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

Paul-Christian Bürkner[†] Jonah Gabry[‡] Aki Vehtari[§]

February 26, 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.

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. 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. Unfortunately, exact LOO-CV is costly, as it requires fitting the model as many times are 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 (Vehtari et al., 2017b). For this reason, approximate versions of LOO-CV have been developed, most notably using Pareto-smoothed importance-sampling (PSIS, Vehtari et al. (2017b,a)), which is applicable to Bayesian models.

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 the fields of 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.

[*] 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.

[‡] Institute for Social and Economic Research and Policy, Columbia University.

[§] Department of Computer Science, Aalto University, Finland.

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. In online supplementary material we provide R code demonstrating how to carry out the method proposed in the paper using an example of spatially correlated crime data.

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

When computing *exact* LOO-CV for a Bayesian model we need to compute the log LOO 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 evaluated at each data point, without leaving any out, and then apply an importance sampling correction (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 $\log p(y_i \,|\, y_{-i}, \theta)$. Most often this is due to the fact that observations depend on other observations from different time periods or different spatial units. 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,

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

where I is the identity matrix of appropriate dimension and $C = K + \sigma^2 I$ is the covariance matrix of the model. These equations were not traditionally intended to be used for LOO-CV of Gaussian process models 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 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 Sundararajan and Keerthi's results 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

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

and expressing this same equation in terms of g_i and $\bar{c}_{ii}$, the log predictive density becomes:

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

(note that Vehtari et al. (2016) has a typo in the corresponding Equation 34).

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

The conditional pointwise log-likelihood (the pointwise log-predictive density evaluated at each data point) 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 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 the formula in (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 supplementary materials we demonstrate this method by computing approximate LOO-CV for a lagged simultaneously autoregressive (SAR) model fit to the 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 to 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 of LOO-CV for the lagged SAR model and show that the approximation works well.

In the case of a Gaussian process that has the marginalization property, exactly LOO-CV is relatively straighforward: we can simply drop the one row and column of the covariance matrix C corresponding to the held out observation when refitting the model. But this does not hold in general for multivariate normal models, and 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 conditional 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

$$y_{\text{mis}(i)} = (y_1, \ldots, y_{i-1}, y_i^{\text{mis}}, y_{i+1}, \ldots, y_N). \tag{6}$$

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 mean and standard deviations as above, but replace y with $y_{\text{mis}(i)}$ in the computation of $\mu_{\tilde{y},-i}$:

$$\mu_{\tilde{y},-i} = y_{\text{mis}(i)} - \bar{c}_{ii}^{-1} g_i, \tag{7}$$

where in this case we have

$$g_i = \left[C^{-1} y_{\text{mis}(i)} \right]_i . \tag{8}$$

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

$$\log p(y_i \,|\, y_{-i}, \theta) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log \sigma^2_{\tilde{y},-i} - \frac{1}{2} \frac{(y_i - \mu_{\tilde{y},-i})^2}{\sigma^2_{\tilde{y},-i}}. \tag{9}$$

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

$$p(y_i \,|\, y_{-i}) \approx \frac{1}{S} \sum_{s=1}^{S} p(y_i \,|\, y_{-i}, \theta_{-i}^{(s)}), \tag{10}$$

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

5. Conclusion

We have provided equations for that enable approximate LOO-CV for non-factorizable multivariate normal models as well as the required equations for exact LOO-CV, which can be used to validate the approximation. Although our motivation is the case of models that cannot be factorized at all, our approach also works for *any* Bayesian model that can be expressed in terms of a multivariate normal likelihood. This may also be useful with 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

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

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., A. Gelman, and J. Gabry (2018). **loo***: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models.* R package version 1.0.0.

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.