SORT 49 (1) January-June 2025, 121-146 DOI: 10.57645/20.8080.02.25

Leave-group-out cross-validation for latent gaussian models

Zhedong Liu[1], Janet Van Niekerk[2] and Håvard Rue[3]

Abstract

Evaluating the predictive performance of a statistical model is commonly done using cross-validation. Among the various methods, leave-one-out cross-validation (LOOCV) is frequently used. Originally designed for exchangeable observations, LOOCV has since been extended to other cases such as hierarchical models. However, it focuses primarily on short-range prediction and may not fully capture long-range prediction scenarios. For structured hierarchical models, particularly those involving multiple random effects, the concepts of short- and long-range predictions become less clear, which can complicate the interpretation of LOOCV results. In this paper, we propose a complementary cross-validation framework specifically tailored for longer-range prediction in latent Gaussian models, including those with structured random effects. Our approach differs from LOOCV by excluding a carefully constructed set from the training set, which better emulates longer-range prediction conditions. Furthermore, we achieve computational efficiency by adjusting the full joint posterior for this modified cross-validation, thus eliminating the need for model refitting. This method is implemented in the R-INLA package (www.r-inla.org) and can be adapted to a variety of inferential frameworks.

MSC: *62-04 62C10 62F15 62J12.*

Keywords: *Bayesian Cross-Validation, Latent Gaussian Models, R-INLA.*

[1] Statistics Program, CEMSE. King Abdullah University of Science and Technology. Kingdom of Saudi Arabia, Thuwal 23955-6900. zhedongliu1@gmail.com

[2] Statistics Program, CEMSE. King Abdullah University of Science and Technology. Kingdom of Saudi Arabia, Thuwal 23955-6900. Department of Statistics, University of Pretoria, South Africa. janet.vanniekerk@kaust.edu.sa

[3] Statistics Program, CEMSE. King Abdullah University of Science and Technology. Kingdom of Saudi Arabia, Thuwal 23955-6900. haavard.rue@kaust.edu.sa

Received: August 2024.
Accepted: March 2025.

1. Introduction

1.1. Rationale and Background

Leave-one-out cross-validation (LOOCV) (Stone, 1974) stands as a popular method to evaluate a statistical model's predictive performance, perform model selections, or estimating some critical parameters in the model. The core concept of LOOCV is elegantly straightforward. Suppose we have data, $\boldsymbol{y} = \{y_i\}$, for $i = 1, \ldots, n$, presumed to be independent and identically distributed (I.I.D.) samples from the true distribution $\pi_T(y)$. Our objective is to determine how well a fitted model can predict a new observation, $\tilde{y}$, sampled from this true distribution. In the Bayesian context, we use the posterior predictive distribution $\pi(y|\boldsymbol{y})$ to predict $\tilde{y}$ sampled from $\pi_T(y)$ as proposed by Geisser and Eddy (1979). Using the logarithmic score (Gneiting and Raftery, 2007), we can compute $E_{\tilde{y}}[\log \pi(\tilde{y}|\boldsymbol{y})]$ as a metric for prediction ability.

Owing to the lack of $\pi_T(y)$, directly computing the expectation becomes infeasible. Nonetheless, since y_i is an exchangeable sample from $\pi_T(y)$, we can estimate this expectation by evaluating

$$u_{\text{LOOCV}} = \frac{1}{n} \sum_{i=1}^{n} \log \pi(y_i|\boldsymbol{y}_{-i}),$$

where y_i is the testing point and $\boldsymbol{y}_{-i}$ is the training set, and $\boldsymbol{y}_{-i}$ are all data except the ith observation.

The informal interpretation of LOOCV is that it mimics "using $\boldsymbol{y}$ to predict $\tilde{y}$" by "using $\boldsymbol{y}_{-i}$ to predict y_i". This intuitive interpretation is then used to justify, often implicitly, the use of LOOCV as a "default" way to evaluate predictive performance.

However, issues can arise in more complex statistical models where the dependency in the model results in the data not being exchangeable (see Vehtari and Ojanen (2012) for a complete discussion of cross-validation (CV) for several types of exchangeability); we describe these kinds of models as "dependent" cases for the purpose of this paper. An intuitive dependent case is a time series. Burman, Chow and Nolan (1994) proposed a block CV method for dependent data from a stationary process, acknowledging the need for a different approach to CV than LOOCV. McQuarrie and Tsai (1998) propose modified cross-validation (MCV) where dependent data chunks are removed together with the relevant point to account for the dependence in a time series (and other dependent data generating models). Bergmeir and Benítez (2012) investigated the properties of blocked CV and other approaches for robust time series model evaluations (see also Bergmeir, Hyndman and Koo (2018) for a study on k-fold CV), while Bürkner, Gabry and Vehtari (2020) proposed a leave-future-out CV strategy. Cerqueira, Torgo and Mozetič (2020) investigated CV and holdout approaches for time series models and concluded that the out-of-sample holdout procedure is more accurate for non-stationary processes than LOOCV.

Besides time series, spatial dependence models come to mind for which Valavi et al. (2018) proposed a buffering strategy by leaving out specific spatial points or areas and

spatial and environmental blocking. Spatial blocking forms clusters of data points according to spatial effects, and environmental blocking forms clusters using K-means (Hartigan and Wong, 1979) on the covariates. Other examples of dependent cases are longitudinal data for multiple subjects in a study (Saeb et al., 2017) and hierarchical models (see Gelman et al. (1995) and Vehtari and Ojanen (2012, Section 5.1.4). Racine (2000) proposed an hv-block CV approach for dependent data while Merkle, Furr and Rabe-Hesketh (2019) considers a multilevel model and shows that marginal WAIC is akin to LOOCV. Roberts et al. (2017) advocate a block cross-validation, partitioning ecological data based on inherent patterns, when the prediction task is not simply short-range prediction. Rabinowicz and Rosset (2022) offers a modification to LOOCV, ensuring an unbiased measure of predictive performance given the correlation between new and observed data, where the unbiasedness is in the sense of randomized both observed and new data. We should note that an assumed prediction task determines the correlation between new and observed data.

In dependent cases, LOOCV can provide a restricted assessment of the models' predictive performance since LOOCV cannot evaluate longer-range prediction. Even in terms of short-range prediction, it is not clear what is short- or longer-range in dependent models that are not purely temporal or spatial models where the range has a physical interpretation. We use the concepts of short-range and longer-range predictions, acknowledging that these concepts can have overlapping meanings.

We thus propose a framework that emulates longer-range prediction scenarios, for hierarchical models, by constructing non-random leave-out sets based on model-based correlations. This can be viewed as a complementary approach to LOOCV for evaluating predictive performance, providing additional informative insights of the predictive ability for dependent cases.

1.2. The prediction task

The critical observation is that the meaning of "prediction" is not clearly defined when we are far away from exchangeability, so that $\boldsymbol{y}$ are *non-exchangeable* samples of $\pi_T(y)$. $\pi_T(\tilde{y}|\boldsymbol{y})$ lacks a unique definition in dependent cases as without a clear *prediction task*, i.e., how we imagine a new data point, $\tilde{y}$, is generated given observed data $\boldsymbol{y}$. This ambiguity extends to the act of "using $\boldsymbol{y}$ to predict $\tilde{y}$" as it is uncertain what our target, $\tilde{y}$, represents. To illustrate these concepts, let us discuss some more concrete examples.

Time-series model

Assume data $\boldsymbol{y} = \{y_1, y_2, \ldots, y_T\}$ is a time-series, observed sequentially at time $1, 2, \ldots, T$. The inherent prediction task is to predict future values, given the temporal nature of the data. We can predict a new observation at $k \geq 1$ steps into the future by $\pi(y_{T+k}|y_1, \ldots, y_T)$.

In this example, the LOOCV will be computed from

$$\pi(y_t|y_1, \ldots, y_{t-1}, y_{t+1}, \ldots, y_T), \qquad t = 1, \ldots, T,$$

which is often referred to as interpolation or imputation of missing values, rather than a prediction. However, the predictive performance of time series models is often assessed through leave-future-out cross-validation (LFOCV) (Bürkner et al., 2020):

$$\sum_{T'=T_0}^{T-k} \log \pi(y_{T'+k}|y_1, \ldots, y_{T'}),$$

where T' starts from time $T_0 > 1$ as we need some data to estimate the model.

The message from this example is that LOOCV, when applied to such models, is essentially evaluating short-range prediction performance rather than longer-range predictive performance.

We acknowledge two issues. First, the distinction between short and longer-range prediction is not always clear-cut, leading to overlapping concepts. For example, a one-step-ahead forecast leans more towards short range than a two-step-ahead prediction. In contrast, a one-step-ahead forecast leans less towards short-range than a missing value imputation. However, this does not deter our discussion. Secondly, while an ideal model succeeds in all prediction tasks, real-world scenarios require us to settle for the definition of the "best fit". Consequently, our choice of evaluation should align with our specific objectives.

Multilevel model

Figure 1 illustrates an example of a multilevel model. Consider observations of student grades or performance. This data exhibits a hierarchical structure: students belong to classes, classes reside within schools, and schools are nested within regions. This hierarchical arrangement is significant because it introduces correlated random effects attributed to the class, school, and region levels, substantially deviating from the exchangeable case.

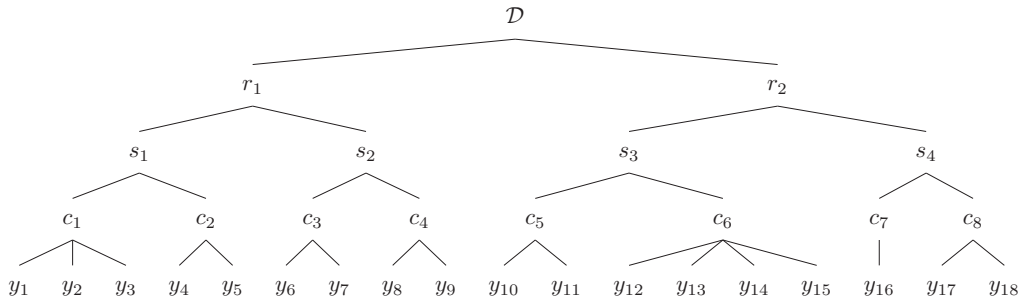

Figure 1. *A nested multilevel model.*

Given such a model, the prediction task becomes ambiguous. Are we aiming to predict the performance of an unobserved student from an observed class? Or are we trying to predict the performance of an unobserved student in an unobserved class, school, or

even region? This difficulty mirrors the challenges in defining asymptotic regimes for these models. As students, classes, schools, and regions can grow indefinitely in various ways, it is unclear whether one of such choices is the most reasonable.

To evaluate predictive performance within this context, users must first explicitly define their prediction task and then evaluate the model according to this definition. It should be noted that applying LOOCV would evaluate the prediction of individual students within observed classes. In our view, this mimics more short-range prediction rather than longer-range prediction, and another framework is needed to quantify the predictive ability for a new student in a new class in a new school in a new region, for example. Our proposal provides some insight into this kind of prediction task.

1.3. LGOCV: Complementing LOOCV for dependent cases

Our discussions illuminate an important insight: when dealing with models that lead to non-exchangeable data, the prediction task implicitly defined through LOOCV may be less appropriate, as it leans more towards assessing imputing qualities and short range predictions than predictive performance for longer range as is usually implied by "out-of-sample" prediction. This prompts the question: What is a suitable approach moving forward?

One observation is the absence of a "one size fits all" solution. Each model may possess a natural prediction task-or several-based on its intended application. Thus, for a specific assessment of predictive performance, we need to define these prediction tasks explicitly. One can then evaluate distinct predictive performance metrics using our proposed leave-group-out cross-validation (LGOCV):

$$u_{\text{LGOCV}} = \frac{1}{n}\sum_{i=1}^{n} \log(\pi(y_i|\boldsymbol{y}_{-I_i})). \tag{1}$$

Here, the *group* (denoted by I_i) is an index set including i. This configuration facilitates that the pair $(y_i, \boldsymbol{y}_{-I_i})$ mimics a specified prediction task, with $\boldsymbol{y}_{-I_i}$ being the data subset excluding the data indexed by I_i. In a multilevel model, as depicted in Figure 1, predicting a student's grade from an unseen class necessitates that I_i includes i and all observations from student i's class. However, more complex models, such as models containing both time series and hierarchical elements, pose challenges when defining a natural prediction task. Therefore, even in complex cases, LOOCV is often applied for its simplicity-even if it leans more towards imputation or short-range prediction.

Developing a framework that evaluates a model's longer-range prediction like the proposed LGOCV, necessitates the construction of the leave-out group I_i for each datapoint y_i. Our approach constructs a model-based group, I_i, for each i by using the prior or posterior correlation among the set of linear predictors. Though we will delve into the construction of I_i in Section 3, an initial understanding is that I_i comprises the data points that correspond to the linear predictors that are most informative for predicting the testing linear predictor, and thus the testing point, y_i. This set ensures that

our LGOCV focuses less on short-range prediction (interpolation) and more on longer-range prediction than LOOCV. In other words, LGOCV tests the model on more difficult prediction tasks since the most influential points are removed together with the testing point, instead of some arbitrary (possibly uncorrelated) point(s). The user needs to only provide a number that indicates the "degree of the independence" between the prediction point and the rest of the data", and we compute these groups for each datapoint in an automated way. In various practical examples, we will show how this model-based procedure produces reasonable groups. Advanced spatial examples applying the proposed method are presented by Adin et al. (2024). For a simple time-series example, our new approach will correspond to evaluating $\pi(y_t|y_1,\ldots,y_{t-k},y_{t+k},\ldots,y_T)$, for fixed $k>1$. This corresponds to removing a sequence of data with length $2k-1$, to predict the central one. As we see, this task mimics a longer-range prediction task. Our interpretation is that LGOCV quantifies the model's ability to predict longer-range more appropriately than LOOCV, when $k>1$, and is similar to the cross-validation procedure proposed by Burman et al. (1994) for stationary processes.

There are two key challenges to address to make our proposal practical. Firstly, we must quantify the information contributed by one data point in predicting another; this is crucial for the group construction. Secondly, we face the computational task of evaluating u_{LGOCV} given a set of groups. The naive computation of LGOCV by fitting models across all potential training sets and evaluating their utility against corresponding testing points is computationally infeasible, especially given the resource-demanding nature of modern statistical models. However, these challenges can be handled elegantly within the framework of latent Gaussian models (LGMs) combined with the integrated nested Laplace approximation (INLA) inference, as detailed in Rue et al. (2009, 2017); Van Niekerk and Rue (2024); Van Niekerk et al. (2023). Throughout this paper, we will assume that our model is an LGM. We will discuss how to integrate the automatic group construction and the fast computation of u_{LGOCV} using the INLA framework. Notably, our proposed methodology has been incorporated into the R-INLA package (www.rinla.org), extending its applicability across all LGMs supported by R-INLA.

1.4. Theoretical aspects

Cross-validation (CV), particularly LOOCV, is frequently considered as an estimator of $E_{\tilde{y}}[\log\pi(\tilde{y}|\boldsymbol{y})]$ or $E_{\tilde{y},\boldsymbol{y}}[\log\pi(\tilde{y}|\boldsymbol{y})]$. The first expectation describes the generalized predictive performance given a specific training set, while the second expectation describes the generalized predictive performance averaged over different identically distributed training sets. These expectations can be evaluated when assuming the existence of the joint density $\pi_T(\tilde{y},\boldsymbol{y})$, representing the true data generation process. Under the assumption of exchangeability and some regularity conditions on the model, the Bernstein-Von-Mises theorem states that $\log\pi(\tilde{y}|\boldsymbol{y})$ converges to a random variable irrelevant to $\boldsymbol{y}$. Consequently, $E_{\tilde{y}}[\log\pi(\tilde{y}|\boldsymbol{y})]$ and $E_{\tilde{y},\boldsymbol{y}}[\log\pi(\tilde{y}|\boldsymbol{y})]$ become equivalent in the limit. If we further assume that $\tilde{y}$ is sampled from the same distribution as all the training data, LOOCV is an asymptotically unbiased estimator of the expectations. Commonly used informa-

tion criteria, such as AIC (Akaike, 1973), WAIC (Watanabe, 2010), are asymptotically equivalent to LOOCV in fully exchangeable cases. This type of analysis is prevalent in the literature with various settings (Stone, 1974, 1977; Yang, 2007; Shao, 1993).

However, a similar analysis does not hold for dependent cases in general. Firstly, the existence of different prediction tasks means that both the model prediction, $\pi(\tilde{y}|\boldsymbol{y})$, and the true data generation process, $\pi_T(\tilde{y}|\boldsymbol{y})$, are not uniquely defined as discussed in Section 1.2. Secondly, the asymptotic scheme is not uniquely defined, even with a specific prediction task. For example, in a temporal model where data $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ is a time-series, observed at time $t_1 < t_2 < \cdots < t_n$ and we denote the last time step as T. Several meanings of of $n \to \infty$ can be considered:

- $T \to \infty$ and $t_i - t_{i-1}$ is a constant
- $t_i - t_{i-1} \to 0$ and T is a constant
- $t_i - t_{i-1} \to 0$ and $T \to \infty$ with $T(t_i - t_{i-1})$ fixed

These scenarios correspond to observing more future data and having higher sample rates within a time frame. As mentioned in Section 1.2, multilevel data can also have various asymptotic regimes. Thirdly, if the data generation process is not stationary, the model will not converge under certain asymptotic regimes, which differentiate $E_{\tilde{y}}[\log \pi(\tilde{y}|\boldsymbol{y})]$ from $E_{\tilde{y},\boldsymbol{y}}[\log \pi(\tilde{y}|\boldsymbol{y})]$ even in asymptotic scenarios. These points highlight that the estimand of CV is not uniquely defined in dependent cases, preventing the establishment of an asymptotic analysis framework.

From the perspective of CV, it is also inappropriate to consider it an estimator since each summand in CV should be viewed as a sample from different distributions due to the relevance of data indexes in dependent cases. For example, if we compute LOOCV in a time series. Each y_t is sampled from a different conditional distribution $\pi_T(y_t|\boldsymbol{y}_{-t})$ and thus the average $\frac{1}{n}\sum_{t=1}^{T} \log \pi(y_t|\boldsymbol{y}_{-t})$ cannot be considered as an estimator in general. Therefore, it is more reasonable to view CV as a predictive measurement rather than an estimator of an expectation. This perspective allows us to interpret the proposed LGOCV as the averaged predictive performance for similar prediction tasks, created systematically by the model.

While the proposal of Merkle et al. (2019) for multilevel model demonstrates that marginal WAIC is akin to LOOCV, we note that conditional WAIC aligns with LGOCV, where a hierarchical level, such as a school, defines the groups. The h-block CV of Burman et al. (1994) is a special case of LGOCV for a stationery model. LFOCV proposed by Bürkner et al. (2020) is similar to LGOCV as shown in Section 5. The spatial buffering proposed by (Valavi et al., 2018) ensures that no test data is spatially next to any training data, and is a special case of LGOCV for model with only spatial effects. LGOCV this provides a framework where no training data is placed next to the test data in terms of the entire model, and not just specific components thereof.

1.5. Plan of paper

We propose the model-based LGOCV to evaluate longer-range prediction performance for latent Gaussian models, as a special case of a hierarchical model. Complementing this, we introduce a computational method to approximate u_{LGOCV} without model refitting, which is crucial for practical implementation of our proposal. Our computational technique also facilitates the calculation of u_{LGOCV} with user-specified groups.

Section 2 introduces LGMs and explains how they can be efficiently inferred using INLA. In Section 3, we discuss the model-based group construction method for LGMs. This method can be implemented in two ways: by using the prior correlation matrix or the posterior correlation matrix of the latent linear predictors. In Section 4, we demonstrate how to approximate the LGOCV predictive density. Finally, in Section 5, we compare the approximated LGOCV with the exact LGOCV computed by Markov chain Monte Carlo (MCMC) and present some applications. We conclude with a general discussion in Section 6.

2. Latent Gaussian models

This section briefly introduces LGMs, as detailed in Rue et al. (2009, 2017); Van Niekerk and Rue (2024); Van Niekerk et al. (2023), since the model-based group construction and fast approximations rely on them. The LGMs can be formulated by

$$\begin{aligned} &y_i|\eta_i,\boldsymbol{\theta} \sim \pi(y_i|\eta_i,\boldsymbol{\theta}),\\ &\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{f}, \quad \boldsymbol{f}|\boldsymbol{\theta} \sim N(0,\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta})), \quad \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}). \end{aligned} \tag{2}$$

In LGMs, each y_i is independent conditioned on its corresponding linear predictor η_i, and hyperparameters $\boldsymbol{\theta}$; $\boldsymbol{\eta}$ is a linear combination of $\boldsymbol{f}$, which is assigned with a Gaussian prior with zero mean and a precision matrix parameterized by $\boldsymbol{\theta}$; $\boldsymbol{A}$ is the design matrix mapping $\boldsymbol{f}$ to $\boldsymbol{\eta}$; $\pi(\boldsymbol{\theta})$ is a prior density of hyperparameters. It is worth mentioning that the prior precision matrix $\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta})$ is very sparse, which is leveraged to speed up the inference.

The model is quite general because $\boldsymbol{f}$ can combine many modeling components, including linear model, spatial components, temporal components, spline components, etc (Wang, Yue and Faraway, 2018; Krainski et al., 2018; Gómez-Rubio, 2020). It is also common with linear constraints on the latent effects $\boldsymbol{f}$ (Rue and Held, 2005).

We can approximate $\pi(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ at some configurations, $\boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_k$. The configurations are located around the mode of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, denoted by $\boldsymbol{\theta}^*$, for numerical integration. Approximations of $\pi(\boldsymbol{\eta}|\boldsymbol{\theta},\boldsymbol{y})$ are computed using the linear relation, $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{f}$. The Gaussian approximation of $\pi(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y})$ plays an essential role, which is outlined as follows.

We have $\pi(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y})$ for a given $\boldsymbol{\theta}$,

$$\pi(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y}) \propto \exp\left\{ -\frac{1}{2}\boldsymbol{f}^{\intercal}\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta})\boldsymbol{f} + \sum_{i=1}^{n} \log(\pi(y_i|\eta_i,\boldsymbol{\theta})) \right\}, \tag{3}$$

whose mode is $\boldsymbol{\mu}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y})$. The Gaussian approximation of $\pi(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y})$ is

$$\pi_G(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y}) \propto \exp\left\{ -\frac{1}{2}\boldsymbol{f}^{\mathsf{T}}(\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta}) + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{C}(\boldsymbol{\theta},\boldsymbol{y})\boldsymbol{A})\boldsymbol{f} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{b}(\boldsymbol{\theta},\boldsymbol{y})\boldsymbol{f} \right\}. \tag{4}$$

In (4), $b_i(\boldsymbol{\theta},\boldsymbol{y}) = g_i'(\eta_i^*) - g_i''(\eta_i^*)\eta_i^*$, and $\boldsymbol{C}(\boldsymbol{\theta},\boldsymbol{y})$ is a diagonal matrix with $C_{ii}(\boldsymbol{\theta},\boldsymbol{y}) = -g_i''(\eta_i^*)$, where $g_i(\eta_i) = \log(\pi(y_i|\eta_i,\boldsymbol{\theta}))$ and $\eta_i^* = \boldsymbol{A}_i\boldsymbol{\mu}_f(\boldsymbol{\theta},\boldsymbol{y})$ with $\boldsymbol{A}_i$ being ith row of $\boldsymbol{A}$. The Gaussian approximation is denoted by,

$$\boldsymbol{f}|\boldsymbol{y},\boldsymbol{\theta} \approx N(\boldsymbol{\mu}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}), \boldsymbol{Q}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y})), \tag{5}$$

where $\boldsymbol{\mu}_f(\boldsymbol{\theta},\boldsymbol{y}) = \boldsymbol{Q}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y})^{-1}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{b}(\boldsymbol{\theta},\boldsymbol{y})$ and $\boldsymbol{Q}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}) = \boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta}) + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{C}(\boldsymbol{\theta},\boldsymbol{y})\boldsymbol{A}$ are the approximated posterior mean and precision matrix of $\boldsymbol{f}$ given $\boldsymbol{\theta}$.

3. Model-based Group Construction

The primary feature of our proposed group construction is that it requires a choice of correlation matrix (prior or posterior) for the linear predictors, and a single mandatory parameter to adjust the difficulty of the prediction task. This parameter is termed "the number of level sets". It can be interpreted as the strength of the non-dependence between the group to leave out and the rest of the data. A higher value would thus ensure that the leave-out group is more independent from the rest of the data, than a lower value. A higher indpeendnece between the leave-out data and the rest of the data simluates a more difficult prediction task for the model. Based on this value and the correlation matrix choice, all other processes are automated. In a multivariate Gaussian distribution, we can quantify the information provided by a data point to predict another data point by the variance reduction of the conditional distribution, and the variance reduction is a function of their correlation coefficient. To elaborate, if X and Y are both Gaussian random variables, the variance reduction resulting from knowing X when predicting Y equates to $\sigma_Y^2\rho^2$, where σ_Y^2 is the marginal variance of Y and ρ is the correlation between X and Y.

In LGMs, the linear predictors, $\boldsymbol{\eta}$, represent the underlining data generation process of data in (2). The linear predictors are designed to have a Gaussian prior and approximated to be Gaussian in posterior given $\boldsymbol{\theta}$ therefore, we can use the absolute value of the correlation matrix of $\boldsymbol{\eta}$ to represent the information provided by one data point to predict another data point. We evaluate those correlation matrices at the mode of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, denoted by $\boldsymbol{\theta}^*$. Then, we have correlation matrices of $\boldsymbol{\eta}$ derived from the prior precision matrix, $\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta}^*)$, and the posterior precision matrix, $\boldsymbol{Q}_{\boldsymbol{f}}(\boldsymbol{\theta}^*,\boldsymbol{y})$. We call the former one prior correlation matrix, denoted by $\boldsymbol{R}_{\text{prior}}$, and the latter one posterior correlation matrix, denoted by $\boldsymbol{R}_{\text{post}}$. Note that the correlation matrices are not fully evaluated and stored to avoid computational burden as they are dense and large; thus, care has to be applied to the implementation to make it feasible. The group would vary with $\boldsymbol{\theta}$. We use $\boldsymbol{\theta}^*$ because it has the highest weight in the posterior. This preference arises because we

frequently employ non-informative priors for hyperparameters. This approach ensures that our focus remains on evidence from the data rather than on arbitrary assumptions about hyperparameters.

Manually constructed groups are often based on prior knowledge and some structured effects, represented by $\boldsymbol{f}$. To imitate this process, we can compute the correlation matrix from a submatrix of $\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta})$. The correlation matrix, $\boldsymbol{R}_{\text{prior}}$, derived from the submatrix of the prior precision matrix, is a correlation matrix conditioning on those unselected effects. The groups constructed using $\boldsymbol{R}_{\text{prior}}$ are viewed to be solely user-defined in the way that it only depends on the priors and not on the data. In some situations this could be motivated, but in general we recommend using $\boldsymbol{R}_{\text{post}}$ to construct groups because the data will be informative to determine the importance of each effect.

When using a correlation matrix $\boldsymbol{R}$, it is natural to select a fixed number of η_j most correlated to η_i and include their index in the group I_i. However, this approach can be problematic as some linear predictors may have identical absolute correlations to η_i, e.g., in a model with only intercept, all the linear predictors are correlated to each other with correlation 1. Instead, we include all indices of η_j's with identical absolute correlations to η_i in I_i if one of them is included. We define a level set as all η_j's with the same absolute correlation to η_i and determine the group size based on the number of level sets, denoted as m. Setting a higher m results in a less dependent training set and testing point. We recommend using a small number of level sets, such as $m = 3$, as a high value of m can result in a large leave-out group size.

The automated group construction process thus involves selecting the number of level sets, m, and the correlation matrix to use, $\boldsymbol{R}_{\text{prior}}$ or $\boldsymbol{R}_{\text{post}}$. For each i, we can associate m level sets with the m largest absolute correlations to η_i, and the union of those level sets forms I_i. As an illustration, we outline the automated group construction procedure in Algorithm 1.

4. Approximation of LGOCV predictive density

In this section, we will explore the process of approximating $\pi(y_i|\boldsymbol{y}_{-I_i})$. The results are straightforward but tedious in implementation; thus, it is crucial to exercise caution to ensure that all potential numerical instabilities are accounted for. Through empirical testing, this new method has shown to be both more accurate and stable compared to the approach outlined in Rue et al. (2009), when $I_i = i$.

We start by writing $\pi(y_i|\boldsymbol{y}_{-I_i})$ as nested integrals,

$$\pi(y_i|\boldsymbol{y}_{-I_i}) = \int_{\boldsymbol{\theta}} \pi(y_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i})d\boldsymbol{\theta} \tag{6}$$

$$\pi(y_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i}) = \int \pi(y_i|\eta_i, \boldsymbol{\theta})\pi(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})d\eta_i. \tag{7}$$

The integral (6) is computed by the numerical integration (Rue et al., 2009), and the integral (7) is computed by Gauss-Hermite quadratures (Liu and Pierce, 1994) as the

conditional posterior density $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ will be approximated by a Gaussian distribution. The key to the accuracy of (7) is that the likelihood, $\pi(y_i|\eta_i,\boldsymbol{\theta})$, is known such that small approximation errors of $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ diminish due to the integration. The accuracy of (6) relies on the accuracy of (7) and the assumption that the removal of $\boldsymbol{y}_{I_i}$ does not have a dramatic impact on the posterior.

Algorithm 1: Find groups for all data points

Input: A correlation matrix choice $\boldsymbol{R}$ (posterior correlation is the default), Number of level sets m;
Output: A list containing the groups for all data points;
Calculate $\boldsymbol{R}$ from the model;
$N \leftarrow$ number of rows in $\boldsymbol{R}$;
groups $\leftarrow$ initialize N empty lists;
for $i = 1$ **to** N **do**
    $\boldsymbol{r} \leftarrow$ absolute values of the i-th row of $\boldsymbol{R}$;
    ordered indices $\leftarrow$ indices of $\boldsymbol{r}$ sorted by value in decreasing order;
    current absolute correlation $\leftarrow 1$;
    $k \leftarrow 1$;
    for $j = 1$ **to** m **do**
        while *current absolute correlation* $==$ $\boldsymbol{r}$[*ordered indices*[k]] **do**
            groups[i].append(ordered indices[k]);
            $k \leftarrow k+1$;
        end
        current absolute correlation $\leftarrow$ $\boldsymbol{r}$[ordered indices[k]];
    end
end
return *groups*;

The computation of the nested integrals reduces to the computation of $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i})$. We will approximate $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ by a Gaussian distribution, denoted by $\pi_G(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$, and $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i})$ by correcting the approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ in Rue et al. (2009). We further improve the mean of $\pi_G(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ using variational Bayes (Van Niekerk and Rue, 2024) in the implementation. In this section, we focus on the explanation of computing $\pi_G(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ and an approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i})$.

Computing $\pi_G(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$

The mean and variance of $\pi_G(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ can be obtained by

$$\begin{aligned}\mu_{\eta_i}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) &= \boldsymbol{A}_i\boldsymbol{\mu}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i}),\\ \sigma^2_{\eta_i}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) &= \boldsymbol{A}_i\boldsymbol{Q}^{-1}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})\boldsymbol{A}_i^{\mathsf{T}}.\end{aligned} \tag{8}$$

The computation of $\pi_G(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ requires the mode of $\pi(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ for each i at each configuration of $\boldsymbol{\theta}$, which is computationally expensive. With the mode at full data, we

use an approximation to avoid the optimization step,

$$\boldsymbol{Q}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) \approx \tilde{\boldsymbol{Q}}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) = \boldsymbol{Q}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}) - \boldsymbol{A}_{I_i}^{\mathsf{T}}\boldsymbol{C}_{I_i}(\boldsymbol{\theta},\boldsymbol{y})\boldsymbol{A}_{I_i}, \tag{9}$$

$$\boldsymbol{\mu}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) \approx \tilde{\boldsymbol{\mu}}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) = \tilde{\boldsymbol{Q}}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})^{-1}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{b}(\boldsymbol{\theta},\boldsymbol{y}) - \boldsymbol{A}_{I_i}^{\mathsf{T}}\boldsymbol{b}_{I_i}(\boldsymbol{\theta},\boldsymbol{y})), \tag{10}$$

where $\boldsymbol{A}_{I_i}$ is a submatrix of $\boldsymbol{A}$ formed by rows of $\boldsymbol{A}$, $\boldsymbol{b}_{I_i}(\boldsymbol{\theta},\boldsymbol{y})$ is a subvector of $\boldsymbol{b}(\boldsymbol{\theta},\boldsymbol{y})$, and $\boldsymbol{C}_{I_i}(\boldsymbol{\theta},\boldsymbol{y})$ is a principal submatrix of $\boldsymbol{C}(\boldsymbol{\theta},\boldsymbol{y})$. When the posterior is Gaussian, the approximation is exact as (9) and (10) define the precision matrix and the mean of the posterior. It seems easy to obtain the moments using (8), but the decomposition of $\tilde{\boldsymbol{Q}}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ is too expensive. To avoid the decomposition of $\tilde{\boldsymbol{Q}}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$, we use the linear relation $\boldsymbol{\eta}_{I_i} = \boldsymbol{A}_{I_i}\boldsymbol{f}$ to map all the computation on $\boldsymbol{f}$ to $\boldsymbol{\eta}_{I_i}$. We compute $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ and $\boldsymbol{\mu}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ through $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y})$ and $\boldsymbol{\mu}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y})$ as shown in the Appendix A using a low rank representation, where $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y})$ is the posterior covariance matrix of $\boldsymbol{\eta}_{I_i}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ is the covariance matrix of $\boldsymbol{\eta}_{I_i}$ with $\boldsymbol{y}_{I_i}$ left out. The computation of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y})$ is non-trivial, especially when linear constraints are applied, which is demonstrated in Appendix B.

The approximation is more accurate when $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ is close to Gaussian. The Gaussianity of $\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i}$ comes from three sources. Firstly, $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ is nearly Gaussian, when η_i is connected to large amount of data (Rue et al., 2009). Secondly, $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ is dominated by the Gaussian prior, which happens when η_i is connected to very few data. Thirdly, the log-likelihood can be close to the log-likelihood of a Gaussian distribution, resulting in the Gaussianity of $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ due to the conjugacy. Thus, $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ is rarely far away from a Gaussian distribution.

Approximating $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i})$

To approximate $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i})$, we use the relation,$\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i}) \propto \frac{\pi(\boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{y}_{I_i}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})}$, where we can approximate $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ at configurations as in Rue et al. (2009). We need to compute $\pi(\boldsymbol{y}_{I_i}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) \approx \int \pi(\boldsymbol{y}_{I_i}|\boldsymbol{\eta}_{I_i},\boldsymbol{\theta})\pi_G(\boldsymbol{\eta}_{I_i}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})d\boldsymbol{\eta}_{I_i}$. A Laplace approximation can be applied to this integral,

$$\pi_{\mathrm{LA}}(\boldsymbol{y}_{I_i}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) = \frac{\pi(\boldsymbol{y}_{I_i}|\boldsymbol{\eta}^*_{I_i},\boldsymbol{\theta})\pi_G(\boldsymbol{\eta}^*_{I_i}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})}{\pi_G(\boldsymbol{\eta}^*_{I_i}|\boldsymbol{\theta},\boldsymbol{y})}, \tag{11}$$

where $\boldsymbol{\eta}^*_{I_i}$ is the mode of $\pi_G(\boldsymbol{\eta}^*_{I_i}|\boldsymbol{\theta},\boldsymbol{y})$. Note that the correction of the hyperparameter reuses $\pi_G(\boldsymbol{\eta}_{I_i}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ and $\pi_G(\boldsymbol{\eta}_{I_i}|\boldsymbol{\theta},\boldsymbol{y})$.

5. Simulations and Applications

This section showcases two simulated examples and two real data applications. The code is available at https://github.com/zhedongliu/LGOCV.

We start with a simulation that tests the approximation accuracy in a multilevel model with various response types. Following this, a time series forecasting simulation is presented. This allows LGOCV results with automatically constructed groups to be compared to the LFOCV. We then delve into disease mapping, contrasting group constructions derived from various strategies. Finally, we apply our methodology to intricate models using a large dataset, as documented by Lowe et al. (2021). For the construction of the leave-out group for LGOCV, we used Algorithm 1 in Section 3. The procedures detailed in Sections 3 and 4 have been integrated into R-INLA, ensuring that all computational tasks in this section are executed through R-INLA.

Simulated Multilevel Model with Various Responses

This example is a simulation that demonstrates the accuracy of the approximation described in Section 4. The main purpose is to compare $\pi(y_i|\boldsymbol{y}_{-I_i})$ computed using an approximation in Section 4 and the same quantity computed using MCMC. Furthermore, we use automatic group construction with the number of level sets equal to 1, corresponding to predicting a data point from a new class.

We simulate data according to the following process. Initially, we simulate 10 class means, denoted as $\boldsymbol{s}$, from a standard normal distribution. Next, we compute 100 linear predictors, $\eta_i = \mu + s_{j(i)}$, where $\mu = \log(10)$ and $j(i)$ is a function mapping data index i to the group index j. For this function, we set $j(i) = \lceil \frac{i}{10} \rceil$, where the ceiling function, $\lceil x \rceil$, rounds a number up to the nearest integer. We generate responses according to the linear predictor and one of three response types; Gaussian, binomial and exponential. The mean of the Gaussian response is η_i, and the standard deviation is 0.1. We generate binomial responses with a success probability of $\frac{1}{1+e^{-\eta_j}}$ for 20 trials. The exponential responses are generated with a mean of e^{η_j}.

We consider the model,

$$\begin{aligned} &\log(\tau_s) \sim N(0, 10^{-4}), \quad \mu \sim N(0, 10^{-4}), \quad s_j|\tau_s \sim N(0, \tau_s), \\ &\eta_i = \mu + s_j, \quad y_i|\eta_i \sim \text{response model}(\eta_i), \end{aligned} \tag{12}$$

where the second parameter of the Gaussian distribution is the precision, and the likelihood is specified according to the data generation process with the given response model.

As a reference, we let the MCMC runs for 10^8 iterations, which makes the Monte Carlo errors negligible. The large size of MCMC samples is required because the predictive distributions are influenced by the tails of $\pi(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$. In Figure 2, (a), (c), and (d) show the data against its group index, which presents a clear group structure; (b), (d), and (e) show the comparison of $\pi(y_i|\boldsymbol{y}_{-I_i})$ obtained from the approximations and MCMC. We use Rstan (Stan Development Team, 2022) for the MCMC.

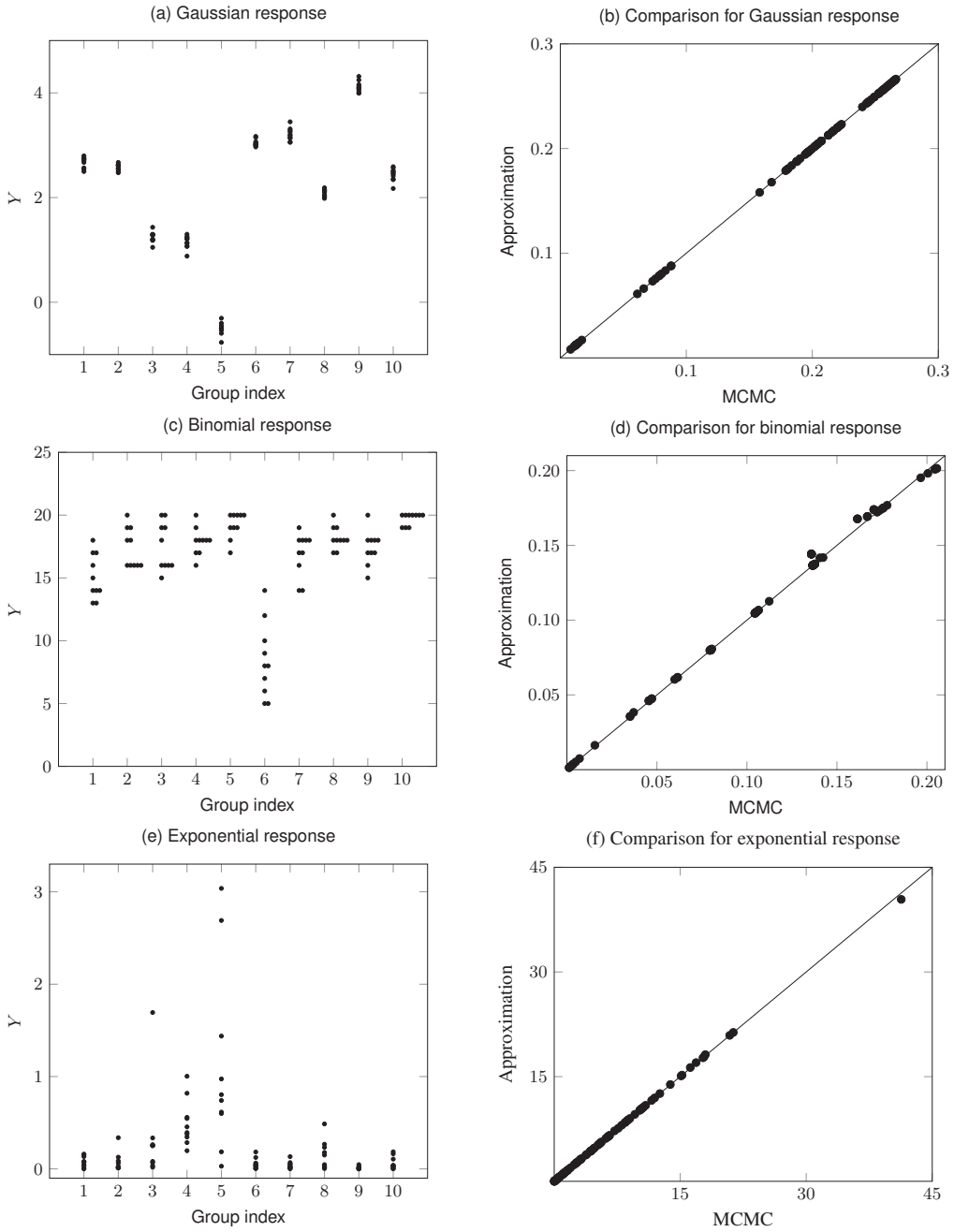

Figure 2. *Comparison of $\pi(y_i|\boldsymbol{y}_{-I_i})$ from approximations and MCMC. First column: y-axis shows response value, x-axis shows group index. Second column: y-axis shows LGOCV from proposed approximation, x-axis shows LGOCV from MCMC.*

This example shows that the approximations are highly accurate. When the response is Gaussian, the approximation almost equals the MCMC results, where the main difference is due to MCMC sampling errors, as our approach is exact up to numerical integration in this case. Also, under both non-Gaussian cases, the results are close to the long-run MCMC results.

Time Series Forecasting

In this example, we will demonstrate how the automatic LGOCV method can measure the forecasting performance of a time series model, while LOOCV is not effective in doing so.

We will first simulate 2000 data points using the following procedure: We will simulate an AR(1) time series by using $u_i = 0.9u_{i-1} + \varepsilon_{u_i}$, where ε_{u_i} follows a standard Gaussian distribution. Next, we will compute linear predictors by calculating $\eta_i = \mu + u_i$, with μ set to 2. Finally, the Gaussian responses have mean η_i, and a standard deviation of 0.1.

We fit a time series model on the simulated data:

$$\begin{aligned} &\mu \sim N(0, 10^{-4}), \quad \boldsymbol{u} \sim N(0, \boldsymbol{Q}_u), \\ &\eta_i = \mu + u_i, \quad y_i|\eta_i \sim N(\eta_i, 100), \end{aligned}$$

where $\boldsymbol{Q}_u$ is determined by an AR(1) model with the true parameters.

The prediction task is k steps forward forecasting for $k = \{1, 2, \ldots, 10\}$ using the true model. The natural cross-validation for these prediction tasks is LFOCV. To replicate the LFOCV, the group in LGOCV for testing point y_i and k steps forward prediction includes $\boldsymbol{y}_{(i-k+1):n}$. We can compute LFOCV for every k, denoted by LFOCV(k). To make the training set similar to the data set, the last 500 data points will be used as testing points, which means $i = \{1501, \ldots 2000\}$ in (1), and the quantity is averaged over 500 data points. We can also compute LGOCV using automatically constructed groups with the number of level sets, $m = \{1, 2, \ldots, 10\}$, denoted by LGOCV(m). In this setting, the automatically constructed group for a testing point y_i with a number of level sets equal to m includes $\boldsymbol{y}_{\max(1,i-m+1):\min(n,i+m-1)}$. Also, LGOCV(1) is equivalent to LOOCV in this model.

To compare LGOCV and LFOCV, we will fit a natural spline to have LFOCV(t) for t as a real number (see Figure 3 (a)) and map the number of level sets in LGOCV to the steps ahead in LFOCV (see Figure 3 (b)). We can see that LOOCV measures approximately 0.4 steps forward forecasting when the simplest prediction task is one step forward forecasting. LGOCV(2) represents roughly a one-step forward forecasting performance of the model. As the number of level sets increases in LGOCV, it represents more steps forward forecasting performance. Note that the specific translation between the automatic LGOCV and LFOCV is only valid in this model and may not be applicable in other models.

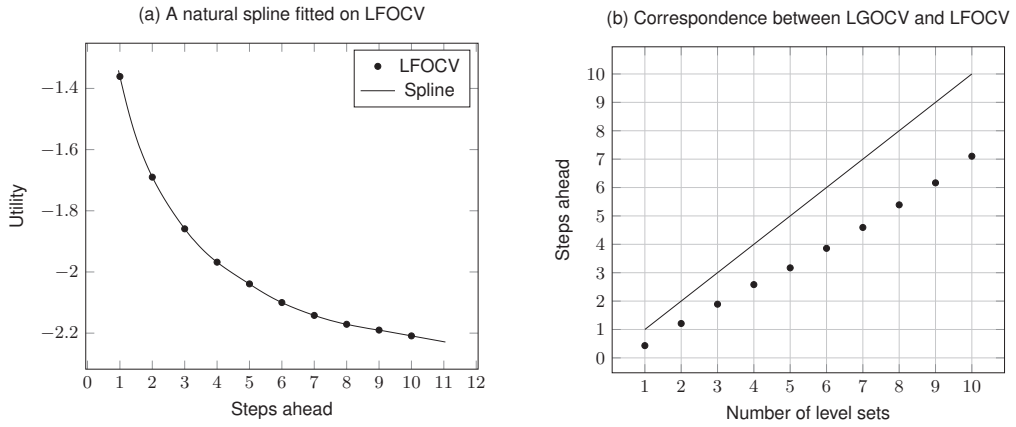

Figure 3. *Comparison of Automatic LGOCV and LFOCV. LOOCV measures approximately 0.4 steps forward forecasting. LGOCV(2) roughly represents a one-step forward forecasting performance.*

Disease Mapping

In this example, we will present groups constructed by different automatic group construction strategies. We will see the differences between those groups and get an idea to choose a proper group construction strategy.

We applied a disease mapping model to data detailing cancer incidence by location (Besag, York and Mollié, 1991; Wakefield, Best and Waller, 2000; Held et al., 2005). This dataset captures oral cavity cancer cases in Germany from 1986-1990 (Held et al., 2005). The response y_i indicates the cases in area i over five years. The case count in each region is influenced by its population and age distribution. The expected case count E_i in the region i is derived from its age distribution and population, ensuring $\sum_i y_i = \sum_i E_i$. Additionally, the covariate x_i represents tobacco consumption in area i.

We fit the following model on the data set:

$$\begin{aligned} y_i|\eta_i &\sim \text{Poisson}(E_i \exp(\eta_i)) \\ \eta_i &= \mu + f_{\text{rw}}(x_i) + u_i + v_i, \end{aligned} \tag{13}$$

where μ is an intercept, $\boldsymbol{u}$ is a spatially structured component, $\boldsymbol{v}$ is an unstructured component (Krainski et al., 2018), and $\boldsymbol{f}_{rw}$ is an intrinsic second-order random-walk model of the covariate x_i (Rue and Held, 2005).

In Figure 4, we illustrate groups formed through various automatic group construction strategies. The testing point is located in the black region, while the data in the group are located in grey areas. As seen in Figure 4 (a) and (b), Groups from $\boldsymbol{R}_{\text{prior}}$ focus solely on spatial effects. Groups from $\boldsymbol{R}_{\text{post}}$ exhibit mostly strong spatial patterns, such as Figure 4 (c). Yet, some points, like in Figure 4 (d), indicate non-spatial patterns. This arises as all model components, including fixed and random effects, priors, and the response variable, are considered.

The spatial patterns in posterior groups may justify incorporating spatial effects into the model, given that data retains this pattern in correlation. In practice, groups from $\boldsymbol{R}_{\text{post}}$ offer a more balanced representation. However, $\boldsymbol{R}_{\text{prior}}$ with selective effects resemble those from manually defined groups.

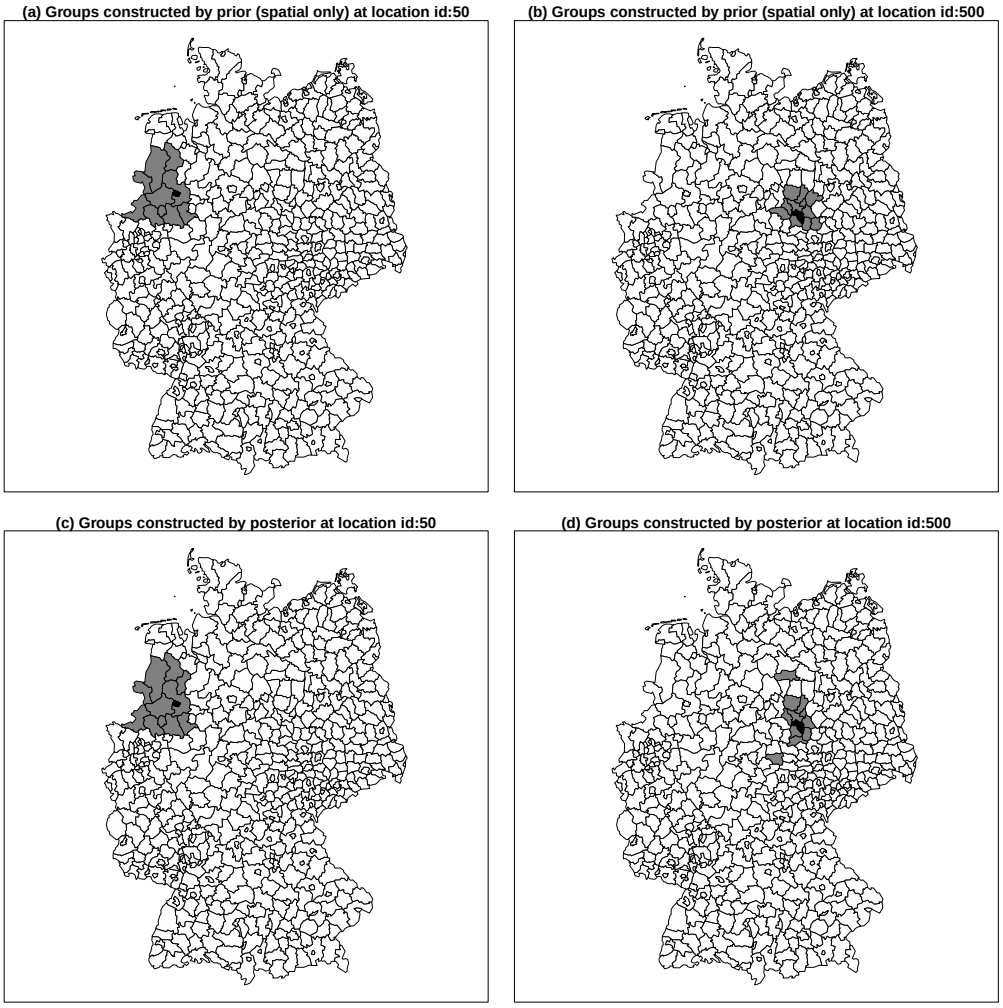

Figure 4. *Groups by different automatic construction strategies. The testing point is located in the black region, and the data in the group are located in the grey regions. In (d), the group constructed by posteriors contains some non-spatial patterns.*

Dengue Risk in Brazil

In this real-world example, we will demonstrate the scalability and adaptability of the automatic LGOCV method in a complex model structure and a large sample size. The automatically constructed model-based groups are consistent with the domain knowledge that dengue disease is prevalent in summer.

We will repeat the variable selection process as shown in Lowe et al. (2021) using the automatic LGOCV. The model chosen by LGOCV is considered to have better predictive power for longer-range predictions, than those selected based on other criteria because the most informative data points for predicting the target are excluded from the training set.

The models study the influence of extreme hydrometeorological hazards on dengue risk, factoring in Brazil's urbanization levels. Our dataset, with $127,224$ samples representing $12,895,293$ dengue cases, covers Brazil's 558 microregions from January 2001 to December 2019. Given the dataset's magnitude and the model's intricacy, LGOCV or LOOCV calculations require the approximation method detailed in Section 4.

Data points include month, year, microregion, and state. The candidate covariates encompass the monthly average of daily minimum (T_{min}) and maximum temperatures (T_{max}), the palmer drought severity index (PDSI), the urbanization levels: overall (u), centered at high (u_1), intermediate (u_2), and more rural levels (u_3) and the access to water supply: overall (w) and centered at high-frequency shortages (w_1), intermediate (w_2), and low-frequency shortages (w_3). To preprocess these covariates' specifics, refer to Lowe (2021).

The data generating model is chosen to be negative binomial, to account for overdispersion. The latent field consists of a temporal component describing a state-specific seasonality using a cyclic first difference prior distribution and a spatial component describing year-specific spatially unstructured and structured random effects using a modified Besag-York-Mollie (BYM2) model with a scaled spatial component (Riebler et al., 2016). The temporal component has replications for each state, and the spatial component has replications for each year. We can express the base model using the INLA-style formula,

```
y ~ 1 + covariates + f(month, model = "rw1", replicate = state, cyclic = TRUE)
    +f(microregion, model = "bym2", replicate = year).
```

In short, we write this model as $\mathrm{y} \sim 1 + \mathrm{covariates} + f_t + f_s$. The number of parameters in this model is $21,567$ with $127,224$ observations for the full model. The appendix of Lowe et al. (2021) and its repository Lowe (2021) provide full details about the models and data.

The model accounts for temporal effects with spatial replicates and spatial effects with temporal replicates, complicated by various constraints. Given its intricacy and the lack of a clear prediction task, crafting groups for LGOCV manually is challenging. Hence, utilizing our automatic group construction through posterior correlation is beneficial. For model comparisons, using the same groups across different models is recommended. The base model, which only incorporates structured components, is chosen for group building. Most automatic groups cluster data from the same year, location, and nearby months to the testing points. Figure 5 displays the relative month frequencies in the group, given the testing points correspond to a specific month. The chart suggests the first half-year data better informs predictions. Even in July and November testing

points, the group frequently includes that data, aligning with the known prevalence of dengue during summer. See Figure 5 (c) and (d) for details.

The results of model selection using deviance information criterion (DIC), LOOCV, and LGOCV ($m = 2,3,4$) are presented in Table 1. The candidate models are those referenced in Lowe (2021). To transform equation 1 into a loss function, we calculated its negative value.

From Table 1 we note that LOOCV prefers the spatio-temporal model that incorporates access to water while the spatio-temporal model with urbanization as a covariate, is preferred by LGOCV. This result is interesting since we can conclude that the same model might not necessarily perform well for short- and longer-range prediction. The practitioner thus needs to decide what the goal of the modeling is, and then choose the model to be used accordingly. If we want to predict dengue risk for a new unobserved area or time point, it seems that urbanization has a better prediction ability than access to water. Note also that as we increase the number of level sets, we are defining a prediction target with an increased range, thus moving further away from LOOCV.

Table 1. *Comparative evaluation of models for predicting variable y based on various environmental factors. This table presents the model selection results, including each model's Deviance Information Criterion (DIC), LOOCV, and LGOCV scores.*
Note: We offset DIC by 826841.66, LOOCV by 3.2721, LGOCV ($m = 2$) by 3.314, LGOCV ($m = 3$) by 3.3763 and LGOCV ($m = 4$) by 3.4372.

Index	Model	DIC	LOOCV	LGOCV		
				($m = 2$)	($m = 3$)	($m = 4$)
1	$y \sim 1 + f_t + f_s$	3615.38	0.0151	0.0158	0.0206	0.0270
2	$y \sim 1 + T_{min} + f_t + f_s$	1562.96	0.0064	0.0067	0.0088	0.0098
3	$y \sim 1 + T_{max} + f_t + f_s$	2228.73	0.0091	0.0098	0.0133	0.0163
4	$y \sim 1 + PDSI + f_t + f_s$	2167.12	0.0092	0.0095	0.0126	0.0184
5	$y \sim 1 + PDSI + T_{min} + f_t + f_s$	160.43	0.0006	0.0006	0.0012	0.0023
6	$y \sim 1 + PDSI + T_{max} + f_t + f_s$	900.65	0.0038	0.0038	0.0057	0.0084
7	$y \sim 1 + PDSI + T_{min} + PDSI * u_1 + u + f_t + f_s$	38.21	0.0002	0*	0*	0*
8	$y \sim 1 + PDSI + T_{min} + PDSI * u_2 + u + f_t + f_s$	39.13	0.0002	0*	0*	0*
9	$y \sim 1 + PDSI + T_{min} + PDSI * u_3 + u + f_t + f_s$	28.64	0.0002	0*	0*	0*
10	$y \sim 1 + PDSI + T_{min} + PDSI * w_1 + w + f_t + f_s$	6.68	0*	0.0005	0*	0.0014
11	$y \sim 1 + PDSI + T_{min} + PDSI * w_2 + w + f_t + f_s$	0*	0*	0.0005	0*	0.0015
12	$y \sim 1 + PDSI + T_{min} + PDSI * w_3 + w + f_t + f_s$	4.55	0*	0.0006	0*	0.0014

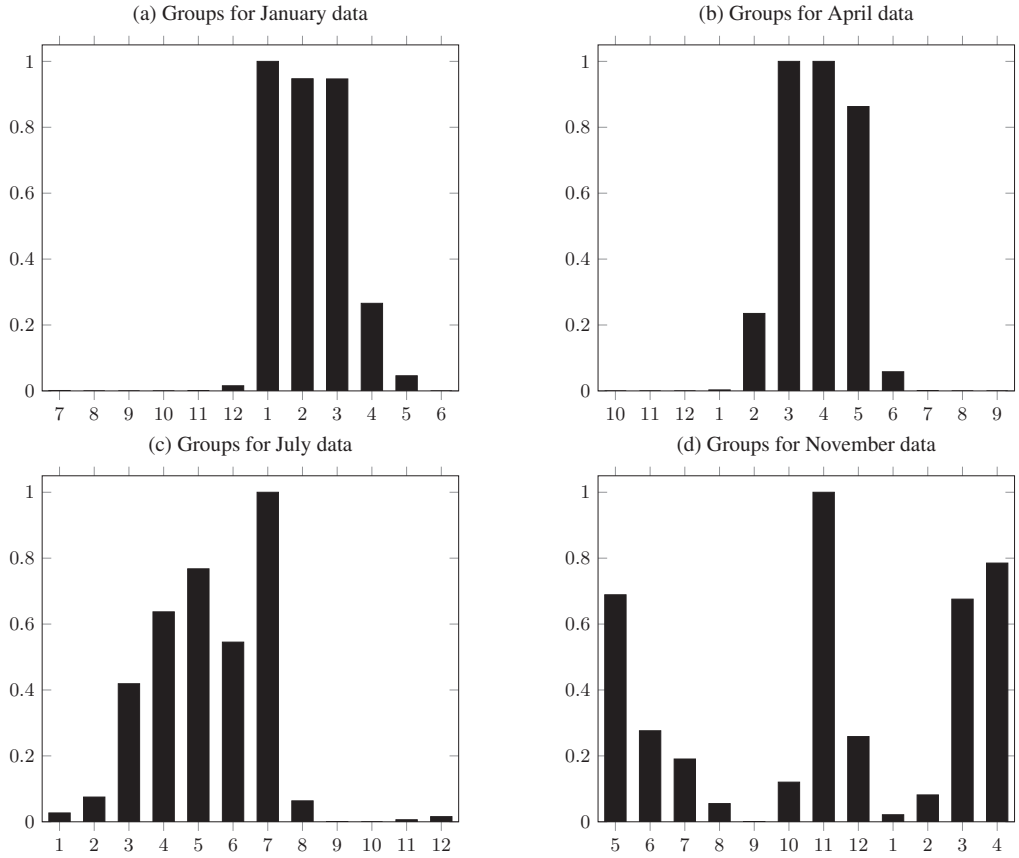

Figure 5. *Groups for testing points from a specific month. y-axis: relative frequency, x-axis: month of data measurement in groups. The first half-year data are more informative for prediction. As shown in (c) and (d), even in July and November, the group often includes data consistent with the known summer prevalence of dengue. Note that dengue is prevalent in the summer months which are approximately November to February.*

6. Discussion

An over-reliance on LOOCV to evaluate predictive capacity in general persists in statistical practice, despite concerns raised in studies such as Roberts et al. (2017); Vehtari et al. (2019). LOOCV can provide an evaluation of short-range predictive ability with well-established asymptotics for some models. On the other hand, what can we do to evaluate the longer-range predictive ability of complex models? Various approaches for specific models, such as time series or spatial models have been proposed, where custom CV procedures are designed to mimic a longer-range prediction task than that of LOOCV. We have introduced an automated approach for evaluating the longer-range prediction ability of any latent Gaussian model, namely LGOCV. LGOCV is designed to be applicable to all models that are latent Gaussian models, and thus provides a framework

for general longer-range predictive ability evaluation without the need for case-by-case considerations. Moreover, we propose a computationally efficient approach to calculate LGOCV scores and metrics based on the INLA methodology. We have shown that our approximate LGOCV implementation is almost exact when compared with the results from MCMC, albeit at a much lower computational cost. This enables practitioners to use the LGOCV approach for complex models and large data.

Our approach is designed for latent Gaussian models and some ideas can thus be extended to the case of non-latent Gausian models with careful consideration of the computational cost associated with this endeavor. LGOCV for LGMs is computationally efficient in INLA, since it is fully parallelizable by computing the necessary quantities only at the mode of the hyperparameters. For huge data ($n > 10^6$) however, the cost will be high since the cost increases linearly in n, albeit much lower than other available approaches. For huge data, performing CV on a subset or constructing the groups manually could be considered. Nonetheless, for LGMs, the proposed LGOCV could be considered as the most feasible approach for longer-range predictive ability evaluation.

The choice of the number of level sets determine the prediction task and thus the degree of independence between the leave-out group and the rest of the data. There is not a one-to-one correspondence between the number of level sets and the number of points to leave out as shown in the simulations and applications, although a higher number of level sets would imply a longer range for the prediction task, than a lower number. The choice of the number of level sets remains arbitrary since it is a user-defined parameter, we recommend a low number like $m = 2$ or $m = 3$ if there is no clear indication of what else m should be. There exists no optimal value of m in general, since it would imply different prediction tasks for different levels of dependency. In our applications, and those of others who have applied the LGOCV framework, it is shown that LGOCV provides the information we need to evaluate longer-range prediction ability, and complements the information from LOOCV.

It is pertinent to note that the proposed LGOCV do not replace a custom CV strategy designed by modelers, tailored for specific applications. We pose it as an alternative default strategy for longer-range prediction ability evaluation, that complements LOOCV in assessing the predictive ability of an LGM, while being computationally efficient and practical for real-world scenarios.

Acknowledgments

The authors thank D. Castro-Camilo, D. Rustand, and E. Krainski for valuable discussions and suggestions.

A. On the computation of $\Sigma_{\eta_{I_i}}(\theta, y_{-I_i})$ and $\mu_{\eta_{I_i}}(\theta, y_{-I_i})$

In this section, we let I_i be I and drop θ to simplify the notation. We have a random vector $\eta_I|y \sim N(\mu_{\eta_I}(y), \Sigma_{\eta_I}(y))$, which can be viewed as a posterior distribution with prior $\eta_I|y_{-I} \sim N(\mu_\eta(y_{-I}), \Sigma_\eta(y_{-I}))$ and likelihood $\pi_G(y_I|\eta_I) \propto \exp\left\{-\frac{1}{2}\eta_I^T C(y_I)\eta_I + b(y_I)\eta_I\right\}$. Now, we need to use the posterior and the likelihood to obtain the prior.

If $\Sigma_{\eta_I}(y)$ is full rank, we have $Q_{\eta_I}(y) = \Sigma_{\eta_I}(y)^{-1}$ and $b_{\eta_I}(y) = Q_{\eta_I}(y)\mu_{\eta_I}(y)$. By conjugacy of Gaussian prior and Gaussian likelihood, $Q_{\eta_I}(y_{-I}) = Q_{\eta_I}(y) - C(y_I)$ and $b_{\eta_I}(y_{-I}) = Q_{\eta_I}(y)\mu_{\eta_I}(y) - b(y_I)$. Then we have desired $\mu_{\eta_I}(y_{-I})$ and $\Sigma_{\eta_I}(y_{-I})$.

If $\Sigma_{\eta_I}(y)$ is singular, we let $\eta|y = Bz|y$, where $B = V\Lambda$ with V containing eigenvectors corresponding to non-zero eigenvalues, Λ containing square root of non-zero eigenvalues on its diagonal, and $z|y \sim N(\mu_z(y), \mathcal{I})$, where $\mathcal{I}$ is an identity matrix and $\mu_z(y) = B^\intercal \mu_{\eta_I}(y)$. By conjugacy, we have $Q_z(y_{-I}) = \mathcal{I} - B^\intercal C(y_I)B$ and $b_z(y_{-I}) = \mu_z(y) - B^\intercal b(y_I)$. It is followed by $\mu_z(y_{-I}) = Q_z(y_{-I})^{-1} b_z(y_{-I})$. Then mean and covariance of $z|y_{-I}$ is $\mu_\eta(y_{-I}) = B\mu_z(y_{-I})$, $\Sigma_\eta(y_{-I}) = B\Sigma_z(y_{-I})B^\intercal$.

B. On the computation of $\Sigma_{\eta_{I_i}}(\theta, y)$ and $\mu_{\eta_{I_i}}(\theta, y)$ with Linear Constraints

We start by illustrating how to compute $\Sigma_{\eta_{I_i}}(\theta, y)$ and $\mu_{\eta_{I_i}}(\theta, y)$ without linear constraints. $\mu_{\eta_{I_i}}(\theta, y)$ is simply obtained by $\mu_{\eta_{I_i}}(\theta, y) = A_{I_i}\mu_f(\theta, y)$. However, we never store large dense matrix like $Q_f(\theta, y)^{-1}$. Thus, $\Sigma_{\eta_{I_i}}(\theta, y)$ cannot be obtained by using matrix multiplication $A_{I_i} Q_f(\theta, y)^{-1} A_{I_i}^\intercal$. Instead, we compute $\Sigma_\eta(\theta, y)$ entry by entry and use the result to fill in entries of $\Sigma_{\eta_{I_i}}(\theta, y)$. We compute $\Sigma_\eta(\theta, y)_{i,j}$ by solving

$$Q_f(\theta, y)x = A_i$$

and $\Sigma_\eta(\theta, y)_{i,j} = A_j x$. The computation is fast because A and $Q_f(\theta, y)$ are sparse, and the factorization of $Q_f(\theta, y)$ is reused.

When linear constraints $\mathcal{C}f = e$ are applied on f, we have

$$\Sigma_f(\theta, y)^* = Q_f(\theta, y)^{-1} - Q_f(\theta, y)^{-1}\mathcal{C}^\intercal(\mathcal{C}Q_f(\theta, y)^{-1}\mathcal{C}^\intercal)^{-1}\mathcal{C}Q_f(\theta, y)^{-1},$$
$$\mu_f(\theta, y)^* = \mu_f(\theta, y) - Q_f(\theta, y)^{-1}\mathcal{C}^\intercal(\mathcal{C}Q_f(\theta, y)^{-1}\mathcal{C}^\intercal)^{-1}(\mathcal{C}\mu_f - e),$$

where $\Sigma_f(\theta, y)^*$ and $\mu_f(\theta, y)^*$ are the mean and the covariance matrix after applying constraints (Rue and Held, 2005). Because $\mu_f(\theta, y)^*$, is always stored, the computation of $\mu_{\eta_{I_i}}(\theta, y)$ is simple. We need to propagate the effects of linear constraints to $\Sigma_\eta(\theta, y)_{i,j}$. This is achieved by computing (Rue and Held, 2005)

$$x^* = x - Q_f(\theta, y)^{-1}\mathcal{C}^\intercal(\mathcal{C}Q_f(\theta, y)^{-1}\mathcal{C}^\intercal)^{-1}\mathcal{C}x,$$

where x solves $Q_f(\theta, y)x = A_i$. Then $\Sigma_\eta(\theta, y)^*_{i,j} = A_j x^*$.

References

Adin, A., Krainski, E. T., Lenzi, A., Liu, Z., Martínez-Minaya, J., and Rue, H. (2024). Automatic cross-validation in structured models: Is it time to leave out leave-one-out? *Spatial Statistics*, page 100843.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó Location Budapest, Hungary.

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Bürkner, P.-C., Gabry, J., and Vehtari, A. (2020). Approximate leave-future-out cross-validation for bayesian time series models. *Journal of Statistical Computation and Simulation*, 90(14):2499–2523.

Burman, P., Chow, E., and Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika*, 81(2):351–358.

Cerqueira, V., Torgo, L., and Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028.

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

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. CRC Press.

Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *Applied statistics*, 28(1):100–108.

Held, L., Natário, I., Fenton, S. E., Rue, H., and Becker, N. (2005). Towards joint disease mapping. *Statistical methods in medical research*, 14(1):61–82.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC.

Liu, Q. and Pierce, D. A. (1994). A note on gauss-hermite quadrature. *Biometrika*, 81(3):624–629.

Lowe, R. (2021). Data and R code to accompany 'Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study'.

Lowe, R., Lee, S. A., O'Reilly, K. M., Brady, O. J., Bastos, L., Carrasco-Escobar, G., de Castro Catão, R., Colón-González, F. J., Barcellos, C., Carvalho, M. S., et al. (2021). Combined effects of hydrometeorological hazards and urbanisation on dengue risk in brazil: a spatiotemporal modelling study. *The Lancet Planetary Health*, 5(4):e209–e219.

McQuarrie, A. D. and Tsai, C.-L. (1998). *Regression and time series model selection*. World Scientific.

Merkle, E. C., Furr, D., and Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, 84:802–829.

Rabinowicz, A. and Rosset, S. (2022). Cross-validation for correlated data. *Journal of the American Statistical Association*, 117(538):718–731.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of econometrics*, 99(1):39–61.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4):1145–1165.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, 4:395–421.

Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., and Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *Gigascience*, 6(5):gix019.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494.

Stan Development Team (2022). RStan: the R interface to Stan. R package version 2.21.5.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2018). blockcv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Biorxiv*, page 357798.

Van Niekerk, J., Krainski, E., Rustand, D., and Rue, H. (2023). A new avenue for bayesian inference with inla. *Computational Statistics & Data Analysis*, 181:107692.

Van Niekerk, J. and Rue, H. (2024). Low-rank variational bayes correction to the laplace method. *Journal of Machine Learning Research*, 25(62):1–25.

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

Vehtari, A., Simpson, D. P., Yao, Y., and Gelman, A. (2019). Limitations of "limitations of bayesian leave-one-out cross-validation for model selection". *Computational Brain & Behavior*, 2(1):22–27.

Wakefield, J. C., Best, N., and Waller, L. (2000). Bayesian approaches to disease mapping. *Spatial epidemiology: methods and applications*, pages 104–127.

Wang, X., Yue, Y. R., and Faraway, J. J. (2018). *Bayesian regression modeling with INLA*. CRC Press.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12).

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473.