Discussion of *Using Stacking to Average Bayesian Predictive Distributions* by Yao et. al

Haakon C. Bakka
CEMSE Division
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Daniela Castro-Camilo
CEMSE Division
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Maria Franco-Villoria
Department of Economics and Statistics
University of Torino, Italy

Anna Freni-Sterrantino
Small Area Health Statistics Unit
Department of Epidemiology and Biostatistics
Imperial College London, United Kingdom

Raphael G. Huser
CEMSE Division
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Thomas Opitz
Biostatistics and Spatial Processes Unit
French National Institute for Agronomic Research
84914 Avignon, France

Håvard Rue*
CEMSE Division
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

June 6, 2018

*Corresponding author. Email: haavard.rue@kaust.edu.sa

We read this paper as a recommendation for researchers conducting leave-one-out cross-validation (LOOCV) for model selection to switch to an approach based on LOOCV-stacking of predictive distributions. We found the paper to be meaningful in this context, due to the fact that computations and underlying assumptions are similar. While we raise below some concerns regarding the LOOCV-stacking approach, we hope that our comments will be helpful for future research in this field. Overall, we found that LOOCV-stacking may not only be difficult to compute, but when computed exactly, it could even be inadequate as a proxy for predictive performance.

Our review is structured as follows: Considering the possible problems when performing LOOCV-stacking in practice, we first discuss a simple non-stationary model dealing with coastlines and other physical barriers to illustrate some of the difficulties when stacking predictive distributions. Then, we underline that although the approach taken by the authors is within a Bayesian perspective, its mechanism is the same in a frequentist setting. Next, we question some choices made by authors for comparing the stacking approach with other popular methods. Finally, we query the stability of Pareto-smoothed importance sampling and the computational time costs of the proposed method.

The problem of estimating the leave-one-out predictive density (LOOPD) for a fixed 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 only relates to η_i 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 $p(\eta_i|y_{-i})$, where

$$p(\eta_i|y_{-i}) \propto \frac{p(\eta_i|y)}{p(y_i|\eta_i)}, \tag{1}$$

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. This can be in the form of a smoothed histogram from Monte Carlo samples, or from deterministic approximations. The question is whether the correction needed for removing y_i (i.e., the term $p(y_i|\eta_i)$ in the denominator of Eq. (1)), 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 must/should/could be computed from a re-run 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 a "failure" (which was not a good choice of name, in retrospect). The failed cases can then be recomputed in parallel after the corresponding observations are removed, and we gain speed by using the joint fit as initial values. In addition to being (much) faster than Markov chain Monte Carlo methods, we also get smooth estimates of the posterior marginals involved, 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.

Assume now that the LOOCV scores (log-LOOPD in our case) can be computed accurately, e.g., by re-running the model many times. The question remains as to whether the collection of the LOOCV scores adequately represents the predictive performance of the model.

A recent example of the use of LOOCV log-scores in spatial modelling can be found in Bakka et al. (2018). That paper introduces the Barrier model, a simple non-stationary model dealing with coastlines and other physical barriers, i.e., not smoothing across land as the stationary models would do. The new model has the same computational complexity and the same implementation effort required to fit the model, compared to the stationary alternative. We aimed to compare this and other spatial and non-spatial models in the application presented in the paper, through LOOCV. When

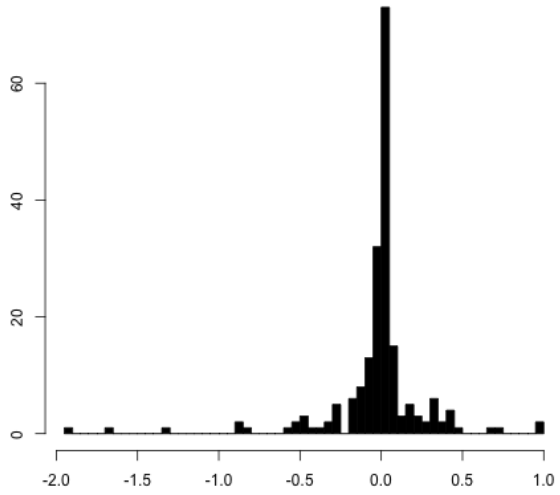

Figure 1: Histogram of differences in LOOCV log predictive density between two spatial models for a dataset on fish larvae.

comparing any two 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 data set, when we examined the individual log-scores, we realised that a small collection of them strongly influenced the model selection result (i.e., taking them out could change the ranking). To illustrate this problem, 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. From this figure, it is clear that we cannot conclude that one model is superior to the other (i.e., that zero is an unreasonable value for the central location of this distribution). In the context of stacking it is clear that we cannot weight one of these two models more than the other with any degree of confidence. To assess the variability inherent to the LOOCV estimate of marginal predictive performance, we then 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 Figure 2 for this computation on the smelt larvae data set. The first interval in this figure corresponds to the histogram in Figure 1. From this figure we note that model 5 (M5) is a bad model (which is no surprise, as this model has only an intercept and an over-dispersed Poisson likelihood). This aside, we are unable to conclude which model is better. In the context of stacking, the five remaining models would be weighted by arbitrary weights to create a stacked model, which we find questionable. As seen from the histogram (Figure 1), the widths of these intervals are due to the fact that only a few of the leave-one-out points have a major contribution to the mean log score difference. We wonder whether combining the bootstrapped uncertainty intervals with the stacking idea (in some way) can lead to a more robust approach to stacking.

Although the paper focuses on the specific technical challenges that the *Bayesian* approach to inference may pose for implementing stacking, stacking predictive distributions could also be promising (modulo the potential issues outlined above) in the setup of *frequentist* estimation, since the only requirement of estimated models is that the conditional distributions of y_i given y_{-i} are tractable. For example, Huser et al. (2017) use the continuous ranked probability score (CRPS) in a cross-validation study to select the "best" model for spatial extremes, by holding out observations from

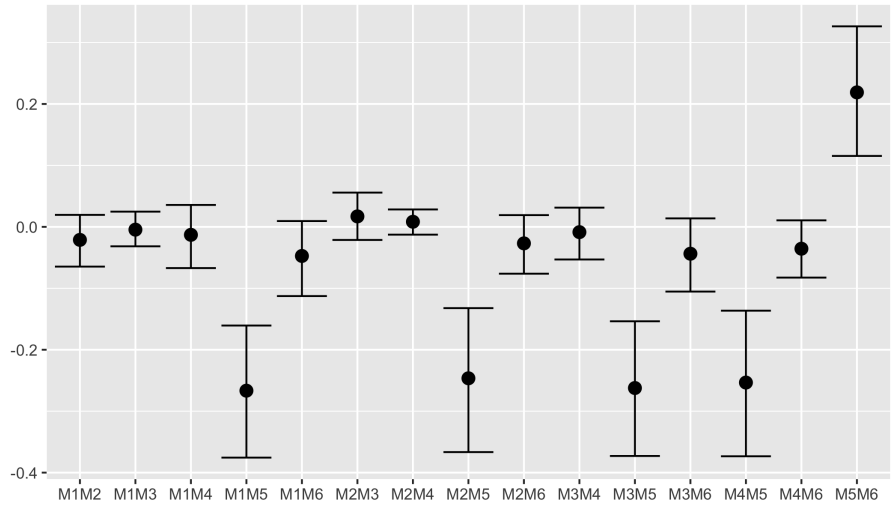

Figure 2: Bootstrapped mean differences in LOOCV negative log predictive density for one non-spatial model (M5) and 5 different spatial models (M1 to M4 and M6) for a dataset on smelt fish larvae. The description of the models can be found in the supplementary material of Bakka et al. (2018).

one station at a time. In this context, the stacking approach would avoid selecting only one of several models with very distinctive features for the prediction of high quantiles, such as asymptotic dependence versus asymptotic independence. We consider that the stacking of predictive distributions in frequentist inference would present a powerful alternative to selecting a single model, in cases where data do not provide clear evidence of one model against the others, while selecting the wrong model may be highly problematic.

We question the authors' choice to compare the stacking approach to the other methods presented in the paper. The probability density function (PDF) obtained from Bayesian model averaging (BMA) can be understood as a weighted average of the PDF obtained under each model, where the weights are determined by the posterior distribution of each model within the training data. As pointed out by the authors, these weights reflect only the fit to the data (i.e., the *within-sample performance*), but they are not obtained in order to maximize the prediction accuracy (i.e., the *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. Observing that the importance ratios $r_{i,k}^s$ (Eq. (6) of the paper) might be highly right-skewed and heavy-tailed, therefore impacting the stability of the proposed inference procedure, the authors suggest fitting the generalized Pareto (GP) distribution to the largest importance ratios and replacing them with the corresponding expected GP order statistics. Following Vehtari et al. (2017)'s guidelines regarding the reliability of this procedure, the authors suggest performing the exact LOOCV instead of adopting the Pareto-smoothed importance sampling approach whenever the estimated GP shape parameter satisfies $\hat{k} > 0.7$. We agree with the authors and want to re-emphasize here that this approach is potentially unstable/invalid with very heavy and "noisy" tails. 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 whenever $k < n - i + 1$, and finite variance whenever $k < (n - i + 1)/2$; 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 is usually 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.

Finally, we have found only few comments in the paper regarding the computational burden of the proposed approach. Calculating stacking weights is computationally demanding as it involves refitting the model n times, unless the Pareto-smoothed importance sampling approximation they propose is used. However, as the authors point out, this is not always possible. Furthermore, to get initial values for these weights, the authors suggest using yet another method (*Pseudo-BMA+ weighting*), which involves the LOOPD and the Bayesian bootstrap. Overall, it seems to us that the computational effort for putting this method into practice might be high. It would be welcome if the authors could provide the running times for the examples in the paper, as they are essential for us to understand the practical implications of all these considerations.

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 (arxiv:1802.06350)*, xx(xx):xx–xx. (Invited extended review, to appear).

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., editors, *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, pages 91–110. Springer Verlag, Berlin.

Hosking, J. R. M. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29:339–349.

Huser, R., Opitz, T., and Thibaud, E. (2017). Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures. *Spatial Statistics*, 21:166–186.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, 67:68–83.

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.

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*. To appear.

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.

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.

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.

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

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