Invited discussion to the paper "Using Stacking to Average Bayesian Predictive Distributions (with Discussion)" by Yao, Vehtari, Simpson and Gelman

Grünwald, P.D.; Heide, R. de

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

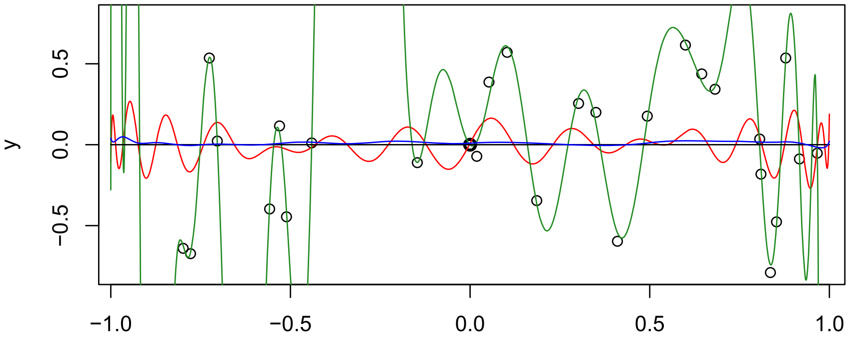

Figure 1: The conditional expectation $\mathbf{E}[Y|X]$ according to the predictive distribution found by stacking (red), standard BMA (green) and SafeBayesian regression (blue), based on models $M_1, \ldots, M_{30}$ with polynomial basis functions, given 50 data points sampled i.i.d. $\sim P^*$, of 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