Bayesian leave-one-out cross-validation for non-factorizable normal models*

Paul-Christian Bürkner† Jonah Gabry‡ Aki Vehtari§

May 7, 2019

Abstract

Cross-validation can be used to measure a model's predictive accuracy for the purpose of model comparison, averaging, or selection. Standard leave-one-out cross-validation (LOO-CV) requires the likelihood to be factorizable, but many important models in temporal and spatial statistics do not have this property. We derive how to compute and validate both exact and approximate LOO-CV for Bayesian non-factorizable models with a multivariate normal likelihood. In a case study, we apply this method to lagged simultaneously autoregressive (SAR) models.

Keywords: cross-validation, Pareto-smoothed importance-sampling, non-factorizable models, SAR models.

1. Introduction

In the absence of new data, cross-validation is a general approach for evaluating a statistical model's predictive accuracy for the purpose of model comparison, averaging, or selection (Geisser and Eddy, 1979; Hoeting et al., 1999; Ando and Tsay, 2010; Vehtari and Ojanen, 2012). One widely used variant of cross-validation is *leave-one-out cross-validation* (LOO-CV), where observations are left out one at a time and then predicted based on the model fit to the remaining data. Predictive accuracy is evaluated by first computing the expected log predictive density of the left-out observation and then taking the sum of these values over all observations to obtain the expected log predictive density (ELPD) as a single measure of predictive accuracy. Exact LOO-CV is costly, as it requires fitting the model as many times as there are observations in the data. Depending on the size of the data, complexity of the model, and estimation method, this can be practically infeasible as it simply requires too much computation time. For this reason, approximate versions of LOO-CV have been developed (Gelfand et al., 1992; Vehtari et al., 2017b), most recently using Pareto-smoothed importance-sampling (PSIS; Vehtari et al., 2017b,a).

A standard assumption of any such LOO-CV approach is that the joint likelihood of the model over all observations has to be factorizable. That is, the observations have to be pairwise conditionally independent given the model parameters. However, many important models do not have this property. Particularly in temporal and spatial statistics it is common to fit models with multivariate normal likelihoods that have structured covariance matrices such that the likelihood does not factorize. This is typically due to the fact that observations depend on other observations

*We thank Daniel Simpson for useful discussion and the Academy of Finland (grants 298742, 313122) for partial support of this work.

†Department of Psychology, University of Münster, Germany.

‡Applied Statistics Center and Institute for Social and Economic Research and Policy, Columbia University, USA.

§Department of Computer Science, Aalto University, Finland.

from different time periods or different spatial units in addition to the dependence on the model parameters.

In this short paper we show how equations derived in Sundararajan and Keerthi (2001) can be repurposed and combined with PSIS to allow for performing efficient approximate LOO-CV for *any* multivariate normal Bayesian model with an invertible covariance matrix, regardless of whether or not the likelihood factorizes. We also provide equations for computing exact LOO-CV for these models, which can be used to validate the approximation. Throughout, a Bayesian model specification and estimation via Markov chain Monte Carlo (MCMC) is assumed. In an online supplementary material we provide R code demonstrating how to carry out the approximation described in the paper.[1]

Although our proposed method makes use of standard multivariate normal theory, we think there is value in explicitly presenting this theory to applied researchers, along with a recommended workflow for implementation in practice.

2. Pointwise log-likelihood for non-factorizable normal models

When computing *exact* LOO-CV for a Bayesian model we need to compute the log leave-one-out predictive densities $\log p(y_i|y_{-i})$ for every response value y_i, $i = 1, \ldots, N$, where y_{-i} denotes all response values except observation i. This requires fitting the model N times. For *approximate* LOO-CV using only a single model fit, we instead calculate the pointwise log-likelihood (the log-predictive density evaluated at each data point), without leaving out any observations, and then apply an importance sampling correction (Gelfand et al., 1992; Vehtari et al., 2017b).

The pointwise log-likelihood is straightforward to compute for *factorizable* models in which response values are conditionally independent given the model parameters θ and the likelihood can be written in the familiar form

$$p(y \,|\, \theta) = \prod_{i=1}^{N} p(y_i \,|\, \theta). \tag{1}$$

When $p(y)$ can be factorized in this way, the conditional pointwise log-likelihood can be obtained easily by computing $\log p(y_i \,|\, \theta)$ for each i.

The situation is more complicated for *non-factorizable* models in which response values are not conditionally independent. When there is residual dependence even after accounting for the model parameters, the conditional pointwise log-likelihood has the general form $\log p(y_i \,|\, y_{-i}, \theta)$. Computing this pointwise log-likelihood for non-factorizable models is often impossible, but there is a large class of multivariate normal models for which an analytical solution is available.

Sundararajan and Keerthi (2001) provide equations for the predictive mean and standard deviation for a zero-mean Gaussian process model with prior covariance K and residual standard deviation σ,

$$y \sim \mathrm{N}(0,\, K + \sigma^2 I), \tag{2}$$

[1] Supplemental materials available at `https://mc-stan.org/loo/articles/loo2-non-factorizable.html`.

where I is the identity matrix of appropriate dimension and $C = K + \sigma^2 I$ is the covariance matrix of the model. In the context of Gaussian process models, these equations did not actually find much practical application because, in most cases, Gaussian processes are combined with a factorizable likelihood so that simpler equations for univariate distributions can be applied. But the derivations of Sundararajan and Keerthi's equations follow from multivariate normal theory and make no use of the special form of C for Gaussian process models and thus immediately generalize to the case of an arbitrary invertible covariance matrix C.

For such models, the LOO predictive mean and standard deviation can be computed using the equations from Sundararajan and Keerthi (2001) as follows:

$$\begin{aligned} \mu_{\tilde{y},-i} &= y_i - \bar{c}_{ii}^{-1} g_i \\ \sigma_{\tilde{y},-i} &= \sqrt{\bar{c}_{ii}^{-1}}, \end{aligned} \tag{3}$$

where $g_i = \left[C^{-1} y\right]_i$ and $\bar{c}_{ii} = \left[C^{-1}\right]_{ii}$. The log predictive density of the ith observation is

$$\log p(y_i \mid y_{-i}, \theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log \sigma_{-i}^2 - \frac{1}{2}\frac{(y_i - \mu_{-i})^2}{\sigma_{-i}^2}, \tag{4}$$

and expressing this same equation in terms of g_i and $\bar{c}_{ii}$, we obtain[2]:

$$\log p(y_i \mid y_{-i}, \theta) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log \bar{c}_{ii} - \frac{1}{2}\frac{g_i^2}{\bar{c}_{ii}}. \tag{5}$$

Evaluating equation (5) for each y_i provides the pointwise log-likelihood values required for the PSIS-LOO-CV approximation. While the computational cost in the factorizable case is only $O(n)$, it is much higher in the non-factorizable case. The latter is usually dominated by the computation of C^{-1}, which is in $O(n^3)$ for dense C. However, if C is sparse (see below for an example) or reduced rank, the computational will be less than $O(n^3)$.

It often requires additional work to take a given multivariate normal model and express it in the same form as the zero mean Gaussian process in (2) in order to apply the equations for the predictive mean and standard deviation. Consider, for example, the lagged simultaneously autoregressive (SAR), which has many applications in both the social sciences (e.g., economics) and natural sciences (e.g., ecology). The model is given by

$$y = \rho W y + \eta + \epsilon, \tag{6}$$

or equivalently

$$(I - \rho W) y = \eta + \epsilon, \tag{7}$$

where ρ is a scalar spatial correlation parameter and W is a user-defined matrix of weights. The matrix W has entires $w_{ii} = 0$ along the diagonal and the off-diagonal entries w_{ij} are larger when units i and j are closer to each other but mostly zero as well. In a linear model, the predictor term

[2]Note that Vehtari et al. (2016) has a typo in the corresponding Equation 34.

is $\eta = X\beta$, with design matrix X and regression coefficients β, but the definition of the lagged SAR model holds for arbitrary η, so these results are not restricted to the linear case.

If we have $\epsilon \sim \mathrm{N}(0, \sigma^2 I)$, it follows that

$$(I - \rho W)y \sim \mathrm{N}(\eta, \sigma^2 I), \tag{8}$$

but this standard way of expressing the model is not compatible with the equations from Sundararajan and Keerthi (2001). To make the lagged SAR model reconcilable with these equations we need to rewrite it as follows (conditional on ρ, η, and σ):

$$y - (I - \rho W)^{-1}\eta \sim \mathrm{N}\left(0, \sigma^2 (I - \rho W)^{-1}(I - \rho W)^{-T}\right), \tag{9}$$

or more compactly, with $\widetilde{W} = (I - \rho W)$,

$$y - \widetilde{W}^{-1}\eta \sim \mathrm{N}\left(0, \sigma^2 (\widetilde{W}^T \widetilde{W})^{-1}\right). \tag{10}$$

Written in this way, the lagged SAR model has the same form as the zero mean Gaussian process from (2). Accordingly, we can compute the leave-one-out predictive densities with the equations from Sundararajan and Keerthi (2001), replacing y with $y - \widetilde{W}^{-1}\eta$ and taking the covariance matrix C to be $\sigma^2(\widetilde{W}^T\widetilde{W})^{-1}$. This implies $C^{-1} = \sigma^{-2}\widetilde{W}\widetilde{W}^T$ and so the overall computational cost is dominated by $\widetilde{W}^{-1}\eta$. In SAR models, W is usally sparse and so is $\widetilde{W}$. Thus, if sparse matrix operations are used, then the computational cost for C^{-1} will be less than $O(n^2)$ and for $\widetilde{W}^{-1}$ it will be less than $O(n^3)$ (depending on number of non-zeros and the fill pattern).

3. Approximate LOO-CV for non-factorizable normal models

The conditional pointwise log-likelihood is the only required input to the PSIS-LOO-CV algorithm from Vehtari et al. (2017b) and thus Sundararajan and Keerthi's repurposed equations allow for approximate LOO-CV for *any* model that can be expressed conditionally in terms of a multivariate normal with invertible covariance matrix C, including those where the likelihood does not factorize. For a Bayesian model fit using MCMC the procedure is as follows:

1. Fit the model using MCMC obtaining S samples from the posterior distribution of the parameters θ.

2. For each of the S draws of θ, compute the pointwise log-likelihood value for each of the N observations in y using (5). The results can be stored in an $S \times N$ matrix.

3. Run the PSIS algorithm from Vehtari et al. (2017b) on the $S \times N$ matrix obtained in step 2. For convenience the `loo` R package (Vehtari et al., 2018) provides this functionality.

In the Section 5, we demonstrate this method by computing approximate LOO-CV for the lagged SAR model fit to spatially correlated crime data.

4. Validation using exact LOO-CV

In order to validate the approximate LOO-CV procedure, and also in order to allow exact computations to be made for a small number of leave-one-out folds for which the Pareto-k diagnostic (Vehtari et al., 2017a) indicates an unstable approximation, we need to consider how we might do *exact* LOO-CV for a non-factorizable model. Here we will provide the necessary equations and in the supplementary materials we provide code for comparing the exact and approximate versions.

In the case of a Gaussian process that has the marginalization property, exact LOO-CV is relatively straightforward: when refitting the model we can simply drop the one row and column of the covariance matrix C corresponding to the held out observation without altering the prior of the other observations. But this does not hold in general for all multivariate normal models. Instead, in order to keep the original prior, we may need to maintain the full covariance matrix C even when one of the observations is left out.

The general solution is to model y_i as a missing observation and estimate it along with all of the other model parameters. For a multivariate normal model $\log p(y_i \,|\, y_{-i})$ can be computed as follows. First, we model y_i as missing and denote the corresponding parameter y_i^{mis}. Then, we define

$$y_{\text{mis}(i)} = (y_1, \ldots, y_{i-1}, y_i^{\text{mis}}, y_{i+1}, \ldots, y_N). \tag{11}$$

to be the same as the full set of observations y but replacing y_i with the *parameter* y_i^{mis}.

Second, we compute the LOO predictive means and standard deviations using the equations from Section 2, but replacing y with $y_{\text{mis}(i)}$ in the computation of $\mu_{\tilde{y},-i}$:

$$\mu_{\tilde{y},-i} = y_{\text{mis}(i)} - \bar{c}_{ii}^{-1} g_i, \tag{12}$$

where in this case we have

$$g_i = \left[C^{-1} y_{\text{mis}(i)}\right]_i. \tag{13}$$

The conditional log predictive density is then computed with the above $\mu_{\tilde{y},-i}$ and the left out observation y_i:

$$\log p(y_i \,|\, y_{-i}, \theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log \sigma^2_{\tilde{y},-i} - \frac{1}{2}\frac{(y_i - \mu_{\tilde{y},-i})^2}{\sigma^2_{\tilde{y},-i}}. \tag{14}$$

Finally, the leave-one-out predictive distribution can be estimated as

$$p(y_i \,|\, y_{-i}) \approx \frac{1}{S}\sum_{s=1}^{S} p(y_i \,|\, y_{-i}, \theta^{(s)}_{-i}), \tag{15}$$

where $\theta^{(s)}_{-i}$ are draws from the posterior distribution $p(\theta \,|\, y_{\text{mis}(i)})$.

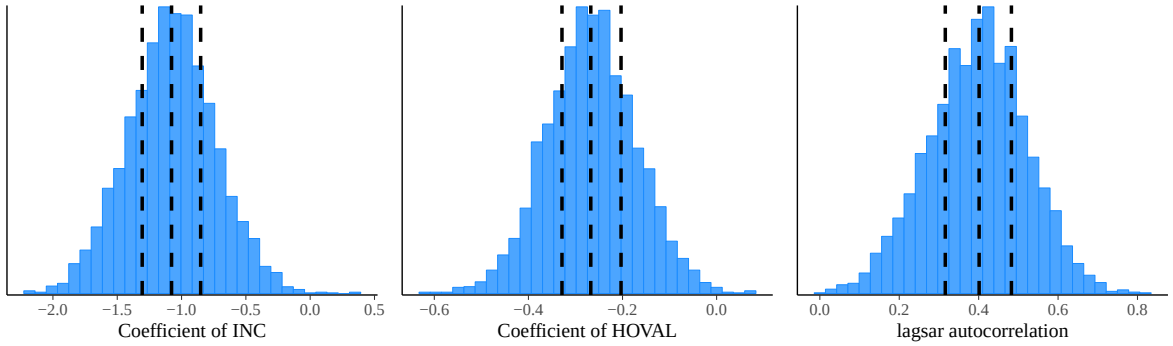

Figure 1: Posterior distribution of selected parameters of the lagged SAR model along with posterior median and 50% central interval.

5. Case Study: Neighborhood Crime in Columbus, Ohio

In order to demonstrate how to carry out the computations implied by these equations, we will first fit a lagged SAR model to data on crime in 49 different neighborhoods of Columbus, Ohio during the year 1980. The data was originally described in Anselin (1988) and ships with the spdep R package (Bivand & Piras, 2015).

The three variables in the data set relevant to this example are: `CRIME`: the number of residential burglaries and vehicle thefts per thousand households in the neighborhood, `HOVAL`: housing value in units of $1000 USD, and `INC`: household income in units of $1000 USD. In addition, we have information about the spatial relationship of neighborhoods from which we can construct the weight matrix to help account for the spatial dependency among the observations. In addition to the loo R package (Vehtari et al., 2018), for this analysis we use the brms interface (Bürkner, 2017) to Stan (Carpenter et al., 2017) to generate a Stan program and fit the model. The complete R code for this case study can be found in the supplemental materials.

In Figure 1, we see that both higher income and higher housing value predict lower crime rates in the neighborhood. Moreover, there seems to be substantial spatial correlation between adjacent neighborhoods, as indicated by the posterior distribution of the `lagsar` parameter.

After fitting the model, the next step is to compute the pointwise log-likelihood values needed for approximate LOO-CV. To do this we use the recipe laid out in Section 2. The quality of the PSIS-LOO approximation can be investigated graphically by plotting the Pareto-k estimate for each observation. Ideally, they should not exceed 0.5, but in practice the algorithm turns out to be robust up to values of 0.7 (Vehtari et al., 2017b,a). In Figure 2, we see that the fourth observation is problematic and so may reduce the accuracy of the LOO-CV approximation.

The PSIS-LOO-CV to approximation of the expected log predictive density for new data reveals $\text{elpd}_{\text{approx}}$ = -187.3. This result still needs to be validated against exact LOO-CV, which is somewhat more involved, as we need to re-fit the model N times each time leaving out a single observations. For the lagged SAR model, we cannot just ignore the held-out observation entirely as

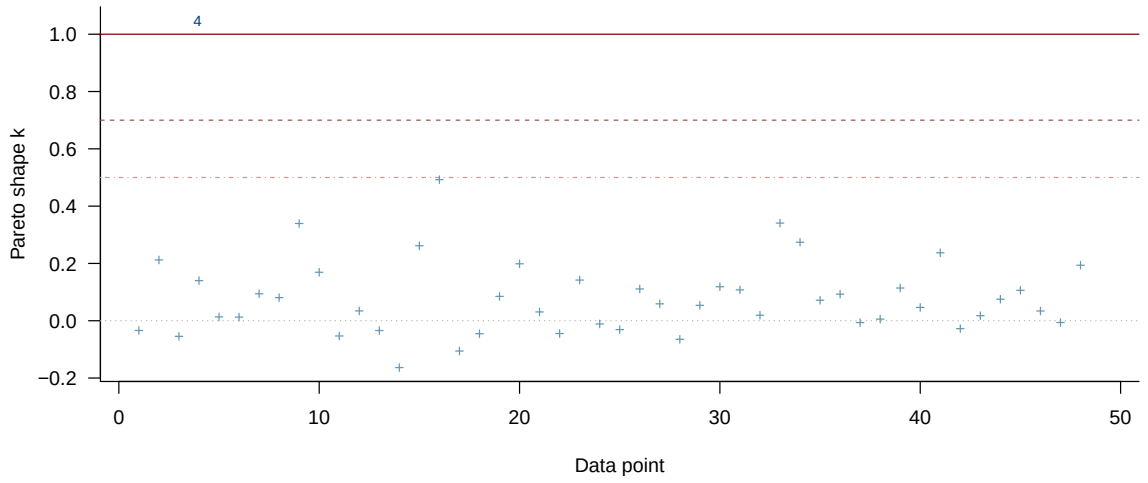

Figure 2: PSIS diagnostic plot showing the Pareto-k-estimate of each observation.

this will change the prior of the other observations. In other words, the lagged SAR model does not have the marginalization property that holds, for instance, for Gaussian process models. Instead, we have to model the held-out observation as a missing value, which is to be estimated along with the other model parameters (see Section 4).

A first step in the validation of the pointwise predictive density is to compare the distribution of the implied response values for the left-out observation using the pointwise mean and standard deviation from (3) to the distribution of the y_i^{mis} posterior-predictive values estimated as part of the model. If the pointwise predictive density is correct, the two distributions should match very closely (up to sampling error). In Figure 3, we overlay these two distributions for the first four observations and see that they match very closely (as is the case for all observations in this example).

In the final step, we compute the pointwise predictive density based on the exact LOO-CV and compare it to the approximate PSIS-LOO-CV result computed earlier. The results of the approximate ($\text{elpd}_{\text{approx}}$ = -187.3) and exact LOO-CV ($\text{elpd}_{\text{exact}}$ = -188.6) are similar but not as close as we would expect if there were no problematic observations. We can investigate this issue more closely by plotting the approximate against the exact pointwise ELPD values.

In Figure 4, the fourth data point – the observation flagged as problematic by the PSIS-LOO approximation – is colored in red and is the clear outlier. Otherwise, the correspondence between the exact and approximate values is strong. In fact, summing over the pointwise ELPD values and leaving out the fourth observation yields practically equivalent results for approximate and exact LOO-CV ($\text{elpd}_{\text{approx},-4}$ = -172.9 vs. $\text{elpd}_{\text{exact},-4}$ = -173.0). From this we can conclude that the difference we found when including *all* observations does not indicate an error in the implementation of the approximate LOO-CV but rather a violation of its assumptions.

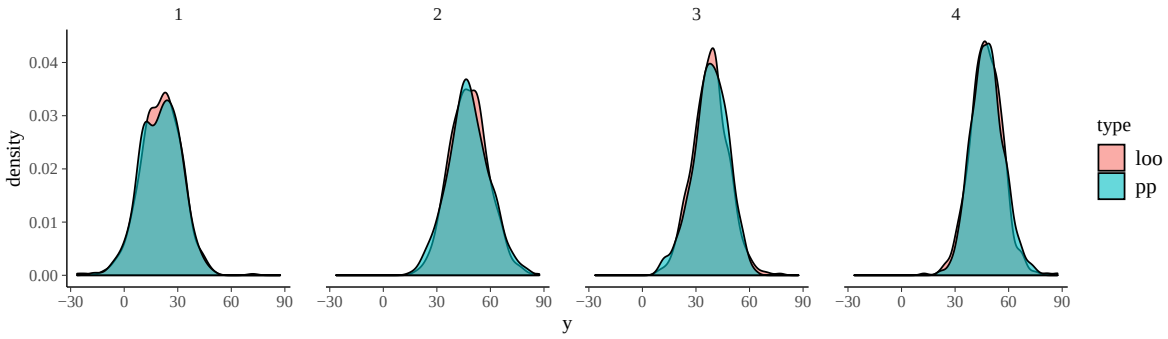

Figure 3: Implied response values of the first four observations computed (a) after model fitting (type = 'loo') and (b) as part of the model in the form of posterior-predictive draws for the missing observation (type = 'pp').

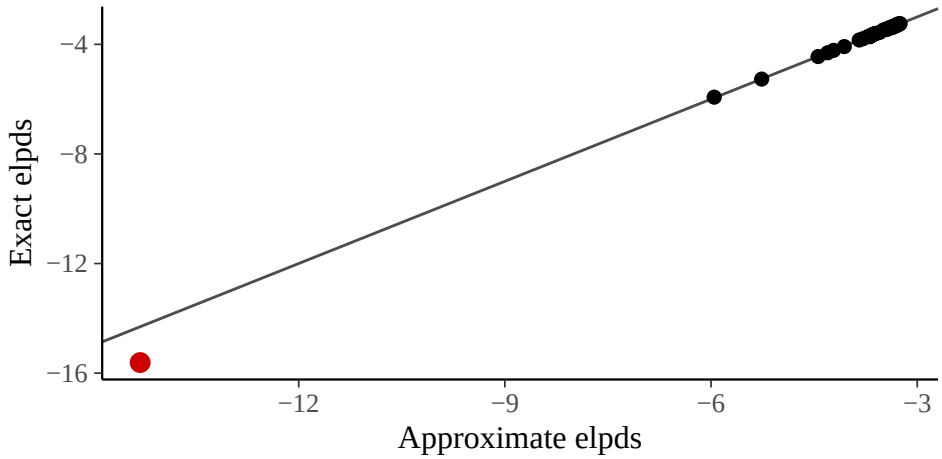

Figure 4: Comparison of approximate and exact pointwise elpd values for the SAR model. Problematic observations are marked as red dots.

6. Conclusion

We have provided equations that enable both approximate and exact LOO-CV for non-factorizable multivariate normal Bayesian models. Although exact LOO-CV is usually impractical, our exact LOO-CV procedure can be used to validate the more efficient PSIS-LOO-CV approximation.

The primary motivation for this paper is to enable approximate LOO-CV for models that cannot be factorized at all, but our approach also works for *any* Bayesian model that can be expressed in terms of a multivariate normal likelihood. Therefore it may also be useful for models that are factorizable but for which the factorized representation is difficult to compute or not available to the researcher for some other reason.

References

Ando, T. and R. Tsay (2010). Predictive likelihood for Bayesian model selection and averaging. *International Journal of Forecasting 26*(4), 744–763.

Anselin, L. (1988). *Spatial econometrics: methods and models*. Dordrecht: Kluwer Academic.

Bürkner, P.-C. (2017, September). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software 80*(1), 1–28.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Ridell (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*.

Geisser, S. and W. F. Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association 74*(365), 153–160.

Gelfand, A., D. Dey, and H. Chang (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian Statistics 4*, 147–167.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: a tutorial. *Statist. Sci. 14*(4), 382–417.

Sundararajan, S. and S. S. Keerthi (2001). Predictive approaches for choosing hyperparameters in Gaussian processes. *Neural Computation 13*(5), 1103–1118.

Vehtari, A., J. Gabry, Y. Yao, and A. Gelman (2018). **loo***: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models.* R package version 2.0.0.

Vehtari, A., A. Gelman, and J. Gabry (2017a). Pareto smoothed importance sampling. *arXiv preprint*.

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

Vehtari, A., T. Mononen, V. Tolvanen, T. Sivula, and O. Winther (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *Journal of Machine Learning Research 17*(103), 1–38.

Vehtari, A. and J. Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys 6*, 142–228.