Efficient leave-one-out cross-validation for Bayesian non-factorizable normal and student-t models

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

[1] Department of Computer Science, Aalto University, Finland
[2] Applied Statistics Center and ISERP, Columbia University, USA
[*] Corresponding author, Email: paul.buerkner@gmail.com

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. A lot of such non-factorizable models fall into the category of multivariate normal models, which can be generalized to multivariate student-t models for increased robustness against outliers. We derive how to efficiently compute and validate both exact and approximate LOO-CV for Bayesian non-factorizable models with a multivariate normal or student-t likelihood. In a case study, we demonstrate this method using lagged simultaneously autoregressive (SAR) importance-sampling, non-factorizable models, Bayesian inference, SAR models models.

Keywords: cross-validation, Pareto-smoothed

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 a pointwise predictive measure and then taking the sum of these values over all observations to obtain a single measure of predictive accuracy (e.g., Vehtari et al., 2017b). In this paper, we focus on the expected log predictive density (ELPD) as the measure of predictive accuracy. The ELPD takes the whole predictive distribution into account and is less focused on the bulk of the distribution compared to other common measures like root mean squared error (RMSE) or mean absolute error (MAE; see Vehtari and Ojanen, 2012, for details). 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 using the ELPD 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 or student-t 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 from different time periods or different spatial units in addition to the dependence on the model parameters.

In this paper we derive how to perform efficient approximate LOO-CV for *any* Bayesian multivariate normal or student-t model with an invertible covariance or scale 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 and to handle problematic observations. Throughout, a Bayesian model specification and estimation via Markov chain Monte Carlo (MCMC) is assumed. We have implemented the developed methods in the R package brms (Bürkner, 2017, 2018). All materials including R source code are available in an online supplement.[1]

2 Pointwise log-likelihood for non-factorizable models

When computing ELPD-based *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. To obtain $p(y_i|y_{-i})$, we need to integrate over the model parameters θ:

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

Here, $p(\theta \,|\, y_{-i})$ is the leave-one-out posterior distribution for θ, that is, the posterior distribution for θ obtained by fitting the model while holding out the ith observation (in Section 3, we will show how refitting the model to data y_{-i} can be avoided). The pointwise log-likelihood $p(y_i \,|\, y_{-i}, \theta)$ is straightforward to compute for *factorizable* models in which response values are conditionally independent given θ, that is, if $p(y_i \,|\, y_{-i}, \theta) = p(y_i \,|\, \theta)$. The full likelihood can then be written in the familiar form

$$p(y \,|\, \theta) = \prod_{i=1}^{N} p(y_i \,|\, \theta). \tag{2}$$

When $p(y \,|\, \theta)$ factorizes this way, the conditional pointwise log-likelihood can be obtained easily by computing $\log p(y_i \,|\, \theta)$ for each i with computational cost $O(n)$.

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 continues to have 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 and student-t models for which we will provide efficient analytical solutions in this paper.

[1] Supplemental materials available at https://github.com/paul-buerkner/psis-non-factorizable-paper.

2.1 Non-factorizable normal models

The density of the N dimensional multivariate normal distribution of vector y is given by

$$p(y|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\left(-\frac{1}{2}(y-\mu)^{\mathrm{T}} \Sigma^{-1} (y-\mu)\right) \tag{3}$$

with mean vector μ and covariance matrix Σ. Often μ and Σ are functions of the model parameters θ, that is, $\mu = \mu(\theta)$ and $\Sigma = \Sigma(\theta)$, but for notational convenience we omit the potential dependence of μ and Σ on θ unless it is relevant. Using standard multivariate normal theory (e.g., Tong, 2012), we know that for the ith observation the conditional distribution $p(y_i|y_{-i}, \theta)$ is univariate normal with mean

$$\tilde{\mu}_i = \mu_i + \sigma_{i,-i} \Sigma_{-i}^{-1} (y_{-i} - \mu_{-i}) \tag{4}$$

and variance

$$\tilde{\sigma}_i = \sigma_{ii} + \sigma_{i,-i} \Sigma_{-i}^{-1} \sigma_{-i,i} \tag{5}$$

In the equations above, μ_{-i} is the mean vector without the ith element, Σ_{-i} is the covariance matrix without the ith row and column (Σ_{-i}^{-1} is its inverse), $\sigma_{i,-i}$ and $\sigma_{-i,i}$ are the ith row and column vectors of Σ without the ith element, and σ_{ii} is the ith diagnonal element of Σ. Equations (4) and (5) can be used to compute the pointwise log-likelihood values as

$$\log p(y_i \,|\, y_{-i}, \theta) = -\frac{1}{2}\log(2\pi\tilde{\sigma}_i) - \frac{1}{2}\frac{(y_i - \tilde{\mu}_i)^2}{\tilde{\sigma}_i}. \tag{6}$$

Evaluating equation (6) for each y_i and each posterior draw θ_s then constitutes the input for the LOO-CV computations. However, the resulting procedure is quite inefficient. Computation is usually dominated by the $O(N^k)$ cost of computing Σ_{-i}^{-1}, where k depends on the structure of Σ. If Σ is dense then $k = 3$. For sparse Σ or reduced rank computations we have $2 < k < 3$. And since Σ_{-i}^{-1} must be computed for each i, the overall complexity is actually $O(N^{k+1})$.

Additionally, if Σ_{-i} also depends on the model parameters θ in a non-trivial manner, which is the case for most models of practical relevance, then it needs to be inverted for each of the S posterior draws. Therefore, in most applications the overall complexity will actually be $O(SN^{k+1})$, which will be impractical in most situations. Accordingly, we seek to find more efficient expressions for $\tilde{\mu}_i$ and $\tilde{\sigma}_i$ that make these computations feasible in practice.

Proposition 1 *If y is multivariate normal with mean vector μ and covariance matrix Σ, then the conditional mean and standard deviation of y_i given y_{-i} for any observation i can be computed as*

$$\tilde{\mu}_i = y_i - \frac{g_i}{\bar{\sigma}_{ii}}, \tag{7}$$

$$\tilde{\sigma}_i = \frac{1}{\bar{\sigma}_{ii}}, \tag{8}$$

where $g_i = \left[\Sigma^{-1}(y-\mu)\right]_i$ and $\bar{\sigma}_{ii} = \left[\Sigma^{-1}\right]_{ii}$.

The proof is based on results from Sundararajan and Keerthi (2001) and is provided in the Appendix. Contrary to the brute force computations in (4) and (5), where Σ_{-i} has to be inverted separately for each i,

equations (7) and (8) require inverting the full covariance matrix Σ only once and it can be reused for each i. This reduces the computational cost to $O(N^k)$ if Σ is independent of θ and $O(SN^k)$ otherwise. If the model is parameterized in terms of the covariance matrix $\Sigma = \Sigma(\theta)$, it is not possible to reduce the complexity further as inverting Σ is unavoidable. However, if the model is parameterized directly through the inverse of Σ, that is $\Sigma^{-1} = \Sigma^{-1}(\theta)$, the complexity goes down to $O(SN^2)$. Note that the latter is not possible in the brute force approach as both Σ and Σ^{-1} are required.

2.2 Non-factorizable student-t models

It is well known that (multivariate) normal models are easily influenced by outliers due to the light tails of the normal distribution (e.g., Schwager et al., 1982). Generalizations of the multivariate normal distribution have been suggested that are more robust to outliers, most notably the multivariate student-t distribution (e.g., Zellner, 1976), which has an additional positive *degrees of freedom* parameter ν that controls the tails of the distribution. If ν is small, the tails are fat and extreme data points are less surprising under the model and thus less influential for the resulting parameter estimates. If ν is large, the multivariate student-t distribution becomes more similar to the corresponding multivariate normal distribution and is equal to the latter for $\nu \to \infty$. As ν can be estimated alongside the other model parameters in student-t models, the fatness of the tails is flexibly adjusted based on information from the observed response values and the prior.

The density of the N dimensional multivariate student-t distribution of vector y is given by

$$p(y|\nu, \mu, \Sigma) = \frac{\Gamma((\nu+N)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{(\nu\pi)^N |\Sigma|}} \left(1 + \frac{1}{\nu}(y-\mu)^{\mathrm{T}} \Sigma^{-1} (y-\mu)\right)^{-(\nu+N)/2} \tag{9}$$

with degrees of freedom ν, location vector μ and scale matrix Σ. The mean of y is μ if $\nu > 1$ and $\frac{\nu}{\nu-2}\Sigma$ is the covariance matrix if $\nu > 2$. Similar to the multivariate normal case, the conditional distribution of the ith observation given all other observations and the model parameters, $p(y_i|y_{-i}, \theta)$, can be computed analytically.

Proposition 2 *If y is multivariate student-t with degrees of freedom ν, location vector μ, and scale matrix Σ, then the conditional distribution of y_i given y_{-i} for any observation i is univariate student-t with parameters*

$$\tilde{\nu}_i = \nu + N - 1, \tag{10}$$

$$\tilde{\mu}_i = \mu_i + \sigma_{i,-i}\Sigma_{-i}^{-1}(y_{-i} - \mu_{-i}), \tag{11}$$

$$\tilde{\sigma}_i = \frac{\nu + \beta_{-i}}{\nu + N - 1}\left(\sigma_{ii} + \sigma_{i,-i}\Sigma_{-i}^{-1}\sigma_{-i,i}\right), \tag{12}$$

where

$$\beta_{-i} = (y_{-i} - \mu_{-i})^{\mathrm{T}} \Sigma_{-i}^{-1} (y_{-i} - \mu_{-i}). \tag{13}$$

A proof based on results of Shah et al. (2014) is given in the Appendix. Here $\tilde{\mu}_i$ is the same as in the normal case and $\tilde{\sigma}_i$ is the same up to the correction factor $\frac{\nu+\beta_{-i}}{\nu+N-1}$, which approaches 1 for $\nu \to \infty$ as one would expect. Based on the above equations, we can compute the pointwise log-likelihood values in the student-t

case as

$$\log p(y_i \mid y_{-i}, \theta) = \log(\Gamma((\tilde{\nu}_i + 1)/2)) - \log(\Gamma(\tilde{\nu}_i/2)) - \frac{1}{2}\log(\tilde{\nu}_i \pi \tilde{\sigma}_i) - \frac{\tilde{\nu}_i + 1}{2} \log\left(1 + \frac{1}{\tilde{\nu}_i} \frac{(y_i - \tilde{\mu}_i)^2}{\tilde{\sigma}_i}\right). \tag{14}$$

This approach has the same overall computational cost of $O(SN^{k+1})$ as the non-optimized normal case and is therefore quite inefficient. Fortunately, the efficiency can again be improved.

Proposition 3 *If y is multivariate student-t with degrees of freedom ν, location vector μ, and scale matrix Σ, then the conditional location and scale of y_i given y_{-i} for any observation i can be computed as*

$$\tilde{\mu}_i = y_i - \frac{g_i}{\bar{\sigma}_{ii}}, \tag{15}$$

$$\tilde{\sigma}_i = \frac{\nu + \beta_{-i}}{\nu + N - 1} \frac{1}{\bar{\sigma}_{ii}}, \tag{16}$$

with

$$\beta_{-i} = (y_{-i} - \mu_{-i})^{\mathrm{T}} \left(\Sigma^{-1} - \frac{\bar{\sigma}_{-i,i} \bar{\sigma}_{-i,i}^{\mathrm{T}}}{\bar{\sigma}_{ii}}\right) (y_{-i} - \mu_{-i}), \tag{17}$$

where $g_i = \left[\Sigma^{-1}(y - \mu)\right]_i$, $\bar{\sigma}_{ii} = \left[\Sigma^{-1}\right]_{ii}$, and $\bar{\sigma}_{-i,i} = \left[\Sigma^{-1}\right]_{-i,i}$ is the ith column vector of Σ^{-1} without the ith element.

The proof is provided in the Appendix. After inverting Σ, computing β_{-i} for a single i is an $O(N^2)$ operation, which needs to be repeated for each observation. So the cost of computing β_{-i} for all observations is $O(N^3)$. The cost of inverting Σ continues to be $O(N^k)$ and so the overall cost is dominated by $O(N^3)$, or $O(SN^3)$ if Σ depends on the model parameters θ in a non-trivial manner. Unlike the normal case, we cannot reduce the computational costs below $O(SN^3)$ even if the model is parameterized directly in terms of $\Sigma^{-1} = \Sigma^{-1}(\theta)$ and so avoids matrix inversion altogether. However, this is still substantially more efficient than the brute force approach, which requires $O(SN^{k+1}) > O(SN^3)$ operations.

2.3 Example: Lagged SAR models

It often requires additional work to take a given multivariate normal or student-t model and express it in the form required to apply the equations for the predictive mean and standard deviation. Consider, for example, the lagged simultaneous autoregressive (SAR) model (Cressie, 1992; Haining and Haining, 2003), a spatial model with many applications in both the social sciences (e.g., economics) and natural sciences (e.g., ecology). The model is given by

$$y = \rho W y + \eta + \epsilon, \tag{18}$$

or equivalently

$$(I - \rho W) y = \eta + \epsilon, \tag{19}$$

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 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 $\epsilon \sim \mathrm{N}(0, \sigma^2 I)$, with residual variance σ^2 and identity matrix I of dimension N, it follows that

$$(I - \rho W)y \sim \mathrm{N}(\eta, \sigma^2 I), \tag{20}$$

but this standard way of expressing the model is not compatible with the requirements of Proposition 1. To make the lagged SAR model reconcilable with this proposition we need to rewrite it as follows (conditional on ρ, η, and σ^2):

$$y \sim \mathrm{N}\left((I - \rho W)^{-1}\eta, \sigma^2 (I - \rho W)^{-1}(I - \rho W)^{-\mathrm{T}}\right), \tag{21}$$

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

$$y \sim \mathrm{N}\left(\widetilde{W}^{-1}\eta, \sigma^2 (\widetilde{W}^{\mathrm{T}}\widetilde{W})^{-1}\right). \tag{22}$$

Written in this way, the lagged SAR model has the required form (3). Accordingly, we can compute the leave-one-out predictive densities with Equations (7) and (8), replacing μ with $\widetilde{W}^{-1}\eta$ and taking the covariance matrix Σ to be $\sigma^2(\widetilde{W}^T\widetilde{W})^{-1}$. This implies $\Sigma^{-1} = \sigma^{-2}\widetilde{W}\widetilde{W}^T$ and so that 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 Σ^{-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. Since $\widetilde{W}$ depends on the parameter ρ in a non-trivial manner, $\widetilde{W}^{-1}$ needs to be computed for each posterior draw, which implies an overall computational cost of less than $O(SN^3)$.

If the errors are student-t distributed, we can apply analogous transformations as above to arrive at the student-t distribution for the responses

$$y \sim \mathrm{t}\left(\nu, \widetilde{W}^{-1}\eta, \sigma^2 (\widetilde{W}^{\mathrm{T}}\widetilde{W})^{-1}\right), \tag{23}$$

with computational cost dominated by the computation of the β_i from Equation (17) which is in $O(SN^3)$.

3 Approximate LOO-CV for non-factorizable models

Exact LOO-CV, requires refitting the model N times, each time leaving out one observation. Alternatively, it is possible to obtain an *approximate* LOO-CV using only a single model fit by instead calculating the pointwise log-predictive density (1), without leaving out any observations, and then applying an importance sampling correction (Gelfand et al., 1992), for example, using Pareto smoothed importance sampling (PSIS; Vehtari et al., 2017b).

The conditional pointwise log-likelihood matrix of dimension $S \times N$ is the only required input to the approximate LOO-CV algorithm from Vehtari et al. (2017b) and thus the equations provided in Section 2 allow for approximate LOO-CV for *any* model that can be expressed conditionally in terms of a multivariate or student-t model with invertible covariance/scale matrix Σ; including those where the likelihood does not factorize.

Suppose we have obtained S posterior draws $\theta^{(s)}$ $(s = 1, \ldots, S)$, from the *full* posterior $p(\theta \mid y)$ using MCMC

or another sampling algorithm. Then, the pointwise log-predictive density (1) can be approximated as:

$$p(y_i \mid y_{-i}) \approx \frac{\sum_{s=1}^{S} p(y_i \mid y_{-i}, \theta^{(s)})\, w_i^{(s)}}{\sum_{s=1}^{S} w_i^{(s)}}, \tag{24}$$

where $w_i^{(s)}$ are importance weights to be computed in two steps. First, we obtain the raw importance ratios

$$r_i^{(s)} \propto \frac{1}{p(y_i \mid y_{-i}, \theta^{(s)})}, \tag{25}$$

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

For Bayesian models fit using MCMC, the whole procedure of evaluating and comparing model fit via PSIS-LOO-CV can be summarized 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 as described in Section 2. 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 to obtain a PSIS-LOO-CV estimate. For convenience, the `loo` R package (Vehtari et al., 2018) provides this functionality.
4. Repeat the steps 1 to 3 for each model under consideration and perform model comparison based on the obtained PSIS-LOO-CV estimates.

In the Section 4, we demonstrate this method by performing approximate LOO-CV for lagged SAR models fit to spatially correlated crime data.

3.1 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 those multivariate normal or student-t models that have 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 Σ 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 or student-t models (in particular it does not hold for SAR models). Instead, in order to keep the original prior, we may need to maintain the full covariance matrix Σ 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 model parameters. For a multivariate normal model $\log p(y_i \mid y_{-i})$ can be computed as follows. First, we model y_i

as missing and denote the corresponding *parameter* y_i^{mis}. Then, we define

$$y_{\text{mis}(i)} = (y_1, \ldots, y_{i-1}, y_i^{\text{mis}}, y_{i+1}, \ldots, y_N). \tag{26}$$

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 log predictive densities as in Equations (6) and (14), but replacing y with $y_{\text{mis}(i)}$ in all computations. Finally, the leave-one-out predictive distribution can be estimated as

$$p(y_i \,|\, y_{-i}) \approx \frac{1}{S} \sum_{s=1}^{S} p(y_i \,|\, y_{-i}, \theta_{-i}^{(s)}), \tag{27}$$

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

4 Case Study: Neighborhood Crime in Columbus, Ohio

In order to demonstrate how to carry out the computations implied by these equations, we will fit and evaluate lagged SAR models 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 and 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, 2018) 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.

We fit a normal SAR model first using the weakly-informative default priors of brms. In Figure 1 (a), 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.

In order to evaluate model fit, the next step is to compute the pointwise log-likelihood values needed for approximate LOO-CV and we apply the method laid out in Section 3. Since this is already implemented in brms[2], we can simply use the built-in `loo` method on the fitted model to obtain the desired results. The quality of the approximation can be investigated graphically by plotting the Pareto-k diagnostic 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 1 (b), we see that the fourth observation is problematic. This has two implications. First, it may reduce the accuracy of the LOO-CV approximation. Second it indicates that the fourth observation is highly influential for the posterior and thus questions the robustness of the inference obtained by means of this model. We will address the former issue first and come back to the latter issue afterwards.

The PSIS-LOO-CV approximation of the expected log predictive density for new data reveals $\text{elpd}_{\text{approx}} =$

[2]Source code is available at https://github.com/paul-buerkner/brms/blob/master/R/log_lik.R.

-186.9. To verify the correctness of our approximate estimates, 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 observation. For the lagged SAR model, we cannot just ignore the held-out observation entirely as this will change the prior distribution. 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 the supplemental material for details on the R code).

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 (see Proposition 1) 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 1 (c), 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 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 ($\text{elpd}_{\text{approx}}$ = -186.9) and exact LOO-CV ($\text{elpd}_{\text{exact}}$ = -188.1) 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 1 (d), 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 ($\text{elpd}_{\text{approx},-4}$ = -173.0 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.

With the correctness of the approximating procedure established for non-problematic observations, we can now go ahead and correct for the problematic observation in the approximate LOO-CV estimate. Vehtari et al. (2017b) recommend to perform exact LOO-CV only for the problematic observations and replace their approximate ELPD contributions with their exact counterparts (see also Paananen et al., 2019, for an alternative method). So this time, we do not use exact LOO-CV for validation of the approximation but rather to improve the latter's accuracy when needed. In the present normal SAR model, only the 4th observation was diagnosed as problematic and so we only need to update the ELPD contribution of this observation. The results of the corrected approximate ($\text{elpd}_{\text{approx}}$ = -188.0) and exact LOO-CV ($\text{elpd}_{\text{exact}}$ = -188.1) are now almost equal for the complete data set as well.

Although we were able to correct for the problematic observation in the approximate LOO-CV estimation, the mere existance of such problematic observations raises doubts about the appropriateness of the normal SAR model for the present data. In particular, since normal models are easily influenced by outliers, we should be worried about the robustness of the parameter estimates in light of the extreme data point. Accordingly, it is sensible to fit a student-t SAR model as a potentially better predicting and more robust alternative. We choose an informative Gamma$(4, 0.5)$ prior (with mean 8 and standard deviation 4) on the degrees of freedom parameter ν to ensure rather fat tails of the likelihood and thus greater robustness against outliers. For all other parameters, we continue to use the weakly-informative default priors of brms. In Figure 2 (a), the marginal posterior distributions of the main model parameters are depicted. Comparing the results

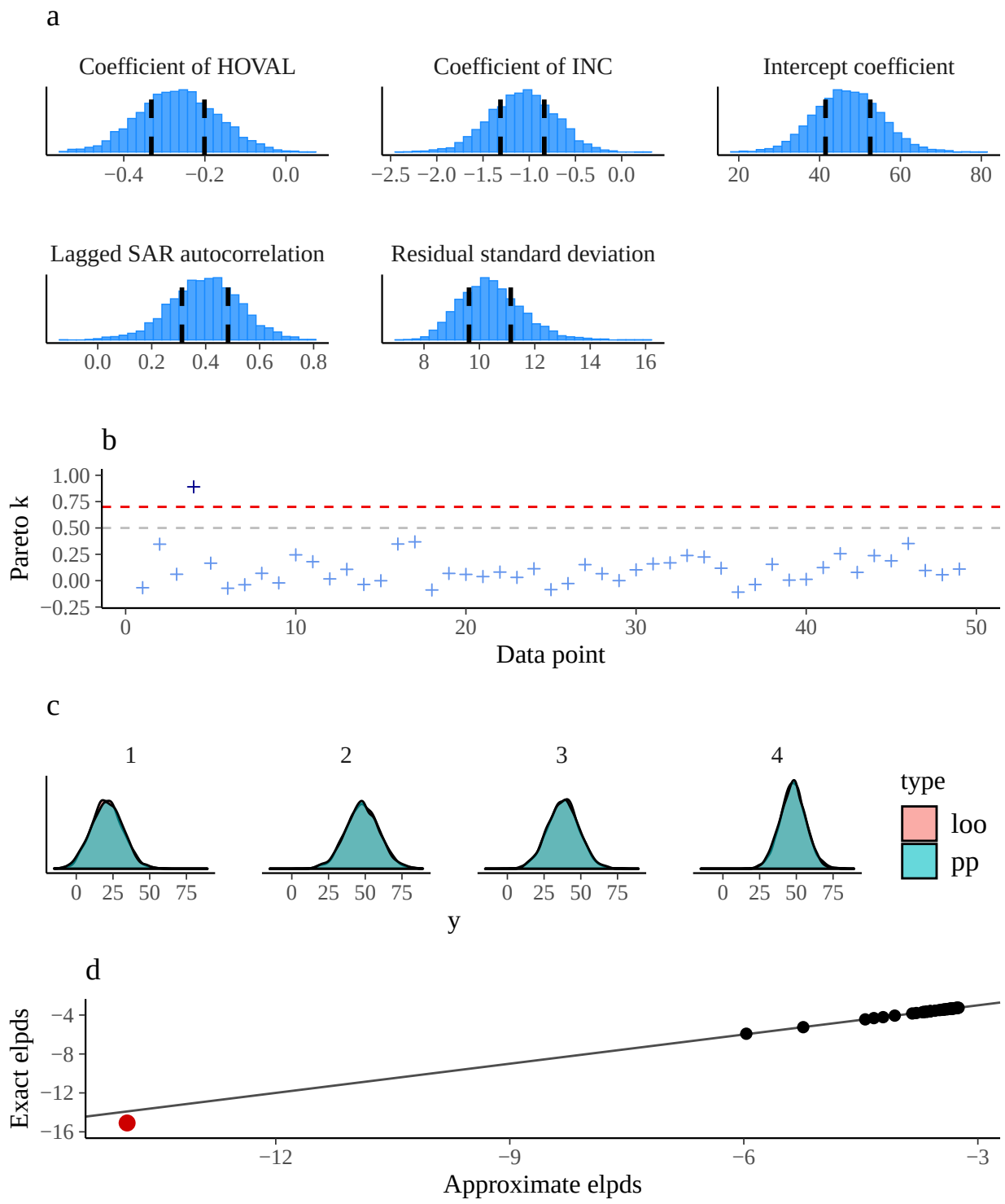

Figure 1: Results of the normal SAR model. 1) Posterior distribution of selected parameters of the lagged SAR model along with posterior median and 50% central interval. 2) PSIS diagnostic plot showing the Pareto-k-estimate of each observation. 3) Implied response values of the first four observations 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'). 4) Comparison of approximate and exact pointwise elpd values. Problematic observations are marked as red dots.

to those shown in Figure 1 (a), we see that the estimates of both the regression parameters and the SAR autocorrelation are quite similar to the estimates obtained from the normal model.

In contrast to the normal case, we see in Figure 2 (b) that the 4th observation is no longer recognized as problematic by the Pareto-k diagnostics. It does exceed 0.5 slightly but does not exceed the more important threshold of 0.7 above which we would stop trusing the PSIS-LOO-CV approximation. Indeed, comparison between the approximate ($\text{elpd}_{\text{approx}}$ = -187.7) and exact LOO-CV ($\text{elpd}_{\text{exact}}$ = -187.9) based on the complete data demonstrates that they are very similar (up to random error due to the MCMC estimation). The results shown in Figure 2 (c) and (d) have the same interpretation as the analagous plots for the normal case and provide further evidence for both the correctness of our (exact and approximate) LOO-CV methods for non-factorizable student-t models and for the quality of the PSIS-LOO-CV approximation for the present student-t SAR model.

Lastly, let us compare the PSIS-LOO-CV estimate of the normal SAR model (after correcting for the problematic observation via refit) to the student-t SAR model. The ELPD difference between the two models is -0.3 (SE = 0.5) in favor of the student-t model, and thus very small and not substantial for any practical purposes. As shown in Figure 3, the pointwise elpd contributions are also highly similar. The student-t model fits slightly but noticably better only for the 4th observation.

5 Conclusion

In this paper we derived how to perform and validate exact and approximate leave-one-out cross-validation (LOO-CV) for non-factorizable multivariate normal and student-t models that are highly relevant to temporal and spatial statistics. The LOO-CV approximations make model fit evaluation and comparison feasible, efficient, and robust for widespread application. 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 or student-t 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.

6 Acknowledgements

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

Appendix

Proof of Proposition 1. In their Lemma 1, Sundararajan and Keerthi (2001) proof for any finite subset z of a zero-mean Gaussian process with covariance matrix Σ that the LOO predictive mean and standard deviation can be computed as

$$\tilde{\mu}_i = z_i - \frac{g_i}{\bar{\sigma}_{ii}}, \tag{28}$$

$$\tilde{\sigma}_i = \frac{1}{\bar{\sigma}_{ii}}, \tag{29}$$

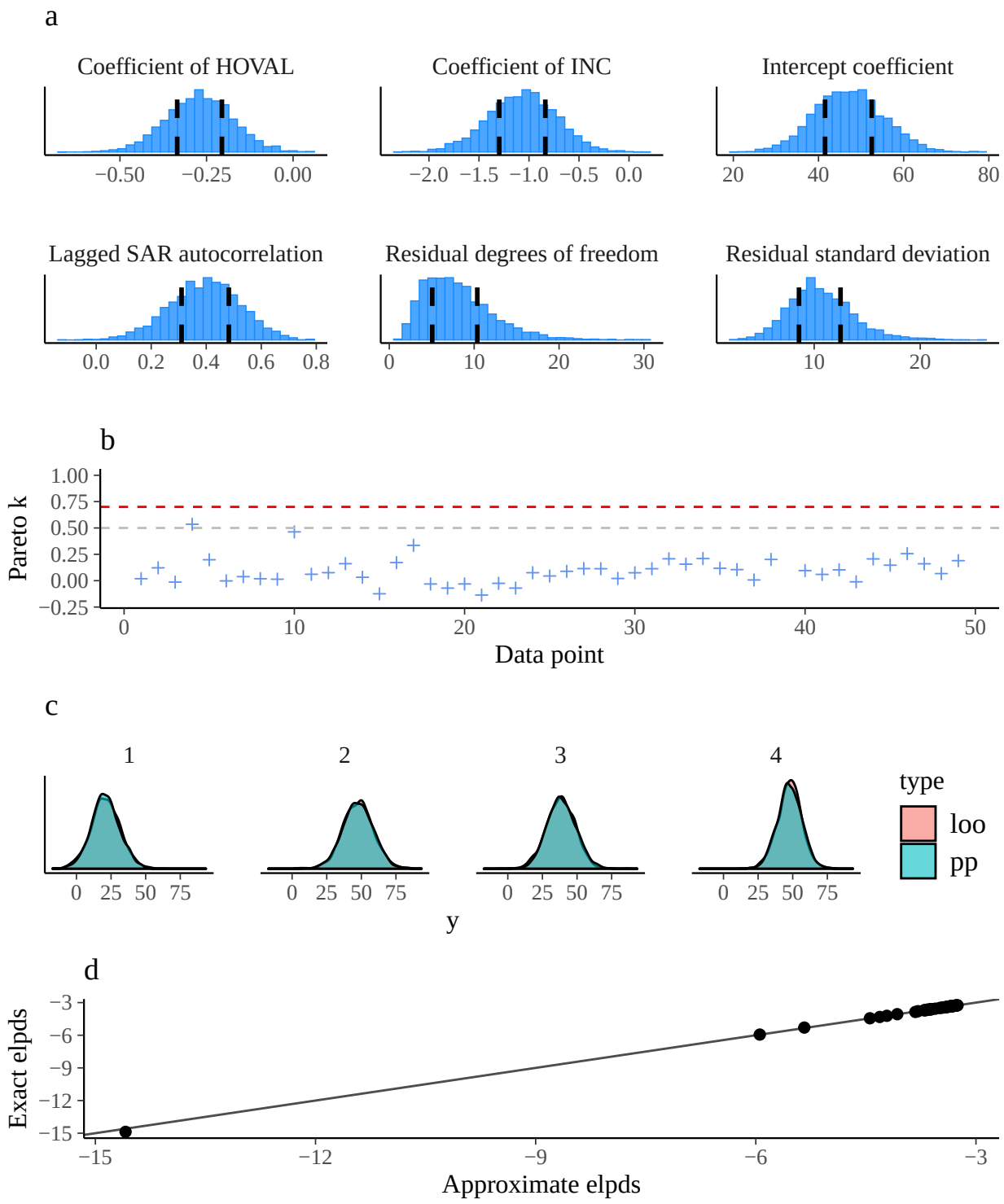

Figure 2: Results of the student-t SAR model. a) Posterior distribution of selected parameters of the lagged SAR model along with posterior median and 50% central interval. b) PSIS diagnostic plot showing the Pareto-k-estimate of each observation. c) Implied response values of the first four observations computed (1) after model fitting (type = 'loo') and (2) as part of the model in the form of posterior-predictive draws for the missing observation (type = 'pp'). d) Comparison of approximate and exact pointwise elpd values. There were no problematic observations for this model.

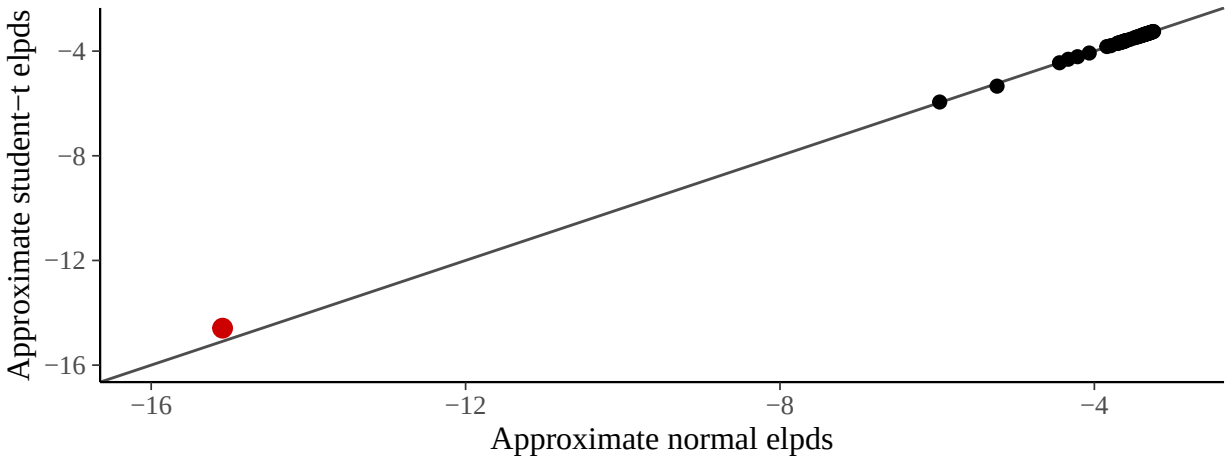

Figure 3: Comparison of approximate pointwise elpd values for the normal SAR model (after refit for the 4th observation) and the student-t SAR model (without refit). Observations with relevant differences are highlighted in red.

where $g_i = \left[\Sigma^{-1} z\right]_i$ and $\bar{\sigma}_{ii} = \left[\Sigma^{-1}\right]_{ii}$. Their proof does not make use of any specific form of Σ and thus directly applies to all zero-mean multivariate normal distributions. If y is multivariate normal with mean μ then $(y - \mu)$ is multivariate normal with mean 0 and unchanged covariance matrix. Thus, we can replace z with $(y - \mu)$ in the above equations. By the same argument we see that, if $(y_i - \mu_i)$ has LOO mean $(y_i - \mu_i) - \frac{g_i}{\bar{\sigma}_{ii}}$, then y has LOO mean $y_i - \frac{g_i}{\bar{\sigma}_{ii}}$ which completes the proof. □

Proof of Proposition 2. Using the parametrization $K := \mathrm{Cov}(y) = \frac{\nu}{\nu-2}\Sigma$ and requiring $\nu > 2$, Shah et al. (2014) proof in their Lemma 3 that, if $y = (y_1, y_2)$ is multivariate student-t of dimension $N = N_1 + N_2$, then y_2 given y_1 is multivaratiate student-t of dimension N_2. Moreover, they provide equations for the parameters of the conditional student-t distribution. When we parameterize for Σ instead of K and allow for $\nu > 0$, we can repeat their proof analogously which yields the following parameters of the conditional student-t distribution of y_2 given y_1:

$$\tilde{\nu}_2 = \nu + N_1, \tag{30}$$

$$\tilde{\mu}_2 = \mu_2 + \Sigma_{2,1}\Sigma_1^{-1}(y_1 - \mu_1), \tag{31}$$

$$\tilde{\sigma}_2 = \frac{\nu + \beta_1}{\nu + N_1}\left(\Sigma_{22} + \Sigma_{2,1}\Sigma_1^{-1}\Sigma_{1,2}\right), \tag{32}$$

with

$$\beta_1 = (y_1 - \mu_1)^{\mathrm{T}}\Sigma_1^{-1}(y_1 - \mu_1). \tag{33}$$

where we use the Subscripts 1 and 2 to refer to the 1st and 2nd subset of y, respectively. Setting $y_1 = y_{-i}$, $y_2 = y_i$ for $i = 1, \ldots, N$ and noting that $N_1 = N_{-i} = N - 1$ completes the proof. □

Proof of Proposition 3. The correctness of Equations (15) and (16) follows directly from Equations (7), and (8). To show (17), we perform a rank-one update of Σ^{-1} as per Theorem 2.1 of Juárez-Ruiz et al. (2016) based on the Sherman-Morrison formula (Bartlett, 1951; Sherman and Morrison, 1950). In general, if we exclude row p and column q from Σ, the inverse $\Sigma_{-p,-q}^{-1}$ of $\Sigma_{-p,-q}$ exists if $\sigma_{pq} \neq 0$ and $\bar{\sigma}_{pq} \neq 0$. The elements

m_{jk} $(j, k = 1, \ldots, N,\ j \neq p,\ k \neq q)$ of $\Sigma^{-1}_{-p,-q}$ are then given by

$$m_{jk} = \bar{\sigma}_{jk} - \frac{\bar{\sigma}_{jp}\bar{\sigma}_{qk}}{\bar{\sigma}_{pq}}. \tag{34}$$

where $\bar{\sigma}_{jk}$ is the (j, k)th element of Σ^{-1}. We now set $p = q = i$ and note that $\sigma_{ii} > 0$ and $\bar{\sigma}_{ii} > 0$ since Σ is a covariance matrix, which leads to

$$m_{jk} = \bar{\sigma}_{jk} - \frac{\bar{\sigma}_{ji}\bar{\sigma}_{ik}}{\bar{\sigma}_{ii}}. \tag{35}$$

for each $i = 1, \ldots, N$. Switching to matrix notation completes the proof. □

References

Ando, T. and Tsay, R. (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.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics*, 22(1):107–111.

Bivand, R. and Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, 63(18):1–36.

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

Bürkner, P.-C. (2018). Advanced bayesian multilevel modeling with the R package brms. *The R Journal*, pages 395–411.

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

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, 4(5):613–617.

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

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

Haining, R. P. and Haining, R. (2003). *Spatial data analysis: theory and practice*. Cambridge university press.

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

Juárez-Ruiz, E., Cortés-Maldonado, R., and Pérez-Rodríguez, F. (2016). Relationship between the inverses of a matrix and a submatrix. *Computación y Sistemas*, 20(2):251–262.

Paananen, T., Piironen, J., Bürkner, P.-C., and Vehtari, A. (2019). Pushing the limits of importance sampling through iterative moment matching. *arXiv preprint arXiv:1906.08850*.

Schwager, S. J., Margolin, B. H., et al. (1982). Detection of multivariate normal outliers. *The annals of statistics*, 10(3):943–954.

Shah, A., Wilson, A., and Ghahramani, Z. (2014). Student-t processes as alternatives to gaussian processes. In *Artificial intelligence and statistics*, pages 877–885.

Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127.

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

Tong, Y. L. (2012). *The multivariate normal distribution*. Springer Science & Business Media.

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

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

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

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

Zellner, A. (1976). Bayesian and non-bayesian analysis of the regression model with multivariate student-t error terms. *Journal of the American Statistical Association*, 71(354):400–405.