# Extending cross-validation for temporal and spatial models

Paul-Christian Bürkner[1], Jonah Gabry[2], & Aki Vehtari[3]

[1] Department of Psychology, University of Münster, Germany
[2] Institute for Social and Economic Research in Policy, Columbia University, USA
[3] Department of Computer Science, Aalto University, Finnland

Abstract

TODO

*Keywords:* cross-validation, pareto-smoothed importance-sampling

## 1 Introduction

### 1.1 Approximate LOO-CV using integrated importance-sampling

To obtain the leave-one-out predictive density $p(y_i \,|\, y_{-i})$ we need to integrate over $\theta$,

$$p(y_i \,|\, y_{-i}) = \int p(y_i \,|\, y_{-i}, \theta)\, p(\theta \,|\, y_{-i})\, d\theta. \tag{1}$$

Here, $p(\theta \,|\, y_{-i})$ is the leave-one-out posterior distribution for $\theta$, that is, the posterior distribution for $\theta$ obtained by fitting the model while holding out the $i$th observation.

To avoid the cost of sampling from $N$ leave-one-out posteriors, it is possible to take the posterior draws $\theta^{(s)}$, $s = 1, \ldots, S$, from the *full* posterior $p(\theta \,|\, y)$, and then approximate the above integral using integrated importance sampling (Vehtari et al. (2016b), Section 3.6.1):

$$p(y_i \,|\, y_{-i}) \approx \frac{\sum_{s=1}^{S} p(y_i \,|\, y_{-i},\, \theta^{(s)})\, w_i^{(s)}}{\sum_{s=1}^{S} w_i^{(s)}}, \tag{2}$$

Correspondence concerning this article should be addressed to Paul-Christian Bürkner, Department of Psychology, University of Münster, Fliednerstrasse 21, 48149 Münster, Germany. E-mail: paul.buerkner@gmail.com where $w_i^{(s)}$ are importance weights. First we compute the raw importance ratios

$$r_i^{(s)} \propto \frac{1}{p(y_i \,|\, y_{-i},\, \theta^{(s)})}, \tag{3}$$

and then stabilize them using Pareto smoothed importance sampling (PSIS) to obtain the weights $w_i^{(s)}$ (Vehtari et al., 2017b, 2017a). The resulting approximation is referred to as PSIS-LOO (Vehtari et al., 2017b).

## 2 M-step ahead predictions for time-series models

## 3 Leave-one-out cross validation for non-factorizable models

When computing approximate leave-one-out cross-validation (LOO-CV) after fitting a Bayesian model, the first step is to calculate the *pointwise* log-likelihood for every response value $y_i$, $i = 1, \ldots, N$. This is straightforward for *factorizable* models in which response values are conditionally independent given the model parameters $\theta$ and the likelihood can be written in the familiar form

$$p(y \,|\, \theta) = \prod_{i=1}^{N} p(y_i \,|\, \theta). \tag{4}$$

The function $p$ will be either a probability density function (PDF) or a probability mass function (PMF) depending on whether we have a continuous or discrete outcome. 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 $i$. We then save each of these individual contributions to the log-likelihood rather than simply summing them to obtain the total log-likelihood.

The situation is more complicated for *non-factorizable* models in which response values are not conditionally independent. When there is residual dependency even after accounting for the model parameters $\theta$, the conditional pointwise log-likelihood has the general form $\log p(y_i \,|\, y_{-i}, \theta)$, where $y_{-i}$ denotes all response values except observation $i$.

### 3.1 LOO-CV for multivariate normal models

Although computing the pointwise log-likelihood for non-factorizable models is often impossible, there is a large class of multivariate normal models for which an analytical solution is available. These equations were initially derived by Sundararajan and Keerthi (2001) with a focus on the special case of a zero-mean Gaussian process model with prior covariance $K$ and residual standard deviation $\sigma$,

$$y \sim \mathrm{N}(0,\, K + \sigma^2 I), \tag{5}$$

where $I$ is the identity matrix of appropriate dimension and $C = K + \sigma^2 I$ is the covariance matrix of the model. Sundararajan and Keerthi's derivations 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 as follows:

$$\begin{aligned} \mu_{\tilde{y},-i} &= y_i - \bar{c}_{ii}^{-1} g_i \\ \sigma_{\tilde{y},-i} &= \sqrt{\bar{c}_{ii}^{-1}}, \end{aligned} \tag{6}$$

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

$$\log p(y_i \,|\, y_{-i}, \theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log \sigma_{-i}^2 - \frac{1}{2}\frac{(y_i - \mu_{-i})^2}{\sigma_{-i}^2}. \tag{7}$$

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

$$\log p(y_i \,|\, y_{-i}, \theta) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log \bar{c}_{ii} - \frac{1}{2}\frac{g_i^2}{\bar{c}_{ii}} \tag{8}$$

(note that Vehtari et al. (2016b) has a typo in the corresponding Equation 34). From these equations we can now derive a recipe for obtaining the conditional pointwise log-likelihood for *all* models that can be expressed conditionally in terms of a multivariate normal with invertible covariance matrix $C$.

### 3.2 Exact LOO-CV with re-fitting

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. In the case of a Gaussian process that has the marginalization property, we could just drop the one row and column of $C$ corresponding to the held out out observation. This does not hold in general for multivariate normal models, however, 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 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 $y_i^{\mathrm{mis}}$. Then, we define

$$y_{\mathrm{mis}(i)} = (y_1, \ldots, y_{i-1}, y_i^{\mathrm{mis}}, y_{i+1}, \ldots, y_N). \tag{9}$$

to be the same as the full set of observations $y$, except replacing $y_i$ with the parameter $y_i^{\mathrm{mis}}$.

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

$$\mu_{\tilde{y},-i} = y_{\mathrm{mis}(i)} - \bar{c}_{ii}^{-1} g_i, \tag{10}$$

where in this case we have

$$g_i = \left[C^{-1} y_{\mathrm{mis}(i)}\right]_i. \tag{11}$$

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

$$\log p(y_i \,|\, y_{-i}, \theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log \sigma^2_{\tilde{y},-i} - \frac{1}{2}\frac{(y_i - \mu_{\tilde{y},-i})^2}{\sigma^2_{\tilde{y},-i}}. \tag{12}$$

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

$$p(y_i \,|\, y_{-i}) \approx \frac{1}{S}\sum_{s=1}^{S} p(y_i \,|\, y_{-i}, \theta_{-i}^{(s)}), \tag{13}$$

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

## 4 Case Studies

To illustrate the above described methods, we use two case studies. In the first case study, we will explain how to perform m-step ahead predictions for an autoregressive model applied to annual measurements of the level of Lake Huron 1875–1972 (Brockwell & Davis, 2016). In the second case study, we will explain how to perform leave-one-out cross-validation for a lagged simultaneously autoregressive model applied to data of Neighborhood Crime in Columbus, Ohio [.

### 4.1 M-step ahead predictions for autoregressive models

### 4.2 LOO-CV for lagged SAR models

A common non-factorizable multivariate normal model is the simultaneously autoregressive (SAR) model, which is frequently used for spatially correlated data. The lagged SAR model is defined as

$$y = \rho W y + \eta + \epsilon \tag{14}$$

or equivalently

$$(I - \rho W)y = \eta + \epsilon, \tag{15}$$

where $\rho$ is the spatial correlation parameter and $W$ is a user-defined weight matrix. The matrix $W$ has entries $w_{ii} = 0$ along the diagonal and the off-diagonal entries $w_{ij}$ are larger when areas $i$ and $j$ are closer to each other. In a linear model, the predictor term $\eta$ is given by $\eta = X\beta$ with design matrix $X$ and regression coefficients $\beta$. However, since the above equation holds for arbitrary $\eta$, these results are not restricted to linear models. If we have $\epsilon \sim \mathrm{N}(0, \sigma^2 I)$, it follows that

$$(I - \rho W)y \sim \mathrm{N}(\eta, \sigma^2 I). \tag{16}$$

For the purpose of computing LOO-CV, it makes sense to rewrite the SAR model in slightly different form. Conditional on $\rho$, $\eta$, and $\sigma$, if we write

$$y - (I - \rho W)^{-1}\eta \sim \mathrm{N}(0, \sigma^2 (I - \rho W)^{-1}(I - \rho W)^{-T}), \tag{17}$$

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

$$y - \widetilde{W}^{-1}\eta \sim \mathrm{N}(0, \sigma^2(\widetilde{W}^T\widetilde{W})^{-1}), \tag{18}$$

then this has the same form as the zero mean Gaussian process from above. 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 $\sigma^2(\widetilde{W}^T\widetilde{W})^{-1}$.

**4.2.1 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 package.

In addition to the loo package (Vehtari et al., 2016a), for this analysis we used the brms interface (Bürkner, 2017b, 2017a) to Stan (Carpenter et al., 2017) to generate a Stan program and fit the model, and also the bayesplot and ggplot2 packages for plotting. 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 neighbood, `HOVAL`: housing value in units of $1000 USD, and `INC`: household income in units of $1000 USD. We will also use the object `COL.nb`, which is a list containing information about which neighborhoods border each other. From this list we will be able to construct the weight matrix to used to

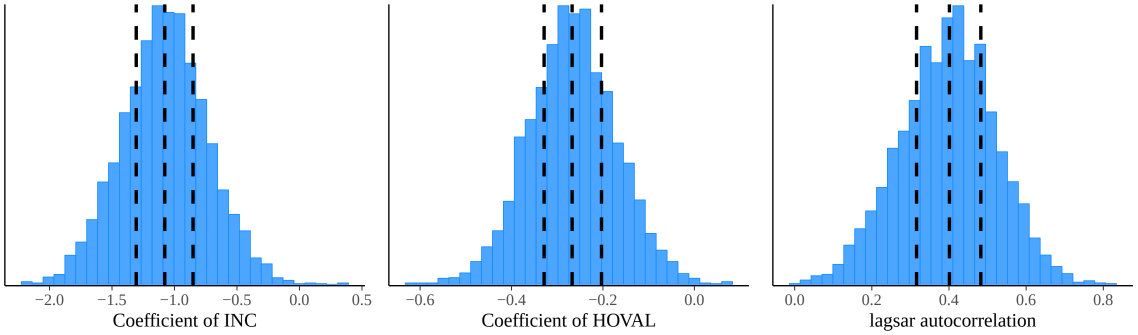

*Figure 1*. Posterior distribution of selected parameters of the lagged SAR model along with posterior median and 50% central interval.

help account for the spatial dependency among the observations. The complete R code for this case study can be found at (loo vignette link).

A model predicting `CRIME` from `INC` and `HOVAL`, while accounting for the spatial dependency via an SAR structure, can be specified in brms as follows:

```
brm(CRIME ~ INC + HOVAL, data = COL.OLD, autocor = cor_lagsar(COL.nb))
```

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.

**4.2.2 Approximate and exact LOO-CV.** 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 3.

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, 2017a). 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 to approximation of the expected log predictive density (ELPD) for new data reveals $\text{elpd}_{\text{approx}}$ = -187.25. 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 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 (cite). 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 (loo vignette link) for details on the R code).

Next, we fit the model $N$ times, each time leaving out a single observation and then computing the log predictive density for that observation. For obvious reasons, this takes

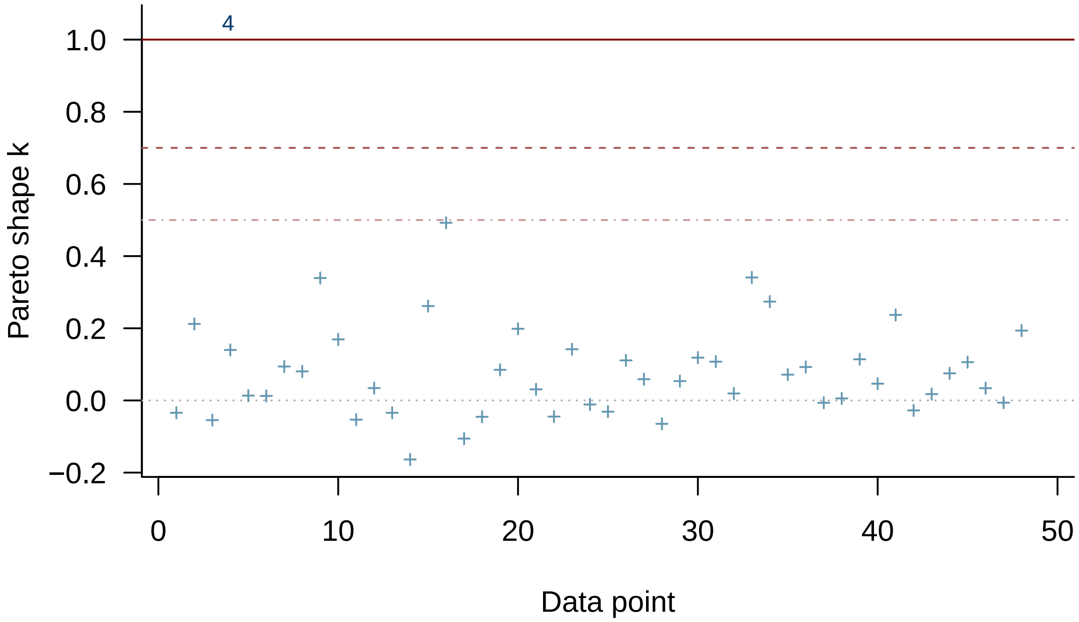

*Figure 2*. PSIS diagnostic plot showing the Pareto-k-estimate of each observation.

much longer than the approximation we computed above, but it is necessary in order to validate the approximate LOO-CV method. Thanks to the PSIS-LOO approximation, in general doing these slow exact computations can be avoided.

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 to the distribution of the $y_i^{\text{mis}}$ 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 49 observations of in this example).

In the final step, we compute the ELPD based on the exact LOO-CV and compare it to the approximate PSIS-LOO result computed earlier. The results of the approximate ($\text{elpd}_{\text{approx}}$ = -187.25) and exact LOO-CV ($\text{elpd}_{\text{exact}}$ = -188.32) 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 ($\text{elpd}_{\text{approx},-4}$ = -172.94 vs. $\text{elpd}_{\text{exact},-4}$ = -173.00). From

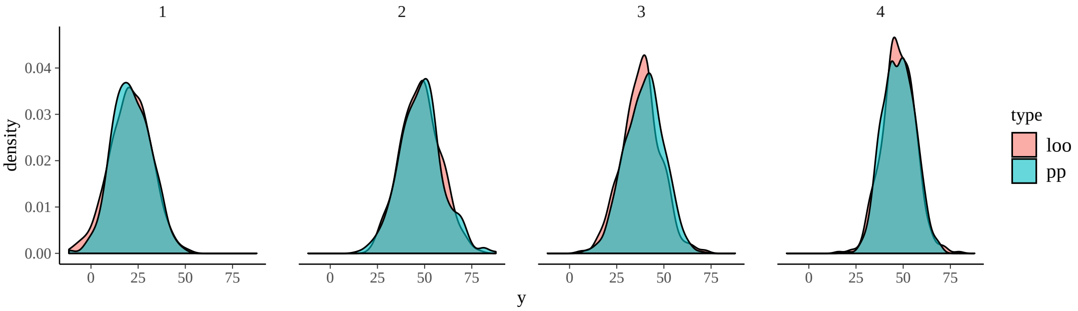

*Figure 3*. Implied response values of the first four observations computed (a) after model fitting (type = 'loo') and (b) as part of the model in the form of posterior-predictive draws (type = 'pp').

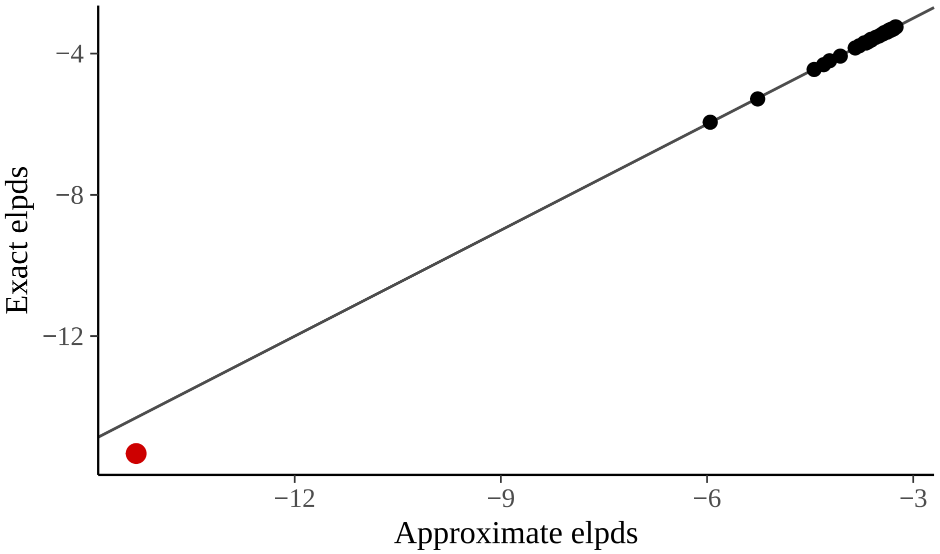

*Figure 4*. Comparison of approximate and exact pointwise elpd values for the SAR model. Problematic observations are marked as red dots.

this we can conclude that the difference we found when including *all* observations does not indicate a bug in our implementation of the approximate LOO-CV but rather a violation of its assumptions.

## 5 Discussion

## References

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

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting*. springer.

Bürkner, P.-C. (2017a). Advanced bayesian multilevel modeling with the r package brms. *arXiv Preprint*, 1–15. Retrieved from https://arxiv.org/abs/1705.11123

Bürkner, P.-C. (2017b). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:10.18637/jss.v080.i01

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

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

Vehtari, A., Gelman, A., & Gabry, J. (2016a). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.* Retrieved from https://github.com/stan-dev/loo

Vehtari, A., Gelman, A., & Gabry, J. (2017a). Pareto smoothed importance sampling. *arXiv Preprint.* Retrieved from https://arxiv.org/abs/1507.02646

Vehtari, A., Gelman, A., & Gabry, J. (2017b). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432. Retrieved from http://link.springer.com/article/10.1007/s11222-016-9696-4

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., & Winther, O. (2016b). Bayesian leave-one-out cross-validation approximations for gaussian latent variable models. *Journal of Machine Learning Research*, *17*(103), 1–38. Retrieved from http://jmlr.org/papers/v17/14-540.html