Efficient leave-one-out cross-validation for Bayesian non-factorized normal and Student-t models

Paul-Christian Bürkner, Jonah Gabry & Aki Vehtari

ONLINE FIRST

Springer

Springer

Efficient leave-one-out cross-validation for Bayesian non-factorized normal and Student-*t* models

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

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 that the observation model can be factorized into simple terms, but a lot of important models in temporal and spatial statistics do not have this property or are inefficient or unstable when forced into a factorized form. We derive how to efficiently compute and validate both exact and approximate LOO-CV for any Bayesian non-factorized model with a multivariate normal or Student-*t* distribution on the outcome values. We demonstrate the method using lagged simultaneously autoregressive (SAR) models as a case study.

Keywords Cross-validation · Pareto-smoothed importance-sampling · Non-factorized models · Bayesian inference · 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 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. 2017). 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

✉ Paul-Christian Bürkner
  paul.buerkner@gmail.com

1   Department of Computer Science, Aalto University, Espoo, Finland

2   Applied Statistics Center and ISERP, Columbia University, New York, USA

 Springer

account and is less focused on the bulk of the distribution compared to other common measures such as the root mean squared error (RMSE) or mean absolute error (MAE; see for details Vehtari and Ojanen 2012). 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, fast approximate versions of LOO-CV have been developed (Gelfand et al. 1992; Vehtari et al. 2017), most recently using Pareto-smoothed importance-sampling (PSIS; Vehtari et al. 2017, 2019).

A standard assumption of any such fast LOO-CV approach using the ELPD is that the model over all observations has to have a factorized form. That is, the overall observation model should be represented as the product of the pointwise models for each observation. However, many important models do not have this factorization property. Particularly in temporal and spatial statistics it is common to fit multivariate normal or Student-t models that have structured covariance matrices such that the model 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. Some of these models are actually non-factorizable, that is, we do not know of any reformulation that converts the observation model into a factorized form. Other non-factorized models could be factorized in theory but it is sometimes more robust or efficient to marginalize out certain parameters, for instance observation-specific latent variables, and then work with a non-factorized model instead.

Conceptually, a factorized model is not required for cross-validation in general or LOO-CV in particular to make sense. This also implies that neither conditional independence nor conditional exchangability are necessary assumptions. However, when using non-factorized observation models in LOO-CV, computational challenges arise. 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 model 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-factorized 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 have access to the pointwise likelihood $p(y_i \mid y_{-i}, \theta)$

---

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

and integrate over the model parameters θ:

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

Here, $p(\theta \mid y_{-i})$ 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 Sect. 3, we will show how refitting the model to data y_{-i} can be avoided).

If the observation model is formulated directly as the product of the pointwise observation models, we call it a *factorized* model. In this case, the likelihood is also the product of the pointwise likelihood contributions $p(y_i \mid y_{-i}, \theta)$. To better illustrate possible structures of the observation models, we formally divide θ into two parts, observation-specific latent variables $f = (f_1, \ldots, f_N)$ and hyperparameters ψ, so that $p(y_i \mid y_{-i}, \theta) = p(y_i \mid y_{-i}, f_i, \psi)$. Depending on the model, one of the two parts of θ may also be empty. In very simple models, such as linear regression models, latent variables are not explicitly presented and response values are conditionally independent given ψ, so that $p(y_i \mid y_{-i}, f_i, \psi) = p(y_i \mid \psi)$ (see Fig. 1a). The full likelihood can then be written in the familiar form

$$p(y \mid \psi) = \prod_{i=1}^{N} p(y_i \mid \psi), \tag{2}$$

where $y = (y_1, \ldots, y_N)$ denotes the vector of all responses. When the likelihood factorizes this way, the conditional pointwise log-likelihood can be obtained easily by computing $p(y_i \mid \psi)$ for each i with computational cost $O(n)$.

If directional paths between consecutive responses are added, responses are no longer conditionally independent, but the model factorizes to simple terms with Markovian dependency. This is common in time-series models. For example, in an autoregressive model of order 1 (see Fig. 1b), the pointwise likelihoods are given by $p(y_i \mid y_{i-1}, \psi)$. In other time series, models may have observation-specific latent variables f_i and conditionally independent responses so that the pointwise log-likelihoods simplify to $p(y_i \mid y_{-i}, f_i, \psi) = p(y_i \mid f_i)$. In models without directional paths between the latent values f (see Fig. 1c), such as latent Gaussian processes (GPs; e.g., Rasmussen 2003) or spatial conditional autoregressive (CAR) models (e.g., Gelfand and Vounatsou 2003), an explicit joint prior over f is imposed. In models with directional paths between the latent values f (see Fig. 1d), such as hidden Markov models (HMMs; e.g., Rabiner and Juang 1986), the joint prior over f is defined implicitly via the directional dependencies. What is more, estimation can make use of the latent Markov property of such models, for example, using the Kalman filter (e.g., Welch et al. 1995). In all of these cases (i.e., Fig. 1a–d), the factorization property is retained and computational cost for the pointwise log-likelihood contributions remains in $O(n)$.

Yet, there are several reasons why a *non-factorized* observation model (see Fig. 1e) may be necessary or preferred. In non-factorized models, the joint likelihood of the response values $p(y \mid \theta)$ is not factorized into observation-specific components, but
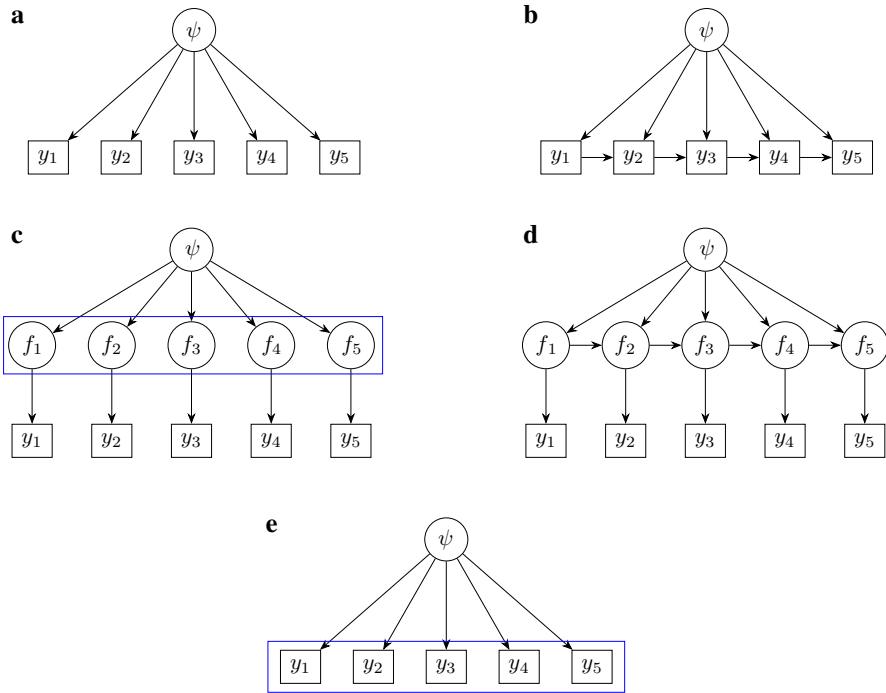
Fig. 1 Directional graphs illustrating common observation model dependency structures schematically. Black rectangles depict manifest variables, that is, observed response values. Black circles depict latent variables and parameters. Blue rectangles indicate a joint prior over the surrounded variables. **a** Conditionally independent responses given hyperparameters (e.g., a linear regression model). **b** Conditionally dependent responses with a Markovian property (e.g., an autoregressive model of order 1). **c** Conditionally independent responses given observation-specific latent variables with a joint prior (e.g., a latent Gaussian process model). **d** Conditionally independent responses given observation-specific latent variables with a Markov property (e.g., a hidden Markov model). **e** Non-factorized model with a joint observation model over all responses (color figure online)

rather given directly as one joint expression. For some models, an analytical factorized formulation is simply not available in which case we speak of a *non-factorizable* model. Even in models whose observation model can be factorized in principle, it may still be preferable to use a non-factorized form. This is true in particular for models with observation-specific latent variables (see Fig. 1c, d), as a non-factorized formulation where the latent variables have been integrated out is often more efficient and numerically stable. For example, a latent GP combined with a Gaussian observation model can be fit more efficiently by marginalizing over f and formulating the GP directly on the responses y (e.g., Rasmussen 2003). Such marginalization has the additional advantage that both exact and approximate leave-one-out predictive estimation become more stable. This is because, in the factorized formulation, leaving out response y_i also implies treating the corresponding latent variable f_i as missing, which is then only identified through the joint prior over f. If this prior is weak, the posterior of f_i is highly influenced by one observation and the leave-one-out predic-

tions of y_i may be unstable both numerically and because of estimation error due to finite MCMC sampling or similar finite approximations.

Whether a non-factorized model is used by necessity or for efficiency and stability, it comes at the cost of having no direct access to the leave-one-out predictive densities (1) and thus to the overall leave-one-out predictive accuracy. In theory, we can express the observation-specific likelihoods in terms of the joint likelihood via

$$p(y_i \mid y_{i-1}, \theta) = \frac{p(y \mid \theta)}{p(y_{-i} \mid \theta)} = \frac{p(y \mid \theta)}{\int p(y \mid \theta) \, dy_i}, \tag{3}$$

but the expression on the right-hand side of (3) may not always have an analytical solution. Computing $\log p(y_i \mid y_{-i}, \theta)$ for non-factorized models is therefore often impossible, or at least inefficient and numerically unstable. However, there is a large class of multivariate normal and Student-t models for which we will provide efficient analytical solutions in this paper.

2.1 Non-factorized normal models

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

$$p(y|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\left(-\frac{1}{2}(y - \mu)^{\mathrm{T}} \Sigma^{-1}(y - \mu)\right) \tag{4}$$

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

$$\tilde{\mu}_i = \mu_i + \sigma_{i,-i} \Sigma_{-i}^{-1}(y_{-i} - \mu_{-i}) \tag{5}$$

and variance

$$\tilde{\sigma}_i = \sigma_{ii} + \sigma_{i,-i} \Sigma_{-i}^{-1} \sigma_{-i,i}. \tag{6}$$

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 diagonal element of Σ. Equations (5) and (6) can be used to compute the pointwise log-likelihood values as

$$\log p(y_i \mid y_{-i}, \theta) = -\frac{1}{2} \log(2\pi \tilde{\sigma}_i) - \frac{1}{2} \frac{(y_i - \tilde{\mu}_i)^2}{\tilde{\sigma}_i}. \tag{7}$$

Evaluating Eq. (7) 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*

$$\tilde{\mu}_i = y_i - \frac{g_i}{\bar{\sigma}_{ii}}, \tag{8}$$

$$\tilde{\sigma}_i = \frac{1}{\bar{\sigma}_{ii}}, \tag{9}$$

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 (5) and (6), where Σ_{-i} has to be inverted separately for each i, Eqs. (8) and (9) 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-factorized Student-t models

Several generalizations of the multivariate normal distribution have been suggested, perhaps most notably the multivariate Student-*t* distribution (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 much fatter than those of the normal distribution. 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 thickness of the tails is flexibly adjusted based on information from the observed response values and the prior. The (multivariate) Student-*t* distribution has been studied in various places (e.g., Zellner 1976; O'Hagan 1979; Fernández and Steel 1999; Zhang and

Yeung [2010]; Piché et al. [2012]; Shah et al. [2014]). For example, Student-t processes which are based on the multivariate Student-t distribution constitute a generalization of Gaussian processes while retaining most of the latter's favorable properties (Shah et al. [2014]).[2]

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

$$p(y|\nu, \mu, \Sigma) = \frac{\Gamma((\nu + N)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{(\nu\pi)^N |\Sigma|}} \left(1 + \frac{1}{\nu}(y - \mu)^T \Sigma^{-1}(y - \mu)\right)^{-(\nu+N)/2} \quad (10)$$

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*

$$\tilde{\nu}_i = \nu + N - 1, \quad (11)$$

$$\tilde{\mu}_i = \mu_i + \sigma_{i,-i} \Sigma_{-i}^{-1} (y_{-i} - \mu_{-i}), \quad (12)$$

$$\tilde{\sigma}_i = \frac{\nu + \beta_{-i}}{\nu + N - 1} \left(\sigma_{ii} + \sigma_{i,-i} \Sigma_{-i}^{-1} \sigma_{-i,i}\right), \quad (13)$$

where

$$\beta_{-i} = (y_{-i} - \mu_{-i})^T \Sigma_{-i}^{-1} (y_{-i} - \mu_{-i}). \quad (14)$$

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

$$\log p(y_i \mid y_{-i}, \theta) = \log(\Gamma((\tilde{\nu}_i + 1)/2)) - \log(\Gamma(\tilde{\nu}_i/2)) - \frac{1}{2} \log(\tilde{\nu}_i \pi \tilde{\sigma}_i)$$

$$- \frac{\tilde{\nu}_i + 1}{2} \log\left(1 + \frac{1}{\tilde{\nu}_i} \frac{(y_i - \tilde{\mu}_i)^2}{\tilde{\sigma}_i}\right). \quad (15)$$

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.

---

[2] A Student-t process is not to be confused with a factorized univariate Student-t likelihood in combination with a Gaussian process on the corresponding latent variables as these models have different properties.

Proposition 3 *If y is multivariate Student-t with degrees of freedom v, 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*

$$\tilde{\mu}_i = y_i - \frac{g_i}{\bar{\sigma}_{ii}}, \tag{16}$$

$$\tilde{\sigma}_i = \frac{v + \beta_{-i}}{v + N - 1} \frac{1}{\bar{\sigma}_{ii}}, \tag{17}$$

with

$$\beta_{-i} = (y_{-i} - \mu_{-i})^{\mathrm{T}} \left(\Sigma^{-1} - \frac{\bar{\sigma}_{-i,i} \bar{\sigma}_{-i,i}^{\mathrm{T}}}{\bar{\sigma}_{ii}} \right) (y_{-i} - \mu_{-i}), \tag{18}$$

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; LeSage and Pace 2009), 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

$$y = \rho W y + \eta + \epsilon, \tag{19}$$

or equivalently

$$(I - \rho W)y = \eta + \epsilon, \tag{20}$$

where ρ is a scalar spatial correlation parameter and W is a user-defined matrix of weights. The matrix W has entries $w_{ii} = 0$ 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. See LeSage and Pace (2009), Sect. 2.3, for a more detailed introduction to SAR models. A general discussion about predictions of SAR models from a frequentist perspective can be found in Goulard et al. (2017).

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

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

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):

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

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

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

Written in this way, the lagged SAR model has the required form (4). Accordingly, we can compute the leave-one-out predictive densities with Eqs. (8) and (9), replacing μ with $\widetilde{W}^{-1}\eta$ and taking the covariance matrix Σ to be $\sigma^2(\widetilde{W}^{\mathrm{T}}\widetilde{W})^{-1}$. 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 usually 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 residuals are Student-t distributed, we can apply analogous transformations as above to arrive at the Student-t distribution for the responses

$$y \sim \mathrm{t}\left(\nu, \; \widetilde{W}^{-1}\eta, \; \sigma^2(\widetilde{W}^{\mathrm{T}}\widetilde{W})^{-1}\right), \tag{24}$$

with computational cost dominated by the computation of the β_i from Eq. (18) which is in $O(SN^3)$.

Studying leave-one-out predictive densities in SAR models is related to considering impact measures, that is, measures to quantify how changes in the predicting variables of a given observation i affect the responses in other observations $j \neq i$ as well as the obtained parameter estimates (see LeSage and Pace 2009, Section 2.7). A detailed discussion of this topic it out of scope of the present paper.

3 Approximate LOO-CV for non-factorized 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. 2017).

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. (2017) and thus the equations provided in Sect. 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:

$$p(y_i \mid y_{-i}) \approx \frac{\sum_{s=1}^{S} p(y_i \mid y_{-i}, \theta^{(s)}) \, w_i^{(s)}}{\sum_{s=1}^{S} w_i^{(s)}}, \qquad (25)$$

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

$$r_i^{(s)} \propto \frac{1}{p(y_i \mid y_{-i}, \theta^{(s)})}, \qquad (26)$$

and then stabilize them using Pareto-smoothed importance-sampling to obtain the weights $w_i^{(s)}$ (Vehtari et al. 2017, 2019). The resulting approximation is referred to as PSIS-LOO-CV (Vehtari et al. 2017).

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 Sect. 2. The results can be stored in an $S \times N$ matrix.
3. Run the PSIS algorithm from Vehtari et al. (2017) 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 1–3 for each model under consideration and perform model comparison based on the obtained PSIS-LOO-CV estimates.

In the Sect. 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. 2019) indicates an unstable approximation, we need to consider how we might do *exact* LOO-CV for a non-factorized 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

$$y_{\text{mis}(i)} = (y_1, \ldots, y_{i-1}, y_i^{\text{mis}}, y_{i+1}, \ldots, y_N). \tag{27}$$

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 Eqs. (7) and (15), but replacing y with $y_{\text{mis}(i)}$ in all computations. Finally, the leave-one-out predictive distribution can be estimated as

$$p(y_i \mid y_{-i}) \approx \frac{1}{S} \sum_{s=1}^{S} p(y_i \mid y_{-i}, \theta_{-i}^{(s)}), \tag{28}$$

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

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 Fig. 2a, 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 Sect. 3. Since this is already implemented in brms,[3] 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. 2017, 2019). In Fig. 2b, 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}} = $ -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 Fig. 2c, 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

---

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

the approximate against the exact pointwise ELPD values. In Fig. 2d, 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. (2017) recommend to perform exact LOO-CV only for the problematic observations and replace their approximate ELPD contributions with their exact counterparts (see also for an alternative method Paananen et al. 2019). 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 existence of such problematic observations raises doubts about the appropriateness of the normal SAR model for the present data. Accordingly, it is sensible to fit a Student-t SAR model as a potentially better predicting alternative due to its fatter tails. We choose an informative $\text{Gamma}(4, 0.5)$ prior (with mean 8 and standard deviation 4) on the degrees of freedom parameter ν to ensure rather fat tails of the likelihood a-priori. For all other parameters, we continue to use the weakly-informative default priors of brms. In Fig. 3a, the marginal posterior distributions of the main model parameters are depicted. Comparing the results to those shown in Fig. 2a, 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 Fig. 3b 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 trusting 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 Fig. 3c, d have the same interpretation as the analogous plots for the normal case and provide further evidence for both the correctness of our (exact and approximate) LOO-CV methods for non-factorized 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

a

b

c

d
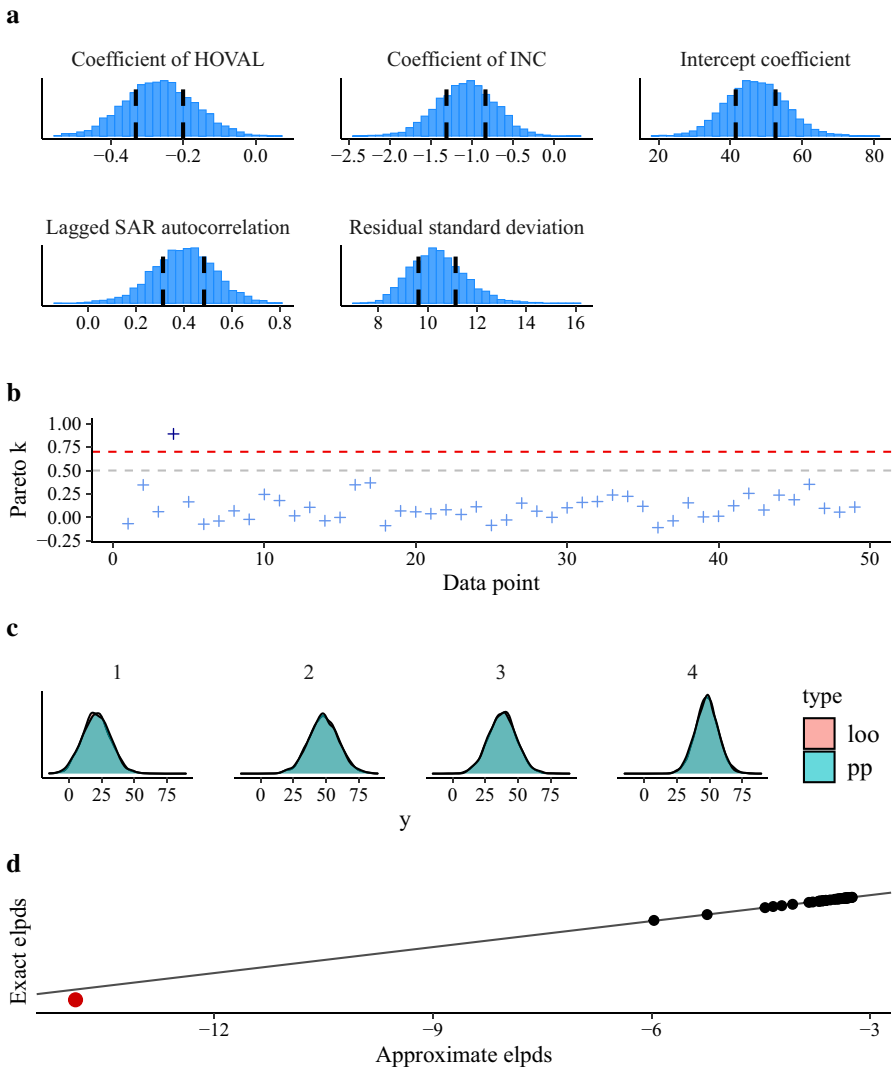
Fig. 2 Results of the normal SAR model. (1) Posterior distribution of selected parameters of the lagged SAR model along with posterior median and 50% central interval. (2) PSIS diagnostic plot showing the Pareto-k-estimate of each observation. (3) Implied response values of the first four observations 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'). As both distributions are almost identical, the 'loo' distribution is hidden behind the 'pp' distribution. (4) Comparison of approximate and exact pointwise elpd values. Problematic observations are marked as red dots (color figure online)

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 Fig. 4, the pointwise elpd contributions are also highly similar. The Student-t model fits slightly but noticeably better only for the 4th observation. Other methods clearly
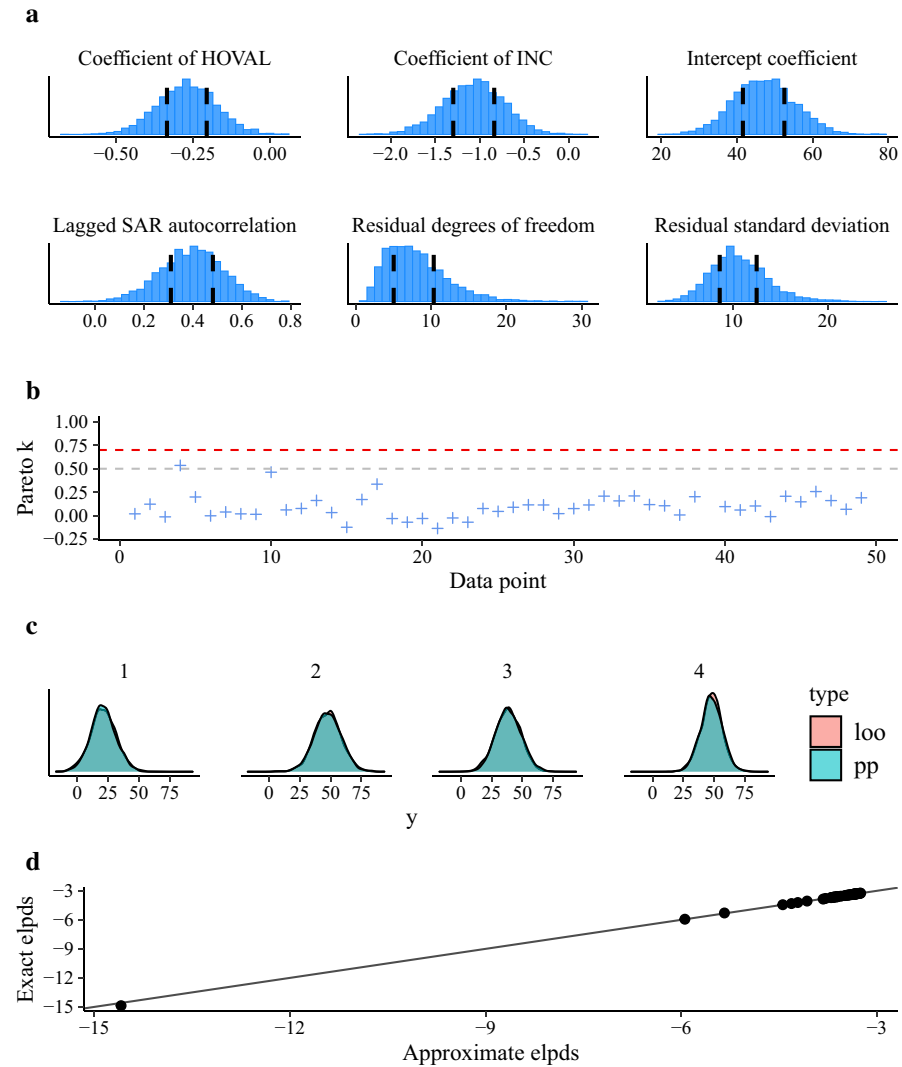
Fig. 3 Results of the Student-*t* SAR model. **a** Posterior distribution of selected parameters of the lagged SAR model along with posterior median and 50% central interval. **b** PSIS diagnostic plot showing the Pareto-*k*-estimate of each observation. **c** Implied response values of the first four 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'). As both distributions are almost identical, the 'loo' distribution is hidden behind the 'pp' distribution. **d** Comparison of approximate and exact pointwise elpd values. There were no problematic observations for this model

identify the 4th observation as problematic as well (e.g., Halleck Vega and Elhorst 2015). However, what exactly causes this outlier remains unclear so far as nobody has been able to verify or spatially register the data set despite many attempts.
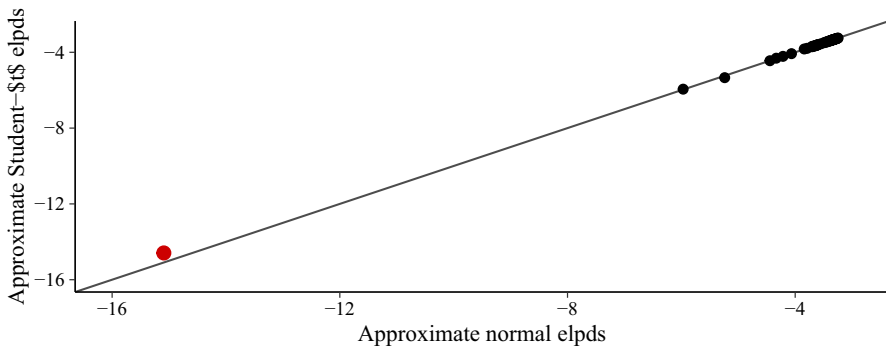
Fig. 4 Comparison of approximate pointwise elpd values for the normal SAR model (after refit for the 4th observation) and the Student-*t* SAR model (without refit). Observations with relevant differences are highlighted in red (color figure online)

5 Conclusion

In this paper we derived how to perform and validate exact and approximate leave-one-out cross-validation (LOO-CV) for non-factorized multivariate normal and Student-*t* models and we demonstrated the practical applicability of our method in a case study using spatial autoregressive models. The proposed LOO-CV approximation makes efficient and robust model fit evaluation and comparison feasible for many models that are widely used in temporal and spatial statistics. Importantly, our method does not only apply to non-factorizable models for which a factorized likelihood is unavailable. It is also useful when factorization is possible in principle but could result in a unstable LOO predictive density, either for numerical reasons or due to the use of finite estimation procedures like Markov-Chain Monte Carlo. In such cases, marginalizing over observation-specific latent variables and using a non-factorized formulation can lead to more robust estimates.

Appendix

Proof of Proposition 1 In their Lemma 1, Sundararajan and Keerthi (2001) prove 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

$$\tilde{\mu}_i = z_i - \frac{g_i}{\bar{\sigma}_{ii}}, \tag{29}$$

$$\tilde{\sigma}_i = \frac{1}{\bar{\sigma}_{ii}}, \tag{30}$$

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. $\qquad\square$

Proof of Proposition 2 Using the parameterization $K := \mathrm{Cov}(y) = \frac{\nu}{\nu-2}\Sigma$ 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 multivariate 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:

$$\tilde{\nu}_2 = \nu + N_1, \tag{31}$$

$$\tilde{\mu}_2 = \mu_2 + \Sigma_{2,1}\Sigma_1^{-1}(y_1 - \mu_1), \tag{32}$$

$$\tilde{\sigma}_2 = \frac{\nu + \beta_1}{\nu + N_1}\left(\Sigma_{22} + \Sigma_{2,1}\Sigma_1^{-1}\Sigma_{1,2}\right), \tag{33}$$

with

$$\beta_1 = (y_1 - \mu_1)^{\mathrm{T}}\Sigma_1^{-1}(y_1 - \mu_1). \tag{34}$$

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. $\qquad\square$

Proof of Proposition 3 The correctness of Eqs. (16) and (17) follows directly from Eqs. (8), and (9). To show (18), 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 $\Sigma_{-p,-q}^{-1}$ are then given by

$$m_{jk} = \bar{\sigma}_{jk} - \frac{\bar{\sigma}_{jp}\bar{\sigma}_{qk}}{\bar{\sigma}_{pq}}. \tag{35}$$

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

$$m_{jk} = \bar{\sigma}_{jk} - \frac{\bar{\sigma}_{ji}\bar{\sigma}_{ik}}{\bar{\sigma}_{ii}}. \tag{36}$$

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

References

Ando T, Tsay R (2010) Predictive likelihood for Bayesian model selection and averaging. Int J Forecast 26(4):744–763

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

Bartlett MS (1951) An inverse matrix adjustment arising in discriminant analysis. Ann Math Stat 22(1):107–111

Bivand R, Piras G (2015) Comparing implementations of estimation methods for spatial econometrics. J Stat Softw 63(18):1–36

Bürkner P-C (2017) brms: an R package for Bayesian multilevel models using Stan. J Stat Softw 80(1):1–28

Bürkner P-C (2018) Advanced Bayesian multilevel modeling with the R package brms. R J 10(1):395–411

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P, Ridell A (2017) Stan: a probabilistic programming language. J Stat Softw

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

Fernández C, Steel MF (1999) Multivariate student-t regression models: pitfalls and inference. Biometrika 86(1):153–167

Geisser S, Eddy WF (1979) A predictive approach to model selection. J Am Stat Assoc 74(365):153–160

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

Gelfand AE, Vounatsou P (2003) Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics 4(1):11–15

Goulard M, Laurent T, Thomas-Agnan C (2017) About predictions in spatial autoregressive models: optimal and almost optimal strategies. Spat Econ Anal 12(2–3):304–325

Haining RP, Haining R (2003) Spatial data analysis: theory and practice. Cambridge University Press, Cambridge

Halleck Vega S, Elhorst JP (2015) The slx model. J Region Sci 55(3):339–363

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. Stat Sci 14(4):382–417

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

LeSage JP, Pace RK (2009) Introduction to spatial econometrics. CRC Press, Boca Raton

O'Hagan A (1979) On outlier rejection phenomena in Bayes inference. J R Stat Soc Ser B (Methodol) 41(3):358–367

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

Piché R, Särkkä S, Hartikainen J (2012) Recursive outlier-robust filtering and smoothing for nonlinear systems using the multivariate student-t distribution. In: 2012 IEEE international workshop on machine learning for signal processing. IEEE, pp 1–6

Rabiner L, Juang B (1986) An introduction to hidden Markov models. IEEE ASSP Mag 3(1):4–16

Rasmussen CE (2003) Gaussian processes in machine learning. In: Summer school on machine learning. Springer, pp 63–71

Shah A, Wilson A, Ghahramani Z (2014) Student-t processes as alternatives to Gaussian processes. Artif Intell Stat 877–885

Sherman J, Morrison WJ (1950) Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. Ann Math Stat 21(1):124–127

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

Tong YL (2012) The multivariate normal distribution. Springer, Berlin

Vehtari A, Gabry J, Yao Y, Gelman A (2018) loo: efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2

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

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

Vehtari A, Simpson D, Gelman A, Yao Y, Gabry J (2019) Pareto smoothed importance sampling. arXiv preprint

Welch G, Bishop G et al (1995) An introduction to the Kalman filter. Technical Report TR 95-041, University of North Carolina, Department of Computer Science

Zellner A (1976) Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. J Am Stat Assoc 71(354):400–405

Zhang Y, Yeung D-Y (2010) Multi-task learning using generalized t process. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp 964–971