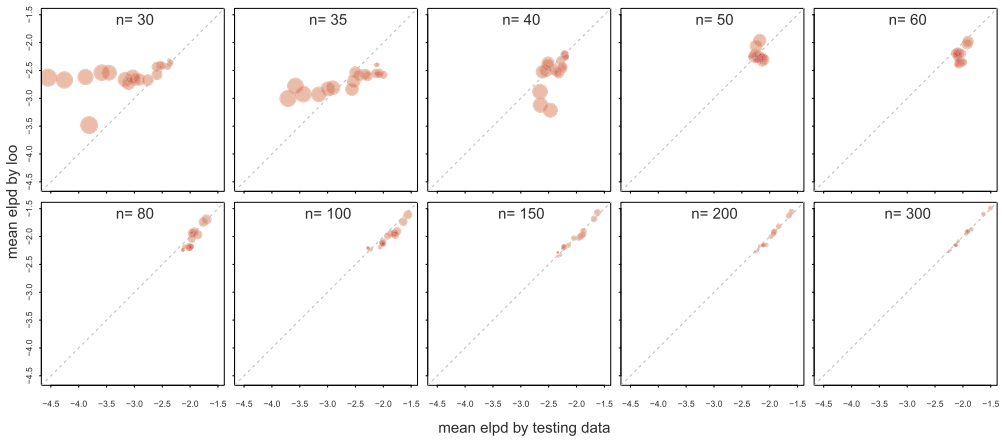

Figure 5: Comparison of the mean elpd estimated by LOO and the mean elpd calculated from test data, for each model and each sample size in the simulation described above. The area of each dot represents the relative complexity of the model as measured by the effective number of parameter divided by sample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

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+ 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 $\mathcal{M} = (M_1, \ldots, M_k)$, 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, Markov chain Monte Carlo (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

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

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

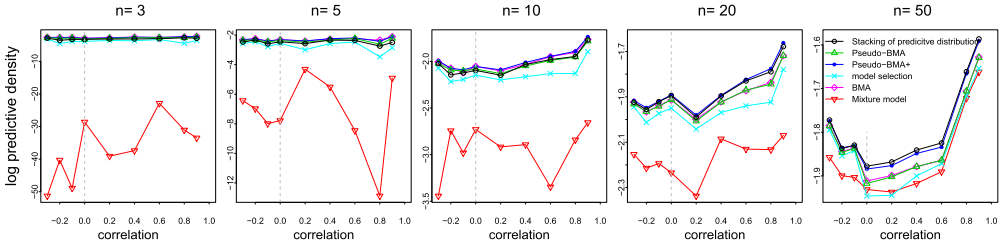

Figure 6: Log predictive densities of the combined distribution obtained by stacking of predictive distributions, BMA, Pseudo-BMA, Pseudo-BMA+, model selection by marginal likelihood, and mixture models. In each case, we evaluate the predictive density by 100 testing data and 100 repeated simulations. 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.

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

In the simulation, we generate the design matrix by $\mathrm{Var}(X_i) = 1$ and $\mathrm{Cor}(X_i, X_j) = \rho$. 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 $\theta = (\theta_1, \ldots, \theta_m)$ given observed data $y = (y_1 \ldots y_n)$ 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:

$$q^* = \arg_{q\in\mathcal{Q}} \min \mathrm{KL}\big(q(\theta), p(\theta|y)\big) = \arg_{q\in\mathcal{Q}} \min\big(\mathrm{E}_q \log q(\theta) - \mathrm{E}_q \log p(\theta, y)\big). \tag{10}$$

One widely used variational family is mean-field family where parameters are assumed to be mutually independent $\mathcal{Q} = \{q(\theta) : q(\theta_1, \ldots, \theta_m) = \prod_{j=1}^m q_j(\theta_j)\}$. Some

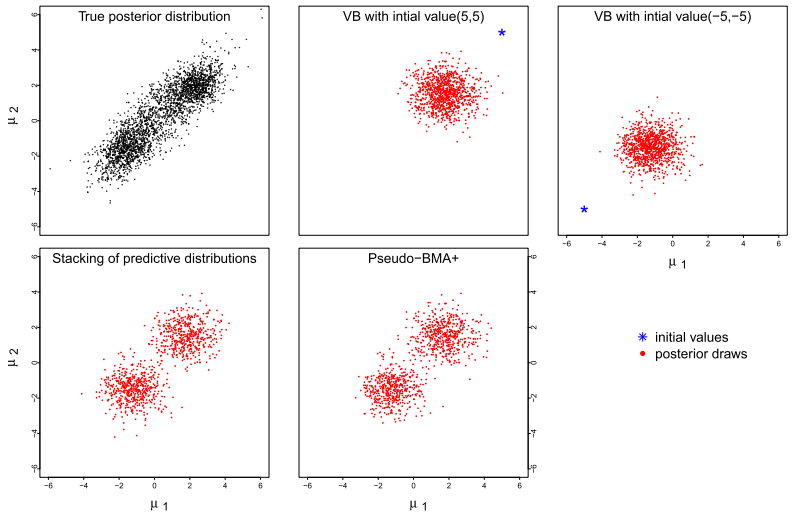

Figure 7: (1) A multi-modal posterior distribution of (μ_1, μ_2). (2–3) Posterior draws from variational inference with different initial values. (4–5) Averaged posterior distribution using stacking of predictive distributions and Pseudo-BMA+ weighting.

recent progress is made to run variational inference algorithm in a black-box way. For example, Kucukelbir et al. (2017) implement *Automatic Differentation 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 $\zeta = T(\theta)$ and uses normal approximation $p(\zeta|\mu, \sigma^2) = \prod_{j=1}^m \mathrm{N}(\zeta_j|\mu_j, \sigma_j^2)$. Plugging this into (10) 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 (10), ending up with an approximation distribution q_k^*. Then we draw S samples $(\theta_k^{(1)}, \ldots, \theta_k^{(S)})$ from $q_k^*(\theta)$ and calculate the importance ratio $r_{i,k}^s$ defined in (6) 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 $\text{N}(0, 1)$ 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 $\mu = (\mu_1, \mu_2)$, 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

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

	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.

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,

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

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:

- `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,

$$\begin{aligned} y \sim & \text{Bernoulli}(\theta), \\ \theta = & \text{logit}^{-1}(\beta_0 + \beta_1 dist + \beta_2 arsenic + \beta_3 assoc + \beta_4 educ). \end{aligned}$$

Model 2 contains the interaction between distances and arsenic levels,

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

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,

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

where B_{dis} is the B-spline basis of distance with the form $(B_{dis,1}(dist), \ldots, B_{dis,q}(dist))$ 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.

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

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

where $B_{dis,ars}$ is the bivariate spline basis with the degree to be $2\times2, 3\times3$ 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

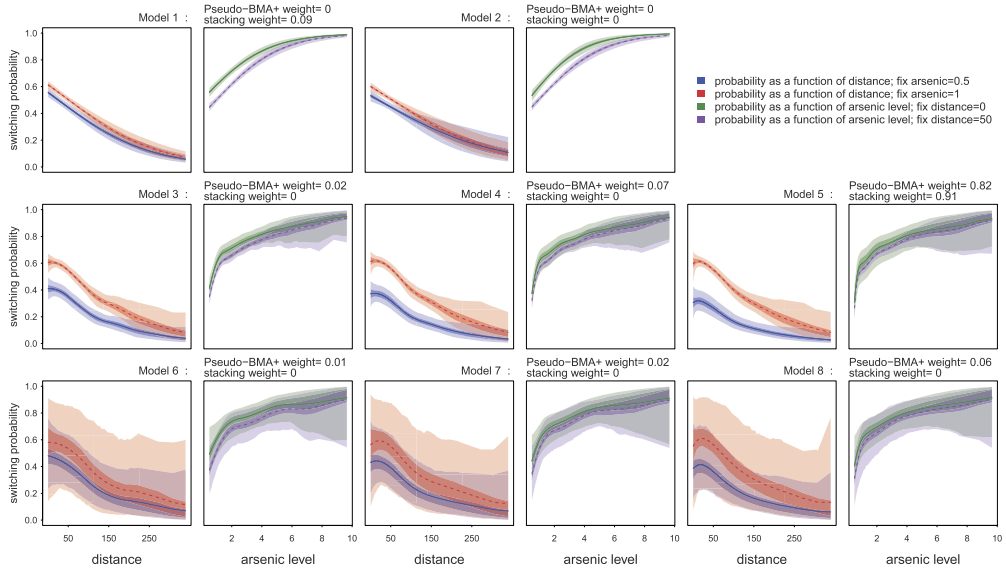

Figure 9: The posterior mean, 50% and 95% confidence interval of the well switching probability in Models 1–8. For each model, the switching probability is shown as a function of (a) the distance to the nearest safe well or (b) the arsenic level of the existing well. In each subplot, other input variables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

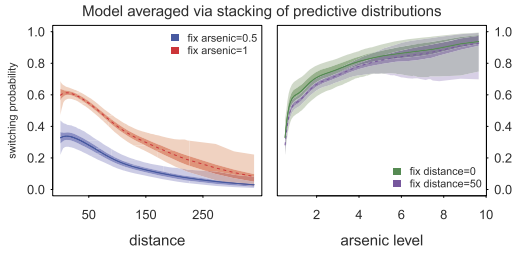

Figure 10: The posterior mean, 50% and 95% confidence interval of the well switching probability in the combined model via stacking of predictive distributions. Pseudo-BMA+ weighting gives a similar result for the combination.

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 con-

tained in Model 5 if setting $\alpha_{dis} = \alpha_{ars} = 0$. 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, $w \sim \text{Dirichlet}(\frac{\alpha}{K^\gamma}, \ldots, \frac{\alpha}{K^\gamma})$, 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,

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

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 (4) 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.

Supplementary Material

Supplementary Material to "Using Stacking to Average Bayesian Predictive Distributions" (DOI: 10.1214/17-BA1091SUPP; .pdf).

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. 936

Akaike, H. (1978). "On the likelihood of a time series model." *The Statistician*, 217–235. 920

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons. 918

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. 934

Breiman, L. (1996). "Stacked regressions." *Machine Learning*, 24(1): 49–64. 919, 930

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer, 2nd edition. MR1919620. 920

Clarke, B. (2003). "Comparing Bayes model averaging and stacking when model approximation error cannot be ignored." *Journal of Machine Learning Research*, 4: 683–712. 919, 934, 940

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. 918, 919, 923

Fokoue, E. and Clarke, B. (2011). "Bias-variance trade-off for prequential model list selection." *Statistical Papers*, 52(4): 813–833. 930

Geisser, S. and Eddy, W. F. (1979). "A Predictive Approach to Model Selection." *Journal of the American Statistical Association*, 74(365): 153–160. 920

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. 920

Gelman, A. (2004). "Parameterization and Bayesian modeling." *Journal of the American Statistical Association*, 99(466): 537–545. 917

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. 937

Gelman, A., Hwang, J., and Vehtari, A. (2014). "Understanding predictive information criteria for Bayesian models." *Statistics and Computing*, 24(6): 997–1016. MR3253850. doi: https://doi.org/10.1007/s11222-013-9416-2. 925

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. MR2798517. 930

Geweke, J. and Amisano, G. (2011). "Optimal prediction pools." *Journal of Econometrics*, 164(1): 130–141. MR2821798. doi: https://doi.org/10.1016/j.jeconom.2011.02.017. 920

Geweke, J. and Amisano, G. (2012). "Prediction with misspecified models." *American Economic Review*, 102(3): 482–486. 920

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102(477): 359–378. 917, 922, 923

Gutiérrez-Peña, E. and Walker, S. G. (2005). "Statistical decision problems and Bayesian nonparametric methods." *International Statistical Review*, 73(3): 309–330. 921

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian model averaging: A tutorial." *Statistical Science*, 14(4): 382–401. MR1765176. doi: https://doi.org/10.1214/ss/1009212519. 918, 919

Key, J. T., Pericchi, L. R., and Smith, A. F. M. (1999). "Bayesian model choice: What and why." *Bayesian Statistics*, 6: 343–370. 918, 923

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. MR3634881. 935

Le, T. and Clarke, B. (2017). "A Bayes interpretation of stacking for M-complete and M-open settings." *Bayesian Analysis*, 12(3): 807–829. MR3655877. doi: https://doi.org/10.1214/16-BA1023. 920, 923

LeBlanc, M. and Tibshirani, R. (1996). "Combining estimates in regression and classification." *Journal of the American Statistical Association*, 91(436): 1641–1650. 919

Li, M. and Dunson, D. B. (2016). "A framework for probabilistic inferences from imperfect models." *ArXiv e-prints:1611.01241*. 921, 925

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. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 937

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. MR1820767. doi: https://doi.org/10.1214/ss/1009212814. 918

Merz, C. J. and Pazzani, M. J. (1999). "A principal components approach to combining regression estimates." *Machine Learning*, 36(1–2): 9–32. 919

Montgomery, J. M. and Nyhan, B. (2010). "Bayesian model averaging: Theoretical developments and practical applications." *Political Analysis*, 18(2): 245–270. 936

Piironen, J. and Vehtari, A. (2017). "Comparison of Bayesian predictive methods for model selection." *Statistics and Computing*, 27(3): 711–735. 917

Rubin, D. B. (1981). "The Bayesian bootstrap." *Annals of Statistics*, 9(1): 130–134. 925

Smyth, P. and Wolpert, D. (1998). "Stacked density estimation." In *Advances in Neural Information Processing Systems*, 668–674. 928

Stan Development Team (2017). *Stan modeling language: User's guide and reference manual*. Version 2.16.0, http://mc-stan.org/. 935

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. MR0501454. 920

Ting, K. M. and Witten, I. H. (1999). "Issues in stacked generalization." *Journal of Artificial Intelligence Research*, 10: 271–289. 919

Vehtari, A., Gelman, A., and Gabry, J. (2017a). "Pareto smoothed importance sampling." *ArXiv e-print:1507.02646*. 920, 924, 932

Vehtari, A., Gelman, A., and Gabry, J. (2017b). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27(5): 1413–1432. 920, 924, 925, 932

Vehtari, A. and Lampinen, J. (2002). "Bayesian model assessment and comparison using cross-validation predictive densities." *Neural Computation*, 14(10): 2439–2468. 925

Vehtari, A. and Ojanen, J. (2012). "A survey of Bayesian predictive methods for model assessment, selection and comparison." *Statistics Surveys*, 6: 142–228. MR3011074. 917, 920, 921, 924

Wagenmakers, E.-J. and Farrell, S. (2004). "AIC model selection using Akaike weights." *Psychonomic bulletin & review*, 11(1): 192–196. 920

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. MR2756194. 920

Wolpert, D. H. (1992). "Stacked generalization." *Neural Networks*, 5(2): 241–259. 919

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. 919

Yang, Y. and Dunson, D. B. (2014). "Minimax Optimal Bayesian Aggregation." *ArXiv e-prints:1403.1345*. 919, 940

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). "Supplementary Material to "Using stacking to average Bayesian predictive distributions"." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1091SUPP. 918

Acknowledgments

We thank the U.S. National Science Foundation, Institute for Education Sciences, Office of Naval Research, and Defense Advanced Research Projects Administration (under agreement number D17AC00001) 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.

Invited Discussion

Bertrand Clarke*

1 Introduction and Summary

Yao, Vehtari, Simpson, and Gelman have proposed a useful and incisive extension to the usual model average predictor based on stacking models. The extension uses score functions as a way to determine weights on predictive densities, thereby giving a full stacking-based distribution for a future outcome. The authors are to be commended for their perceptivity and I am grateful to the Editor-in-Chief of *Bayesian Analysis* for the opportunity to contribute to the discussion.

The authors develop their ideas logically: Initally, they review the concept of statistical problem classes, namely $\mathcal{M}$-closed, -complete, and -open. These classes are defined by the relative position of the unknown true model (assuming it exists) to the models on a model list that are available for use. Then, they recall the definitions of various model averaging techniques, including the original form of stacking due to Wolpert (1992). Typically, model averaging techniques are most useful for large $\mathcal{M}$-closed problems (where model selection may not be effective) and $\mathcal{M}$-complete problems.

The authors then state the definition of a proper scoring rule and define a model averaging procedure with respect to one. Intuitively, a scoring rule S is a real valued function of two variables: One is a distribution (usually assumed to have a density) and the other is an outcome of a data generator (DG). When a DG is stochastic, i.e., its outcomes are drawn according to a probability distribution, it makes sense to regard its outcomes as corresponding to a random variable. The scoring rule is meant to encapsulate how far a P chosen to generate predictions is from a realized outcome y. The idea is that our P can be (and probably is) wrong whereas by definition y is 'right' because it came from the DG. Hence, loosely, $S(P, y)$ is small, possibly negative, when P is poorly chosen and large when P is well-chosen, both relative to y.

The definition of a scoring rule can be extended to include the case that $Y = y$ has a distribution. This gives a real valued function that behaves somewhat like a distance between two distributions, say P and Q, and is of the form $S(P, Q) = \int S(P, y)dQ(y)$.

To state the authors' central definition, which defines their new methodology as an extension of Wolpert (1992), write

$$\max_{w \in \mathcal{S}^K} S\left(\sum_{k=1}^{K} w_k p(\cdot \mid Y = y, M_k), p_T(\cdot \mid Y = y)\right), \tag{1}$$

where: i) $\mathcal{S}^K$ is the simplex in K dimensions with generic element $w = (w_1, \ldots, w_K)$, ii) the M_k's are candidate models and p_T denotes the density of the true model, and, iii)

*Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583 USA, bclarke3@unl.edu

$y = (y_1, \ldots, y_n)$ is an outcome of $Y = (Y_1, \ldots, Y_n) \sim p_T^n$ in which the Y_i's are IID. It is understood that the integration in S is with respect to p_T. The first and second entries of S in (1) use the predictive densities for a new $Y_{n+1} = y_{n+1}$, given $Y = y$, under the K candidate models and p_T, respectively.

Since (1) is intractable as written, the authors replace $p(y_{n+1} \mid Y = y, M_k)$ with

$$\hat{p}_{k,-i}(y_i) = \int p(y_i \mid \theta_k, M_k) p(\theta_k \mid y_{-i}, M_k) \mathrm{d}\theta_k \tag{2}$$

in which the subscript $-i$ indicates that the i-th data point, $i = 1, \ldots, n$, has been left out. It is also assumed that model M_k is defined by a conditional density for Y_i and includes a prior $p(\theta_k)$ where θ_k is the parameter for model M_k. Using (2) in (1) and reverting to the initial definition of the score function, the stacking weights are

$$(\hat{w}_i, \ldots, \hat{w}_K) = \arg \max_{w \in \mathcal{S}^K} \frac{1}{n} \sum_{i=1}^{n} S \left(\sum_{k=1}^{K} w_k \hat{p}_{k,-i}(\cdot), y_i \right), \tag{3}$$

assuming they exist and are unique. It is important to note the interchangeability of p_T and $Y_i = y_i$ which is 'true' in the sense that it is a valid outcome of p_T. Now, the 'stack' of predictive densities is

$$\hat{p}(y_{n+1} \mid Y = y) = \sum_{k=1}^{K} \hat{w}_k p(y_{n+1} \mid Y = y, M_k), \tag{4}$$

where the coefficients come from (3). Expression (4) can be used to obtain point and interval predictors for Y_{n+1}.

2 Prequentialism, Problem Classes, and Comparisons

The central methodological contribution of the paper is a general technique, essentially Expression (4), for using a score function to find coefficients that can be used to form a stack of densities that happen to be the predictive densities from K models. Thus, it is a method for producing predictors and it can be compared to other methods that produce predictors. One natural way to do this is to invoke the Prequential Principle as enunciated in Dawid (1984): Any criterion for assessing the agreement between the predictor and the DG should depend only on the predictions and outcomes. There are two key features to this: i) There should be no conflict/confluence of interest between the assessment and the generation of the values fed into the assessment, and ii) The values of either the DG or predictor that were not used are not relevant. At root, Prequentiality primarily requires that the comparison of predictions with outcomes be disjoint from how the predictions were generated.

The Prequential Principle is very general: It does not require that the outcomes of the DG even follow a distribution. So, the Prequential Principle accommodates any sort of streaming data or data that can be regarded as having a time-ordering – even if the

ordering is imposed by the analysis. Parallel to this generality, and given the ubiquity of data that is not plausibly stochastic, it makes sense to merge the definitions of $\mathcal{M}$-complete and $\mathcal{M}$-open and redefine $\mathcal{M}$-open as those DG's that are not meaningfully described by any stochastic process, i.e., no true probability model exists. Indeed, the definition of $\mathcal{M}$-complete and $\mathcal{M}$-open offered originally in Bernardo and Smith (2000) p. 384-5 allows for this but permits $\mathcal{M}$-complete and $\mathcal{M}$-open problems to overlap. Redefining the $\mathcal{M}$-open class of problems so it is disjoint from the $\mathcal{M}$-complete (and $\mathcal{M}$-closed) classes of problems seems logical. Doing this also helps focus on model list selection which deserves more attention; see the comparisons in Clarke et al. (2013).

In the context of the Prequential Principle, the assumption built into the paper (and comments) is that stacking the posterior predictive densities is done using only the first n data points and the goal is to predict the $n+1$. That is, n data points (x_i, y_i) for $i = 1, \ldots, n$ are available to form a predictor such as $\hat{p}$ and that the predictor is evaluated at x_{n+1} to predict $Y_{n+1}(x_{n+1})$ and its performance assessed in some way that does not involve S. Implicitly, it is assumed the prediction problem will be repeated many times and we are looking only at the n-th stage so as to examine the updating of the predictor. In this way the authors' framework may be seen as Prequential (although this is a point on which reasonable people might disagree). Aside from the formula (4), updating can include changing the models, the $\mathcal{M}_k$'s, or even the model averaging technique itself. This is done chiefly by the comparing the predictions with the realized values. Here, the x_i's are regarded as deterministic explanatory variables and the Y_i's are random. In much of the paper, the x_i's are suppressed in the notation.

In their Gaussian mixture model example (Section 4.1), the authors treat their problem as $\mathcal{M}$-open because the true model is not on the model list. While this is reasonable for the sake of argument, it actually underscores the importance of model list selection because choosing a better model list would make the problem $\mathcal{M}$-closed. Nevertheless, in this example, the authors make a compelling point by comparing three different predictors: Bayes model averaging (BMA), stacking of means (under squared error), and stacking of predictive distributions by using a log score as in (4). Figure 1 shows that BMA converges to the model on the model list closest to the true model i.e., BMA has unavoidable (and undetectable) bias. By contrast, stacking of means and stacking of predictive distributions both do well in terms of their means (Figure 2, middle panel) and stacking of predictive densities outperforms both BMA and stacking of means in other senses (Figure 2, left and right panels) because, as shown in Figure 1, it converges to the correct predictive distribution. The distribution associated with the stacking of means converges to a broad lump that does not seem useful. This example shows that matching whole distributions is feasible and sensible. It also shows BMA does not routinely perform well despite its asymptotic optimality; see Raftery and Zheng (2003).

In the example of Section 4.2, Figures 3 and 4 show that, again, stacking means and stacking predictive densities are the best among seven model averaging methods while BMA ties for fourth place or is in last place. Other comparisons have found BMA to have similarly disappointing finite sample behavior. (There is good evidence that a technique called Pseudo-BMA+ is competitive with the two versions of stacking.)

In the example of Section 4.3 where the goal is to obtain a density, the authors show stacking predictive densities and Pseudo-BMA+ outperform four other techniques for

obtaining a density; one is BMA. The remaining examples show further properties of stacking predictive densities (Section 4.4) or use the method to do predictive inference (Sections 4.5 and 4.6). The results on real data seem reasonable.

A remaining question is the relationship between using predictions from stacked predictive distributions and the Prequential Principle. Obviously, point predictors from stacked predictive distributions can be compared directly to outcomes and hence the Prequential Principle is satisfied. However, one of the authors' arguments is that obtaining a full distribution, as can be done by using their method, is better than simply using point predictors; this is justified by using what appear to be non-Prequential assessments such as the mean log density; see Figures, 2, 3, 4, and 6. First, the log-score was used to form the stacks. So, is log really the right way to assess performance? Second, all too often, taking a mean requires a distribution to exist and so the evaluation of the predictor may therefore depend on the true distribution if only to define the mode of convergence. It's hard to tell if this is the case with the present predictor; these are points on which the authors should comment. Moreover, effectively, the new method gives a prediction distribution (4) that leaves us with two questions: i) How should we use the distributional information, including that from the smoothing and Bayesian bootstrap, assuming it's valid? and ii) Is the distributional information associated with the stacking of means or densities valid, i.e., an accurate representation of its variability?

The answer to i) might simply be the obvious: It's a distribution and therefore any operation we might wish to perform on it, e.g., extract interval or percentile predictors, is feasible and it can be assessed in the score function of our choice. Of course, in practice, we do not know the actual distribution of the future outcome; we have only an estimate of it that we hope is good. Perhaps the consistency statement in Section 3.2 is enough. The answer to ii) seems to require more thought on what exactly the Pareto smoothing and Bayesian bootstrap are producing. This is important because an extra quantity, the score function, has been introduced and the solution in (4) – and hence the distribution assigned to stacked means or predictive densities – can depend on it, possibly delicately. The effect of the score function and the validity of the distribution the authors have associated to stacking of means are points for which the authors might be able to provide some insight.

3 What About Score Functions in the $\mathcal{M}$-Open Case?

One can plausibly argue that the authors' methodology really only applies to $\mathcal{M}$-closed and -complete problems. In other words, the examples they use are simulated and so are $\mathcal{M}$-closed or real data for which one might believe a stochastic model exists even if it is tremendously complex. Of course, even if such arguments are accepted, one can use the authors' techniques in $\mathcal{M}$-open problems – it just might not work as well as methods that are designed for $\mathcal{M}$-open problems.

One technique that was invented with $\mathcal{M}$-open problems in mind is due to Shtarkov (1987). The analogous Bayesian problem and solution was given in Le and Clarke (2016). In both cases the log score was used; however, the authors' work suggests that this technique can be generalized to other score functions.

Recall the idea behind Shtarkov's original formulation is that prediction can be treated as a game in which a Forecaster chooses a density q (for prediction) and Nature chooses an outcome y not constrained by any rule. The payoff to Nature (or loss to the Forecaster) is $\log q(y)$, i.e., log loss. Naturally, the Forecaster wants to minimize the loss. So, assume the Forecaster has 'Advisors' represented by densities p_θ; each Advisor corresponds to a θ. The advisors announce their densities before Nature issues a y. If the Forecaster has a pre-game idea about the relative abilities of the advisors to give good advice, this may be formulated into a prior $p(\theta)$. Now it makes sense for the Forecaster to minimize the maximum regret, i.e., to seek the smallest loss (incurred by the best Advisor). This means finding the q that minimizes

$$\sup_y \left[\log \frac{1}{q(y)} - \inf_\theta \log \frac{1}{p(\theta)p(y \mid \theta)} \right]. \tag{5}$$

The solution exists in closed form and can be computed; see Le and Clarke (2016). Following the authors, consider replacing the log loss in (5) by a general score function, S. Now, the Forecaster wants the $q \in \mathcal{Q}$, say $q_{\mathsf{opt},\mathsf{S}}$, that minimizes

$$\sup_y \left[S(q(\cdot), y) - \inf_\theta S(p(\theta)p(\cdot \mid \theta), y) \right], \tag{6}$$

where $\mathcal{Q}$ is a collection of densities. Expression (6) may be converted to a form analogous to (3), possibly releasing the sum-to-one constraint that some have argued is not appropriate for $\mathcal{M}$-open problems. It is not obvious that a closed form for (6) can be given; $q_{\mathsf{opt},\mathsf{S}}$ might only be available computationally. In either case, $q_{\mathsf{opt},\mathsf{S}}$ would depend on S, give an alternative solution to Shtarkov's problem, and might perform better for $\mathcal{M}$-open data than score based stacking. If the authors had any insight on these points, many readers would be glad to hear them.

References

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. Chichester, West Sussex, UK: John Wiley and Sons. MR1274699. doi: https://doi.org/10.1002/9780470316870. 947

Clarke, J., Yu, C.-W., and Clarke, B. (2013). "Prediction in M-complete problems with limited sample size." *Bayesian Analysis*, 8: 647–690. MR3102229. doi: https://doi.org/10.1214/13-BA826. 947

Dawid, A. P. (1984). "Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach." *Journal of the Royal Statistical Society. Series A*, 147: 278–292. MR0763811. doi: https://doi.org/10.2307/2981683. 946

Le, T. and Clarke, B. (2016). "Using the Bayesian Shtarkov solution for predictions." *Computational Statistics & Data Analysis*, 104: 183–196. MR3540994. doi: https://doi.org/10.1016/j.csda.2016.06.018. 948, 949

Raftery, A. and Zheng, Y. (2003). "Performance of Bayes model averaging." *Journal of the American Statistical Association*, 98: 931–938. 947

Shtarkov, Y. M. (1987). "Universal Sequential Coding of Single Messages." *Problems of Information Transmission*, 23: 3–17. MR0914346. 948

Wolpert, D. (1992). "Stacked generalization." *Neural Networks*, 5: 241–259. 945

Invited Discussion

Meng Li*†

1 Introduction

I would like to offer my congratulations to Yao et al. for a welcome and interesting addition to the growing literature on model averaging. Earlier papers on stacking cited by the authors have mostly focused on averaging models to improve point estimation. The authors now demonstrate that the same idea can be extended to the combination of predictive distributions in the $\mathcal{M}$-open case. In the next several pages, I will first review the paper connecting it to the related literature (Section 2), then comment on the $\mathcal{M}$-complete case (Section 3) and an application that may be favorable to the proposed method (Section 4). Section 5 concludes this comment.

2 Overview

Suppose we have a list of parametric models under consideration $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_K\}$ for the observations $y^{(n)} = \{y_1, \ldots, y_n\} \in \mathcal{Y}^n$ with $\mathcal{Y}$ the sample space. Yao et al. (2017) address a general problem of model aggregation from an interesting perspective: how to average the multiple models in $\mathcal{M}$ such that the resulting model combination has an optimal predictive *distribution*. This distinguishes its goal from two areas that have been well studied, i.e., weighing models targeted to an optimal point prediction and selecting a single model possibly with uncertainty quantification. Yao et al. (2017) particularly focus on the $\mathcal{M}$-open case (Bernardo and Smith, 1994) to allow the true model to fall outside of $\mathcal{M}$.

One of the most popular approaches is to use Bayesian model probabilities $\text{pr}(\mathcal{M}_k \mid y^{(n)})$ as weights, with these weights forming the basis of Bayesian Model Averaging (BMA). Philosophically, in order to interpret $\text{pr}(\mathcal{M}_j \mid y^{(n)})$ as a model *probability*, one must rely on the assumption that one of the models in the list $\mathcal{M}$ is exactly true, known as the $\mathcal{M}$-closed case. This assumption is arguably always flawed, although one can still use $\text{pr}(\mathcal{M}_k \mid y^{(n)})$ as a model weight from a pragmatic perspective, regardless of the question of interpretation. In the case of $\mathcal{M}$-complete or $\mathcal{M}$-open, an alternative approach is to formulate the model selection problem in a decision theoretic framework, selecting the model in $\mathcal{M}$ that maximizes expected utility. Yao et al. (2017) adopt a stacking approach along the line of this decision theoretic framework (Bernardo and Smith, 1994; Gutiérrez-Peña et al., 2009; Clyde and Iversen, 2013).

There are various scoring rules available when defining the unity function in a decision theoretic framework. The choice can and probably should depend on the specific

*This work was partially supported by the Ralph E. Powe Junior Faculty Enhancement Award by ORAU.

†Noah Harding Assistant Professor, Department of Statistics, Rice University, Houston, TX, U.S.A., meng@rice.com

application—similar principle has been demonstrated by Claeskens and Hjort (2003) that model selection may focus on the parameter singled out for interest. The classical stacking method uses quadratic loss (or *energy score* with $\beta = 2$ in the paper), targeting at an optimal point prediction. Yao et al. (2017) consider a range of scoring rules generalizing the stacked density estimation by Smyth and Wolpert (1998). The authors recommend to use the log scoring rule, which essentially finds the Kullback–Leibler divergence projection of the true data generating density to the convex hull $\mathcal{C} = \{\sum_{k=1}^{K} w_k p(\cdot \mid \mathcal{M}_k) : \sum_{k=1}^{K} w_k = 1, w_k \geq 0\}$ where $p(\cdot \mid \mathcal{M}_k)$ is the predictive density under model $\mathcal{M}_k$, targeting at an optimal predictive density.

The decision theoretic framework used in the paper bypasses the need to philosophically interpret or calibrate model weights, but has to evaluate the expected utility. Expected utility can be approximated either via cross-validation (Clyde and Iversen, 2013) or using a nonparametric prior (Gutiérrez-Peña and Walker, 2005; Gutiérrez-Peña et al., 2009). The authors use leave-one-out cross-validation to construct an approximation of the expected utility, while one may generally consider a k-fold cross-validation as in Clyde and Iversen (2013). While Bayesian model probabilities often have analytical forms available thus are computationally appealing (Liang et al., 2008), the computational burden in cross-validation is unfavorably intensive. The authors overcome this difficulty by using the Pareto smoothed importance sampling (Vehtari and Lampinen, 2002; Vehtari et al., 2012) to approximate leave-one-out predictive densities, which self diagnoses the reliability of the approximation based on some estimated parameter and leads to manageable computation. The authors thoughtfully design simple but effective simulations to illustrate and compare how stacking of distributions and selected existing methods behave, and provide R and Stan code for routine implementation.

3 $\mathcal{M}$-complete and nonparametric references

The nonparametric Bayes literature provides a rich menu of possibilities to approximate the true data generating scheme, ranging from Dirichlet processes to Gaussian processes; refer to Hjort et al. (2010) for a review. There is a rich literature showing that Bayesian nonparametric models often have appealing frequentist asymptotic properties, such as appropriate notions of consistency (Schwartz, 1965) and optimal rates of convergence (van der Vaart and van Zanten, 2009; Bhattacharya et al., 2014; Castillo, 2014; Shen and Ghosal, 2015; Li and Ghosal, 2017; Ghosal and van der Vaart, 2017). When an optimal predictive density is the goal, one may ask why not pursue the direction of flexible modeling utilizing the rapidly developed nonparametric Bayes literature?

While I look forward to open discussions about the question above, one possible argument is that although Bayesian nonparametric models are appealing from a prediction viewpoint, they also have disadvantages in terms of not only interpretability but also in involving large numbers of parameters, which increase automatically as the sample size increases. This may lead to daunting memory, storage and communication issues in modern applications involving large data sets. It is thus often preferable from a variety of viewpoints to approximate the performance of a very rich and provably flexible nonparametric model by taking a weighted average of much simpler parametric models.

Keeping the interpretability in mind while being aware of flexible nonparametric models, we indeed reach to the case of $\mathcal{M}$-complete which "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)", according to the definition from Bernardo and Smith (1994). As a result, we can use a nonparametric Bayes surrogate for the true model and calculate the expected utility based on this surrogate, a general approach that fits into the $\mathcal{M}$-complete case. Li and Dunson (2018) use a nonparametric Bayesian reference to assign weights to models based on Kullback–Leibler divergence to define a model weight that can be used in goodness-of-fit assessments, comparing imperfect models, and providing model weights to be used in model aggregation. It seems promising to migrate this idea to the two stacking approaches used in the paper, one based on optimization using proper scoring rules and the other called *pseudo-BMA* in Section 3.4 in a form of exponential weighting (Rigollet and Tsybakov, 2012):

- Stacking using proper scoring rules. One may obtain the weights by optimizing an approximated version of (3):

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

  where $\hat{p}_t(\tilde{y}_i)$ is a nonparametric Bayes model and other notations are defined in Yao et al. (2017).

- Pseudo-BMA. We replace the empirical observations used in Section 3.4 by the nonparametric surrogate. Specifically, the quantity elpd^k can be estimated by

$$\widehat{\text{elpd}}^k = \sum_{i=1}^n \int \hat{p}_t(\tilde{y}_i) \log p(\tilde{y}_i|y,M_k) d\tilde{y}_i \quad (2)$$

  instead of $\widehat{\text{elpd}}^k_{\text{loo}}$ used in the paper, and the final weights become

$$w_k = \frac{\exp(\widehat{\text{elpd}}^k)}{\sum_{k=1}^K \exp(\widehat{\text{elpd}}^k)}. \quad (3)$$

The use of nonparametric reference models eliminates the need of cross-validation that gives rise to the main computational hurdle in stacking of distributions. In addition, this estimate based on nonparametric reference potentially induces an inherent complexity penalty, a phenomenon observed in Li and Dunson (2018), thus the log-normal approximation for weights regularization in Section 3.4 may be not necessary.

Although we here focus on Bayesian machinery, one can approximate the expected utility using any method that estimates $d(\sum_{k=1}^K w_k p(\cdot|y,M_k), p_t(\cdot|y))$. For example, if we use the recommended Kullback–Leibler divergence, i.e., $d(p,q) = \text{KL}(q,p)$, then there is substantial work that has focused on estimating the Kullback–Leibler divergence

between two unknown densities based on samples from these densities (Leonenko et al., 2008; Pérez-Cruz, 2008; Bu et al., 2018). Here the setting is somewhat different as there is only one sample, but the local likelihood methods of Lee and Park (2006) and the Bayesian approach of Viele (2007) can potentially be used, among others.

Furthermore, all of these methods in a decision theoretic framework or BMA are focused on providing weights for model aggregation, and are not useful for goodness-of-fit assessments of (absolute) model adequacy. The nonparametric reference models in the $\mathcal{M}$-complete case enables the assessment the quality of each individual model in $\mathcal{M}$ as well as the entire model list. One of course needs to specify an absolute scale to define what is adequate, but rules of thumb such as the one provided by Li and Dunson (2018) based on the convention of Bayes factors may be obtainable.

4 Data integration

Section 5.3 makes a great point that an ideal case for stacking is that the K models in the model list are orthogonal. This ideal case is not fully demonstrated by the paper, but it insightfully points to a possible solution to a challenging problem—data integration.

Modern techniques enable researchers to acquire rich data from multiple platforms, thus it becomes possible to combine various data types of fundamental differences to make a single decision, hopefully more informative than any decision based on an individual data resource. In response to this demand, there has been a recent surge of interest in data integration expanding into a variety of emerging areas, for example, imaging genetics, omics data, and analysis of covariate adjusted complex objects (such as functions, images, trees, and networks). One concrete example that I have been working on comprises a cohort of patients with demographic, clinical and omics data; the omics data include single nucleotide polymorphisms (SNPs), expression, and micro-ribonucleic acids (miRNAs). In these cases, the dramatic heterogeneity across data structures, which is one of root causes that fail many attempts especially those trying to map various data structures to a common space such as the Euclidean space, seems to be a characteristic favorable to the stacking approach. The sample size is usually not large, thus even the cross-validation approach without approximation may be computationally manageable.

5 Summary

To summarize, Yao et al. (2017) have tackled the model averaging problem that is one of fundamental tasks in statistics. They have proposed improvements to existing stacking methods for stacking of densities. This method requires intensive leave-one-out posterior distributions to approximate the expected utility, and the authors propose to use Pareto smoothed importance sampling to scale up the implementation.

I would like to thank Yao et al. for writing an interesting paper. I appreciated that the paper has several detailed and thoughtful demonstrations to compare the proposed methods to existing model weights and help readers understand how stacking and BMA behave differently. The integration with R and Stan makes the method immediately

available to practitioners. I find the work useful and expect the proposed methods positively impact model averaging and its application to a wide range of problems in practice. I hope the comments on $\mathcal{M}$-complete and a possible application to data application add some useful insights to a paper already rich in content.

References

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, New York, NY. MR1274699. doi: https://doi.org/10.1002/9780470316870. 951, 953

Bhattacharya, A., Pati, D., and Dunson, D. (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *The Annals of Statistics*, 42(1):352–381. MR3189489. doi: https://doi.org/10.1214/13-AOS1192. 952

Bu, Y., Zou, S., Liang, Y., and Veeravalli, V. V. (2018). Estimation of KL divergence: optimal minimax rate. *IEEE Transactions on Information Theory*, 64(4):2648–2674. MR3782280. doi: https://doi.org/10.1109/TIT.2018.2805844. 954

Castillo, I. (2014). On Bayesian supremum norm contraction rates. *The Annals of Statistics*, 42(5):2058–2091. MR3262477. doi: https://doi.org/10.1214/14-AOS1253. 952

Claeskens, G. and Hjort, N. L. (2003). The Focused Information Criterion. *Journal of the American Statistical Association*, 98(464):900–916. MR2041482. doi: https://doi.org/10.1198/016214503000000819. 952

Clyde, M. A. 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., editors, *Bayesian Theory and Applications*, pages 483–498. Oxford University Press. MR3221178. doi: https://doi.org/10.1093/acprof:oso/9780199695607.003.0024. 951, 952

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge. MR3587782. doi: https://doi.org/10.1017/9781139029834. 952

Gutiérrez-Peña, E., Rueda, R., and Contreras-Cristán, A. (2009). Objective parametric model selection procedures from a Bayesian nonparametric perspective. *Computational Statistics & Data Analysis*, 53(12):4255–4265. MR2744322. doi: https://doi.org/10.1016/j.csda.2009.05.018. 951, 952

Gutiérrez-Peña, E. and Walker, S. G. (2005). Statistical decision problems and Bayesian nonparametric methods. *International Statistical Review*, 73(3):309–330. 952

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., editors (2010). *Bayesian Nonparametrics*. Cambridge University Press, Cambridge. MR2722988. doi: https://doi.org/10.1017/CBO9780511802478.002. 952

Lee, Y. K. and Park, B. U. (2006). Estimation of Kullback–Leibler divergence by local likelihood. *Annals of the Institute of Statistical Mathematics*, 58(2):327–340. MR2246160. doi: https://doi.org/10.1007/s10463-005-0014-8. 954

Leonenko, N., Pronzato, L., and Savani, V. (2008). A class of Rényi information estimators for multidimensional densities. *The Annals of Statistics*, 36(5):2153–2182. MR2458183. doi: https://doi.org/10.1214/07-AOS539. 954

Li, M. and Dunson, D. B. (2018). Comparing and weighting imperfect models using D-probabilities. *arXiv preprint arXiv:1611.01241v3*. 953, 954

Li, M. and Ghosal, S. (2017). Bayesian detection of image boundaries. *The Annals of Statistics*, 45(5):2190–2217. MR3718166. doi: https://doi.org/10.1214/16-AOS1523. 952

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. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 952

Pérez-Cruz, F. (2008). Kullback–Leibler divergence estimation of continuous distributions. In *IEEE International Symposium on Information Theory – Proceedings*, pages 1666–1670. IEEE. 954

Rigollet, P. and Tsybakov, A. B. (2012). Sparse Estimation by Exponential Weighting. *Statistical Science*, 27(4):558–575. MR3025134. doi: https://doi.org/10.1214/12-STS393. 953

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26. 952

Shen, W. and Ghosal, S. (2015). Adaptive Bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, 42(4):1194–1213. MR3426318. doi: https://doi.org/10.1111/sjos.12159. 952

Smyth, P. and Wolpert, D. (1998). Stacked density estimation. In *Advances in neural information processing systems*, pages 668–674. 952

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675. MR2541442. doi: https://doi.org/10.1214/08-AOS678. 952

Vehtari, A. and Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2439–2468. 952

Vehtari, A., Ojanen, J., et al. (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228. MR3011074. doi: https://doi.org/10.1214/12-SS102. 952

Viele, K. (2007). Nonparametric estimation of Kullback–Leibler information illustrated by evaluating goodness of fit. *Bayesian Analysis*, 2(2):239–280. MR2312281. doi: https://doi.org/10.1214/07-BA210. 954

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2017). Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*, pages 1–28. 951, 952, 953, 954

Invited Discussion

Peter Grünwald* and Rianne de Heide†

Yao et al. (2018) aim to improve Bayesian model averaging (BMA) in the $\mathcal{M}$-open (misspecified) case by replacing it with *stacking*, which is extended to combine predictive distributions rather than point estimates. We generally applaud the program to adjust Bayesian methods to better deal with $\mathcal{M}$-open cases and we can definitely see merit in stacking-based approaches. Yet, we feel that the main method advocated by Yao et al. (2018), which stacks based on the log score, while often outperforming BMA, fails to address a crucial problem of the $\mathcal{M}$-open-BMA setting. This is the problem of *hypercompression* as identified by Grünwald and Van Ommen (2017), and shown also to occur with real-world data by De Heide (2016). We explore this issue in Section 2; first, we very briefly compare stacking to a related method called *switching*.

1 Stacking and Switching

Standard BMA can already be viewed in terms of minimizing a sum of log score prediction errors via Dawid's (1984) *prequential interpretation* of BMA. Based on this interpretation, Van Erven et al. (2012) designed the *switch distribution* as a method for combining Bayes predictive densities with asymptotics that coincide, up to a $\log \log n$ factor, with those of the Akaike Information Criterion (AIC) and leave-one-out cross validation (LOO). It can vastly outperform standard BMA (see Figure 1 from their paper), yet is designed in a manner that stays closer to the Bayesian ideal than stacking. It has the additional benefit that *if* one happens to be so lucky to unknowingly reside in the $\mathcal{M}$-closed (correctly specified) case after all, the procedure becomes statistically consistent, selecting asymptotically the smallest model M_k that contains the data generating distribution P^*. We suspect that in this $\mathcal{M}$-closed case, stacking will behave like AIC, which, in the case of nested models, even asymptotically will select an overly large model with positive probability (for theoretical rate-of-convergence and consistency results for switching see Van der Pas and Grünwald (2018)). Moreover, by its very construction, switching, like stacking, should resolve another central problem of BMA identified by (Yao et al., 2018, Section 2), namely its sensitivity to the prior chosen within the models M_k. On the other hand, in the $\mathcal{M}$-open case, switching will asymptotically concentrate on the single, smallest M_k that contains the distribution $\tilde{P}$ closest to P^* in KL-divergence; stacking will provide a weighted predictive distribution that may come significantly closer to P^*, as indicated by (Yao et al., 2018, Section 3.2). To give a very rough idea of 'switching': in the case of just two models $\mathcal{M} = \{M_1, M_2\}$, switching can be interpreted as BMA applied to a modified set of models $\{M_{\langle j \rangle} : j \in \mathbb{N}\}$ where $M_{\langle j \rangle}$ represents a model that follows the Bayes predictive density of model M_1 until time j and then switches to the Bayes predictive density corresponding to model M_2; dynamic programming allows for efficient implementation even when the number

*CWI, Amsterdam and Leiden University, The Netherlands, pdg@cwi.nl
†CWI, Amsterdam and Leiden University, The Netherlands, r.de.heide@cwi.nl

of models K is larger than 2. It would be of interest to compare stacking to switching, and compile a list of the pros and cons of each.

2 Standard BMA, Stacking and SafeBMA

Grünwald and Van Ommen (2017) give a simple example of BMA misbehaving in an $\mathcal{M}$-open regression context. We start with a set of $K+1$ models $\mathcal{M} = \{M_1, \ldots, M_K\}$ to model data $(Z_1, Y_1), (Z_2, Y_2), \ldots$. Each model $M_k = \{p_{\beta,\sigma^2} : \beta \in \mathbb{R}^{k+1}, \sigma^2 > 0\}$ is a standard linear regression model, i.e. a set of conditional densities expressing that $Y_i = \sum_{j=0}^{k} \beta_j X_{ij} + \xi_i$. Here X_{ij} is the j-th degree Legendre polynomial applied to one-dimensional random variable Z_i with support $[-1, 1]$ (i.e. $X_{i1} = Z_i, X_{i2} = (3Z_i^2 - 1)/2$, and so on), and the ξ_i are i.i.d. $N(0, \sigma^2)$ noise variables. We equip each model with standard priors, for example, a $N(0, \sigma^2)$ prior on the β's conditional on σ^2 and an inverse Gamma on σ^2. We put a uniform or a decreasing prior on the models M_k themselves. The actual data Z_i, Y_i are i.i.d. $\sim P^*$. Here P^* is defined as follows: at each i, a fair coin is tossed. If the coin lands heads, then Z_i is sampled uniformly from $[-1, 1]$, and Y_i is sampled from $N(0, 1)$. If it lands tails, then (Z_i, Y_i) is simply set to $(0, 0)$. Thus, M_1, the simplest model on the list, already contains the density in $\bigcup_{k=1..K} M_k$ that is closest to $P^*(Y \mid X)$ in KL divergence. This is the density $p_{\tilde{\beta},1/2}$ with $\tilde{\beta} = 0$, which is incorrect in that it assumes homoskedastic noise while in reality noise is heteroskedastic; yet $p_{\tilde{\beta},1/2}$ does give the correct regression function $\mathbf{E}[Y \mid X] \equiv 0$. M_1 is thus 'wrong but highly useful'. Still, while M_1 receives the highest prior mass, until a sample size of about $2K$ is reached, BMA puts nearly all of its weight on models $M_{k'}$ with k' close to the maximum K, leading to rather dreadful predictions of $\mathbf{E}[Y \mid X]$. Figure 1 (green) shows $\mathbf{E}[Y \mid X]$ where the expectation is under the Bayes predictive distribution arrived at by BMA at sample size 50, for $K = 30$. On the other hand, *SafeBayesian* model averaging, a simple modification of BMA that employs likelihoods raised to an empirically determined power $\eta < 1$, performs excellently in this experiment; for details we refer to Grünwald and Van Ommen (2017). We also note that other common choices for priors on (β, σ^2) lead to the same results; also, we can take the $X_{i0}, X_{i1}, \ldots, X_{iK}$ to be trigonometric basis functions or i.i.d. Gaussians rather than polynomials of Z_i, still getting essentially the same results. De Heide (2016) presents various real-world data sets in which a similar phenomenon occurs.

Given these problematic results for BMA in an $\mathcal{M}$-open scenario, it is natural to check how Yao et al. (2018)'s stacking approach (based on log score) fares on this example. We tried (implementation details at the end of this section, and obtained the red line in Figure 1. While the behaviour is definitely better than that of BMA, we do see a milder variation of the same overfitting phenomenon. We still regard this as undesirable, especially because another method (SafeBMA) behaves substantially better. To be fair, we should add that (Yao et al., 2018, Section 3.3.) advise that for extremely small n, their current method can be unstable. The figure reports the result on a simulated data sequence, for which, according to the diagnostics in their software, their method should be reasonably accurate (details at the end of this section). Since, moreover, results (not shown) based on the closely related LOO model selection with log

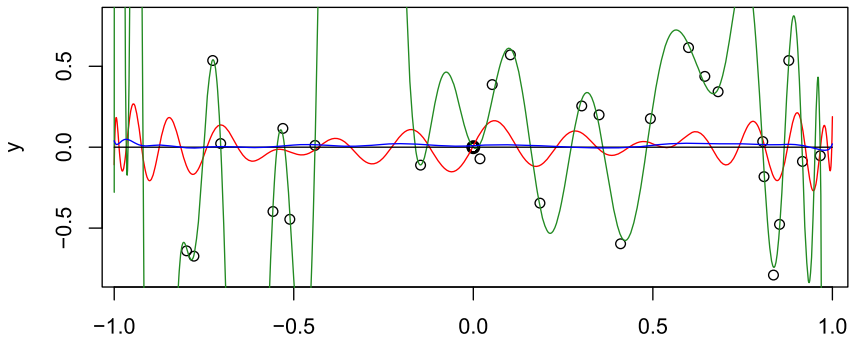

Figure 1: The conditional expectation $\mathbf{E}[Y|X]$ according to the predictive distribution found by stacking (red), standard BMA (green) and SafeBayesian regression (blue), based on models $M_1, \ldots, M_{30}$ with polynomial basis functions, given 50 data points sampled i.i.d. $\sim P^*$, of which approximately half are placed in $(0, 0)$. The true regression function is depicted in black. Behaviour of stacking and standard BMA slowly improves as sample size increases and becomes comparable to SafeBMA around $n = 80$ for stacking and $n = 120$ for BMA. Implementation details are given at the end of the section.

score yield very similar results, we do think that there is an issue here – stacking in itself is not sufficient to get useful weighted combinations of Bayes predictive distributions in some small sample situations where such combinations do exist.

Hypercompression The underlying problem is best explained in a simplified setting without random covariates: let $Y_1, Y_2, \ldots$ i.i.d. $\sim P^*$ and each model M_k a set of densities for the Y_i. Denote by $\tilde{p}$ the density in $\bigcup_{k=1..K} M_k$ that minimizes KL divergence to P^*. Then, under misspecification, we can have for some $k = 1..K$ that

$$\mathbf{E}_{Y^n \sim P^*}\left[-\log p(y_1, \ldots, y_n \mid M_k)\right] \ll \mathbf{E}_{Y^n \sim P^*}\left[-\log \tilde{p}(y_1, \ldots, y_n)\right]. \tag{1}$$

This can happen even for a k such that $\tilde{p} \notin M_k$. (1) is possible because $p(y_1, \ldots, y_n \mid M_k)$ is a mixture of distributions in M_k, and may thus be closer to P^* than any single element of M_k. This phenomenon, dubbed *hypercompression* and extensively studied and explained by Grünwald and Van Ommen (2017), has the following effect: if M_j for some $j \neq k$ contains $\tilde{p}$ and, at the given sample size, has its predictive distribution $p(y_n \mid y^{n-1}, M_j)$ already indistinguishable from $\tilde{p}$, yet the posterior based on M_k has not concentrated on anything near $\tilde{p}$ (or M_k does not even contain $\tilde{p}$), then M_k might still be preferred in terms of log score and hence chosen by BMA. The crucial point for the present discussion is that with stacking *based on the log score*, the preferred method of Yao et al. (2018) (see Section 3.1.), the same can happen: (1) implies that for a substantial fraction of outcomes y_i in $y_1, \ldots, y_n$, one will tend to have, with $y_{-i} := (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$, that

$$-\log p(y_i \mid y_{-i}, M_k) \ll -\log \tilde{p}(y_i), \tag{2}$$

hence also giving an advantage to M_k compared to the KL-best $\tilde{p}$ and $M_{k'}$.

But why would this be undesirable? It turns out that the predictive distribution $p(\cdot \mid y_{-i}, M_k)$ in (2) achieves being significantly closer to P^* in terms of KL divergence than any of the elements inside M_k, by *being a mixture of elements of M_k which themselves are all very 'bad', i.e. very far from P^* in terms of KL divergence* (see in particular Figure 7 and 8 of Grünwald and Van Ommen (2017)). As a result, using a log score oriented averaging procedure, whether it be BMA or stacking, one can select an M_k whose predictive is good, at sample size i, in log score, but quite bad in terms of just about any other measure. For example, consider a linear model M_k as above. For such models, for fixed σ^2, as a function of β, the KL divergence $D(P^* \| p_{\beta,\sigma^2}) := \mathbf{E}_{X\sim P^*}\mathbf{E}_{Y\sim P^*|X}[\log p^*(Y \mid X)/p_{\beta,\sigma^2}(Y \mid X)$ is linearly increasing in the mean squared error $\mathbf{E}_{X,Y\sim P^*}(Y-\beta^T X)^2$. Therefore, one commonly associates a predictive distribution $p(y_i \mid x_i)$ that behaves well in terms of log score (close in KL divergence to P^*) to be also good in predicting y_i as a function of the newly observed x_i in terms of the squared prediction error. Yet, this is true only if p is actually of the form $p_{\beta,\sigma^2} \in M_k$; the Bayes predictive distribution, being a mixture, is simply not of this form and can be good at the log score yet very bad at squared error predictions.

Now it might of course be argued that none of this matters: stacking for the log score was designed to come up with a predictive that is good in terms of log score... and it does! Indeed, if one really deals with a practical prediction problem in which one's prediction quality will be *directly* measured by log score, then stacking with the log score should work great. But to our knowledge, the only such problems are data compression problems in which log score represents codelength. In most applications in which log score is used, it is rather used for its generic properties, and then the resulting predictive distributions may be used in other ways (they may be plotted to give insight in the data, or they may be used to make predictions against other loss functions, which may not have been specified in advance). For example (Yao et al., 2018, end of Section 3.1) cite the generic properties that log score is local and proper as a reason for adopting it. Our example indicates that in the $\mathcal{M}$-open case, such use of log score for its generic properties only can give misleading results. The SafeBayesian method overcomes this problem by exponentiating the likelihood to the power η that minimizes a variation of log-score for predictive densities (the *R-log loss*, Eq. (23) in Grünwald and Van Ommen (2017)) in which loss cannot be made smaller by mixing together bad densities.

Some Details Concerning Figure 1 The conditional expectations $\mathbf{E}[Y \mid X]$ in Figure 1 are based on a simulation in which the models are trained with 30 Legendre polynomial basis functions on 50 data points, as described in Section 2. The green curve represents $\mathbf{E}[Y \mid X]$ according to the predictive distribution resulting from BMA with a uniform prior on the models, where we used the function `bms` of the `R`-package `BMS`. The red curve is based on stacking of predictive distributions, where we used the implementation with `Stan` and `R` exactly as described in the appendix of Yao et al. (2018). The black line depicts the true regression function $Y = 0$. The blue curve is `SBRidgeIlog`, which is an implementation of I-log-SafeBayesian Ridge Regression (see Grünwald and Van Ommen (2017) for details) from the `R`-package `SafeBayes` (De Heide, 2016), based on the largest model M_K. The regression functions based on M_k for all $k < K$ are even closer to $Y = 0$

(not shown). The regression function according to the Safe BMA predictive distribution is a mixture of all these Ridge-based regression functions hence also close to 0.

As Yao et al. (2018) note, the implementation of their method can be unstable when the ratio of relative sample size to the effective number of parameters is small. We encountered this unstable behaviour for a large proportion of the simulations when the sample size was relatively small, and the Pareto-k-diagnostic (indicating stability) was above 0.5, though mostly below 0.7, for some data points. In those cases the method did not give sensible outputs, irrespective of the true regression function (which we set to, among others, $Y_i = 0.5X_i + \xi_i$ and $Y_i = X_i^2 + \xi_i$, and we also experimented with a Fourier basis). Thus, we re-generated the whole sample of size $n = 50$ many times and only considered the runs in which the k-diagnostic was below 0.5 for all data points. In all those cases, we observed the overfitting behaviour depicted in Figure 1. This 'sampling towards stable behaviour' may of course induce bias. Nevertheless, the fact that we get very similar results for model selection rather than stacking (mixing) based on LOO with log-score indicates that the stacking curve in Figure 1 is representative.

References

Dawid, A. (1984). "Present Position and Potential Developments: Some Personal Views, Statistical Theory, The Prequential Approach." *Journal of the Royal Statistical Society, Series A*, 147(2): 278–292. MR0763811. doi: https://doi.org/10.2307/2981683. 957

Van Erven, T., Grünwald, P., and de Rooij, S. (2012). "Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC–BIC dilemma." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3): 361–417. With discussion, pp. 399–417. MR2925369. doi: https://doi.org/10.1111/j.1467-9868.2011.01025.x. 957

Grünwald, P. and Van Ommen, T. (2017). "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it." *Bayesian Analysis*, 12(4): 1069–1103. MR3724979. doi: https://doi.org/10.1214/17-BA1085. 957, 958, 959, 960

De Heide, R. (2016). "The Safe–Bayesian Lasso." Master's thesis, Leiden University. 957, 958, 960

Van der Pas, S. and Grünwald, P. (2018). "Almost the Best of Three Worlds: Risk, Consistency and Optional Stopping for the Switch Criterion in Nested Model Selection." *Statistica Sinica*, 28(1): 229–255. MR3752259. 957

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). "Using Stacking to Average Bayesian Predictive Distributions." *Bayesian Analysis*. 957, 958, 959, 960, 961

Contributed Discussion

A. Philip Dawid*

1 Scoring rules

The paper extends the method of stacking to allow probabilistic rather than point predictions. It does this by applying a scoring rule, which is a loss function $S(y, Q)$[1] for critiquing a quoted probabilistic forecast Q for a quantity Y in the light of Nature's realised outcome y. In retrospect this simple extension has been a long time coming. Although theory and applications of proper scoring rules have been around for at least 70 years, this fruitful and versatile concept has lain dormant until quite recently, and is still woefully unfamiliar to most statisticians. See Dawid (1986); Dawid and Sebastiani (1999); Grünwald and Dawid (2004); Parry et al. (2012); Dawid and Musio (2014, 2015); Dawid et al. (2016) for a variety of theory, examples and applications.

A scoring rule $S(y, Q)$ is a special case of a loss function $L(y, a)$, measuring the negative worth of taking an act a in some action space $\mathcal{A}$, when the variable Y turns out to have value y. In this special case, $\mathcal{A}$ is set of probability distributions. Conversely, given any action space and loss function we can construct an associated proper scoring rule $S(y, Q) := L(y, a_Q)$, where a_Q is a *Bayes act* against Q, thus minimising $L(Q, a) := \mathrm{E}_{Y\sim Q}\, L(Y, a)$. This extends stacking to general decision problems.

2 Prequential strategies

However, when we take probability forecasting seriously, there are more principled ways to proceed. Rather than assessing forecasts using a cross-validatory approach—which, though popular, has little foundational theory and does not easily extend beyond the context of independent identically distributed ("IID") observations—we can conduct *prequential* (predictive sequential) assessment (Dawid, 1984). This considers the observations in sequence,[2] and at any time t constructs a probabilistic prediction P_{t+1} for the next observable Y_{t+1}, based on the currently known outcomes $\mathbf{y}_t = (y_1, \ldots, y_t)$. Any method of doing this is a *prequential strategy*. Many (though by no means all) strategies can be formed by applying some principle to a parametric statistical model $M = \{P_\theta : \theta \in \Theta\}$ for the sequence of observables $(Y_1, Y_2, \ldots)$—which need not incorporate independence, and (in contrast to all the approaches mentioned in § 2 of the paper) not only need not be considered as containing the "true generating process", but does not even require that such a process exist (the "$\mathcal{M}$-absent" case). Example M-based strategies are the "plug-in" density forecast $p_{t+1} = p(y_{t+1} \mid \mathbf{y}_t; \widehat{\theta}_t)$, with $\widehat{\theta}_t$ the

*University of Cambridge, apd@statslab.cam.ac.uk

[1] Various notations occur in the literature. Mine here, which, compared with the paper, interchanges the arguments and takes the negative, is perhaps the most traditional.

[2] If there is no natural sequence, they can be ordered arbitrarily—the specific ordering typically makes little or no difference.

maximum likelihood estimate based on the current data $\mathbf{y}_t$; and the Bayesian density forecast $p_{t+1} = \int p(y_{t+1} \mid \mathbf{y}_t; \theta)\, \pi_t(\theta) d\theta$, with $\pi_t(\theta)$ the posterior density of θ given $\mathbf{y}_t$.

2.1 $\mathcal{M}$-closed case

Purely as a sanity check, it can be shown (Dawid, 1984) that the above M-based strategies are typically consistent for hypothetical data generated from some distribution in M (a fictitious M-closed scenario). Straightforward extensions base forecasts on a discrete collection $\mathcal{M}$ of parametric statistical models, and exhibit consistency for fictitious data generated from any distribution in any of these. Another possible $\mathcal{M}$-based prequential strategy in this case (though computationally demanding, and restricted to IID models) might be based on forecasting Y_{t+1} by applying the paper's stacking methodology (using a suitable scoring rule) to the partial data $\mathbf{y}_t$. It would be interesting to investigate its sanity under the fictitious $\mathcal{M}$-closed assumption.

2.2 $\mathcal{M}$-open case

Consider a model $M = \{P_\theta\}$, and a hypothetical generating distribution $Q \notin M$. All these distributions can exhibit dependence between observations. For data $\mathbf{y}_T$ the "best-fitting" value of θ is θ^*_T, minimising the cumulative discrepancy $\sum_{t=1}^T d(Q_t, P_{\theta,t})$, where $P_{\theta,t}$ [resp., Q_t] is the forecast distribution for Y_t given $\mathbf{y}_{t-1}$ under P_θ [resp., Q], and d is the discrepancy function associated with proper scoring rule S.[3] In the $\mathcal{M}$-closed case that $Q = P_{\theta_0} \in M$, for any choice of S we will have $\theta^*_T = \theta_0$; but otherwise θ^*_T will typically depend on S—which is why it is important to choose S appropriate to the context, and not, say, to assess an estimate based on log-score using a quadratic criterion. Moreover, in non-IID cases θ^*_T is typically data-dependent. In any case θ^*_T depends on the unknown (even fictional) Q. However, we could estimate θ^*_T by the "best-performing" value θ^{**}_T, minimising the empirical score $\sum_{t=1}^T S(y_t, P_{\theta,t})$. Under very broad conditions (typically not requiring *e.g.* ergodicity) this will be consistent even in the M-open case, in the sense that, no matter what Q may be, $\theta^{**}_T - \theta^*_T \to 0$ as $T \to \infty$, almost surely under Q (Skouras and Dawid, 2000). Again, this M-open sanity check (which subsumes the M-closed sanity check) extends to the case of a collection $\mathcal{M}$ of models; and again it would be interesting to check whether a prequential stacking strategy could preserve this property.

2.3 $\mathcal{M}$-absent case

The most incisive test of a forecasting method is its performance in the $\mathcal{M}$-absent case. The overall (negative) performance of any strategy, on a full data-set $\mathbf{y}_T$, can be assessed by its total prequential score $\sum_{t=1}^T S(y_t, P_t)$, which judges forecasts relative to actual data, so allowing direct data-driven comparison of strategies.

Consider now a model $M = \{P_\theta\}$, and (for data $\mathbf{y}_T$) the best-performing value θ^{**}_T as defined in § 2.2. There is a wealth of theory (again, woefully ignored by most

[3] Again, my notation transposes the arguments of d compared with that of the paper.

statisticians), developed in the discipline of "prediction with expert advice" (Vovk, 2001), which shows that, under suitable conditions, we can construct a prequential strategy P which *for any data whatsoever* will perform essentially as well as $P_{\theta^{**}}$; for example (Cesa-Bianchi and Lugosi, 2006; Rakhlin et al., 2015) such that, for any infinite data-sequence $\mathbf{y} = (y_1, y_2, \ldots)$,

$$\lim_{T\to\infty} T^{-1}\left\{\sum_{t=1}^{T} S(y_t, P_t) - \sum_{t=1}^{T} S(y_t, P_{\theta^{**}_T, t})\right\} = 0.$$

By adding probabilistic assumptions about the origin of $\mathbf{y}$ we can then (if so desired) recover some of the consistency results in § 2.1 and § 2.2 above. And again extension to forecasts based on a collection $\mathcal{M}$ of models is straightforward. I wonder how well stacking would perform by this criterion?

References

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning and Games*. Cambridge: Cambridge University Press. MR2409394. doi: https://doi.org/10.1017/CBO9780511546921. 964

Dawid, A. P. (1984). "Statistical Theory. The Prequential Approach (with Discussion)." *Journal of the Royal Statistical Society, Series A*, 147: 278–292. MR0763811. doi: https://doi.org/10.2307/2981683. 962, 963

Dawid, A. P. (1986). "Probability Forecasting." In Kotz, S., Johnson, N. L., and Read, C. B. (eds.), *Encyclopedia of Statistical Sciences*, volume 7, 210–218. Wiley-Interscience. MR0892738. 962

Dawid, A. P. and Musio, M. (2014). "Theory and Applications of Proper Scoring Rules." *Metron*, 72: 169–183. MR3233147. doi: https://doi.org/10.1007/s40300-014-0039-y. 962

Dawid, A. P. and Musio, M. (2015). "Bayesian Model Selection Based on Proper Scoring Rules (with Discussion)." *Bayesian Analysis*, 10: 479–521. MR3420890. doi: https://doi.org/10.1214/15-BA942. 962

Dawid, A. P., Musio, M., and Ventura, L. (2016). "Minimum Scoring Rule Inference." *Scandinavian Journal of Statistics*, 43: 123–138. MR3466997. doi: https://doi.org/10.1111/sjos.12168. 962

Dawid, A. P. and Sebastiani, P. (1999). "Coherent Dispersion Criteria for Optimal Experimental Design." *Annals of Statistics*, 27: 65–81. MR1701101. doi: https://doi.org/10.1214/aos/1018031101. 962

Grünwald, P. D. and Dawid, A. P. (2004). "Game Theory, Maximum Entropy, Minimum Discrepancy, and Robust Bayesian Decision Theory." *Annals of Statistics*, 32: 1367–1433. MR2089128. doi: https://doi.org/10.1214/009053604000000553. 962

Parry, M. F., Dawid, A. P., and Lauritzen, S. L. (2012). "Proper Local Scoring Rules." *Annals of Statistics*, 40: 561–92. MR3014317. doi: https://doi.org/10.1214/12-AOS971. 962

Rakhlin, A., Sridharan, K., and Tiwari, A. (2015). "Online Learning Via Sequential Complexities." *Journal of Machine Learning Research*, 16: 155–186. MR3333006. 964

Skouras, K. and Dawid, A. P. (2000). "Consistency in Misspecified Models." Research Report 218, Department of Statistical Science, University College London. 963

Vovk, V. (2001). "Competitive On-Line Statistics." *International Statistical Review*, 69: 213–248. 964

Contributed Discussion

William Weimin Yoo*

In this paper, Yao et al. (2018) consider the problems of model selection and aggregation of different candidate models for inference. Inspired by the stacking of means method proposed in the frequentist literature, the authors generalize this idea by developing a procedure to stack Bayesian predictive distributions. Given a list of candidate models with their corresponding predictive distributions, the goal here is to find a linear combination of these distributions such that it is as close as possible to the true date generating distribution, under some score criterion. To find the linear combination (model) weights, they replace the full predictive distributions with their leave-one-out (LOO) versions in the objective function, and proceed to solve this convex optimization problem. The authors then propose a further approximation to LOO computation by importance sampling, with the importance weights obtained by fitting a generalized Pareto distribution. To showcase the benefits of the newly proposed stacking method, the authors conducted extensive simulation studies and data analyses, by comparing with other techniques in the literature such as BMA (Bayesian Model Averaging), BMA with LOO (Psedo-BMA), BMA with LOO and Bayesian Bootstrap, and others.

We can take a graphical modeling perspective on LOO. For example, replacing marginal likelihoods $p(y|M_k)$ with $\prod_{i=1}^{n} p(y_i|y_j : j \neq i, M_k)$ is akin to simplifying a complete (fully connected) graph linking observations to one where the Markov assumption holds, i.e., the node corresponding to y_i is independent conditioned on its neighbors $\{y_j : j \neq i\}$. In the proposed stacking method, the full predictive distribution $p(\widetilde{y}|y)$ is approximated by the LOO $p(y_i|y_j : j \neq i)$, and further approximation is needed because the LOO is typically expensive to compute. However if there are some structures in the data, such as clusters of data points/nodes, then one can take advantage of this by conditioning on a smaller cluster $\mathcal{B}$ of nodes around y_i and compute instead $p(y_i|y_j : j \neq i, j \in \mathcal{B})$. This would then further speed up computations as one can fit models using only local data points.

Another point I would like to make is that the superb performance of the stacking method warrants further theoretical investigations. Figure 2(c) in the simulations shows that the proposed method is robust to incorrect/bad models, in the sense that its performance stays unchanged even if more incorrect models are added to the list. It would be nice if we will also have some theoretical guarantees that the stacking method will concentrate on the correct ($\mathcal{M}$-closed) or the best models ($\mathcal{M}$-complete) in the model list. In addition, Figure 9 shows that this method is able to "borrow strength" across different models, by using some aspects of a model to enhance performance of a different model. Therefore aggregation by stacking adds value by bringing the best out of each individual model component, and it would be interesting to characterize through theory what this added value is. This then invites us to reflect on how the quality of individual model component affects the final stacked distribution. For example, given posterior

*Mathematical Institute, Leiden University, The Netherlands, yooweimin0203@gmail.com

contraction rates for the different posterior models, what would be the aggregated rate for the stacked posterior? Also, how does prediction/credible sets constructed using the stacked posterior compare with those constructed from each model component, will it be bigger, smaller or something in between?

As most existing methods including the newly proposed stacking method use some form of linear combinations, it would be interesting to find other ways of aggregation. As pointed out by the authors, linear combinations of predictive distributions will not be able to reproduce truths that are generated through some other means, e.g., convolutions. To apply stacking in the convolutional case, I think one way is to do everything in the Fourier domain, by stacking log Fourier transforms (i.e., log characteristic functions) of the predictive posterior densities, exponentiate and then apply inverse Fourier transform to approximate the truth generated through convolutions.

I think another possible area of application for the stacking method is in distributed computing. In this modern big data era, data has grown to such size that it is sometimes infeasible or impossible to analyze them using standard Markov Chain Monte Carlo (MCMC) algorithms on a single machine. Hence this gave rise to the divide and conquer strategy where data is first divided into batches and (sub)posterior distribution is computed for each batch. The final posterior for inference is then obtained by aggregating these (sub)posteriors. To this end, I think the present stacking method can be deployed after some modifications, with potential to yield superior performance when compared with existing weighted average-type approaches.

I find the paper to be very interesting and the stacking method proposed is a key contribution to the Bayesian model averaging/selection literature. It is shown to be superior than the golden standard, i.e., BMA and its finite sample performance is tested comprehensively though a series of numerical experiments and real data analyses. I think this is a very promising research direction, and any future contributions are very welcomed.

References

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). "Using stacking to average Bayesian predictive distributions." *Bayesian Analysis*, 1–28. Advance publication. 966

Contributed Discussion

Robert L. Winkler*, Victor Richmond R. Jose†,
Kenneth C. Lichtendahl Jr.‡, and Yael Grushka-Cockayne§

Yao et al. (2018) propose a stacking approach for aggregating predictive distributions and compares its approach with several alternatives, such as Bayesian model averaging and mean stacking. Because distributions are more informative than point estimates and aggregating distributions can increase the information upon which decisions are made, any paper that encourages the use of distributions and their aggregation is important and most welcome. Therefore, we applaud the authors for bringing attention to these issues and for their careful approach to distribution stacking using various scoring rules.

To stack distributions, the authors propose the use of convex combinations of competing models' cdfs. The different cdf stackers they consider vary in the (non-negative) weights assigned to each model in the combination. This cdf stacking approach is framed by the logic of Bayesian model averaging: there is a prior distribution over which model is correct and the data are generated by this one true model. The result is a convex combination of the competing models, where the weights follow from the prior distribution over which model is correct.

The idea of a "true model" is in the spirit of hypothesis testing in the sense that it is black and white–one truth that we try to identify. In most real-world situations, however, the existence of a "true model" is highly doubtful at best. A better approach may be to think in terms of information aggregation from multiple inputs/forecasters/models. Under this approach, no base model is true, but all provide some useful information that is worth combining. The following example illustrates this point and motivates the idea of aggregating information with a quantile stacker.

Suppose there are $k \geq 2$ forecasters. Forecaster i privately observes the sample $\boldsymbol{x}_i = (x_{N_{i-1}+1}, \ldots, x_{N_i})$ of size n_i for $i = 1, \ldots, k$, where $N_i = \sum_{j=1}^{i} n_j$. Each forecaster will report their quantile function Q_i for the uncertain quantity of interest x_{N_k+1}. The data are drawn from the normal-normal model: $\mu \sim N(\mu_0, m\lambda)$ and $(x_j|\mu) \sim_{iid} N(\mu, \lambda)$ for $j = 1, \ldots, N_k + 1$, where λ denotes the precision.

Forecaster i's posterior-predictive beliefs are $(x_{N_k+1}|\boldsymbol{x}_i) \sim N(\mu_i, \lambda_i)$, where $\mu_i = m/(m+n_i)\mu_0 + n_i/(m+n_i)\bar{x}_i$, $\bar{x}_i = (1/n_i)\sum_{j=1}^{n_i} x_{N_{i-1}+j}$, and $\lambda_i = (m+n_i)/(m+n_i+1)\lambda$ (Bernardo and Smith, 2000, p. 439). The corresponding quantile function is $Q_i(u) = \mu_i + \lambda_i^{-1/2}\Phi^{-1}(u)$, where Φ is the standard normal cdf. Once the decision maker hears the forecaster's updated quantile functions, his updated beliefs are $(x_{N_k+1}|Q_1, \ldots, Q_k) \sim N(\mu_{dm}, \lambda_{dm})$, where $\mu_{dm} = m/(m+N_k)\mu_0 + N_k/(m+N_k)\bar{x}$, $\bar{x} = (1/N_k)\sum_{i=1}^{k} n_i\bar{x}_i$, and $\lambda_{dm} = (m+N_k)/(m+N_k+1)\lambda$. The decision maker's posterior-predictive quantile

*The Fuqua School of Business, Duke University, rwinkler@duke.edu
†McDonough School of Business, Georgetown University, vrj2@georgetown.edu
‡Darden School of Business, University of Virginia, lichtendahlc@darden.virginia.edu
§Darden School of Business, University of Virginia, grushkay@darden.virginia.edu

function $Q_{dm}(u) = \mu_{dm} + \lambda_{dm}^{-1/2}\Phi^{-1}(u)$ can be rewritten as

$$\begin{aligned} Q_{dm}(u) &= \frac{m}{m+N_k}\mu_0 + \frac{N_k}{m+N_k}\sum_{i=1}^{k}\frac{n_i\bar{x}_i}{N_k} + \lambda_{dm}^{-1/2}\sum_{i=1}^{k}\frac{\Phi^{-1}(u)}{k} \\ &= \frac{m-km}{m+N_k}\mu_0 + \sum_{i=1}^{k}\left(\frac{m+n_i}{m+N_k} - \frac{1}{k}\frac{\lambda_i^{1/2}}{\lambda_{dm}^{1/2}}\right)\mu_i + \sum_{i=1}^{k}\frac{1}{k}\frac{\lambda_i^{1/2}}{\lambda_{dm}^{1/2}}Q_i(u). \end{aligned} \tag{1}$$

According to the example above, the result of aggregating the forecasters' information is mean/quantile stacking. The mean/quantile stacker in (1) is a linear combination of the prior-predictive mean, the forecasters' posterior-predictive means, and the forecaster's posterior-predictive quantiles. Interestingly, only the weights on the quantiles are necessarily positive. Choosing this linear combination to optimize a quantile scoring rule (Jose and Winkler, 2009; Gneiting, 2011; Grushka-Cockayne et al., 2017), such as the pinball loss function, may be a useful way to choose the weights.

In fact, a quantile stacker solves the example in Yao et al's Section 4.1 exactly and would be perfectly calibrated. The quantile stacker $0.6Q_3(u)+0.4Q_4(u)$, where $Q_i(u) = \mu_i+\Phi^{-1}(u)$ and $\mu_i = i$, yields the aggregate quantile function $3.4+\Phi^{-1}(u)$, which is the quantile function of the true data-generating process in Section 4.1. A quantile stacker may also work well in the regression example of Section 4.2.

When aggregating model forecasts of a continuous quantity of interest, each model may be well-calibrated to the training data, although each model may have different sharpness. In this case, a result in Hora (2004) kicks in. The convex combination of well-calibrated models' cdfs cannot be well-calibrated (unless they are all identical). Typically, the convex combination of cdfs will lead to an underconfident aggregate cdf (Lichtendahl Jr et al., 2013). The same holds for forecasts of a binary event; see Ranjan and Gneiting (2010). Because the average quantile forecast is always sharper than the average probability forecast (Lichtendahl Jr et al., 2013), the average quantile may be better calibrated when the average cdf is underconfident.

In the case of aggregating binary-event forecasts, such as the example in Yao et al's Section 4.6, it might be optimal to transform the probabilities prior to combining them in a generalized linear model. Lichtendahl Jr et al. (2018) propose a Bayesian model in the spirit of the example given here. The model results in a generalized additive model for combining the base model's binary-event forecasts.

We are thankful that the authors highlight the importance of combining predictive distributions, and we hope this paper stimulates further work in this area.

References

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. Chichester, England: John Wiley & Sons, Ltd. MR1274699. doi: https://doi.org/10.1002/9780470316870. 968

Gneiting, T. (2011). "Quantiles as optimal point forecasts." *International Journal of Forecasting*, 27: 197–207. 969

Grushka-Cockayne, Y., Lichtendahl Jr, K. C., Jose, V. R. R., and Winkler, R. L. (2017). "Quantile evaluation, sensitivity to bracketing, and sharing business payoffs." *Operations Research*, 65(3): 712–728. MR3655435. doi: https://doi.org/10.1287/opre.2017.1588. 969

Hora, S. C. (2004). "Probability judgments for continuous quantities: Linear combinations and calibration." *Management Science*, 50: 597–604. 969

Jose, V. R. R. and Winkler, R. L. (2009). "Evaluating quantile assessments." *Operations Research*, 57: 1287–1297. MR2583517. doi: https://doi.org/10.1287/opre.1080.0665. 969

Lichtendahl Jr, K. C., Grushka-Cockayne, Y., Jose, V. R. R., and Winkler, R. L. (2018). "Bayesian ensembles of binary-event forecasts: When is it appropriate to extremize or anti-extremize?" URL https://arxiv.org/abs/1705.02391. 969

Lichtendahl Jr, K. C., Grushka-Cockayne, Y., and Winkler, R. L. (2013). "Is it better to average probabilities or quantiles?" *Management Science*, 59(7): 1594–1611. 969

Ranjan, R. and Gneiting, T. (2010). "Combining probability forecasts." *Journal of the Royal Statistical Society Series B*, 72(1): 71–91. MR2751244. doi: https://doi.org/10.1111/j.1467-9868.2009.00726.x. 969

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). "Using stacking to average Bayesian predictive distributions." *Bayesian Analysis*. 968

Contributed Discussion

Kenichiro McAlinn*, Knut Are Aastveit†, and Mike West‡

1 Combination of Predictive Densities

The authors address model combination when the data generating process is unavailable ($\mathcal{M}$-complete) or unattainable ($\mathcal{M}$-open). A starting-point is that Bayesian model averaging (BMA) is widely applied in contexts that violate the $\mathcal{M}$-closed assumptions– the assumptions under which it is the "right way" to combine predictive densities. The failure of BMA is highlighted by the authors in examples. Some of our own interests are in time series forecasting, with multiple examples of BMA degenerating fast to the wrong model, and failing to adapt to temporal shifts in the data. This is not a failure of Bayesian thinking, of course; the message remains follow-the-rules, but understand when assumptions fail to justify blind application of the rules. The authors' methodology relates to other methods responding to this. For example, optimal pooling (Geweke and Amisano, 2011) as referenced by the authors, and the DeCo approach (Billio et al., 2013; Aastveit et al., 2018), address the problem explicitly in the $\mathcal{M}$-incomplete setting. Focused on time series examples, the conceptual basis of these works is of course much broader.

The authors development of density combination methods based on prior uses of stacking in point estimation also parallels the historical development of Bayesian density combination methods following the literature on combination of point forecasts (e.g. Hall and Mitchell, 2007; Geweke and Amisano, 2011; Billio et al., 2013; Aastveit et al., 2014; Kapetanios et al., 2015; Aastveit et al., 2018; Del Negro et al., 2016, and references therein). Then, much of the density combination literature has grown from somewhat algorithmic and empirical model-based perspectives. We regard the stacking approach as largely adopting this perspective. While the authors argue cogently for the construction, connect with what they term "pseudo"-Bayesian thinking and aspects of asymptotics, the import of the work is largely– and strongly– based on the examples and empirical evaluation. We are led to ask, is this new approach to "averaging Bayesian predictive distributions"... really Bayesian?

2 Bayesian Predictive Synthesis

The recently developed approach of Bayesian predictive synthesis (BPS: McAlinn and West, 2018; McAlinn et al., 2018) is explicitly founded on subjective Bayesian principles and theory, and defines an over-arching framework within which many existing density (and other) combination "rules" can be recognized as special cases. Critically, this provides opportunity to understand the implicit Bayesian assumptions underlying special

*Booth School of Business, University of Chicago, kenichiro.mcalinn@chicagobooth.edu
†Norges Bank, BI Norwegian Business School, knut-are.aastveit@norges-bank.no
‡Department of Statistical Science, Duke University, mike.west@duke.edu

cases. BPS links to past literature on subjective Bayesian "agent/expert opinion analysis" (e.g. Genest and Schervish, 1985; West and Crosse, 1992; West, 1992) and provides a formal Bayesian framework that regards predictive densities from multiple models (or individuals, agencies, etc) as data to be used in prior-posterior updating by a Bayesian observer (see also West and Harrison, 1997, Sect 16.3.2). The approach allows for the integration of other sources of information and explicitly provides the ability to deal with $\mathcal{M}$-incompleteness. A main theoretical component of BPS is a general theorem describing a subset of Bayesian analyses showing how densities can be "synthesized". Special cases include traditional BMA, most existing forecast pooling rules, and– in terms of theoretical construction– the stacking approach in the article.

In McAlinn and West (2018) and McAlinn et al. (2018) BPS is developed for time series forecasting where the underlying Bayesian foundation defines a class of dynamic latent factor models. The sequences of predictive densities define time-varying priors for inherent latent factor processes linked to the time series of interest. BPS is able to learn and adapt to the biases, aspects of mis-calibration, and– critically– inter-dependences among predictive densities. A further practical key point is that BPS can– and should– be defined with respect to specific predictive goals; this is a point of wider import presaged in the earlier Bayesian macroeconomics literature and illustrated in McAlinn and West (2018) and McAlinn et al. (2018) through separate forecast combination models for multiple different goals (multiple-step ahead forecasting). Applications in macroeconomic forecasting in these papers demonstrate that a class of proposed BPS models can significantly improve over conventional methods (including BMA and other pooling/weighting schemes). Further, as BPS is a fully-specified Bayesian model within which the information from each of the sources generating predictive density are treated as (complicated) "covariates," posterior inferences on (time-varying or otherwise) parameters weighting and relating the sources provides direct access to inferences on their biases and inter-dependencies.

It is of interest to consider how the current stacking approach relates to BPS through an understanding of how the resulting rule for predictive density combination can be interpreted in BPS theory (see equation (1), and the discussion thereafter, in McAlinn and West 2018). As with other combination rules, an inherent latent factor interpretation is implied and this may provide opportunity for further development. In related work with BPS based on mixture models, Johnson and West (2018) highlight the opportunities to improve both resulting predictions and generate insights about the practical impact of model inter-dependencies that are largely ignored by other approaches. This can be particularly important in dealing with larger numbers of predictive densities when the underlying models generating the densities are known or expected to have strong dependencies (a topic touched upon in Section. 5.3 in the article).

References

Aastveit, K. A., Gerdrup, K. R., Jore, A. S., and Thorsrud, L. A. (2014). "Nowcasting GDP in real time: A density combination approach." *Journal of Business & Economic Statistics*, 32: 48–68. MR3173707. doi: https://doi.org/10.1080/07350015.2013.844155. 971

Aastveit, K. A., Ravazzolo, F., and van Dijk, H. K. (2018). "Combined density Nowcasting in an uncertain economic environment." *Journal of Business & Economic Statistics*, 36: 131–145. MR3750914. doi: https://doi.org/10.1080/07350015.2015.1137760. 971

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. (2013). "Time-varying combinations of predictive densities using nonlinear filtering." *Journal of Econometrics*, 177: 213–232. MR3118557. doi: https://doi.org/10.1016/j.jeconom.2013.04.009. 971

Del Negro, M., Hasegawa, R. B., and Schorfheide, F. (2016). "Dynamic prediction pools: an investigation of financial frictions and forecasting performance." *Journal of Econometrics*, 192(2): 391–405. MR3488085. doi: https://doi.org/10.1016/j.jeconom.2016.02.006. 971

Genest, C. and Schervish, M. J. (1985). "Modelling expert judgements for Bayesian updating." *Annals of Statistics*, 13: 1198–1212. MR0803766. doi: https://doi.org/10.1214/aos/1176349664. 972

Geweke, J. F. and Amisano, G. G. (2011). "Optimal prediction pools." *Journal of Econometrics*, 164: 130–141. MR2821798. doi: https://doi.org/10.1016/j.jeconom.2011.02.017. 971

Hall, S. G. and Mitchell, J. (2007). "Combining density forecasts." *International Journal of Forecasting*, 23: 1–13. 971

Johnson, M. C. and West, M. (2018). "Bayesian predictive synthesis: Forecast calibration and combination." Technical report. ArXiv:1803.01984. MR3755068. 972

Kapetanios, G., Mitchell, J., Price, S., and Fawcett, N. (2015). "Generalised density forecast combinations." *Journal of Econometrics*, 188: 150–165. MR3371665. doi: https://doi.org/10.1016/j.jeconom.2015.02.047. 971

McAlinn, K., Aastveit, K. A., Nakajima, J., and West, M. (2018). "Multivariate Bayesian predictive synthesis in macroeconomic forecasting." *Submitted*. ArXiv: 1711.01667. 971, 972

McAlinn, K. and West, M. (2018). "Dynamic Bayesian predictive synthesis in time series forecasting." *Journal of Econometrics*, to appear. ArXiv:1601.07463. MR3664859. 971, 972

West, M. (1992). "Modelling agent forecast distributions." *Journal of the Royal Statistical Society (Series B: Methodological)*, 54: 553–567. MR1160482. 972

West, M. and Crosse, J. (1992). "Modelling of probabilistic agent opinion." *Journal of the Royal Statistical Society (Series B: Methodological)*, 54: 285–299. MR1157726. 972

West, M. and Harrison, P. J. (1997). *Bayesian Forecasting & Dynamic Models*. Springer Verlag, 2nd edition. MR1482232. 972

Contributed Discussion

Minsuk Shin*

Overview

I congratulate the authors on an improvement in predictive performance of Bayesian model averaging. This improvement is considerably significant under $\mathcal{M}$-open settings (Bernardo and Smith, 2009) where we know that the true model is not one of the considered candidate models. As the authors pointed out, it is well-known that the weights of standard Bayesian model averaging (BMA) asymptotically concentrates on a single model that is closest to the true model in Kullback–Leibler (KL) divergence. This would be problematic under $\mathcal{M}$-open settings, because a single (wrong) model would dominate the predictive inference so that its predictive performance might be attenuated.

To address this issue, the authors bring the *stacking* idea to BMA. Instead of minimizing the mean square error of the point estimate as in original stacking procedures, they propose a procedure to evaluate the model weights that maximizes the empirical scoring rule based on the leave-one-out (LOO). This results in a convex combination of models that is empirically close (in terms of the considered score rule or the divergence function) to the true model that generates the observed data. The authors also circumvent the computational intensity of the LOO procedure by adopting Pareto smoothed importance sampling (Vehtari et al., 2017). This importance sampling procedure uses the importance sampling weights approximated by the order statistics of the fitted generalized Pareto distribution, so refitting each model n times can be avoided.

Some Issues in Extensions to High-dimensional Model Selection

I think that the proposed work is very interesting under a situation where the number of models is fixed as n increases. It would be also interesting to investigate theoretical properties of the stacking procedure under high-dimensional model (or variable) selection regime (Narisetty et al., 2014; Castillo et al., 2015). In theory of high-dimensional variable selections, it is not uncommon to assume that the number of predictors, saying p, increases at a certain rate of n. When p is fixed and only n increases, the asymptotic results in Section 3.2 should hold. However, when p increases faster than n, the uniform convergence (over models) of the estimator of the score rule may not hold, because the total number of models increases at an exponential rate of p, that is 2^p. It is well-known that *Akaike Information Criterion* (AIC) is asymptotically equivalent to LOO, and AIC is not consistent in model selection even under low-dimensional settings. So, the stacking procedure based on LOO might not be optimal in selecting the best model under high-dimensional settings, and might suffer from high false discovery rates.

*Department of Statistics, Harvard University, Cambridge, MA, mshin@fas.harvard.edu

Also, in computational aspects for high-dimensional variable selection (for linear regression models), it would be interesting to consider a stochastic search algorithm such as shotgun stochastic search (SSS) (Hans et al., 2007) that is developed for standard Bayesian variable selection. By slightly modifying the SSS algorithm, one might be able to explore models having high predictive densities. However, the number of candidate models is enormous under high-dimensional settings, and the approximation of the predictive densities for all the candidate models is computationally intensive even when the Pareto importance sampling technique is used. A possible option might be a two-step procedure that first collects a number of models by using a standard Bayesian variable selection procedure such as (George and McCulloch, 1993; Raftery et al., 1997; Hans et al., 2007), then the stacking weights can be evaluated based on the pre-specified models. However, standard Bayesian variable selection procedures choose models with large posterior model probability that is proportional to the marginal likelihood, so the pre-selected models might not be optimal in prediction.

Conclusion

Once again, I would like to congratulate the authors of this paper, and I think that it would be interesting to extend the idea to high-dimensional model selection problems.

References

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian Theory*, volume 405. John Wiley & Sons. MR1274699. doi: https://doi.org/10.1002/9780470316870. 974

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al. (2015). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 974

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 975

Hans, C., Dobra, A., and West, M. (2007). "Shotgun stochastic search for "large p" regression." *Journal of the American Statistical Association*, 102(478): 507–516. MR2370849. doi: https://doi.org/10.1198/016214507000000121. 975

Narisetty, N. N., He, X., et al. (2014). "Bayesian variable selection with shrinking and diffusing priors." *The Annals of Statistics*, 42(2): 789–817. MR3210987. doi: https://doi.org/10.1214/14-AOS1207. 974

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association*, 92(437): 179–191. MR1436107. doi: https://doi.org/10.2307/2291462. 975

Vehtari, A., Gelman, A., and Gabry, J. (2017). "Pareto smoothed importance sampling." *arXiv preprint arXiv:1507.02646*. 974

Contributed Discussion

Tianjian Zhou*

I congratulate the authors for an insightful discussion of combining predictive distributions using stacking. Consider a set of predictive distributions built from multiple models, $p(\tilde{y} \mid y, M_k)$ for $k = 1, \ldots, K$. A combination of these predictive distributions has the form $\hat{p}(\tilde{y} \mid y) = \sum_{k=1}^{K} w_k p(\tilde{y} \mid y, M_k)$. The coherent Bayesian method of choosing the weights (w_k's) is Bayesian model averaging (BMA) with posterior model probabilities. However, this approach has several limitations. The authors therefore propose stacking of predictive distributions, which chooses the weights by minimizing divergence of $\hat{p}(\tilde{y} \mid y)$ from the true distribution.

Model Space and Computational Complexity Some considerations of BMA still apply also for stacking. For example, the choice of candidate models. For a regression analysis with p predictors, there are 2^p possible linear subset regression models. One could potentially consider most or all of them, as in Section 4.5. Sometimes it is also necessary to consider interaction and nonlinear terms, as in Section 4.6. Thus, the size of the model space can be enormous, making computation infeasible for reasonably large p. Similar strategies as for implementing BMA could potentially also be used for stacking.

Compared to BMA, stacking is more appropriate for the $\mathcal{M}$-open case. That is, the true model need not be in the model space $\mathcal{M}$. Still, it would be better if the true model or a model close to the true model were included in the set of candidate models. For example, comparing Figures 3 and 4 the performance of stacking is better under the $\mathcal{M}$-closed case compared to the $\mathcal{M}$-open case, in terms of predictive density. Thus, from a practical perspective, it is still desirable to work with a reasonably large class of models.

In all examples, the candidate models belong to the same model family (e.g., of regression models). It would be interesting to see if performance can be further improved by combining models that belong to different model families. For example, combining linear regression models and regression tree models.

Alternatives As discussed by the authors, sometimes it is preferable to fit a supermodel that includes the separate models as special cases. For example, a regression model that includes all predictors and possible interaction and nonlinear terms, using, for example, a spike-and-slab prior for the regression coefficients. I would like to point out another interesting alternative, Bayesian nonparametric (BNP) models. BNP models do not necessarily include the separate models as special cases, but can be chosen to put positive prior probability mass everywhere, that is, within arbitrary neighborhoods of any model. In that sense, BNP models are "always right". Importantly, BNP priors make fewer parametric assumptions and are highly flexible. For example, for Section 4.1,

*Research Institute, NorthShore University HealthSystem, tjzhou95@gmail.com

a natural choice would be the widely used Dirichlet process mixture model. For Section 4.6, a natural choice could be a Bayesian additive regression tree (BART) model or a Gaussian process prior. Of course, BNP models have their own limitations, such as often difficult implementation, complex computation and possibly poor performance with small sample size. A possible way out could be to include BNP models as candidate models in stacking. Little would change in the overall setup.

Applications I would like to suggest potential applications of the proposed method to biomarker discovery and the prediction of patient outcomes. A biomarker refers to a measurable biological signal whose presence is indicative of some outcome or disease. Examples of biomarkers include physical characteristics such as height, weight, blood pressure, as well as genomic level measurements such as gene expression. Outcomes of interest include, for example, the presence or progression of cancer and risk of mortality or risk of readmission. The goals are: (1) finding a subset of biomarkers that are highly predictive of the outcome of interest, and (2) predicting the outcome of interest for a future patient. Traditional variable selection approaches ignore model uncertainty and use a single set of selected biomarkers for prediction. Model combination approaches, on the other hand, report a set of possible models and average over these models for prediction. Therefore, model combination approaches might give more sensible biomarker selections and have better prediction accuracy. There are some existing works using BMA (e.g. Volinsky et al. 1997 and Yeung et al. 2005). Considering the better performance of stacking relative to BMA, stacking should be considered for the same applications and could improve existing analyses.

There are some challenges for using stacking, specific to this application. First, the number of candidate biomarkers can be large, especially for genomic level biomarkers. Biomarkers can also have non-additive and nonlinear effects on the outcomes. Second, missing data are common. Third, it is very common that the patient population is heterogeneous, and different subgroups of patients might call for different outcome-predictive biomarkers. I have discussed the first issue in the second paragraph. For the second issue, the candidate models should be able to handle missing data. To reduce the sensitivity of the analysis with respect to missing data, we could average over multiple models with different strategies of dealing with the missing data, and different assumptions about the missingness. Similarly, for the third issue, some candidate models should be included that account for the possibility of subgroup effects.

References

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A. (1997). "Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(4): 433–448. MR1765176. doi: https://doi.org/10.1214/ss/1009212519. 977

Yeung, K. Y., Bumgarner, R. E., and Raftery, A. E. (2005). "Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data." *Bioinformatics*, 21(10): 2394–2402. 977

Contributed Discussion*

Lennart Hoogerheide† and Herman K. van Dijk‡

Basic practice and probabilistic foundation A basic practice in macroeconomic and financial forecasting[1] is to make use of a weighted combination of forecasts from several sources, say models, experts and/or large micro-data sets. Let y_t be the variable of interest, and assume that some form of predictive values $\tilde{y}_{1t}, \ldots, \tilde{y}_{nt}$ is available for y_t with a set of weights $w_{1t}, \ldots, w_{nt}$ where n is also a decision variable. Then, basic practice in economics and finance is to make use of:

$$\sum_{i=1}^{n} w_{it} \tilde{y}_{it}. \tag{1}$$

A major purpose of academic and professional forecasting is to give this practice a probabilistic foundation in order to quantify the uncertainty of such predictive density features as means, volatilities and tail behaviour. A leading example of a forecast density being produced and used in practice is the Bank of England's fan chart for GDP growth and inflation, which has been published each quarter since 1996. For a survey on the evolution of density forecasting in economics, see Aastveit et al. (2018) and for a related formal Bayesian foundational motivation, see McAlinn and West (2018).

Proposal by Yao, Vehtari, Simpson and Gelman In recent literature and practice in statistics as well as in econometrics, it is shown that Bayesian Model Averaging (BMA) has its limitations for forecast averaging, see the earlier reference for a summary of the literature in economics. The authors focus in their paper on the specific limitation of BMA when the true data generating process is not in the set and also indicate the sensitivity of BMA in case of weakly or non-informative priors. As a better approach in terms of forecast accuracy and robustness, the authors propose the use of *stacking*, which is used in point estimation, and extend it to the case of combinations of predictive densities. A key step in the stacking procedure is that an optimisation step is used to determine the weights of a mixture model in such a way that the averaging method is then relatively robust for misspecified models, in particular, in large samples.

Dynamic learning to average predictively We fully agree that BMA has the earlier mentioned restrictions. However, we argue that a static approach to forecast averaging,

*This comment should not be reported as representing the views of Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank. An extended version is available as Tinbergen DP.

†VU University Amsterdam and Tinbergen Institute, l.f.hoogerheide@vu.nl

‡Erasmus University Rotterdam, Norges Bank and Tinbergen Institute, hkvandijk@ese.eur.nl

[1]In applied statistics and economics the usual terminology is forecasting instead of prediction. In more theoretical work the usage of the word prediction is common. In this comment we use both terminologies interchangeably.

as suggested by the authors, will in many cases remain sensitive for the presence of a bad forecast and extremely sensitive to a very bad forecast. We suggest to extend the approach of the authors to a setting where learning about features of predictive densities of possibly incomplete or misspecified models can take place. This extension will improve the process of averaging over good and bad forecasts. To back-up our suggestion, we summarise how this has been developed in empirical econometrics in recent years by Billio et al. (2013), Casarin et al. (2018), and Baştürk, Borowska, Grassi, Hoogerheide, and Van Dijk (2018). Moreover, we show that this approach can be extended to combining not only forecasts but also policies. The technical tools necessary for the implementation refer to filtering methods from the nonlinear time series literature and we show their connection with dynamic machine learning.

The fundamental predictive density combination Let the predictive probability distribution of the variable of interest y_t of (1), given the set $\tilde{\boldsymbol{y}}_t = (\tilde{y}_{1t}, \ldots, \tilde{y}_{nt})'$, be specified as a large discrete mixture of conditional probabilities of y_t given $\tilde{\boldsymbol{y}}_t$ coming from n different models with weights $\boldsymbol{w}_t = (w_{1t}, \ldots, w_{nt})'$ that are interpreted as probabilities and form a convex combination. One can then give (1) a stochastic interpretation using mixtures. Such a probability model, in terms of densities, is given as:

$$f(y_t|\tilde{\boldsymbol{y}}_t) = \sum_{i=1}^{n} w_{it} f(y_t|\tilde{y}_{it}). \tag{2}$$

Let the predictive densities from the n models be denoted as $f(\tilde{y}_{it}|I_i), i = 1, \ldots, n$, where I_i is the information set of model i. Given the *fundamental* density combination model of (2) and the predictive densities from the n models, one can specify, given standard regularity conditions about existence of sums and integrals, that the marginal predictive density of y_t is given as a discrete/continuous mixture,

$$f(y_t|I) \sim \sum_{i=1}^{n} w_{it} \int f(y_t|\tilde{y}_{it}) f(\tilde{y}_{it}|I_i) d\tilde{y}_{it} \tag{3}$$

where I is the joint information set of all models. The numerical evaluation of this equation is simple when all distributions have known simulation properties. An important research line in economics and finance has been to make this approach operational to more realistic environments by allowing for model incompleteness and dynamic learning where the densities have no known simulation properties; see the earlier cited references.

Mixtures with model incompleteness and dynamic weight learning A first step is to introduce, possibly, time-varying model incompleteness by specifying a Gaussian mixture combination model with time varying volatility which controls the potential size of the misspecification in all models in the mixture. When the uncertainty level tends to zero then the mixture of experts or the smoothly mixing regressions model is recovered as limiting case, see Geweke and Keane (2007), Jacobs et al. (1991). The weights can be interpreted as a convex set of probabilistic weights of different models which are updated periodically using *Bayesian learning* procedures. One can write the model in the form of a nonlinear state space which allows to make use of algorithms

Data driven time series models for asset returns			
Model forecasts	VAR	SV	DFM
	↓	↓	↓
Hidden layer 1 of unobserved weights	w_{11}	w_{12}	w_{13}
	↘	↓	↙
Forecast combination	Mixture of predictive densities		
Data driven portfolio strategies			
Strategies	Model Momentum		Residual Momentum
	↓		↓
Hidden layer 2 of unobserved weights	w_{21}		w_{22}
	↘		↙
Policy combination	Mixture of policies		

Figure 1: Data driven density combinations and machine learning.

based on sequential Monte Carlo methods such as particle filters in order to approximate combination weight and predictive densities.

Forecast combinations, policy combinations and machine learning To extend the predictive density combination approach to a policy combination one with time-varying learning weights is a very natural step in economics. We summarise in Figure 1 how to combine forecasts and policies using a two-layer mixture. That is, we start with a mixture of predictive densities of three data driven time series models, *i.e.* a Vector-Autoregressive model (VAR), a Stochastic Volatility model (SV) and a Dynamic Factor Model (DFM). These are combined with a mixture of two data driven portfolio strategies that are known as momentum strategies. For background on the model and residual momentum strategies we refer to Baştürk, Borowska, Grassi, Hoogerheide, and Van Dijk (2018). It is noteworthy that this graphical representation is similar to the one used in machine learning. In our procedure the unobserved weights are integrated out using (particle) filtering. Our empirical results, see Baştürk, Borowska, Grassi, Hoogerheide, and Van Dijk (2018), show that the choice of a model set in a mixture is important for effective policies. We emphasise that this approach is fully Bayesian and does not contain an optimisation step as is used in stacking approach. However, the optimisation can be easily made dynamic. For a similar technique used in optimal pooling of forecasts we refer to Geweke and Amisano (2011).

References

Aastveit, K. A., Mitchell, J., Ravazzolo, F., and van Dijk, H. K. (2018). "The evolution of forecast density combinations in economics." Forthcoming in the Oxford Research Encyclopaedia in Economics and Finance. 978

Baştürk, N., Borowska, A., Grassi, S., Hoogerheide, L., and Van Dijk, H. K. (2018). "Learning Combinations of Bayesian Dynamic Models and Equity Momentum Strategies." *Journal of Econometrics*, Forthcoming. 979, 980

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. (2013). "Time-varying combinations of predictive densities using nonlinear filtering." *Journal of Econometrics*, 177: 213–232. MR3118557. doi: https://doi.org/10.1016/j.jeconom.2013.04.009. 979

Casarin, R., Grassi, S., Ravazzolo, F., and van Dijk, H. K. (2018). "Predictive Density Combinations with Dynamic Learning for Large Data Sets in Economics and Finance." Technical report. 979

Geweke, J. and Amisano, G. (2011). "Optimal prediction pools." *Journal of Econometrics*, 164(1): 130–141. MR2821798. doi: https://doi.org/10.1016/j.jeconom.2011.02.017. 980

Geweke, J. and Keane, M. (2007). "Smoothly mixing regressions." *Journal of Econometrics*, 138: 252–290. MR2380699. doi: https://doi.org/10.1016/j.jeconom.2006.05.022. 979

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). "Adaptive mixtures of local experts." *Journal of Neural Computation*, 3: 79–87. 979

McAlinn, K. and West, M. (2018). "Dynamic Bayesian predictive synthesis for time series forecasting." *Journal of Econometrics*. Forthcoming. MR3664859. 978

Contributed Discussion

Haakon C. Bakka[*], Daniela Castro-Camilo[†], Maria Franco-Villoria[‡], Anna Freni-Sterrantino[§], Raphaël Huser[¶], Thomas Opitz[‖], and Håvard Rue[**]

The problem of estimating the leave-one-out predictive density (LOOPD) for a model can be clarified when considering a regression-type setup. Let η be the linear predictor for conditionally independent data y, so that y_i relates to η_i only through the likelihood $p(y_i|\eta_i)$. For simplicity, we fix and then ignore the remaining variables (see Rue et al., 2009, Sec. 6.3, for a more general treatment). We can compute the LOOPD from $p(\eta_i|y_{-i}) \propto p(\eta_i|y)/p(y_i|\eta_i)$, noting that $p(y_i|\eta_i)$ is a known function of η_i. Suppose that we can estimate $p(\eta_i|y)$ well in the region $[\mu_i - \gamma\sigma_i, \mu_i + \gamma\sigma_i]$ (with obvious notation), and that this region contains most of the probability mass. The question is whether the correction needed for removing y_i (i.e., the denominator $p(y_i|\eta_i)$) is "small enough" so that also $p(\eta_i|y_{-i})$ has most of its probability mass in the same region. If so, computing $p(\eta_i|y_{-i})$ by correcting $p(\eta_i|y)$ in this way is stable; otherwise, it is potentially unreliable and should be computed from a rerun without y_i. Depending on the inference algorithm, initial values can be extracted from the full model to speed up the corrected run. Following this rationale, (R-)INLA (Rue et al., 2009; Martins et al., 2013; Rue et al., 2017; Bakka et al., 2018) compute LOOPDs using integrated nested Laplace approximations. Cases where the above test does not hold are marked as "failures". The failed cases can then be recomputed after the corresponding observations are removed, and we gain speed by using the joint fit as initial values. In addition to being faster than Markov chain Monte Carlo methods, we also get smooth estimates of the posterior marginals, which helps the optimisation step for the weights. Held et al. (2010) discuss this approach in more details and compare it with estimates obtained by Markov chain Monte Carlo.

Recently, Bakka et al. (2018) used leave-one-out cross-validation (LOOCV) log-scores in spatial modelling. That paper introduces the Barrier model, a non-stationary model dealing with coastlines and other physical barriers. The goal was to compare several spatial and non-spatial models through LOOCV. When comparing several models using the mean LOOCV log-score, we always end up choosing one model as "the best". However, such a way to rank models ignores uncertainty. With our dataset, a

[*]CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, haakon.bakka@kaust.edu.sa

[†]CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, daniela.castro@kaust.edu.sa

[‡]Department of Economics and Statistics, University of Torino, Italy, maria.francovilloria@unito.it

[§]Small Area Health Statistics Unit, Department of Epidemiology and Biostatistics, Imperial College London, United Kingdom, a.freni-sterrantino@imperial.ac.uk

[¶]CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, raphael.huser@kaust.edu.sa

[‖]Biostatistics and Spatial Processes Unit, French National Institute for Agronomic Research, 84914 Avignon, France, thomas.opitz@inra.fr

[**]CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, haavard.rue@kaust.edu.sa

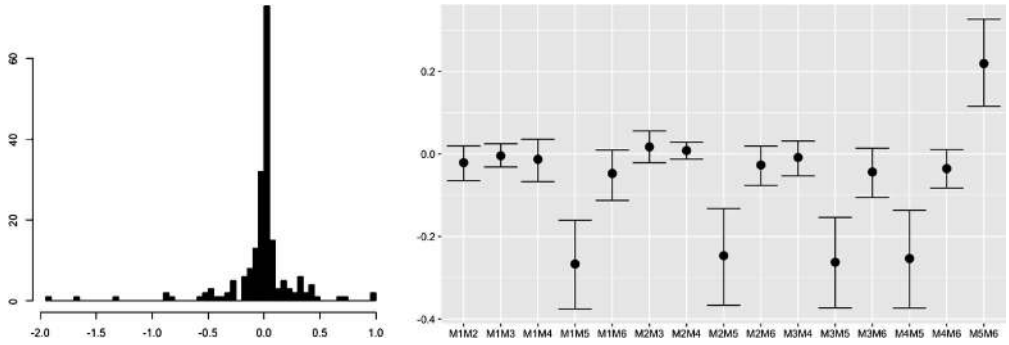

Figure 1: *Left:* Histogram of differences in LOOCV log-predictive density between two spatial models for a dataset on fish larvae. *Right:* Bootstrapped mean differences for one non-spatial model (M5) and for 5 different spatial models (M1 to M4 and M6). See the supplementary material of Bakka et al. (2018) for more details.

small subset of the individual log-scores strongly influenced the model selection result. For illustration, the left panel of Figure 1 depicts the histogram of log-score differences between two example models (for each held-out point in the leave-one-out procedure), from which the mean (or sum) is usually computed. It is clear that we cannot conclude that one model is superior to the other (i.e., that zero is a bad choice for the center of this distribution). In the context of stacking, we cannot give more weight to one of these two models with any degree of confidence. To further assess the variability inherent to the LOOCV estimate of marginal predictive performance, we bootstrapped the mean-differences to compute uncertainty intervals, and decided to conclude that one model was better than another only if this interval did not include zero; see the right panel of Figure 1 for this computation on our dataset. The first interval in this figure corresponds to the histogram in Figure 1. The non-spatial model 5 (M5) performs poorly, but we cannot conclude that there is a best model. In the context of stacking, the five "equivalent" models would be weighted by highly arbitrary weights to create a stacked model, which we find questionable. We wonder whether combining the bootstrapped uncertainty intervals with the stacking idea (in some way) could lead to a more robust approach to stacking.

We question the authors' choice to compare the stacking approach to the other methods presented in the paper. Indeed, they point out that Bayesian model averaging (BMA) weights reflect only the fit to the data (i.e., *within-sample performance*)without maximizing the prediction accuracy (i.e., *out-of-sample performance*). Thus, the comparison of BMA (or its modified versions, Pseudo-BMA and Pseudo-BMA+) against the stacking of distributions, which is conveniently constructed to improve prediction accuracy, does not seem fair. To highlight the gains and the pitfalls of stacking predictive distributions, it would be more reasonable to compare the prediction ability of the stacking approach against the prediction ability of each one of the stacked models.

In Section 3.3, the authors advocate the use of Pareto-smoothed importance sampling as a cheap alternative to exact LOOCV, which can be computationally expensive

for large sample sizes. We agree with the authors' guidelines, and we here re-emphasize that this approach is potentially unstable/invalid when the importance ratios have a very heavy or "noisy" tail. Indeed, it is well-known that the ith order statistics $X_{(i)}$ in a sample of size n from the GP distribution with shape parameter k has finite mean for $k < n - i + 1$, and finite variance for $k < (n - i + 1)/2$; see, e.g., Vännman (1976). In particular, this implies that the maximum $X_{(n)}$ has infinite mean for $k \geq 1$. As the GP shape parameter is usually estimated with high uncertainty, especially with heavy tails, a conservative decision rule is preferred in practice. Moreover, we want to stress that the estimation of the shape parameter k via maximum likelihood may be strongly influenced by the largest observations. Therefore, more robust approaches might be preferred. Possibilities include using methods based on probability weighted moments, which were found to have good small sample properties (Hosking and Wallis, 1987; Naveau et al., 2016), or using a Bayesian approach with strong prior shrinkage towards light tails. Opitz et al. (2018) recently developed a penalized complexity (PC) prior (Simpson et al., 2017) for k, designed for this purpose.

References

Bakka, H., Rue, H., Fuglstad, G. A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D., and Lindgren, F. (2018). "Spatial modelling with R-INLA: A review." *WIREs Computational Statistics*, MR1535567. doi: https://doi.org/10.2307/wics.1443. 982, 983

Held, L., Schrödle, B., and Rue, H. (2010). "Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA." In Kneib, T. and Tutz, G. (eds.), *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, 91–110. Berlin: Springer Verlag. MR2664630. doi: https://doi.org/10.1007/978-3-7908-2413-1_6. 982

Hosking, J. R. M. and Wallis, J. R. (1987). "Parameter and quantile estimation for the generalized Pareto distribution." *Technometrics*, 29: 339–349. MR0906643. doi: https://doi.org/10.2307/1269343. 984

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). "Bayesian computing with INLA: New features." *Computational Statistics & Data Analysis*, 67: 68–83. MR3079584. doi: https://doi.org/10.1016/j.csda.2013.04.014. 982

Naveau, P., Huser, R., Ribereau, P., and Hannart, A. (2016). "Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection." *Water Resources Research*, 52: 2753–2769. 984

Opitz, T., Huser, R., Bakka, H., and Rue, H. (2018). "INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles." *Extremes*. doi: https://doi.org/10.1007/s10687-018-0324-x. 984

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion)." *Journal of the Royal Statistical Society, Series B*, 71(2): 319–392. MR2649602. doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x. 982

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). "Bayesian Computing with INLA: A Review." *Annual Reviews of Statistics and Its Applications*, 4(March): 395–421. MR3634300. doi: https://doi.org/10.1214/16-STS576. 982

Simpson, D. P., Rue, H., Riebler, A., Martins, T., and Sørbye, S. H. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors." *Statistical Science*, 32: 1–28. MR3634300. doi: https://doi.org/10.1214/16-STS576. 984

Vännman, K. (1976). "Estimators based on order statistics from a Pareto distribution." *Journal of the American Statistical Association*, 71: 704–708. MR0440779. 984

Contributed Discussion

Marco A. R. Ferreira*

I congratulate the authors on delivering a stimulating and thought-provoking article. In this comment, in the context of data observed over time, in cases when one model has a posterior probability close to one but a stacking weight much smaller than one, I suggest a way to investigate the causes of the disagreement.

Here we focus on the case when data arrives over time. To simplify the discussion, let us assume that data are observed at discrete time points. Let y_t be a vector that contains all the data observed at time t, $t = 1, \ldots, T$. Further, let $y_{1:t} = (y_1', \ldots, y_t')'$. Then, instead of using the leave-one-out predictive density $p(y_i|y_{-i}, M_k)$, we may consider the one-step ahead predictive density $p(y_t|y_{1:(t-1)}, M_k)$ which is given by

$$p(y_t|y_{1:(t-1)}, M_k) = \int p(y_t|y_{1:(t-1)}, \theta_k, M_k) p(\theta_k|y_{1:(t-1)}, M_k) d\theta_k.$$

We note that after y_t has been observed, comparing $p(y_t|y_{1:(t-1)}, M_1), \ldots, p(y_t|y_{1:(t-1)}, M_K)$ allows one to evaluate the relative ability of each model to predict at time $t-1$ the vector of observations y_t. Hence, in the context of data observed over time, instead of $p(y_i|y_{-i}, M_k)$, it seems more natural to use the one-step ahead predictive density $p(y_t|y_{1:(t-1)}, M_k)$. Thus, for data observed over time the stacking of predictive distributions would choose weights

$$\widehat{w} = \text{argmax}_{w \in \mathcal{S}_1^K} \sum_{t=t^*+1}^{T} \log \sum_{k=1}^{K} w_k p(y_t|y_{1:(t-1)}, M_k), \tag{1}$$

where the summation on t starts at $t^* + 1$ because the first t^* observations are used to train the models to reduce dependence on priors for parameters. We note that the above equation is very similar to that of the optimal prediction pools of Geweke and Amisano (2011, 2012), except that they start the summation at $t = 1$.

It is also helpful to consider the formula for the posterior probability for each model. To keep exposition simple, let us assume equal prior probabilities for the competing models. Further, we assume that the first t^* observations are used for training. Then, the posterior probability for model M_k is

$$P(M_k|y_{1:T}) = \frac{\prod_{t=t^*}^{T} p(y_t|y_{1:(t-1)}, M_k)}{\sum_{k=1}^{K} \prod_{t=t^*}^{T} p(y_t|y_{1:(t-1)}, M_k)}. \tag{2}$$

Keeping (1) and (2) in mind, what can we infer when a model $\widetilde{M}$ has posterior probability close to one but its weight $\widetilde{w}$ in the stacking of predictive distributions is much smaller than one? The posterior probability being close to one means that $\widetilde{M}$

*Department of Statistics, Virginia Tech, Blacksburg, VA 24061, marf@vt.edu

is probably, amongst the K models being considered, the model closest in Kullback–Leibler sense to the true data generating mechanism. But its weight $\widetilde{w}$ being much smaller than one means that there are important aspects of the true data generating mechanism that have not been incorporated in $\widetilde{M}$.

We note that both (1) and (2) depend on the data only through the one-step ahead predictive densities $p(y_t|y_{1:(t-1)}, M_k)$. Thus, for data observed over time, when there are disagreements between the posterior probabilities of models and the stacking weights, an examination of the one-step ahead predictive densities $p(y_t|y_{1:(t-1)}, M_k)$ such as plotting them over time as in Vivar and Ferreira (2009) may help identify what aspects of the true data generating mechanism are being neglected by model $\widetilde{M}$.

For example, an examination of $p(y_t|y_{1:(t-1)}, M_k)$ may indicate that model $\widetilde{M}$ provides better probabilistic predictions 95% of the time, but that in the remaining 5% of the time the observations are outliers with respect to $\widetilde{M}$ but are not outliers with respect to a model M^* that has fatter tails than $\widetilde{M}$. In that situation, the outlying observations would prevent $\widetilde{w}$ from being close to one. Further examination of the outlying observations could possibly suggest ways to improve model $\widetilde{M}$ to get it closer to the true data generating mechanism.

As another example, an examination of $p(y_t|y_{1:(t-1)}, M_k)$ may indicate that $\widetilde{M}$ and another model M^* take turns at providing better probabilistic predictions. For example, say that for a certain environmental process, $\widetilde{M}$ provides better predictions during a certain period of time, and then after that M^* provides better predictions, and after that $\widetilde{M}$ provides better predictions, and so on. In that case, probably the environmental process has different regimes, and thus for example a Markov switching model (Frühwirth-Schnatter, 2006) may be adequate to model such environmental process.

I would imagine that a sensibly estimated leave-one-out predictive density $p(y_i|y_{-i}, M_k)$ could also be used for diagnostics. I would appreciate if the authors can comment on advantages and difficulties associated with such use.

Finally, in the $\mathcal{M}$-closed case, will the stacking weight of the true model converge to one as the sample size increases?

References

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer. MR2265601. 987

Geweke, J. and Amisano, G. (2011). "Optimal prediction pools." *Journal of Econometrics*, 164(1): 130–141. MR2821798. doi: https://doi.org/10.1016/j.jeconom.2011.02.017. 986

Geweke, J. and Amisano, G. (2012). "Prediction with misspecified models." *American Economic Review*, 102(3): 482–86. 986

Vivar, J. C. and Ferreira, M. A. R. (2009). "Spatio-temporal models for Gaussian areal data." *Journal of Computational and Graphical Statistics*, 18: 658–674. MR2751645. doi: https://doi.org/10.1198/jcgs.2009.07076. 987

Contributed Discussion

Luis Pericchi[*†]

This article gave me a Déja Vu, of 25 years ago in London when the book by Bernardo and Smith (1994) was being finished and Key et al. (1999) was starting. With the formalization of the complement of the M-closed perspective, the end of Bayes Factors and Bayesian Model Averaging (BMA) was predicted or at least confined to a very small corner of statistical practice. However, re-sampling Bayes Factors, particularly the Geometric Intrinsic Bayes Factors were being invented around the same years and these re-sampling schemes changed completely the perspective. For some historical reasons however, non re-sampling Intrinsic Bayes Factors were much more developed along the years. Perhaps this will be one of the positive consequences of this paper and the previous in this line, to recover the thread of development, theoretical and practical, of the rich mine vein of re-sampling Bayes Factors.

Just two illustrations of the fundamentally different behavior of re-sampling Bayes Factors, more in tune with open perspectives are in Bernardo and Smith (1994) p. 406 (Lindley's paradox revisited) and in Key et al. (1999) p. 369 on which the Intrinsic Implicit Priors were first named. However this paper seems to restrict its scope to non re-sampling Bayes Factors which is insufficient. On the other hand the paper interestingly relinquishing marginal likelihoods appears to get away, at least to some extend by using K-L loss. This should be study also theoretically, but it should be noted that the loss function, even restricted to K-L functional form, changes also whenever the training sample and validation sample sizes change, and the change is huge. This paper seems also restricted to $n-1$ cross validation. Also it seems that the only goal of statistics is to make good predictions, but good explanations are also of paramount importance. In that direction Key et al. (1999) define different combinations of training sizes that show the differences.

I finish with a list of questions and a conjecture: The questions are,

1. Are these stacking solutions asymptotically efficient estimators? Is $L2$ convergence sufficient in this context?
2. Are these approximations really Bayesian? in the sense that: is there a prior that would produce asymptotically equivalent inference with stacking? In other words Intrinsic Implicit Priors exists here?
3. Is there an optimal training sample size or combination of global and Local utility functions, when the objective is prediction? Identification?

[*]I am grateful to Adrian Smith and Jim Berger for many discussions on the subject of Bayesian Model Selection and Hypothesis Testing along the years.

[†]Department of Mathematics, University of Puerto Rico Rio Piedras, PR, USA, luis.pericchi@upr.edu

4. The difficult problem of calculation of posterior model probabilities in the open perspectives can be inversely solved as an optimization problem maximizing utility functions?

Finally I conjecture that casual choice of estimators within models would lead to un-Bayesian inefficient solutions, and the authors seem to agree with this conjecture in 5.3. Careful consideration of all the entertained models and admissible estimators for parameters should be considered prior to the optimization procedures. In fact this may solve the old conundrum of whether the same priors should or not be used for estimation and Selection.

References

Bernardo, J. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons, 1st. edition. MR1274699. doi: https://doi.org/10.1002/9780470316870. 988

Key, J., Pericchi, L. R., and Smith, A. F. M. (1999). "Bayesian model choice: what and why?" In *Bayesian Statistics 6*, 343–370. Oxford University Press. MR1723504. 988

Contributed Discussion

Christopher T. Franck*

I congratulate the authors on a fascinating article which will positively impact statistical research and practice in the years to come. The authors' procedure for stacking Bayesian predictive distributions differs from Bayesian model averaging (BMA) in two important ways. First, the stacking procedure chooses weights based directly on the predictive distribution of new data while BMA chooses weights based on fit to observed data. Second, the stacking procedure is not sensitive to priors on parameters to the extent that BMA is. The leave-one-out approach in the stacking procedure bears some resemblance to intrinsic Bayes factors, which made me curious as to whether intrinsic Bayesian model averaging (iBMA) can make up some of the reported performance difference between stacking and BMA. In this note, I restrict attention to the authors' linear subset regressions example, adopt the authors' Bayesian model, and compare BMA with iBMA. Since iBMA does not improve prediction over BMA in this initial study, I ultimately suggest that the stacking procedure is superior to iBMA for prediction.

It appears that the stacking procedure's replacement of the full predictive distribution with the leave-one-out predictive distribution $\int p(y_i|\theta_k, M_k)p(\theta_k|y_{-i}, M_k)d\theta_k$ is the major mechanism that bestows invariance to priors. This approach makes priors resemble posterior distributions by conditioning on a subset of the data. Further, a simple re-expression of $p(\theta_k|y_{-i}, M_k)$ via Bayes' rule reveals that nuisance constants which accompany priors (discussed by the authors in the BMA segment of Section 2) cancel out when any portion of the data is used to train priors. This is the same tactic that partial Bayes factors (Berger and Pericchi, 1996) use to cancel the unspecified constants which accompany improper priors and contaminate resulting Bayes factors. Briefly, a partial Bayes factor takes a training sample from the observed data, uses the likelihood of the training sample to update the prior, and forms a Bayes factor as the ratio of marginal likelihoods that adopt the trained prior alongside the remainder of the likelihood. An intrinsic Bayes factor is an average across some or all possible training samples. Where the original motivation for intrinsic Bayes was to enable model selection using improper priors, the approach is also used for model selection that is robust to vague proper priors. For improper priors, the goal is usually to choose a minimal training sample size to render the prior proper. As the training sample size increases for fixed n, the prior exerts less influence on the posterior model probabilities, but the method becomes less able to discern competitive models (Fulvio and Fulvio, 1997).

Using the authors' Bayesian model and data generating process and a similar out-of-sample testing procedure, I compared standard BMA with iBMA. In the iBMA case, I formed intrinsic Bayes factors which I then translated to posterior model probabilities for use in model averaging. I obtained 50 Monte Carlo replicates with 10 test points. I considered iBMA training sample sizes of 1, 5, and $n-1$. The $\mathcal{M}$-open case (not shown) favored the iBMA setting with $n-1$ training samples, leaving only one data point for

*Department of Statistics, Virginia Tech, chfranck@vt.edu

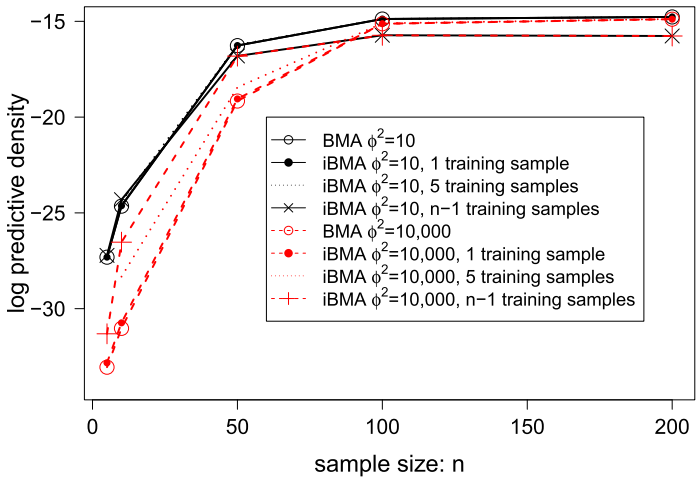

Figure 1: Averages of log predictive densities for 10 test points based on 50 Monte Carlo replicates. Black corresponds to authors' prior, red corresponds to a more vague prior on coefficients. Increasing the training sample size diminishes the effect of the prior but also reduces log predictive density at higher sample sizes. None of the iBMA settings produce predictive densities substantially higher than the BMA approach (open circles).

the likelihood. This setting barely changed the posterior model probabilities from their uniform prior, which works well only in this specific case where a near uniform mixture of the 15 candidate models performs adequately. The $\mathcal{M}$-closed results shown in Figure 1 confirm that (i) iBMA diminishes influence of the prior as the size of the training sample increases (note overlap in $n-1$ training lines), (ii) an excessively large training sample proportion erodes the predictive density especially at larger sample sizes, and more importantly suggests that (iii) endowing BMA with prior-invariance machinery that resembles the stacking procedure's does not appear to offer any advantage in predictive density. Hence, I second the authors' conclusion that stacking is a superior approach for prediction. The present study suggests that the stacking procedure's prior invariance property is a convenient bonus but not the major reason for its impressive performance.

References

Berger, J. O. and Pericchi, L. R. (1996). "The Intrinsic Bayes Factor for Model Selection and Prediction." *Journal of the American Statistical Association*, 91(433): 109–122. URL http://www.jstor.org/stable/2291387. MR1394065. doi: https://doi.org/10.2307/2291387. 990

Fulvio, D. S. and Fulvio, S. (1997). "Alternative Bayes factors for model selection." *Canadian Journal of Statistics*, 25(4): 503–515. MR1614347. doi: https://doi.org/10.2307/3315344. 990

Contributed Discussion

Eduard Belitser[†] and Nurzhan Nurushev[*‡]

We thank the authors for an interesting paper that takes the idea of *stacking* from the point estimation problem and generalize to the combination of predictive distributions. Let us first mention some key ides of the present paper. One of the main problems in statistics is prediction in the presence of multiple candidate models or learning algorithms $\mathcal{M} = (M_1, \dots, M_K)$. 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. These mentioned problems in the present paper are addressed by providing new stacking method. The authors compare this method to several alternatives: stacking of means, Bayesian model averaging (BMA), Pseudo-BMA, and a variant of Pseudo-BMA that is stabilized using the Bayesian bootstrap. Based on simulations and real-data applications, they recommend stacking of predictive distributions, with bootstrapped-Pseudo-BMA as an approximate alternative when computation cost is an issue.

We enjoyed reading the paper and would like to make three comments/question. First, the methodology of the present paper relies on the knowledge of K, the number of models in the list $\mathcal{M} = (M_1, \dots, M_K)$. Without loss of generality assume $K \in (0, n]$. We wonder whether the authors could come up with a general idea of how to extend the stacking method to the unknown number of models, i.e., K is unknown. Perhaps the problem can be addressed by adding prior on K in author's framework, but it might lead to big computational costs.

Second, all problems discussed in the present paper are examples of *supervised learning*. In other words, for each observation of the predictor measurements $x_i, i = 1, \dots, n$, there is an associated response measurement y_i. The authors wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference). In contrast, *unsupervised learning* describes the somewhat more difficult situation in which for every observation $i = 1, \dots, n$, we observe a vector of measurements x_i but no associated response y_i. For instance, it is not possible to fit linear or logistic regression models, since there is no response variable to predict. This situation is referred to as unsupervised because we lack a response variable that can supervise our analysis. One of the popular examples of unsupervised learning is *cluster* analysis. The goal of cluster analysis is to ascertain, on the basis of $x_1, \dots, x_n$, whether the observations fall into relatively distinct groups(e.g., stochastic block model, see Holland et al. (1983)). We wonder whether it is possible to extend the stacking method to the examples of unsupervised learning (e.g., stochastic block model).

*Research funded by the Netherlands Organisation for Scientific Research NWO.
[†]Department of Mathematics, VU Amsterdam, e.n.belitser@vu.nl
[‡]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, n.nurushev@uva.nl

The third comment is related to the simulation subsection 4.2. In that subsection the authors consider some $K = 15$ different models of linear regression $Y = \beta_1 X_1 + \ldots + \beta_J X_J + \epsilon$, $\epsilon \sim \mathrm{N}(0, 1)$, where the number of predictors J is 15. However, the total number of all possible linear regressions with at most $J = 15$ predictors is 2^{15}. We wonder whether the methods studied in the present paper with all $K = 2^{15}$ possible models would be computationally costly. For instance, LASSO and Ridge methods solve this problem without any big computational costs. It would be also interesting to know for the future research whether the corresponding estimators of the present paper can achieve the minimax rate for sparse linear regression problem studied in Bunea et al. (2007).

We hope these comments will inspire the authors and other people to work on these interesting problems in the future.

References

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). "Aggregation for Gaussian regression." *Annals of Statistics*, 35(4): 1674–1697. MR2351101. doi: https://doi.org/10.1214/009053606000001587. 993

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). "Stochastic blockmodels: First steps." *Social Networks*, 5(2): 109–137. MR0718088. doi: https://doi.org/10.1016/0378-8733(83)90021-7. 992

Contributed Discussion

Matteo Iacopini[*,†] and Stefano Tonellato[‡]

We congratulate the authors for their excellent research which leaded to the development of a new statistical method for model comparison. The procedure is computationally fast and can be applied in a variety of settings ranging from mixture models to variable selection in regression frameworks.

Pseudo Bayesian model averaging and reference pseudo Bayesian model averaging

The authors mentioned the contribution by Li and Dunson (2016) as a possible alternative for weighting competing models. In this discussion, we present a small simulation study comparing the performance of the pseudo Bayesian model averaging (Pseudo-BMA) introduced in Section 3.4 with the performance of the reference pseudo Bayesian model averaging (Reference-Pseudo-BMA), mentioned in Section 2, and based on $\widetilde{KL}_2$, as in Li and Dunson (2016).

The data are generated from the following model: $Y_i|x_i \sim e^{-2x_i}\mathcal{N}(y|x, 0.1) + (1 - e^{-2x_i})N(y|x^4, 0.1)$, $X_i \sim \mathcal{U}(0, 1)$, $i = 1, \ldots, N$. We estimated $K = 5$ different linear regression models, where model M_k is defined as: $Y_i|x_i = \beta_k x^k + \varepsilon_i$, $\varepsilon_i \sim N(0, 0.1)$, $\beta_k \sim N(0, 10)$ and $\sigma \sim Ga(0.1, 0.1)$.

Reference-Pseudo-BMA requires the preliminary estimation of the predictive density via a Bayesian nonparametric approach, which we computed by using a weight dependent Dirichlet process prior for the estimation of a fully nonparametric Bayesian density regression (Müller et al., 2013, ch. 4).

We run 100 simulations for each different value of the sample size in the grid $N \in \{5, 10, 20, 30, 40, 50\}$ and for each N we computed the mean log-predictive densities of the two combination methods. The results are shown in Figure 1, which plots the posterior log-predictive density (averaged over simulations) for each value of the sample size, for the two cases. As expected, the Pseudo-BMA performs better than the Reference-Pseudo-BMA, and the difference of performance decreases with N.

For one of the previously run simulations (similar results were found in the other cases), Figure 2 shows the true conditional density $p(y|x)$ (red curve) at some fixed values of the covariate x together with the predictive densities provided by Pseudo-BMA (blue) and by Reference-Pseudo-BMA (black), respectively. The unsatisfactory approximation of the true predictive density is due to the inadequacy of the parametric

[*]Ca' Foscari University of Venice, Cannaregio 873, 30121, Venice, Italy
[†]Université Paris I – Panthéon-Sorbonne, 106-112 Boulevard de l'Hôpital, 75642 Paris Cedex 13, France, matteo.iacopini@unive.it
[‡]Ca' Foscari University of Venice, Cannaregio 873, 30121, Venice, Italy, stone@unive.it

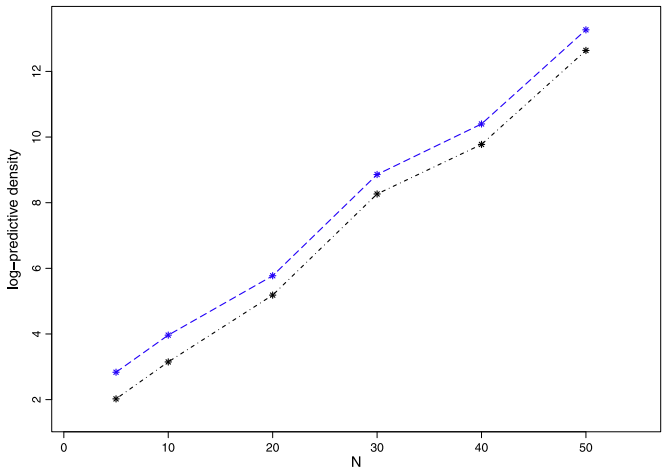

Figure 1: Posterior log-predictive density of model (a) (*blue*) and model (b) (*black*), for different values of the sample size $N \in \{5, 10, 20, 30, 40, 50\}$.

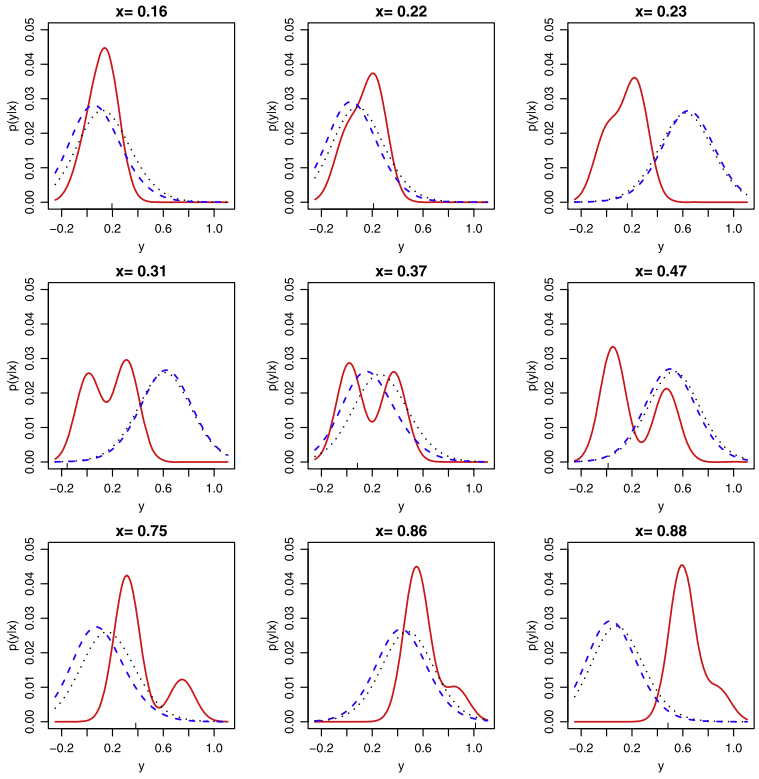

Figure 2: Conditional densities $p(y|x)$ for several values of x, $N = 20$. True function (*red*), model (b) estimate (*black*) and model (a) estimate (*blue*).

models under comparison, not to the methods used in order to produce stacking. What is interesting to notice is that despite the different weights computed according to the two schemes, the combined conditional predictive densities are rather similar for all the values of the conditioning variable x. This feature has been proved to hold also when the sample size increases and similar results (not reported here) have been obtained for different model specifications.

The main insight from this small simulation study is twofold: first, the approach proposed by the authors outperforms the alternative weighting schemes, both in fitting and in computational efficiency. Second, the scheme of Li and Dunson (2016) yields comparable results in terms of conditional density estimation. This might suggest that coupling Pseudo-BMA and Reference-Pseudo-BMA might be a successful strategy in those circomstances when leave-one-out or Pareto smoothed importance sampling leave-one-out cross-validation are suspected to produce unstable results, due to small sample size or large values of $\hat{k}$.

References

Li, M. and Dunson, D. B. (2016). "Comparing and weighting imperfect models using D-probabilities." 994, 996

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2013). *Bayesian nonparametric data analysis*. Springer. MR3309338. doi: https://doi.org/10.1007/978-3-319-18968-0. 994

Contributed Discussion

Merlise Clyde *

Stacking (Wolpert, 1992; Breiman, 1996b,a) has seen renewed attention in recent years as an ensemble learning method for prediction, particularly among kaggle competitions! In the M-open context, Clyde and Iversen (2013) arrived at stacking of densities from a decision-theoretic solution using a log scoring rule (see also Walker et al. (2001) for an alternative computational approach to the same problem). I hope the authors contributions to predictive density estimation will encourage more practitioners to look beyond point estimation and consider quantification of uncertainty. Examples where uncertainty quantification is being addressed quite successfully with ensemble learning of densities include computer models, such as those used in weather forecasting. Raftery et al. (2005); Gneiting and Raftery (2007) and related papers propose log-scoring rules among other utilities for ensemble learning with implementation in the `R` package `ensembleBMA` (Fraley et al., 2018). Other examples include forecasting with economic time series (see discussion by McAlinn et al. of Yao et al.). A key distinguishing feature in density ensemble learning methods in the M-open context is whether the predictive densities or parameters of the densities are independent of the data used to learn the weights (computer models) or that the available data must be used to both learn the predictive densities and the optimal weights for the ensemble. The former case, as used in examples in section 4.1 of the paper, does not require any data splitting such as Leave-One-Out cross-validation to learn the predictive densities and then solve the predictive optimization problem and can help illustrate potential problems.

Faraway, So Close

Let's reconsider the example of section 4.1 of the authors, but change the true data generating distribution to a $\mathsf{N}(42, 1)$ while keeping the candidate densities for ensemble learning to be the same as in the paper: $\mathsf{N}(\mu_k, 1)$ with $\mu_k = k$ for $k = 1, \ldots, 8$. Traditional Bayesian Model Averaging (BMA) will degenerate to the model closest to the true model in Kullback–Leibler divergence, hence, the $\mathsf{N}(8, 1)$. However, stacking of densities, will fair no better and could be worse if the weights do not degenerate to zero for $k < 8$. Raftery et al. (2005) and following papers, explicitly accommodate potential bias in computer models used in the means of each of the predictive densities in the ensembles; applying that approach rather than using a $N(\mu_k, 1)$ we could correct for bias using components that are $N(a_k + b_k\mu_k, 1)$ where a_k and b_k are learned from the data. In our simple example, it would appear that there would be no unique solution to the weights or a_k and b_k. This is not a concern if one is only interested in predictive analytics without the desire to interpret weights as a measure of importance, etc.

Stacking can also fail if the component densities are too simple. Consider the situation where the true data generating model is $\mathbf{Y} = \sum_{k=1}^{K} \mathbf{X}_k\beta_k + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_n)$ as

*Department of Statistical Science, Duke University, Durham, NC 27708-0251 USA, clyde@duke.edu
url: http://stat.duke.edu/~clyde

in example 4.2. We will take the candidate densities to be Gaussian densities centered at $\mathbf{X}_k\hat{\beta}_k$ where $\hat{\beta}_k$ is the ordinary least squares estimate of β from the simple linear regression of $\mathbf{Y}$ on $\mathbf{X}_k$. In the case that $\mathbf{X}_j$ and $\mathbf{X}_k$ are all mutually orthogonal vectors, $\hat{\beta}_k$ is as unbiased estimate of β_k, however, the stacked predictive mean, $\sum_k x_k^*\hat{\beta}_k$ using either squared error or log-scoring rules will be asymptotically biased as individual weights will be less than 1. There is the potential that the bias will actually grow with k, as more predictors will dilute the weights. In this case, the bias corrections posed in Raftery et al. (2005) would have no effect. The more realistic case where the $\mathbf{X}_k$ are not orthogonal or the component densities are centered at Bayesian estimates may suffer from the same problems.

Arguments in Clyde and Iversen (2013) showed that BMA would fail when the true model was outside the class of models used in BMA, however, these examples illustrate that stacking can suffer from similar problems as BMA. While both BMA and stacking of densities are guaranteed to be "close" in some measure, they may still be far from the truth.

Choice of Weights Revisited

In the original papers on stacking (Breiman, 1996a,b; Wolpert, 1992), the constraints that the weights sum to one and be non-negative did not follow from any first principles, but appeared to work well in practice. For the above regression example using squared error loss, the non-negativity constraint on weights is in fact not binding, however, the sum to one constraint is the source of the problem. By relaxing that constraint, one can recover the true predictive mean asymptotically, however, it is not obvious what the weights should sum to as part of the optimization problem. In terms of the dual Lagrangian formulation of the constrained optimization problem, the choice of constraint for the sum of the weights is related to the pre-specification of the penalty in lasso regression.

For stacking with densities, the solution for the weights based on an EM algorithm to find the optimal weights under the log-scoring rule (Clyde and Iversen, 2013) demonstrates that the weights always satisfy the constraints of non-negativity and sum to one so that relaxation is not a possible avenue to address the problems identified above.

The arguments above suggest that stacking of densities, like BMA, may be sensitive to the choice of models that are used in the ensemble.

References

Breiman, L. (1996a). "Heuristics of instability and stabilization in model selection." *Annals of Statistics*, 24: 2350–2383. MR1425957. doi: https://doi.org/10.1214/aos/1032181158. 997, 998

Breiman, L. (1996b). "Stacked Regressions." *Machine Learning*, 24: 49–64. 997, 998

Clyde, M. A. and Iversen, E. S. (2013). *Bayesian Model Averaging in the M-Open framework*, chapter Bayesian Theory and Applications, 484–498. Oxford University Press.

MR3221178. doi: https://doi.org/10.1093/acprof:oso/9780199695607.003.0024. 997, 998

Fraley, C., Raftery, A. E., Sloughter, J. M., Gneiting, T., and of Washington., U. (2018). *ensembleBMA: Probabilistic Forecasting using Ensembles and Bayesian Model Averaging*. R package version 5.1.5. URL https://CRAN.R-project.org/package=ensembleBMA 997

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction and estimation." *Journal of the American Statistical Association*, 102: 359–378. MR2345548. doi: https://doi.org/10.1198/016214506000001437. 997

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). "Using Bayesian model averaging to calibrate forecast ensembles." *Monthly Weather Review*, 133: 1155–1174. 997, 998

Walker, S. G., Guti/'errez-Pena, E., and Muliere, P. (2001). "A decision theoretic approach to model averaging." *The Statistician*, 50: 31–39. MR1821897. doi: https://doi.org/10.1111/1467-9884.00258. 997

Wolpert, D. H. (1992). "Stacked generalization." *Neural Networks*, 5: 241–259. 997, 998

Rejoinder

Yuling Yao*, Aki Vehtari†, Daniel Simpson‡, and Andrew Gelman§

We thank the editorial team for organizing the discussion. We are pleased to find so many thoughtful discussants who agree with us on the advantage of having stacking in the toolbox for combining Bayesian predictive distributions. In this rejoinder we will provide further clarifications and discuss some limitations and extensions of Bayesian stacking.

1 When is leave-one-out cross validation appropriate?

1.1 Exchangeability

Stacking maximizes the weighted leave-one-out (LOO) scoring rule using all observations:

$$\max_{w\in\mathcal{S}_1^K} \frac{1}{N}\sum_{i=1}^{N} S\Big(\sum_{k=1}^{K} w_k \hat{p}_{k,-i}, y_i\Big), \tag{1}$$

where S_1^K denotes the simplex space $\{w : \sum_{k=1}^K w_k = 1, w_k \in [0,1]\}$ and $\hat{p}_{k,-i}(\cdot) = \int p(\cdot|x_i, y_{-i}, \theta_k, M_k) p(\theta_k|y_{-i}, x_{-i}, M_k) d\theta_k$ is the leave-i-out predictive density for model M_k. To emphasize the data generating models, we include here covariates x, which were suppressed in the main paper for simplicity.

The asymptotic optimality of stacking relies on the consistency of the leave-one-out scoring rule:

$$\frac{1}{N}\sum_{i=1}^{N} S\left(\hat{p}_{k,-i}, y_i\right) - \mathrm{E}_{(\tilde{x},\tilde{y})|(x,y)}\, S\left(p(\cdot|\tilde{x}, x, y, M_k), \tilde{y}\right) \to 0. \tag{2}$$

The conditional iid assumption of y given x is typically sufficient and is used in many proofs, but it is not necessary. The key assumption we need is *exchangeability* (Bernardo and Smith, 1994, Chapter 6).

As discussed by **Dawid** and **Clarke**, it is not clear what happens to LOO or LOO-stacking in the asymptotic limit, if there is no true data-generating mechanism, $p(\tilde{y}|\tilde{x})$. Assuming such a true or underlying distribution is equivalent to assuming stationarity of the data-generating process. From a Bayesian standpoint, it is appropriate to model a non-stationary process with a non-stationary model. Realistically, though, whatever model we use, stationary or not, when applied to real data will encounter unmodeled trends; hence we any asymptotic results can only be considered as provisional.

*Department of Statistics, Columbia University, yy2619@columbia.edu
†Helsinki Institute of Information Technology, Department of Computer Science, Aalto University, Aki.Vehtari@aalto.fi
‡Department of Statistical Sciences, University of Toronto, simpson@utstat.toronto.edu
§Department of Statistics and Department of Political Science, Columbia University, gelman@stat.columbia.edu

1.2 Data-generating mechanisms and sample reweighting

If we have not yet observed the new data point (y_{N+1}, x_{N+1}), we can use LOO to approximate the expectation over different possible values for (y_{N+1}, x_{N+1}). Instead of making a model $p(y, x)$, we re-use the observation as a pseudo-Monte Carlo sample from $p(y_{N+1}, x_{N+1})$, while not using it for inference about θ.

For the predictive performance estimate, however, we need to model know how the future x_{N+1} will be generated. In standard LOO we implicitly assume future x_{N+1} comes from the empirical distribution $\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$.

If we assume that the future distribution is different from the past, then the covariate shift can be taken into account by weighting (e.g. Shimodaira, 2000; Sugiyama and Müller, 2005; Sugiyama et al., 2007). When we know exactly or have estimated from extra information $p_{N+1}(x_{N+1}) = p(x_{N+1}|x)$, we can use importance weighting to adjust the sample weights $r_i \propto p_{N+1}(x_i)/p(x_i)$ and replace (1) by $\sum_{i=1}^{N} r_i S(\sum_{k=1}^{K} w_k \hat{p}_{k,-i}, y_i)$. The LOO consistency asks for

$$\frac{\sum_{i=1}^{N} r_i S\left(\hat{p}_{k,-i}, y_i\right)}{\sum_{i=1}^{N} r_i} - \mathrm{E}_{\tilde{x}}\, \mathrm{E}_{\tilde{y}|(\tilde{x},x,y)}\, S\left(p(\cdot|\tilde{x}, x, y, M_k), \tilde{y}\right) \to 0.$$

In particular, the convergence of importance sampling does not require independence of covariates $x_1, \ldots, x_N$.

1.3 Fixed design

If x is fixed or chosen by design, we can still make a conditional model $p(y|x, \theta)$ and analyze the posterior distribution, $p(\theta|x, y)$. We can reinterpret r_i as the weight for the i-th observation. Standard LOO will assign equal weights on each x_i. If we care about the performance for some fixed x_i than for others, we can use different and weighting schemes to adjust.

1.4 Non-factorizable models

Our fast LOO approximation (PSIS-LOO) generally applies to factorizable models $p(y|\theta, x) = \prod_{i=1}^{N} p(y_i|\theta, x_i)$ such that the pointwise log-likelihood can be obtained easily by computing $\log p(y_i|\theta, x_i)$.

Non-factorizable models can sometimes be factorized by re-parametrization. For example, consider a multilevel model with M groups, if we denote the group level parameter and global parameter as θ_m and ψ, then the joint density is

$$p(y|x, \theta, \psi) = \prod_{j=1}^{J} \left[\prod_{n=1}^{N_j} p(y_{jn}|x_{jn}, \theta_j) p(\theta_j|\psi) \right] p(\psi), \tag{3}$$

where y are partially exchangeable, i.e. y_{mn} are exchangeable in group j, and θ_m are exchangeable. Rearrange the data and denote the group label of (x_i, y_i) by z_i, then (3) can be reorganized as $\prod_{i=1}^{N'} p(y_i|x_i, z_i, \theta, \psi)$ so the previous results follow. Furthermore, de-

pending on whether the prediction task is to predict a new group, or a new observations within a particular group j, we should consider leave-one-point-out or leave-one-group-out, corresponding to modeling the new covariate by $p(\tilde{x}, \tilde{z}) \propto \delta(\tilde{z} = j) \sum_{z_i = j} \delta(\tilde{x} = x_i)$ or $\sum_{z=1}^{J} \sum_{z_i = j} \delta(\tilde{z} = j, \tilde{x} = x_i)$.

When there is no obvious re-parametrization making the model conditional factorizable, the pointwise log-likelihood has the general form $\log p(y_i|y_{-i}, \theta)$. It is still possible to use PSIS-LOO in some special non-factorizable forms. Vehtari et al. (2018a) provide a marginalization strategy of PSIS-LOO to evaluate simultaneously autoregressive normal models. **Bakka et al.** express concerns about the reliability of PSIS. We refer to Vehtari et al. (2017) and Vehtari et al. (2018b) for computation and diagnostic details.

Lastly, although we used LOO, other variations of cross-validation could be used in stacking. Roberts et al. (2017) review cross-validation strategies for data with temporal, spatial, hierarchical, and phylogenetic structure. Many of these can also be computed fast by PSIS as demonstrated for m-step-ahead cross-validation for time series (Buerkner et al., 2018).

2 Stacking in time series

2.1 Prequentialism

When observation y_t come in sequence, there is no reason in general to use a conditionally independent or exchangeable for them. Nevertheless, LOO and LOO-stacking can still be applicable if the concern is the whole structure of the observed time points. For example, we might be interested analyzing whether more or less babies would be born on some special days of the year.

If the main purpose is to make prediction for the next not-yet-observed data, we can utilize the prequential principle (Dawid, 1984):

$$p(y_{1:N}|\theta) = \prod_{t=1}^{N} p(y_t|y_{1:t-1}, \theta),$$

and replace the LOO density in (1) by the sequential predictive density leaving out all future data: $p(y_t|y_{<t}) = \int p(y_t|y_{1:t-1}, \theta) p(\theta|y_{1:t-1}) d\theta$ in each model, and then stacking follows. This is similar to the approach developed by Geweke and Amisano (2011, 2012). **Ferreira** and **Dawid** suggest similar ideas. The ergodicity of y will yield,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} S\left(p(\cdot|y_{<t}), y_t\right) - \lim_{N\to\infty} \frac{1}{N} \mathrm{E}_{Y_{1:N}} \sum_{t=1}^{N} S\left(p(\cdot|Y_{<t}), Y_t\right) \to 0. \tag{4}$$

When there is a particular horizon of interest for prediction, $p(y_t|y_{<t})$ above is generalized to m-step-ahead predictive density $p(y_{t<m}|y_{<t}) = p(y_t, \ldots, y_{t+m-1}|y_1, \ldots, y_{t-1}) = \int p(y_{t<m}|y_{<t}, \theta) p(\theta|y_{<t}) d\theta$.

However, the prequential approach introduces a different prediction task. Unless some stationarity of the true data generating mechanism is assumed (e.g., $P(y_t|y_{<t}) =$

$P(y_{t'}|y_{<t'})$), the *average* cumulative performance (the second term in (4)) is different from the *one-step-ahead* assessment in (2), which is only evaluated at next unseen observation $t = N + 1$.

2.2 Dynamic approximation of posterior densities

The exact prequential evaluation requires refitting each model for each t, which can be approximated by PSIS as, $p(y_t|y_{<t}) = \int p(y_t|\theta, y_{<t}) \frac{p(\theta|y_{<t})}{p(\theta|y)} p(\theta|y) d\theta$. We then start from the full data inference $p(\theta|y)$ and dynamically update $p(\theta|y_{<t})$ using PSIS approximation. When $p(\theta|y_{<t})$ reveals large discrepancy from $p(\theta|y)$ for some small t, which can be diagnosed by PSIS-$\hat{k}$, we refit the model $p(\theta|y_{<t})$ and update the proposal. Buerkner et al. (2018) verify such approximation gives stable and accurate results with minimal number of refits in an auto regressive model.

2.3 Dynamic stacking weights

Hoogerheide and Dijk and **McAlinn, Aastveit and West** point out that static weighting is not desired in time series, as a model good at making short-term prediction might not do well in the long run. Stacking can be easily made dynamic, allowing the explanation power of models to change over time. A quick fix is to replace model weights w_t in (1) by time-varying $w_{t,k}$ in the t-th term. To incorporate historical information, we can add regularization term $-\tau \sum_{t=2}^{N} ||w_{t,\cdot} - w_{t-1,\cdot}||$ in the stacking objective function. The heterogeneity of stacking weights can also be generalized to other hierarchical data structures, and this can be seen as related to a generalization of the mixture formulation of Kamary et al. (2014).

Bayesian predictive synthesis (BPS, McAlinn and West, 2017; McAlinn et al., 2017) has been developed for dynamic Bayesian combination of time series forecasting. The prediction in BPS takes the form $\int \alpha(y|z) \prod_{k=1:K} h_k(z_k) dz$ where $z = z_{1:K}$ is the latent vector generated from predictive densities $h_k(\cdot)$ in each model and $\alpha(y|x)$ is the distribution for y given z that is designed to calibrate the model-specific biases and correlations. We agree that BPS is more flexible in its combination form, for stacking is restricted to linear weighting $\alpha(y|z) = \sum_k \alpha_k(z) \delta_{z_k}(y)$. On the other hand, stacking has its flexibility that can be tailored for decision makers' specific utility, and the convex optimization is more computationally feasible for complicated models. It will be interesting to make further comparisons in the future.

3 Response to other discussions

3.1 Reliance on the list of models and the restriction to linear combinations

We agree with **Clyde** and **Zhou** that the performance of stacking depends on the choice of model list, as stacking can do nothing better than the optimal linear combination

from the model list. Stacking is not strongly sensitive to the misspecified models (see Section 4.1 of our paper), but it will be sensitive to how good an approximation is possible given the ensemble space.

We discuss the concern of inflexibility of linear-additive-form of density combination in Section 5.2, and construct the same orthogonal regression example as **Clyde**, in which stacking will not work to approximate the true model that is a convolution of individual densities. By optimizing the leave-one-out performance of combined prediction, the stacking framework can be extended to more general combination forms, such as the posterior family used in the BPS literature. Furthermore, simplex constraints will be unnecessary if it goes beyond the linear combination of densities. We are interested in testing such approaches. **Yoo** proposes another way to obtain convolutional combinations by stacking in the Fourier domain.

3.2 Model expansion as an alternative

One setting where stacking can be used, but full model expansion could be more difficult, is when some set of different sorts of models have been separately fit. The same idea is summarized by **Pericchi** as "careful consideration of all the entertained models and admissible estimators for parameters should be considered prior to the optimization procedures." We are less concerned about the situation described by **Belitser and Nurushev**, **Shin**, and **Zhou**, in which the number of models are so large that stacking can be both computationally expensive and theoretically inconsistent, because in that setting we would recommend moving to a continuous model space that encompasses the separate models in the list.

Stacking is not designed for model selection, but for model averaging to get good predictions. We do not recommend to use it as model selection, although models with zero weights could be discarded from the average. For large p and small n, instead of stacking or other model averaging methods, we recommend using an encompassing model with all variables and prior information about the desired level of sparsity (Piironen and Vehtari, 2017b,c). For example, the regularized horseshoe prior can be considered as a continuous extension of the spike-and-slab prior with discrete model averaging over models with different variable combinations (Piironen and Vehtari, 2017c). For high-dimensional variable selection we recommend a projection predictive approach (Piironen and Vehtari, 2016, 2017a), which has a smaller variance in selection process due to the use of the encompassing model as a reference model and has better predictive performance due to making the inference conditional on the selection process and the encompassing model.

3.3 Nonparametric approaches

Li and **Iacopini and Tonellato** suggest the use of nonparametric reference models to eliminate the need of cross-validation. If we are able to make a good nonparametric model there is probably no need for model averaging. Although model averaging might be used as part of model reduction, instead of using component models $p(\cdot|y, M_k)$ we

would prefer to form the component models using a projection predictive approach which projects the information from the reference model to the restricted models (Piironen and Vehtari, 2016, 2017a).

Zhou suggests Bayesian nonparametric (BNP) models as an alternative to model averaging. Indeed, the spline models used in the experiments in Section 4.6 of our paper can be considered as BNP models. We can compute fast LOO-CV also for Gaussian processes and other Gaussian latent variable models (Vehtari et al., 2016).

3.4 Logarithmic scoring rules

Finally, we emphasize that the choice of scoring rules in stacking depends on the underlying application, and it is unlikely to give one optimal result that is applicable to any situation in advance. As **Winkler, Jose, Lichtendahl and Grushka-Cockayne** and **Grüwald and Heide** point out, there is no need to use log score if the focus is some other utility. Our proposed stacking framework is applicable to any scoring rule. We are particularly interested in interval stacking that optimizes the interval score, which is likely to provide better interval estimation and posterior uncertainties.

We thank **Franck** for numerically verifying that stacking outperforms intrinsic Bayesian model averaging (iBMA) in simulations. This result suggests that the stacking procedure's prior invariance property is a convenient bonus but not the only reason for its impressive performance.

References

Bernardo, J. M. and Smith, A. F. (1994). *Bayesian theory*. John Wiley & Sons. MR1274699. doi: https://doi.org/10.1002/9780470316870. 1001

Buerkner, P., Vehtari, A., and Gabry, J. (2018). "PSIS assisted m-step-ahead predictions for time-series models." Technical report. URL http://mc-stan.org/loo/articles/m-step-ahead-predictions.html 1003, 1004

Dawid, A. P. (1984). "Present position and potential developments: Some personal views: Statistical theory: The prequential approach." *Journal of the Royal Statistical Society. Series A*, 278–292. MR0763811. doi: https://doi.org/10.2307/2981683. 1003

Geweke, J. and Amisano, G. (2011). "Optimal prediction pools." *Journal of Econometrics*, 164(1): 130–141. MR2821798. doi: https://doi.org/10.1016/j.jeconom.2011.02.017. 1003

Geweke, J. and Amisano, G. (2012). "Prediction with misspecified models." *American Economic Review*, 102(3): 482–486. 1003

Kamary, K., Mengersen, K., Robert, C. P., and Rousseau, J. (2014). "Testing hypotheses via a mixture estimation model." *arXiv preprint arXiv:1412.2044*. 1004

McAlinn, K., Aastveit, K. A., Nakajima, J., and West, M. (2017). "Multivariate Bayesian Predictive Synthesis in Macroeconomic Forecasting." *arXiv preprint arXiv:1711.01667*. 1004

McAlinn, K. and West, M. (2017). "Dynamic Bayesian predictive synthesis in time series forecasting." *arXiv preprint arXiv:1601.07463*. MR3664859. 1004

Piironen, J. and Vehtari, A. (2016). "Projection predictive model selection for Gaussian processes." In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. 1005, 1006

Piironen, J. and Vehtari, A. (2017a). "Comparison of Bayesian predictive methods for model selection." *Statistics and Computing*, 27(3): 711–735. 1005, 1006

Piironen, J. and Vehtari, A. (2017b). "On the hyperprior choice for the global shrinkage parameter in the horseshoe prior." In *Artificial Intelligence and Statistics*, 905–913. 1005

Piironen, J. and Vehtari, A. (2017c). "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics*, 11(2): 5018–5051. 1005

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." *Ecography*, 40(8): 913–929. 1003

Shimodaira, H. (2000). "Improving predictive inference under covariate shift by weighting the log-likelihood function." *Journal of Statistical Planning and Inference*, 90(2): 227–244. MR1795598. doi: https://doi.org/10.1016/S0378-3758(00)00115-4. 1002

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007). "Covariate shift adaptation by importance weighted cross validation." *Journal of Machine Learning Research*, 8(May): 985–1005. 1002

Sugiyama, M. and Müller, K.-R. (2005). "Input-dependent estimation of generalization error under covariate shift." *Statistics & Decisions*, 23(4/2005): 249–279. MR2255627. doi: https://doi.org/10.1524/stnd.2005.23.4.249. 1002

Vehtari, A., Buerkner, P., and Gabry, J. (2018a). "Leave-one-out cross-validation for non-factorizable models." Technical report. URL http://mc-stan.org/loo/articles/loo2-non-factorizable.html 1003

Vehtari, A., Gabry, J., Yao, Y., and Gelman, A. (2018b). "loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models." R package version 2.0.0. 1003

Vehtari, A., Gelman, A., and Gabry, J. (2017). "Pareto smoothed importance sampling." *arXiv preprint arXiv:1507.02646*. 1003

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., and Winther, O. (2016). "Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models." *Journal of Machine Learning Research*, 17(1): 3581–3618. MR3543509. 1006