Fast and scalable computations for Gaussian hierarchical models with intrinsic conditional autoregressive spatial random effects

Marco A. R. Ferreira[a,*], Erica M. Porter[a], Christopher T. Franck[a]

[a]*Department of Statistics, Virginia Tech, USA*

Abstract

Fast algorithms are developed for Bayesian analysis of Gaussian hierarchical models with intrinsic conditional autoregressive (ICAR) spatial random effects. To achieve computational speed-ups, first a result is proved on the equivalence between the use of an improper CAR prior with centering on the fly and the use of a sum-zero constrained ICAR prior. This equivalence result then provides the key insight for the algorithms, which are based on rewriting the hierarchical model in the spectral domain. The two novel algorithms are the Spectral Gibbs Sampler (SGS) and the Spectral Posterior Maximizer (SPM). Both algorithms are based on one single matrix spectral decomposition computation. After this computation, the SGS and SPM algorithms scale linearly with the sample size. The SGS algorithm is preferable for smaller sample sizes, whereas the SPM algorithm is preferable for sample sizes large enough for asymptotic calculations to provide good approximations. Because the matrix spectral decomposition needs to be computed only once, the SPM algorithm has computational advantages over algorithms based on sparse matrix factorizations (which need to be computed for each value of the random effects variance parameter) in situations when many models need to be fitted. Three simulation studies are performed: the first simulation study shows improved performance in computational speed in estimation of the SGS algorithm compared to an algorithm that uses the spectral decomposition of the precision matrix; the second simulation study shows that for model selection computations with 10 regressors and sample sizes varying from 49 to 3600, when compared to the current fastest state-of-the-art algorithm implemented in the R package INLA, SPM computations are 550 to 1825 times faster; the third simulation study shows that, when compared to default INLA settings, SGS and SPM combined with reference priors provide much more adequate uncertainty quantification. Finally, the application of the novel SGS and SPM algorithms is illustrated with a spatial regression study of county-level median household income for 3108 counties in the contiguous United States in 2017.

Keywords: Areal data; ICAR random effects; Markov random fields; Singular Gaussian distribution; Spatial statistics.

*Corresponding author. Department of Statistics, Virginia Tech, Blacksburg, Virginia, 24061, USA.
Email address: marf@vt.edu (Marco A. R. Ferreira)

1. Introduction

Bayesian hierarchical models with conditional autoregressive (CAR) (Besag, 1974) spatial random effects are used in a wide variety of fields such as economics, environmental science, and neuroscience. One of the most widely used spatial hierarchical models has intrinsic CAR (ICAR) random effects (Besag et al., 1991). The distribution of each ICAR random effect typically conditions on the ICAR effects of its neighbors. While this formulation has become ubiquitous in practice, two main difficulties arise when using ICAR spatial random effects. First, the joint density of ICAR effects is improper, which complicates inference on the spatial random effects. Second, the priors placed on the hyperparameters of the hierarchical model may unduly influence the analysis. To address these issues, Keefe et al. (2018, 2019) have proposed sum-zero constrained ICAR models and a corresponding reference prior for hierarchical models with sum-zero constrained ICAR random effects. These methods have been implemented in the R package ref.ICAR (Porter et al., 2019). Despite these advances, it has remained unclear whether Bayesian inferences based on the sum-zero constrained ICAR differ from the historical practice of centering the spatial random effects at each iteration of the Markov chain Monte Carlo (MCMC) algorithm. Here we show that the resulting analyses are equivalent.

A second practical contribution of this work is a huge speed up in computations for ICAR models. The proof of the equivalence result mentioned in the above paragraph provides the key insight for the speed up. Specifically, the spectral decomposition of the ICAR precision matrix in the equivalency proof is used to transform both the dependent variable and the regressors from their respective domains to the spectral domain. We call these new algorithms the Spectral Gibbs Sampler (SGS) and the Spectral Posterior Maximizer (SPM). We note that a similar spectral decomposition transformation was first proposed by Crainiceanu and Ruppert (2004) in the context of likelihood ratio tests for linear mixed models and was used by Kang et al. (2008) for the analysis of genomic wide association studies. To the best of our knowledge, similar spectral decomposition transformations have not been used for the analysis of hierarchical models with ICAR spatial random effects.

A decision on whether to use algorithms based on the spectral decomposition or algorithms based on numerical linear algebra for sparse matrices (e.g., as implemented in the R package INLA (Rue et al., 2009; Martins et al., 2013)) depends on the problem to be solved. Specifically, even though the computation of matrix spectral decomposition scales cubically with sample size, for a given neighborhood structure this decomposition needs to be computed only once to transform the data to the spectral domain, and then the SGS and SPM algorithm are applied directly to the spectral-domain data. In contrast, numerical decompositions for sparse matrices implemented in the R package INLA (the current fastest state-of-the-art) need to be performed for each different value of the random effects variance parameter. Thus, in situations when many models need to be fitted, the SPM algorithm may be substantially faster than INLA. For example, as shown in a simulation study in Section 7, for model selection computations with 10 regressors and sample sizes varying from 49 to 3600, when compared to INLA, SPM computations are 550 to 1825 times faster.

In addition, another simulation study shows that, when compared to default INLA settings, SGS and SPM combined with reference priors provide much more adequate uncertainty quantification.

As shown in Section 7, when the sample size is too small for the asymptotic approximations used in SPM and INLA to be valid, MCMC computations such as in SGS provide more reliable inference. Thus, we have performed a simulation study to compare the SGS algorithm with the MCMC algorithm suggested by Keefe et al. (2019). Of note, the methods in Keefe et al. (2019) use a spectral decomposition without rewriting the model in the spectral domain, and so only achieve a reduction to $O(n^2)$ computational cost. We call the algorithm suggested by Keefe et al. (2019) the Spectral Decomposition of the Precision (SDP) algorithm. When compared to the SPD algorithm, our SGS algorithm reduces the computational time by 25.48% for datasets with 49 regions and 99.95% for datasets with 3600 regions. Therefore, the reduction in computational time is substantially more important for larger spatial datasets.

Finally, we illustrate the application of the new SGS and SPM algorithms with an analysis of median household income for 3108 counties in the contiguous United States in 2017. In this analysis, we consider 8 possible socio-economic covariates and one of the objectives is to select the best model amongst the possible $2^8 = 256$ models. In a standard laptop, a full exploration of the model space with INLA takes 3267 seconds (54.45 minutes) whereas SPM takes 11.7 seconds with 8.2 seconds spent in the computation of the spectral decomposition and transformation of the data to the spectral domain, and 3.5 seconds spent in the estimation and computation of model selection criteria for all $2^8 = 256$ possible models.

The remainder of the article is organized as follows. Section 2 reviews ICAR models including specification of priors. Section 3 proves the equivalency between the sum-zero constrained ICAR prior and the improper ICAR prior with centering on the fly. Section 4 reviews the hierarchical sum-zero constrained ICAR model. Section 5 proposes the fast and scalable SGS algorithm for posterior simulation. Section 6 presents the spectral posterior maximizer (SPM) approach to accelerate computations for large sample sizes. Section 7 presents three simulation studies: the first simulation study compares computational speed in estimation of the MCMC algorithms SGS and SDP; the second simulation study compares computational speed in estimation and model selection of SPM and INLA; and the third simulation study compares statistical properties of default INLA statistical procedures with those of SGS and SPM combined with reference priors. Section 8 illustrates the use of the SGS and SPM algorithms with a spatial data analysis on county level median income in the United States. Section 9 provides a discussion and concluding remarks. For convenience of exposition, all proofs appear in the Appendix.

2. Hierarchical model specification

2.1. Model for the spatial data

We consider spatial areal data, that is, data observed on a geographical region of interest that is partitioned into n disjoint subregions. We denote these subregions by numbers $1, \ldots, n$ and consider that we have one observation for each subregion. For ease of reference, we use a similar notation to that used by

Keefe et al. (2018). Further, we assume there is a neighborhood structure where N_j is the set of subregions that are neighbors of subregion j, $j = 1, \ldots, n$. Specifically, we consider the model

$$\mathbf{y} = \alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\phi} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}$ is an n-dimensional vector that contains the observed response variable, α is an intercept, $\mathbf{1}$ is the n-dimensional vector of ones, $\mathbf{X}$ is an $n \times p$ design matrix that has one row for each subregion and one column for each regressor, and $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients. Further, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ is an n-dimensional error vector sometimes referred to as the vector of unstructured random effects. We assume that $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed $N(0, \sigma^2)$. Furthermore, $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)'$ is a vector of spatial random effects that is independent of $\boldsymbol{\epsilon}$. In Section 2.2.1 we assume that $\boldsymbol{\phi}$ is assigned an improper ICAR prior (Besag and Kooperberg, 1995); the underlying problem with such specification is that when the distribution of $\boldsymbol{\epsilon}$ is proper and that of $\boldsymbol{\phi}$ is improper, the distribution of $\mathbf{y}$ is also improper. To address that difficulty, in Section 2.2.2 we assume $\boldsymbol{\phi}$ is assigned a sum-zero constrained ICAR prior (Keefe et al., 2018, 2019).

2.2. ICAR priors for spatial random effects

Here we introduce the improper ICAR model and the sum-zero constrained ICAR model.

2.2.1. Improper ICAR model

Consider a set of uncentered spatial random effects $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)'$ over the n subregions. First, we assume that $\boldsymbol{\omega}$ follows an improper ICAR specification (Besag and Kooperberg, 1995) with a joint density that is defined up to a constant of proportionality as

$$p(\boldsymbol{\omega}) \propto \exp\left\{-\frac{\tau}{2\sigma^2}\boldsymbol{\omega}'\mathbf{H}\boldsymbol{\omega}\right\}. \tag{2}$$

Here, $\tau\sigma^{-2}\mathbf{H}$ is the precision matrix where $\tau > 0$ is a noise-to-signal ratio parameter. The matrix $\mathbf{H}$ is symmetric and positive semidefinite with elements

$$(H)_{ij} = \begin{cases} h_i, & \text{if } i = j, \\ -g_{ij}, & \text{if } i \in N_j, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Here, $g_{ij} \geq 0$ is a measure of similarity between subregions i and j. Further, by symmetry we have that $g_{ij} = g_{ji}$. Finally, the ith diagonal element is equal to the sum of the off-diagonal elements in row i, that is, $h_i = \sum_{j \neq i} g_{ij}$. For example, a widely used measure of similarity is a binary indicator equal to $g_{ij} = 1$ if subregions i and j share a border, and $g_{ij} = 0$ otherwise.

The fact that $\mathbf{H}$ is positive semidefinite results from the equality $h_i = \sum_{j \neq i} g_{ij}$. With this equality in mind, it is easy to verify that $n^{-1/2}\mathbf{1}$ is a normalized eigenvector of $\mathbf{H}$ corresponding to a null eigenvalue (Ferreira

and De Oliveira, 2007; De Oliveira and Ferreira, 2011). Further, we assume that there are no islands in the region of interest, that is, any two subregions are connected by a path. As a consequence, the null eigenvalue of $\mathbf{H}$ has multiplicity one and $\mathbf{H}$ is singular. This leads the ICAR "density" in Equation (2) to be improper, as $\int_{\mathbb{R}^n} p(\boldsymbol{\omega}) d\boldsymbol{\omega} = \infty$. When the improper ICAR specification is used as a prior for spatial random effects, practitioners implementing MCMC algorithms usually simulate the spatial random effects from a working full conditional distribution obtained by ignoring any constraint. Then, to guarantee propriety of the joint posterior distribution, practitioners usually center on the fly, recentering the simulated spatial random effects at each MCMC iteration.

2.2.2. Sum-zero constrained ICAR model

Now consider the sum-zero constrained ICAR specification proposed by Keefe et al. (2018, 2019). Let $\mathbf{K}$ be a symmetric positive semi-definite matrix such that $\mathbf{1}'\mathbf{K}\mathbf{1} > 0$, that is, the sum of the elements of $\mathbf{K}$ is positive. In addition, let $\mathbf{H}^+$ be the Moore-Penrose generalized inverse of $\mathbf{H}$. Keefe et al. (2018) have derived a sum-zero constrained ICAR model by using a limit argument in three steps. The first step considers a proper CAR model (Ferreira and De Oliveira, 2007) with a positive definite precision matrix equal to $\tau\sigma^{-2}(\kappa\mathbf{K} + \mathbf{H})$, where $\kappa > 0$ is a scalar. The second step centers the vector of proper CAR spatial random effects so that its elements sum to zero. The third and last step takes the limit when κ approaches zero to obtain the sum-zero constrained ICAR model. Keefe et al. (2018) have shown that for any symmetric positive semi-definite matrix $\mathbf{K}$ such that $\mathbf{1}'\mathbf{K}\mathbf{1} > 0$, the resulting sum-zero constrained ICAR model is the singular Gaussian distribution $N(\mathbf{0}, \tau^{-1}\sigma^2\mathbf{H}^+)$.

Singular Gaussian distributions are very useful because their mean vectors and singular covariance matrices implicitly encode linear constraints. In particular, singular Gaussian distributions can be used both as prior distributions for unknown quantities in Bayesian hierarchical models as well as distributions for data. For example, Ferreira et al. (2010, 2011) have used singular Gaussian distributions to build dynamic multiscale spatiotemporal models. Keefe et al. (2019) have used the singular Gaussian distribution $N(\mathbf{0}, \tau^{-1}\sigma^2\mathbf{H}^+)$ as a prior distribution for spatial random effects.

To elicit the linear constraints implicit in a singular Gaussian distribution, we need to consider the spectral decomposition of its singular covariance matrix. Specifically, let $\mathbf{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a singular covariance matrix with dimension n and rank $n - k$. Let $\boldsymbol{\Sigma} = \mathbf{P}\mathbf{D}\mathbf{P}'$ be the spectral decomposition of $\boldsymbol{\Sigma}$ where $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_n)$ is a matrix with columns $\mathbf{p}_1, \ldots, \mathbf{p}_n$ equal to the normalized eigenvectors of $\boldsymbol{\Sigma}$, and $\mathbf{D} = \text{diag}(d_1, \ldots, d_n)$ with $d_1 \geq \cdots \geq d_{n-k} > d_{n-k+1} = \cdots = d_n = 0$ being the respective ordered eigenvalues of $\boldsymbol{\Sigma}$. Then, for any index $i = n - k + 1, \ldots, n$ of an eigenvector corresponding to a null eigenvalue of $\boldsymbol{\Sigma}$, we have that $E(\mathbf{p}_i'\mathbf{Z}) = \mathbf{p}_i'\boldsymbol{\mu}$ and $Var(\mathbf{p}_i'\mathbf{Z}) = \mathbf{p}_i'\boldsymbol{\Sigma}\mathbf{p}_i = 0$. Thus, $P(\mathbf{p}_i'\mathbf{Z} = \mathbf{p}_i'\boldsymbol{\mu}) = 1$, $i = n - k + 1, \ldots, n$. Therefore, the singular Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ implicitly encodes k linear

constraints $\mathbf{p}_i'\mathbf{Z} = \mathbf{p}_i'\boldsymbol{\mu}, i = n-k+1, \ldots, n$. The density of the singular Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$p(\mathbf{z}) = (2\pi)^{-(n-k)/2} \left(\prod_{i=1}^{n-k} d_i \right)^{1/2} \exp\left\{ -\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{+}(\mathbf{z}-\boldsymbol{\mu}) \right\} \prod_{i=n-k+1}^{n} \mathbb{1}(\mathbf{p}_i'\mathbf{z} = \mathbf{p}_i'\boldsymbol{\mu}), \tag{4}$$

where $\boldsymbol{\Sigma}^{+}$ is the Moore-Penrose generalized inverse of $\boldsymbol{\Sigma}$ and $\mathbb{1}(.)$ is the indicator function.

Computation of probabilities for singular Gaussian distributions with density given by Equation (4) are integrals over $\mathbb{R}^{n-k}$ in the spectral domain. Consider the spectral transformation of $\mathbf{Z}$ to the spectral domain, that is, $\mathbf{R} = (R_1, \ldots, R_n)' = \mathbf{P}'\mathbf{Z}$. Then, $R_1, \ldots, R_n$ are independent such that for $i = 1, \ldots, n-k$ the distribution of R_i is univariate $N(\mathbf{p}_i'\boldsymbol{\mu}, \mathbf{p}_i'\boldsymbol{\Sigma}\mathbf{p}_i)$, and for $i = n-k+1, \ldots, n$ the distribution of R_i is degenerate $P(R_i = \mathbf{p}_i'\boldsymbol{\mu}) = 1$. Let us partition $\mathbf{R}$ into a non-degenerate part $\mathbf{R}_1 = (R_1, \ldots, R_{n-k})'$ and a degenerate part $\mathbf{R}_2 = (R_{n-k+1}, \ldots, R_n)'$. Correspondingly, partition $\mathbf{P} = (\mathbf{P}_1, \mathbf{P}_2)$ such that $\mathbf{R}_1 = \mathbf{P}_1'\mathbf{Z}$ and $\mathbf{R}_2 = \mathbf{P}_2'\mathbf{Z}$. Then, to compute $P(\mathbf{Z} \in A)$, we first find the equivalent set $A^* = \{\mathbf{R}_1 : \mathbf{Z} = \mathbf{P}\mathbf{R} \in A \text{ and } \mathbf{R}_2 = \mathbf{P}_2'\boldsymbol{\mu}\}$. Thus, $A^* \in \mathbb{R}^{n-k}$. After that, we compute $P(\mathbf{Z} \in A) = P(\mathbf{R}_1 \in A^*)$ as an integral in $\mathbb{R}^{n-k}$ with respect to the distribution of $\mathbf{R}_1 = (R_1, \ldots, R_{n-k})$. Such integrals are easily approximated by Monte Carlo methods.

Now we connect these general results for singular Gaussian distributions to the particular case of the sum-zero constrained ICAR distribution $N(\mathbf{0}, \tau^{-1}\sigma^2\mathbf{H}^{+})$. Because the rank of $\mathbf{H}^{+}$ is $n-1$, this distribution implicitly encodes one linear constraint. Further, because $n^{-1/2}\mathbf{1}$ is the normalized eigenvector of $\mathbf{H}$ corresponding to its null eigenvalue, $n^{-1/2}\mathbf{1}$ is also the eigenvector of $\mathbf{H}^{+}$ corresponding to the null eigenvalue of $\mathbf{H}^{+}$. Thus, $\boldsymbol{\phi} \sim N(\mathbf{0}, \tau^{-1}\sigma^2\mathbf{H}^{+})$ implicitly encodes the constraint $\mathbf{1}'\boldsymbol{\phi} = 0$. Finally, the sum-zero constrained ICAR prior distribution for the spatial random effects $\boldsymbol{\phi}$ has density equal to

$$p(\boldsymbol{\phi}) = (2\pi\sigma^2)^{-(n-1)/2}\tau^{(n-1)/2} \left(\prod_{i=1}^{n-1} s_i \right)^{1/2} \exp\left\{ -\frac{\tau}{2\sigma^2}\boldsymbol{\phi}'\mathbf{H}\boldsymbol{\phi} \right\} \mathbb{1}(\mathbf{1}'\boldsymbol{\phi} = 0), \tag{5}$$

where $s_1 \geq \cdots \geq s_{n-1} > s_n = 0$ are the ordered eigenvalues of $\mathbf{H}$. Note that the sum-zero constraint explicitly appears in the expression of the density in Equation (5).

Importantly, the density in Equation (5) is completely specified including the constant of proportionality. This may facilitate computation of Bayes factors for Bayesian model selection. Further, this complete specification shows a sharp distinction between the sum-zero constrained ICAR density and the improper ICAR density that appears in Equation (2), where the latter does not have a well-defined constant of proportionality.

Additionally, and again in contrast to the improper ICAR density in Equation (2), the exponent for τ/σ^2 is well-defined and equal to $(n-1)/2$ in the sum-zero constrained ICAR density in Equation (5). The fact that this exponent is equal to $(n-1)/2$ is crucial because it allows for specification of unequivocal full conditional distributions for σ^2 and τ. This fact supports the current practice in applications of the improper ICAR prior that use the exponent $(n-1)/2$ (e.g., see Hodges et al., 2003; Rue and Held, 2005; Banerjee et al., 2014). We also note that earlier publications that implemented the improper ICAR prior used a similar exponent equal to $n/2$ (Besag and Kooperberg, 1995; Besag et al., 1995). Further, Lavine and Hodges (2012)

have shown that different ways to take a limit to obtain an improper ICAR model from a proper CAR model may lead to different exponents, including the $(n-1)/2$ and $n/2$ exponents. In addition, Lavine and Hodges (2012) criticized previous attempts to obtain the exponent as being mathematically incorrect.

Fortunately, Keefe et al. (2018) proposed a formal mathematical way to obtain a sum-zero constrained ICAR model as a limit of a proper CAR model that leads to a unique distribution. The key to the construction proposed by Keefe et al. (2018) is to first project the proper CAR spatial random effects onto the subspace where the sum of the random effects is zero, and to take the limit after this projection. As shown by Keefe et al. (2018), two of the ways to approach the limit considered by Lavine and Hodges (2012) correspond to two distinct $\mathbf{K}$ matrices. While in Lavine and Hodges (2012) these two ways to approach the limit lead to two distinct exponents equal to $(n-1)/2$ and $n/2$, Keefe et al. (2018) show that if, before taking the limit, the random effects are projected onto the subspace where the sum of the random effects is zero, then the limit leads to a unique exponent equal to $(n-1)/2$.

Henceforth, consistent with the results of Keefe et al. (2018) and with existing applications of improper ICAR models with centering on the fly (Hodges et al., 2003; Rue and Held, 2005; Banerjee et al., 2014), we will assume that the constant of proportionality in the improper ICAR density in Equation (2) is proportional to $(\tau\sigma^{-2})^{(n-1)/2}$.

3. Equivalence between sum-zero constrained ICAR and improper ICAR

Next, we introduce some notation. Let $\mathbf{H} = \mathbf{QSQ}'$ be the spectral decomposition of $\mathbf{H}$ where $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n)$ is a $n \times n$ matrix with columns that are the normalized eigenvectors of $\mathbf{H}$ and $\mathbf{S} = \text{diag}(s_1, s_2, \ldots, s_n)$ where $s_1 \geq s_2 \geq \cdots \geq s_{n-1} > s_n = 0$ are the ordered eigenvalues of $\mathbf{H}$. We assume for the vector of spatial random effects the sum-zero constrained ICAR prior $\boldsymbol{\phi}|\sigma^2, \tau \sim N(\mathbf{0}, \sigma^2\tau^{-1}\mathbf{H}^+)$ proposed by Keefe et al. (2018) with density given by Equation (5). Then, the following proposition gives the full conditional distribution of $\boldsymbol{\phi}$.

Proposition 3.1. *If the prior for the vector of spatial random effects $\boldsymbol{\phi}$ is the sum-zero constrained ICAR prior $\boldsymbol{\phi}|\sigma^2, \tau \sim N(\mathbf{0}, \sigma^2\tau^{-1}\mathbf{H}^+)$, then the full conditional distribution of $\boldsymbol{\phi}$ is $\boldsymbol{\phi}|\mathbf{y}, \alpha, \boldsymbol{\beta}, \sigma^2, \tau \sim N(\mathbf{Q}^*\mathbf{m}, \sigma^2\mathbf{Q}^*\mathbf{S}^*\mathbf{Q}^{*\prime})$, where $\mathbf{Q}^* = (\mathbf{q}_1, \ldots, \mathbf{q}_{n-1})$, $\mathbf{S}^* = \text{diag}((1+\tau s_1)^{-1}, \ldots, (1+\tau s_{n-1})^{-1})$, and $\mathbf{m} = \mathbf{S}^*\mathbf{Q}^*(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.*

Now consider the case when someone applies Bayes' Theorem using the improper ICAR prior with density given by Equation (2) while ignoring the constraint $\mathbf{1}'\boldsymbol{\phi} = 0$. They obtain a working intermediate full conditional distribution for uncentered spatial random effects $\boldsymbol{\omega}$. Then they simulate $\boldsymbol{\omega}$ from this working intermediate full conditional distribution and center $\boldsymbol{\omega}$ on the fly to obtain a simulated realization of $\boldsymbol{\phi}$ that satisfies the constraint $\mathbf{1}'\boldsymbol{\phi} = 0$. The next proposition provides the resulting implied full conditional distribution for the simulated $\boldsymbol{\phi}$.

Proposition 3.2. *Assume for a vector of uncentered spatial random effects $\boldsymbol{\omega}$ the improper ICAR prior with density given by Equation (2). In addition, consider the likelihood based on the hierarchical model given in Equation (1) replacing $\boldsymbol{\phi}$ with $\boldsymbol{\omega}$. Assume that $\boldsymbol{\omega}$ is simulated from the resulting full conditional distribution and a vector of centered spatial random effects is obtained as $\boldsymbol{\phi} = (\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}')\boldsymbol{\omega}$. Then:*

1. *The full conditional distribution for $\boldsymbol{\omega}$ is $\boldsymbol{\omega}|\mathbf{y}, \alpha, \boldsymbol{\beta}, \sigma^2, \tau \sim N(\mathbf{k}, \mathbf{C})$, where $\mathbf{C} = \sigma^2(\mathbf{I} + \tau\mathbf{H})^{-1}$ and $\mathbf{k} = \mathbf{C}\sigma^{-2}(\mathbf{y} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{I} + \tau\mathbf{H})^{-1}(\mathbf{y} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta})$.*
2. *The implied full conditional distribution for $\boldsymbol{\phi}$ is $\boldsymbol{\phi}|\mathbf{y}, \alpha, \boldsymbol{\beta}, \sigma^2, \tau \sim N(\mathbf{R}\mathbf{k}, \mathbf{R}\mathbf{C}\mathbf{R}')$, where $\mathbf{R} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'$.*

The following theorem states the equivalence between the full conditional distribution for $\boldsymbol{\phi}$ obtained using the sum-zero constrained ICAR prior and the implied full conditional distribution for $\boldsymbol{\phi}$ obtained by using the improper ICAR prior and centering on the fly.

Theorem 3.1. *The full conditional distribution for $\boldsymbol{\phi}$ under the conditions of Proposition 3.1 and the implied full conditional distribution for $\boldsymbol{\phi}$ under the conditions of Proposition 3.2 are equivalent.*

Finally, we note that Ferreira (2019) showed that the singular Gaussian distribution $N(\mathbf{0}, \tau^{-1}\sigma^2\mathbf{H}^+)$ is the limiting distribution of a one-at-a-time Gibbs sampler applied to the intrinsic CAR prior in Equation (2) with centering on the fly. Unfortunately, that result was not directly applicable to posterior analyses. In contrast, by showing the equivalence of posterior analyses, Theorem 3.1 implies that methods and computations developed for sum-zero constrained ICAR models are directly applicable to ICAR-based Gaussian hierarchical models widely used in practice.

4. Hierarchical model specification

Since Gaussian hierarchical models with sum-zero constrained ICAR random effects and Gaussian hierarchical models with improper ICAR centered-on-the-fly random effects are equivalent, we can explore the well-defined sum-zero constrained ICAR distribution to accelerate computations.

We now combine the column corresponding to the intercept and the design matrix to form $\mathbf{F} = [\mathbf{1}|\mathbf{X}]$. Likewise, we expand the vector $\boldsymbol{\beta}$ to include the intercept, that is, let $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$. Then, in the original spatial domain, the hierarchical model we consider is given by

$$\mathbf{y} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\phi} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{6}$$

$$\boldsymbol{\phi} \mid \sigma^2, \tau \sim N(\mathbf{0}, \sigma^2\tau^{-1}\mathbf{H}^+). \tag{7}$$

The following two sections discuss two key decisions that need to be made for the practical application of the above hierarchical model: Section 4.1 discusses the assignment of priors for $\boldsymbol{\theta}$, σ^2, and τ; and Section 4.2 discusses the choice of the neighborhood structure that implies a choice of the matrix $\mathbf{H}$.

4.1. Assignment of priors

Bayesian analysis requires the assignment of a joint prior density for $\boldsymbol{\theta}$, σ^2, and τ. While researchers should specify proper informative priors if legitimate prior information about these parameters is available, we find that in practice many researchers will use vague priors for $\boldsymbol{\theta}$ and σ^2. In addition, most researchers (including many Bayesian statisticians) will have a difficult time specifying an informative prior for τ.

Previous literature have focused on gamma priors for precision parameters such as τ (e.g., see Bernardinelli et al., 1995; Best et al., 1999; Lee, 2013; Sørbye and Rue, 2014). In particular, Bernardinelli et al. (1995) proposed to assign priors based on the concept of fairness in the sense that the marginal variance implied for the spatial random effects be of the same magnitude as the variance of the unstructured random error. Using a similar idea, Sørbye and Rue (2014) proposed to assign priors for distinct competing neighborhood structures based on the implied geometric mean of the ICAR marginal variance. Both gamma prior proposals by Bernardinelli et al. (1995) and by Sørbye and Rue (2014) require some subjective choices, and to the best of our knowledge there are no systematic studies about the frequentist statistical properties of procedures based on these gamma priors. Such statistical properties are important if the procedures are to be used repeatedly by many different users (e.g., see Berger, 2006; Efron, 2015). Further, Keefe et al. (2019) have shown that some other gamma priors previously proposed in the literature and widely used (Best et al., 1999; Lee, 2013) may have undue strong influence in the analysis. Specifically, gamma priors for τ have exponentially decaying tails that may dominate the posterior analysis in undesirable ways leading to highly biased point estimates and credible intervals that have frequentist coverage very far from the nominal level.

Thus, for the case when no prior information is available, we recommend the use of the reference prior proposed by Keefe et al. (2019). This reference prior leads to posterior analyses with good frequentist properties such as small mean squared estimation errors and credible intervals with frequentist coverage close to nominal. In addition to their Bayesian interpretation, Bayesian credible intervals that attain frequentist coverage close to nominal can be interpreted as frequentist confidence intervals. Since frequentist properties convey the average behavior of a method when it is used multiple times, good frequentist properties are particularly important for methods that are to be used automatically by many researchers. With these considerations in mind, the reference prior proposed by Keefe et al. (2019) is implemented in the package ref.ICAR (Porter et al., 2019) that is available for the statistical programming language R (R Core Team, 2014) from the Comprehensive R Archive Network (CRAN, https://cran.r-project.org/).

To specify the reference prior, let $\mathbf{G} = \mathbf{I}_n - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'$ be the matrix that projects vectors in $\mathbb{R}^n$ onto the space orthogonal to the space spanned by the columns of $\mathbf{F}$. Further, consider the spectral decomposition $\mathbf{G} = \mathbf{MLM}'$, where $\mathbf{L}$ is a diagonal matrix with diagonal elements equal to the eigenvalues of $\mathbf{G}$ ordered in decreasing order, and $\mathbf{M}$ is a matrix with columns equal to the corresponding eigenvectors of $\mathbf{G}$. Furthermore, let $\mathbf{M}^*$ be the n by $(n-p-1)$ matrix that contains the columns of $\mathbf{M}$ corresponding to the nonzero eigenvalues of $\mathbf{G}$. Finally, let $\lambda_1 \geq \cdots \geq \lambda_{n-p-1}$ be the ordered eigenvalues of the matrix $\mathbf{M}^{*\prime}\mathbf{H}^{+}\mathbf{M}^{*}$. Then, the reference

prior proposed by Keefe et al. (2019) is given by

$$p(\boldsymbol{\theta}, \sigma^2, \tau) \propto \frac{1}{\tau\sigma^2}\left[\sum_{j=1}^{n-p-1}\left(\frac{\lambda_j}{\tau+\lambda_j}\right)^2 - \frac{1}{n-p-1}\left\{\sum_{j=1}^{n-p-1}\left(\frac{\lambda_j}{\tau+\lambda_j}\right)\right\}^2\right]^{1/2}. \tag{8}$$

We note that the reference prior for $\boldsymbol{\theta}$ is uniform on $\mathbb{R}^{p+1}$ and the prior for σ^2 is proportional to its reciprocal. In addition, the decay of the prior for τ as τ goes to ∞ is proportional to τ^{-2}. Further, as τ goes to 0 the prior for τ converges to a positive constant. Hence, the reference prior for τ is a proper density.

Computation of the eigenvalues $\lambda_1, \ldots, \lambda_{n-p-1}$ may be time consuming for large sample sizes. Thus, for those situations we propose the use of an approximate reference prior that has the same tail behavior of the reference prior given in Equation (8). Specifically, results from Keefe et al. (2019) show that the reference prior for τ behaves as $O(1)$ as $\tau \to 0$ and as $O(\tau^{-2})$ as $\tau \to \infty$. Thus, here we propose the use of an approximate reference prior of the form

$$p^{(a)}(\boldsymbol{\theta}, \sigma^2, \tau) \propto \frac{1}{\sigma^2(a_\tau + \tau)^2}, \tag{9}$$

where $a_\tau > 0$ is a hyperparameter. Our experience shows that $a_\tau = 0.5$ works well in practice.

Keefe et al. (2019) shows that analysis based on the reference prior (8) provides parameters' estimates with good statistical properties and credible intervals with correct quantification of uncertainty. Therefore, in the absence of relevant prior information, we recommend as an automatic safe choice the reference prior.

4.2. Choice of neighborhood structure

The choice of neighborhood structure is a crucial modeling decision that depends on the substantive applied problem at hand. We note that there are algorithms such as the one implemented in the function poly2nb from the R package spdep (Bivand et al., 2013) that greatly facilitate the analysis of spatial areal data by automatically obtaining a neighborhood structure from a map of the region of interest. However, we strongly believe that such automatically obtained neighborhood structures should not be used as is, but instead they should be carefully examined and revised to make sure they are adequate for answering the scientific questions at hand. In particular, automatic algorithms often provide a neighborhood structure that is not fully connected. However, our practical experience is that if the data analyst thinks carefully, then in most applied problems she/he will decide that the neighborhood structure should be fully connected.

Take for example the widely analyzed dataset on lip cancer in Scotland initially analyzed by Clayton and Kaldor (1987) and Breslow and Clayton (1993). There are two main objectives in the analysis of this dataset: to produce a smoothed map of risk, and to estimate the effect on lip cancer risk of the percentage of the work force employed in forestry, fishing, or agriculture. In this application, the spatial random effects play two main roles: to account for possible spill over effects (e.g., in case lip cancer was contagious), and to account for spatially varying regressors not included in the analysis such as population smoking habits and level of solar ultraviolet radiation. To account for spatial dependence due to regressors not included in

the analysis, it would be inadequate to define as neighbors only counties that share a border. For example, counties located in geographic islands in the north of Scotland would clearly be more related to counties in the north of mainland Scotland than to counties in the south. Accordingly, in their analyses Clayton and Kaldor (1987) and Breslow and Clayton (1993) used a fully connected neighborhood structure for the lip cancer dataset. Statistical evidence in favor of a fully connected neighborhood structure for this dataset is provided in Freni-Sterrantino et al. (2018).

Specifically, Freni-Sterrantino et al. (2018) discuss how to define and scale ICAR models for disconnected graphs. In addition, they consider the Scotland lip cancer dataset with a disconnected graph using a scaled ICAR and an unscaled ICAR, and provide for these two disconnected-graph models DICs equal to 299.4 and 298.5, respectively (Table 1, Freni-Sterrantino et al., 2018). Further, for a scaled ICAR with the connected graph considered by Clayton and Kaldor (1987) and Breslow and Clayton (1993), the DIC is equal to 297.1 (p. 32, Freni-Sterrantino et al., 2018). Thus, for the Scotland lip cancer dataset the data provides more support for the connected neighborhood than for the disconnected neighborhood.

Finally, we note that the use of an automatic statistical procedure to deal with an automatically obtained disconnected neighborhood structure could have unwarranted consequences such as spatial random effects for geographic islands being set to zero (Thomas et al., 2004) or being unduly shrank to zero (Freni-Sterrantino et al., 2018). According to our practical experience, in the vast majority of applied spatial areal data analyses the most adequate neighborhood structure will be fully connected.

5. Fast and scalable posterior simulation

In this section, we propose the SGS algorithm for the simulation from the posterior distribution of $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$, σ^2, and τ that is fast and scalable. Specifically, the SGS algorithm is a Metropolis-within-Gibbs sampler based on transforming the hierarchical model from the spatial domain to the spectral domain.

In particular, we premultiply the terms in Equation (6) by the matrix $\mathbf{Q}$ that contains the eigenvectors of $\mathbf{H}$. Let $\mathcal{Y} = \mathbf{Q}'\mathbf{y}$, $\mathcal{X} = \mathbf{Q}'\mathbf{F}$, $\boldsymbol{\xi} = \mathbf{Q}'\boldsymbol{\phi}$, $\boldsymbol{\zeta} = \mathbf{Q}'\boldsymbol{\epsilon}$, and $\mathbf{u} = (\mathbf{0}'_{n-1}, \sqrt{n})'$. Note that because $Cov(\boldsymbol{\zeta}) = Cov(\mathbf{Q}'\boldsymbol{\epsilon}) = \sigma^2\mathbf{Q}'\mathbf{Q} = \sigma^2\mathbf{I}$, then $\boldsymbol{\zeta} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. In addition, note that $Cov(\boldsymbol{\xi}) = Cov(\mathbf{Q}'\boldsymbol{\phi}) = \sigma^2\tau^{-1}\mathbf{Q}'\mathbf{H}^{+}\mathbf{Q} = \sigma^2\tau^{-1}\mathbf{S}^{+} = \sigma^2\tau^{-1}\text{diag}(s_1^{-1}, \ldots, s_{n-1}^{-1}, 0)$. Therefore, the hierarchical model given by Equations (6) and (7) can be written in the spectral domain as

$$\mathcal{Y} = \mathcal{X}\boldsymbol{\theta} + \boldsymbol{\xi} + \boldsymbol{\zeta}, \quad \boldsymbol{\zeta} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{10}$$

$$\boldsymbol{\xi} \mid \sigma^2, \tau \sim N(\mathbf{0}, \sigma^2\tau^{-1}\mathbf{S}^{+}). \tag{11}$$

Analysis based on the spectral domain model given by (10) and (11) is much simpler, faster, and more scalable than analysis based on the original spatial domain model because both the error vector $\boldsymbol{\zeta}$ and the spectral random effects $\boldsymbol{\xi}$ have diagonal covariance matrices. As a result, computations that in the original spatial domain would involve matrix multiplications with computational cost that increases quadratically

with the sample size n become Hadamard vector and matrix multiplications that have computational cost increasing linearly with n.

To develop the SGS algorithm, we first integrate out the spectral random effects $\boldsymbol{\xi}$ from the model. We do this because $\boldsymbol{\xi}$ is highly correlated *a posteriori* with σ^2 and τ. Thus, we find that integrating out $\boldsymbol{\xi}$ analytically allows us to develop an MCMC algorithm that converges much faster. Integrating out $\boldsymbol{\xi}$ we obtain the model in the spectral domain:

$$\mathcal{Y} = \mathcal{X}\boldsymbol{\theta} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim N(\mathbf{0}, \sigma^2\{\mathbf{I} + \tau^{-1}\mathbf{S}^+\}). \tag{12}$$

Our spectral Metropolis-within-Gibbs sampler includes a Gibbs step for $\boldsymbol{\theta}$, as well as a joint Metropolis-Hastings step for τ and σ^2. From a posterior sample of these parameters, a posterior sample of the spectral random effects $\boldsymbol{\xi}$ can be easily obtained by composite sampling, and then a posterior sample of the spatial random effects $\boldsymbol{\phi}$ can be obtained using the expression $\boldsymbol{\phi} = \mathbf{Q}\boldsymbol{\xi}$.

Let $\odot$ denote the Hadamard product (p. 45, Magnus and Neudecker, 1999) that returns the matrix of element-wise products. Further, let $\otimes$ be the Kronecker product. In addition, let

$$\mathbf{b}(\tau) = \left(\frac{\tau s_1}{\tau s_1 + 1}, \cdots, \frac{\tau s_{n-1}}{\tau s_{n-1} + 1}, 1\right)'. \tag{13}$$

Note that $\mathcal{X}'(\mathbf{I} + \tau^{-1}\mathbf{S}^+)^{-1} = \left\{\mathcal{X} \odot (\mathbf{1}_p' \otimes \mathbf{b}(\tau))\right\}'$, where the operation on the left costs about $2pn^2$ products and $2pn^2$ sums, whereas the operation on the right is much faster and costs just $2np$ products.

Then, one iteration of our SGS algorithm proceeds as follows.

1. Simulate $\boldsymbol{\theta}$ from its full conditional distribution $N(\mathbf{m}_\theta, \mathbf{C}_\theta)$, where

$$\mathbf{C}_\theta = \sigma^2\{\mathcal{X}'(\mathbf{I} + \tau^{-1}\mathbf{S}^+)^{-1}\mathcal{X}\}^{-1} = \sigma^2\left[\left\{\mathcal{X} \odot (\mathbf{1}_p' \otimes \mathbf{b}(\tau))\right\}'\mathcal{X}\right]^{-1},$$

   and

$$\begin{aligned} \mathbf{m}_\theta &= \left\{\mathcal{X}'(\mathbf{I} + \tau^{-1}\mathbf{S}^+)^{-1}\mathcal{X}\right\}^{-1}\mathcal{X}'(\mathbf{I} + \tau^{-1}\mathbf{S}^+)^{-1}\mathcal{Y} && (14)\\ &= \left[\left\{\mathcal{X} \odot (\mathbf{1}_p' \otimes \mathbf{b}(\tau))\right\}'\mathcal{X}\right]^{-1}\left\{\mathcal{X} \odot (\mathbf{1}_p' \otimes \mathbf{b}(\tau))\right\}'\mathcal{Y}. && (15)\end{aligned}$$

2. Propose $\sigma^{2(prop)}$ from Lognormal$(\log(\sigma^{2(curr)}), \delta_\sigma)$.
3. Propose $\tau^{(prop)}$ from Lognormal$(\log(\tau^{(curr)}), \delta_\tau)$.
4. Accept $(\sigma^{2(prop)}, \tau^{(prop)})$ with probability equal to $\min(1, a)$ where

$$\begin{aligned} a &= \left(\frac{\sigma^{2(prop)}}{\sigma^{2(current)}}\right)^{-(n+2)/2}\left(\frac{\tau^{(prop)}}{\tau^{(curr)}}\right)^{(n-1)/2}\left(\frac{\prod_{i=1}^{n-1}(s_i\tau^{(curr)} + 1)}{\prod_{i=1}^{n-1}(s_i\tau^{(prop)} + 1)}\right)^{1/2}\frac{p(\tau^{(prop)})}{p(\tau^{(curr)})}\\ &\quad \times \exp\left[-\frac{1}{2}\left\{(\mathcal{Y} - \mathcal{X}\boldsymbol{\theta}) \odot \left(\sigma^{-2(prop)}\mathbf{b}(\tau^{(prop)}) - \sigma^{-2(curr)}\mathbf{b}(\tau^{(curr)})\right)\right\}'(\mathcal{Y} - \mathcal{X}\boldsymbol{\theta})\right]. \end{aligned}$$

We note that when using the reference prior analysis proposed by Keefe et al. (2019), one needs to compute the spectral decomposition of the precision matrix $\mathbf{H}$. Thus, the use of the reference prior and the

use of our SGS algorithm are synergistic. Further, we note that because the vector of spectral random effects $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)' = \mathbf{Q}'\boldsymbol{\phi}$ has both prior and posterior covariance matrices that are diagonal, simulation of $\xi_1, \ldots, \xi_n$ can be done very fast and be implemented in parallel.

The SGS and the SDP algorithms have many things in common, and one crucially important difference that makes the SGS algorithm much faster. Specifically, the SDP algorithm computes the spectral decomposition $\mathbf{H} = \mathbf{QSQ}'$ of the $\mathbf{H}$ matrix before the MCMC iterations. After that, the SDP algorithm uses the identity $(\mathbf{I}+\tau^{-1}\mathbf{H}^{+})^{-1} = \mathbf{Q}(\mathbf{I}+\tau^{-1}\mathbf{S}^{+})^{-1}\mathbf{Q}'$ to perform matrix inversions of $\mathbf{I}+\tau^{-1}\mathbf{H}^{+}$ within the MCMC iterations, where the inversion of the diagonal matrix $\mathbf{I}+\tau^{-1}\mathbf{S}^{+}$ can be performed with $O(n)$ operations. Thus, instead of the usual $O(n^3)$ operations needed to invert a matrix, in the SDP algorithm the inversion of $\mathbf{I}+\tau^{-1}\mathbf{H}^{+}$ costs $O(n^2)$ operations per MCMC iteration due to matrix multiplications.

In contrast, we have realized that we can compute equivalent matrix multiplications before the start of the MCMC algorithm. These equivalent matrix multiplications are the spectral transformations of the dependent variable and of the regressors that the SGS algorithm computes before the MCMC iterations. As a result, in the SGS algorithm the corresponding matrix inversions within each MCMC iteration cost $O(n)$ operations. Thus, the main difference between the SGS and the SDP algorithms is that the MCMC iterations in the SDP algorithm are applied to the original variables whereas the SGS algorithm first spectral transforms the variables and then runs specialized MCMC iterations for the spectral transformed variables. All other aspects of the SDP and SGS algorithms are the same; so much so that if you start the two algorithms with the same pseudo-random seed for the pseudo-random numbers generator, the posterior samples generated by the two algorithms will be exactly the same.

6. Computation acceleration for large sample sizes

For large sample sizes, further computational speed ups may be achieved through the use of large sample approximations. In particular, when the sample size is large enough for the use of an asymptotic normality approximation to the joint posterior density of the unknown parameters, we may estimate the parameters using the maximum a posteriori (MAP) (that is, the posterior mode) and we may compute credible intervals using the MAP and the Fisher information matrix. These computations require optimization of the posterior density, which can be performed orders of magnitude faster than MCMC algorithms. We call this the spectral posterior maximizer (SPM) approach.

The computational acceleration through spectral transformations that we propose is particularly useful for model selection problems. For those problems, after the spectral decomposition of the matrix $\mathbf{H}$ is performed and the data are transformed onto the spectral domain, no additional matrix decompositions need to be performed. The computations for each model can then be performed in the spectral domain and grow linearly with sample size. In contrast, the currently fastest state-of-the-art computational tool for hierarchical models with CAR priors, implemented in the R package INLA (Rue et al., 2009; Martins et al., 2013), is based on fast matrix decompositions that, nevertheless, have to be performed at each iteration of

an optimization procedure. Thus, in INLA the matrix decompositions, albeit fast, need to be performed multiple times for each fitted model.

Take for example a problem where the researcher is interested in fitting a hierarchical model with ICAR random effects to data with one dependent variable and 10 regressors from all counties in the United States. There are over 3000 such counties. With 10 regressors, there are $2^{10} = 1024$ possible models. As we show in a simulation study in Section 7, in such an example our proposed approach computes information criteria such as AIC and/or DIC for all 1024 models very fast. Specifically, in a MacBook Pro with a 2.7 GHz Intel Core i7 processor and MacOS Mojave operating system, with computations implemented in R version 4.0.2 optimized with Intel's Math Kernel Library, INLA version 20.03.17 takes about 5 hours while SPM takes about 30 seconds.

Additional acceleration of computations can be obtained by using the fact that for large sample sizes the posterior density may be well approximated by a Gaussian density. In addition, further speed up can be obtained by the use of the approximate reference prior (9). In that case, the joint posterior density of $(\boldsymbol{\theta}, \sigma^2, \tau)$ is $p(\boldsymbol{\theta}, \sigma^2, \tau|\mathcal{Y}, \mathcal{X}) \propto p^{(a)}(\boldsymbol{\theta}, \sigma^2, \tau)p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}, \sigma^2, \tau)$.

Further, we note that because σ^2 and τ are positive quantities, convergence to normality will happen faster (that is, for smaller sample sizes) for the corresponding logarithm reparameterizations $\gamma = \log \sigma^2$ and $\psi = \log \tau$. Multiplying the approximate reference prior (9) by the Jacobian of the transformation equal to $e^{\gamma}e^{\psi}$, in this new parameterization the approximate reference prior becomes

$$p^{(a)}(\boldsymbol{\theta}, \gamma, \psi) \propto \frac{e^{\psi}}{(a_\tau + e^{\psi})^2}. \tag{16}$$

Thus, we estimate $(\boldsymbol{\theta}, \gamma, \psi)$ with the posterior mode

$$(\widehat{\boldsymbol{\theta}}, \widehat{\gamma}, \widehat{\psi}) = \max_{(\boldsymbol{\theta},\gamma,\psi)} \log[p^{(a)}(\boldsymbol{\theta}, \gamma, \psi)p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \gamma, \psi)] = \max_{(\boldsymbol{\theta},\gamma,\psi)} \log[p^{(a)}(\boldsymbol{\theta}, \gamma, \psi)p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}, \gamma, \psi)]. \tag{17}$$

Specifically, for a given ψ the posterior mode of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}}(\psi) = \left[\left\{\mathcal{X} \odot (\mathbf{1}_p' \otimes \mathbf{b}(e^{\psi}))\right\}' \mathcal{X}\right]^{-1} \left\{\mathcal{X} \odot (\mathbf{1}_p' \otimes \mathbf{b}(e^{\psi}))\right\}' \mathcal{Y}. \tag{18}$$

where $\mathbf{b}(\tau)$ is defined in Equation (13). In addition, for given $(\boldsymbol{\theta}, \psi)$ the posterior mode of γ is

$$\widehat{\gamma}(\boldsymbol{\theta}, \psi) = -\log(n) + \log\left\{\sum_{i=1}^{n-1} \frac{(\mathcal{Y}_i - \mathcal{X}_i'\boldsymbol{\theta})^2}{1 + e^{-\psi}s_i^{-1}} + (\mathcal{Y}_n - \mathcal{X}_n'\boldsymbol{\theta})^2\right\}. \tag{19}$$

Thus, we use Equations (18) and (19) to write the joint posterior density of $(\boldsymbol{\theta}, \gamma, \psi)$ as a function of ψ only. After that, we use a one-dimensional numerical maximizer to find the MAP $\widehat{\psi}$ of ψ, and then recursively substitute ψ by $\widehat{\psi}$ in Equations (18) and (19) to find the MAPs $\widehat{\boldsymbol{\theta}}$ and $\widehat{\gamma}$. Finally, the corresponding estimates of σ^2 and τ are $\widehat{\sigma}^2 = e^{\widehat{\gamma}}$ and $\widehat{\tau} = e^{\widehat{\psi}}$.

For large samples, uncertainty quantification may be performed with the posterior information matrix. Let $\eta(\psi, c) = \sum_{i=1}^{n-1}(s_i e^{\psi} + 1)^{-c}$. For the parameterization $(\boldsymbol{\theta}, \gamma, \psi)$, the posterior information matrix for

the model in the spectral domain given in Equation (12) is

$$I(\boldsymbol{\theta},\gamma,\psi)=\begin{pmatrix} e^{-\gamma}\left\{\mathcal{X}\odot(\mathbf{1}_p'\otimes\mathbf{b}(\tau))\right\}'\mathcal{X} & \mathbf{0}_p & \mathbf{0}_p \\ \mathbf{0}_p' & \frac{n}{2} & -\frac{1}{2}\eta(\psi,1) \\ \mathbf{0}_p' & -\frac{1}{2}\eta(\psi,1) & \frac{1}{2}\eta(\psi,2)+\frac{2a_\tau e^{\psi}}{(a_\tau+e^{\psi})^2} \end{pmatrix}. \tag{20}$$

Hence, the asymptotic posterior covariance matrix of $(\boldsymbol{\theta},\gamma,\psi)$ based on the approximate reference prior is the inverse of the posterior information matrix given by

$$Cov_{asymp}(\boldsymbol{\theta},\gamma,\psi)=\begin{pmatrix} e^{\gamma}\left[\left\{\mathcal{X}\odot(\mathbf{1}_p'\otimes\mathbf{b}(\tau))\right\}'\mathcal{X}\right]^{-1} & \mathbf{0}_{p,2} \\ \mathbf{0}_{2,p} & Cov_{asymp}(\gamma,\psi) \end{pmatrix}, \tag{21}$$

where $\mathbf{0}_{p,q}$ is the $p\times q$ matrix of zeros and

$$Cov_{asymp}(\gamma,\psi)=2\left[n\eta(\psi,2)+\frac{4na_\tau e^{\psi}}{(a_\tau+e^{\psi})^2}-\{\eta(\psi,1)\}^2\right]^{-1}\begin{pmatrix} \eta(\psi,2)+\frac{4a_\tau e^{\psi}}{(a_\tau+e^{\psi})^2} & \eta(\psi,1) \\ \eta(\psi,1) & n \end{pmatrix}. \tag{22}$$

Hence, asymptotic credible intervals for elements of $\boldsymbol{\theta}$ as well as for γ and ψ may be trivially computed from the asymptotic covariance matrix. In addition, corresponding credible intervals for σ^2 and τ can be computed in a straightforward manner by exponentiating the limits of the credible intervals for γ and ψ, respectively. Section 8 illustrates the use of these asymptotic approximations with a real dataset.

7. Simulation studies

This section presents simulation studies to compare the computational cost as well as statistical properties of the SGS and SPM algorithms with competing algorithms.

7.1. Computational time

In this section we present two simulation studies: the first simulation study compares computational speed in estimation of the MCMC algorithms SGS and SDP; the second simulation study compares computational speed in estimation and model selection of SPM and INLA. In the two simulation studies, we consider regular square grids with sizes: 49, 100, 400, 900, 1600, 2500, and 3600. The spatial random effects follow a sum-zero constrained ICAR specification with first-order neighborhood structure. All computations have been performed in R version 4.0.2 optimized with Intel's Math Kernel Library running on a MacBook Pro with a 2.7 GHz Intel Core i7 processor and MacOS Mojave operating system. Finally, we note that INLA is an R package with core computations implemented in C and Fortran, whereas we have implemented the SGS, SPM, and SDP algorithms exclusively in R.

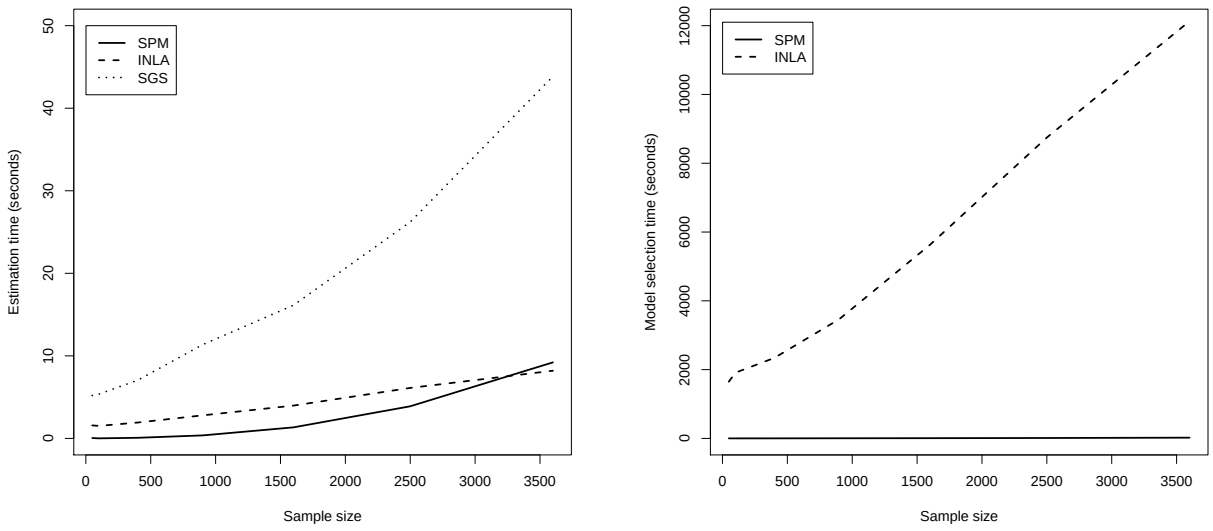
For estimation with MCMC, the computational time depends mostly on the sample size. Thus, in the first simulation study we have considered for the regression part of the model an intercept and one explanatory variable simulated independently and identically distributed from the standard normal distribution. The values of the parameters in the first simulation study are $\tau=1$, $\sigma^2=2$, and $\boldsymbol{\beta}=(1,5)'$. We note that

Table 1: Computational time in seconds of each MCMC method to run 15000 iterations.

Grid size	**SGS**	**SDP**
49	5.00	6.71
100	4.91	17.86
400	7.43	242.60
900	11.46	1,531.03
1600	15.70	6,824.96
2500	25.05	24,268.77
3600	42.47	102,639.10

often two competing MCMC algorithms will have very different computational costs per MCMC iteration but, due to their distinct autocorrelation functions for the traces of the simulated parameters, they will also have distinct effective sample sizes. However, as explained at the end of Section 5, the difference between the SGS and SPM algorithms is that the SPM algorithm is applied to the original variables whereas the SGS algorithm is applied to the spectral transformed variables. As a result, for a given number of iterations the SGS and SPM algorithms will have the exact same effective sample size. Hence, we compare the SGS and SPM algorithms in terms of computational time. Table 1 presents the computational time in seconds of the SGS and SDP algorithms to run 15000 MCMC iterations. When compared to the SDP algorithm, our novel SGS algorithm produces a substantial decrease in computational time. For sample sizes from 49 to 3600 subregions computational times vary respectively from 6.71 seconds to 28.5 hours for the SDP algorithm and from 4.91 seconds to 42.47 seconds for the SGS algorithm. Therefore, when compared to the SDP algorithm, our SGS algorithm provides substantial speed-ups.

In addition, as discussed in Section 6, for larger sample sizes approximations based on Gaussian approximations such as SPM and INLA may provide accelerated computations. Figure 1 presents computational time of SPM, SGS, and INLA for sample sizes varying from 49 to 3600 for hierarchical models with 10 possible regressors, with Panel (a) presenting computational time for estimation and Panel (b) presenting computational time for model selection with full model space search. All INLA computations presented here have used INLA version 20.03.17. For estimation, computational times of SPM and INLA are comparable whereas SGS is somehow slower. For model selection with full model space search, SGS is not competitive, thus Panel (b) of Figure 1 presents results only for SPM and INLA. As pointed out in Section 6, while in SPM the spectral decomposition needs to be computed only once and then used for all models, in INLA the matrix decomposition has to be computed at each iteration of optimization procedures and for each model. This difference in numerical matrix algebra computations shows up remarkably in Panel (b) of Figure 1. Specifically, while model selection computations based on INLA increase from 27.5 minutes for $n = 49$ to 3.4 hours for $n = 3600$, computations for model selection based on SPM take from 0.9 seconds for $n = 49$ to



(a) (b)

Figure 1: Computational time of SPM, INLA, and SGS for different sample sizes for hierarchical model with 10 regressors: (a) Estimation; (b) Model selection with full model space search.

21.8 seconds for $n = 3600$. Therefore, from a computational time point of view for model selection SPM is the preferred algorithm.

7.2. Statistical properties

In this section we present a simulation study that compares properties of the statistical procedures implemented in INLA, SGS, and SPM. Because these methods are to be used by many users, we consider four frequentist properties: bias, mean squared error (MSE), frequentist coverage and mean width of 95% credible intervals. We consider these frequentist properties for the variance of the error σ^2, the variance of the spatial random effect σ^2/τ, the signal-to-noise ratio τ, the regression coefficient β_j and the intercept β_0.

We consider square grids with sample sizes equal to 49, 100, 400, 900, 1600, 2500, and 3600. We assume a first-order neighborhood structure. In all settings, we assume $\beta = (1, 2, 5)'$ and the explanatory variables are a random sample from the standard Gaussian distribution. We note that default methods used by a wide range of researchers should be invariant to the measurement units used for the dependent and explanatory variables. To partially check for invariance to measurement units, we consider two values for σ^2: 1 and 100. Finally, our experience shows that datasets with moderate to strong spatial dependence have estimated values of τ between 0.1 and 0.5. Based on that information, we consider for τ values that range from strong spatial dependence to practically independent observations: 0.1, 0.5, 1, and 10.

We have performed this simulation study from the perspective of a usual practicioner who performs her/his analysis with the default choices implemented in statistical R packages. The prior implemented in

SGS is the reference prior proposed by Keefe et al. (2019) given in Equation (8). The default prior in INLA for the parameters of the hierarchical model given in Equation (6) and (7) is a noninformative uniform prior for the intercept, independent normal priors with mean 0 and variance 1000 for the regression coefficients, and gamma priors with parameters 1 and 5×10^{-5} (that is, prior mean and prior standard deviation equal to 0.2×10^5) for σ^{-2} and for $\tau\sigma^{-2}$ (Rue, 2021). Finally, the prior implemented in SPM is the approximate reference prior given in Equation (9) with hyperparameter $a_\tau = 0.5$ which works well for the settings of this simulation study. For reasons of space, in this manuscript we focus on MSE, frequentist coverage and mean width of 95% credible intervals for σ^2, σ^2/τ, and β_j for sample sizes 49, 400, 1600, and 3600, for the case when $\sigma^2 = 100$. Plots for all parameters, frequentist properties, and sample sizes can be found in the supplementary material.

Figures 2, 3, and 4 present plots of $\sqrt{\text{MSE}}$, frequentist coverage and mean width of 95% credible intervals as functions of the true value of $\log_{10}(\tau)$ for σ^2, σ^2/τ, and β_j. Black lines represent SGS, red lines represent INLA, and blue lines represent SPM. In terms of $\sqrt{\text{MSE}}$, methods SGS and SPM are fairly close to each other, with SPM having a smaller $\sqrt{\text{MSE}}$ for the estimation of the variances σ^2 and σ^2/τ. Meanwhile, INLA has much higher $\sqrt{\text{MSE}}$ for the estimation of σ^2 and σ^2/τ, pointing to inappropriate uncertainty quantification. This incorrect estimation of variances leads INLA to have higher $\sqrt{\text{MSE}}$ than SGS and SPM in the estimation of the regression coefficients β_j.

With respect to frequentist properties of 95% credible intervals, SGS is the safest method providing coverage close to nominal for all parameters under all settings considered. SPM provides credible interval with coverage close to nominal for small values of τ that correspond to stronger spatial dependence, and the coverage deteriorates for cases of smaller sample sizes combined with larger values of τ. But the coverage of SPM credible intervals improves with larger sample sizes and is close to nominal for the three considered parameters for sample sizes larger or equal to $n = 400$ and moderate to strong spatial dependence with $\tau < 1$. In contrast, INLA 95% credible intervals have very low coverage for the variances σ^2 and σ^2/τ, with coverage close to zero for the smallest sample size $n = 49$. Even for sample size $n = 3600$, INLA 95% credible intervals have lower coverage than nominal for σ^2 and σ^2/τ. With respect to the regression coefficient β_j, all three methods provide credible intervals with frequentist coverage close to nominal. However, for smaller sample sizes INLA credible intervals for β_j are on average much wider than SGS and SPM credible intervals.

Based on these results, we recommend the use of SGS for the analysis of smaller spatial data with sample size less than 400 observations. And we recommend the use of SPM for larger sample sizes. Recall that SGS is based on the reference prior (8) and SPM is based on an approximate reference prior (9). The appropriate uncertainty quantification of SGS and SPM concurs with previously published results for objective Bayesian analysis of other inferential problems (e.g., see Berger et al., 2001; Ferreira and De Oliveira, 2007; Sun and Berger, 2007; Ferreira and Suchard, 2008; Fonseca et al., 2008; Salazar et al., 2012; Fonseca et al., 2012; Ferreira and Salazar, 2014; Keefe et al., 2019; He et al., 2021). Therefore, for the analysis of the models considered here, we recommend the use of either the reference prior (8) or the approximate reference prior (9).

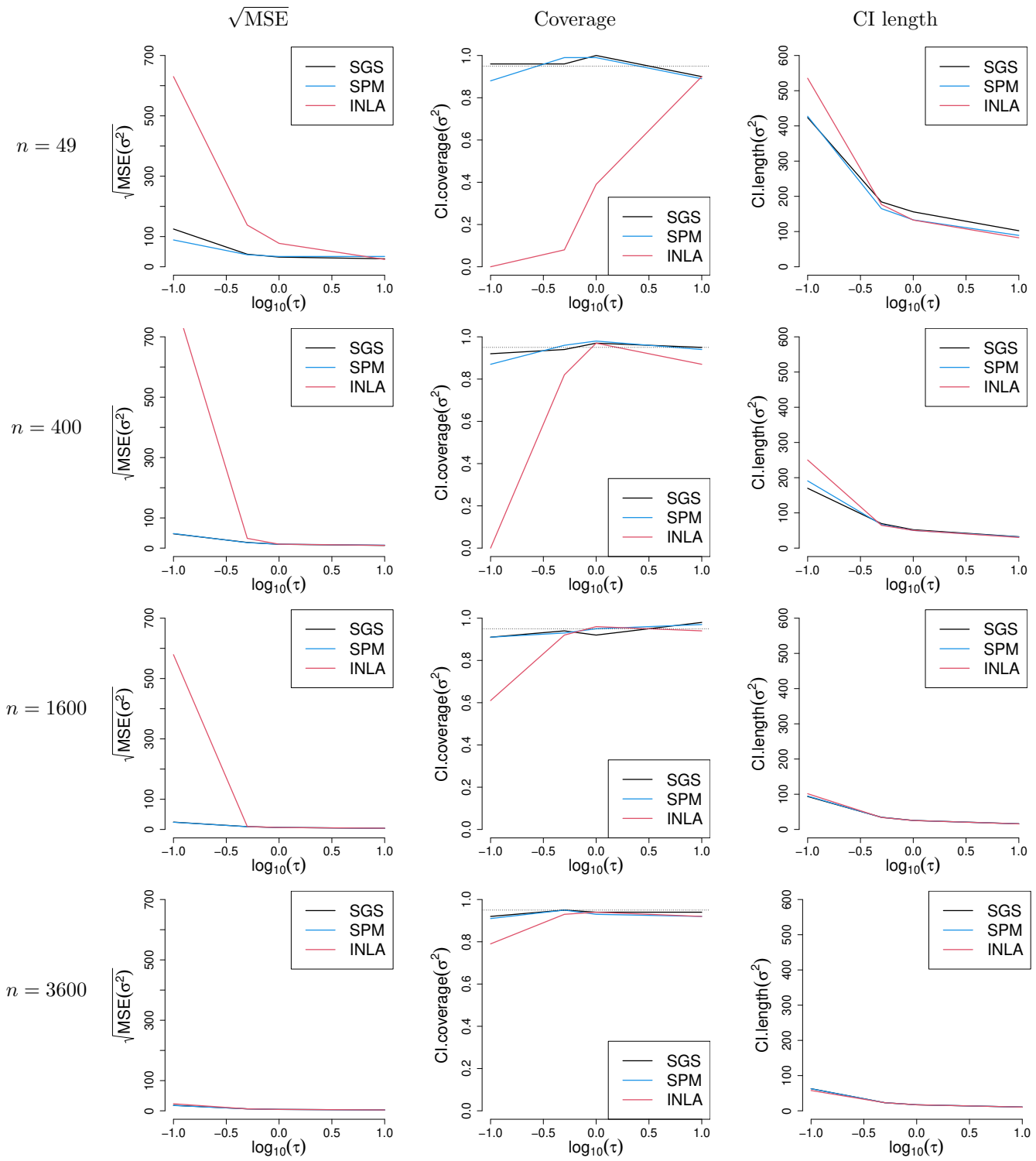


Figure 2: Square root MSE of $\hat{\sigma}^2$, as well as frequentist coverage and mean length of 95% credible intervals for σ^2 as a function of true value of $\log(\tau)$ for true value of $\sigma^2 = 100$. Frequentist coverage close to or above the nominal 0.95 indicates appropriate quantification of uncertainty. If two methods have frequentist coverage above nominal level, then the method with shorter credible intervals is preferable.

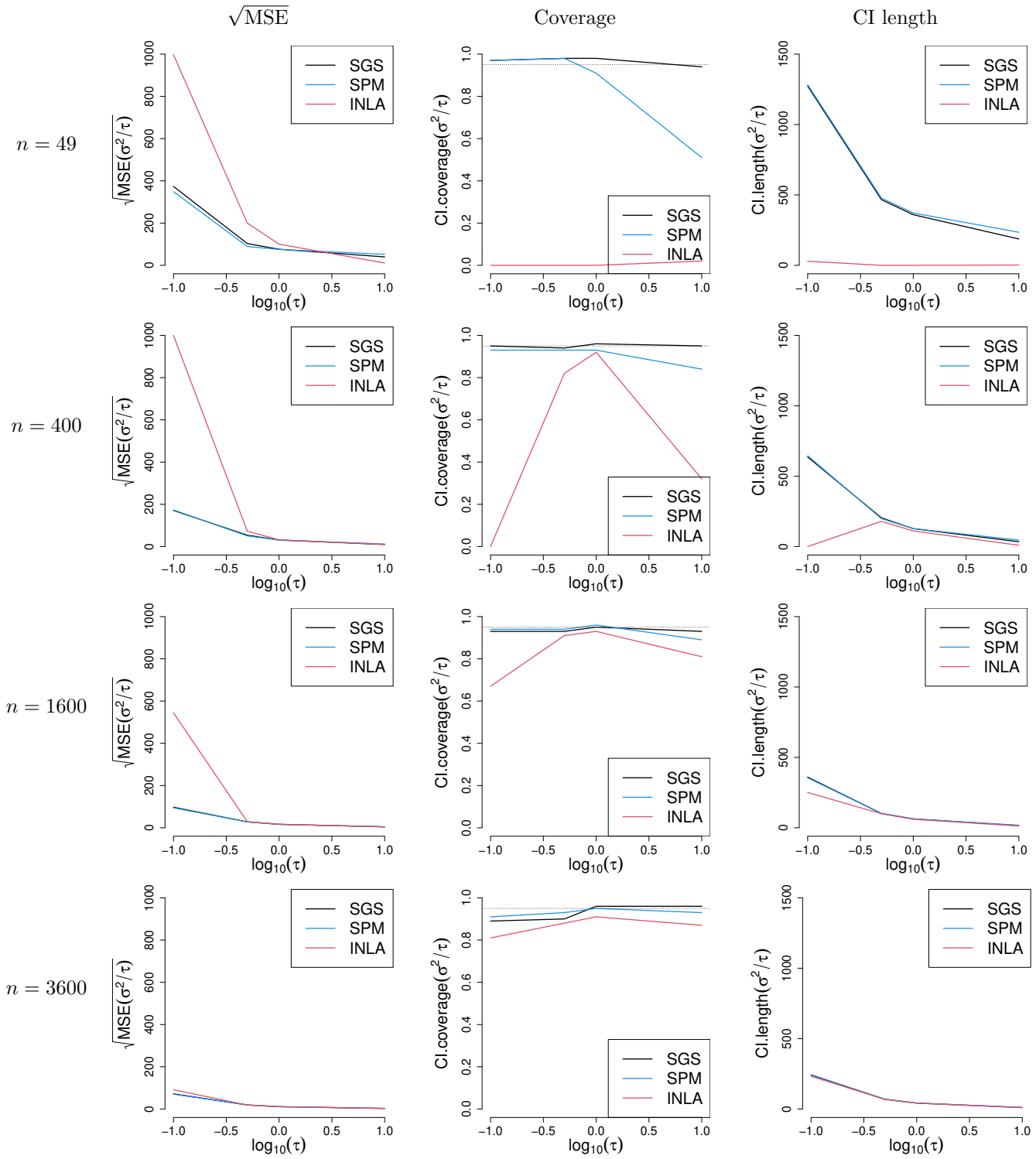


Figure 3: Square root MSE of $\widehat{\sigma^2/tau}$, as well as frequentist coverage and mean length of 95% credible intervals for σ^2/tau as a function of true value of $\log(\tau)$ for true value of $\sigma^2 = 100$. Frequentist coverage close to or above the nominal 0.95 indicates appropriate quantification of uncertainty. If two methods have frequentist coverage above nominal level, then the method with shorter credible intervals is preferable.

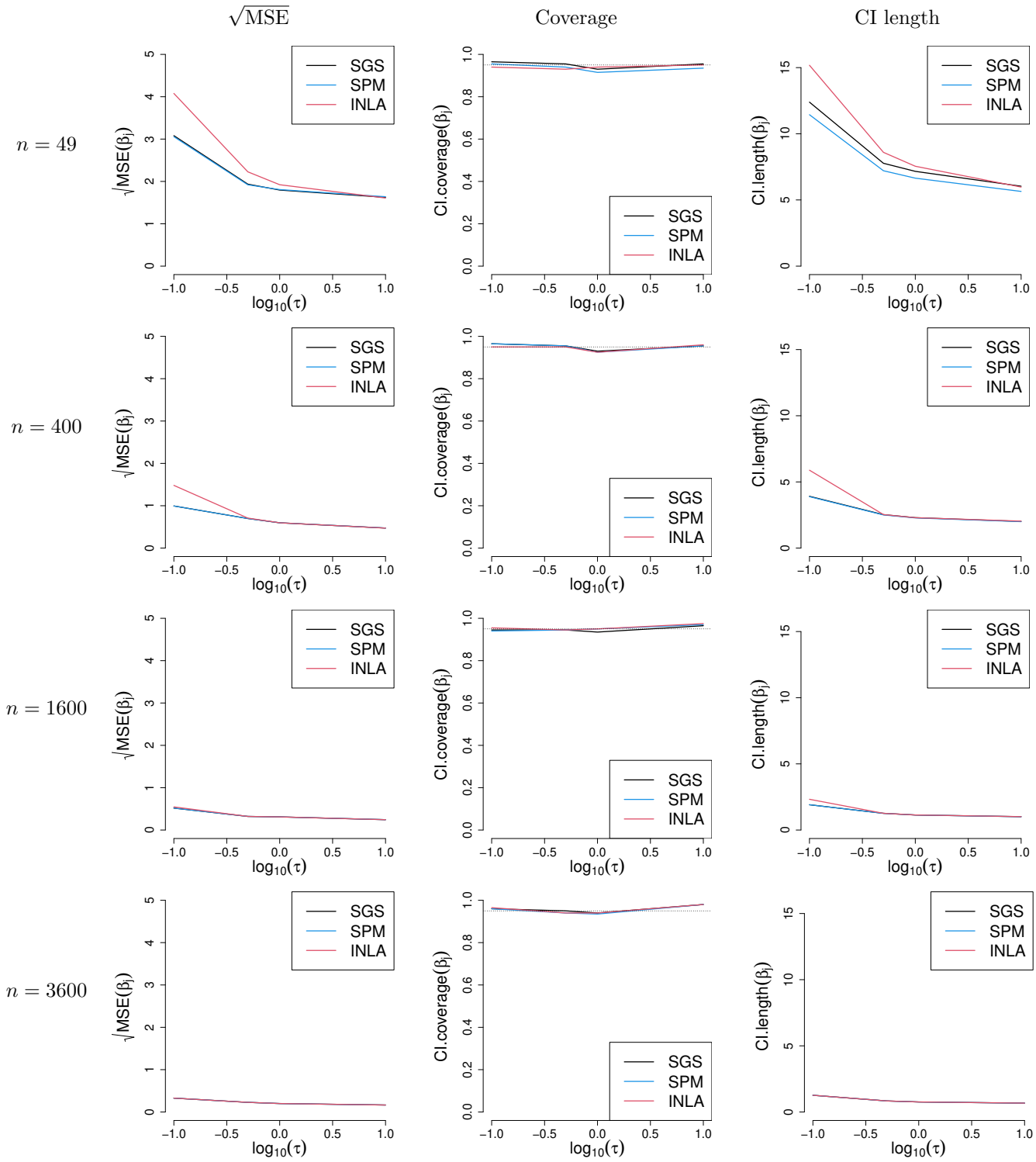


Figure 4: Square root MSE of $\widehat{\beta}_j$, as well as frequentist coverage and mean length of 95% credible intervals for β_j as a function of true value of $\log(\tau)$ for true value of $\sigma^2 = 100$. Frequentist coverage close to or above the nominal 0.95 indicates appropriate quantification of uncertainty. If two methods have frequentist coverage above nominal level, then the method with shorter credible intervals is preferable.

8. Application: Household income in the United States

To illustrate the possibilities of our novel SGS algorithm to handle large spatial datasets, we analyze median household income in the contiguous United States in 2017 per county for a total of 3108 counties (or similar geopolitical entities). The data were downloaded on May 1, 2019, from the website of the Economic Research Service of the United States Department of Agriculture.

Specifically, we consider the logarithm of the median household income per county as the dependent variable in the hierarchical model given by Equations (6) and (7). In addition, we consider 8 regressors: logarithm of the county population in 2017; logarithm of unemployment rate 2017; three indicator variables for whether the county belongs to a large metropolitan area, a medium metropolitan area, or a small metropolitan area; and three level of education variables: logarithm of percent adults with less than high school education, with a high school degree, and with a bachelor's degree or higher. Thus, the intercept corresponds to a baseline of a non-metropolitan area and a level of education of some bachelor's or associate degree.

Panels (a) through (e) of Figure 5 present maps of the United States with data per county on the logarithm scale of median household income, percent of adults with a bachelor's degree or higher, population size, unemployment, and metro areas. Each of the variables considered presents some level of spatial dependence. In addition, there seems to be correlation between the dependent variable and the regressors; however this dependence is somehow difficult to determine in the face of the spatial dependence. Scatterplots (not shown) in the logarithm scale of the dependent variable versus each of the regressor variables indicate that it is reasonable to assume linear relationships after logarithmic transformation. Finally, a linear model without spatial dependence applied to the logarithm-transformed variables results in the residuals mapped in panel (f) of Figure 5. Visual inspection of the map of residuals and a Moran-I test (Moran-I statistic = 0.39, p-value $< 10^{-15}$) indicate presence of spatial dependence.

Thus, to account for spatial dependence we consider the hierarchical linear model with ICAR random effects to analyze these data. We have performed the analysis with our SGS and SPM algorithms, as well as INLA, implemented in R version 4.0.2 optimized with Intel's Math Kernel Library running on a MacBook Pro with a 2.7 GHz Intel Core i7 processor and MacOS Mojave operating system. Specifically, 15000 iterations of the SGS algorithm including the initial time to compute the eigenvalue decomposition of the matrix $\mathbf{H}$ and without the simulation of the spatial random effects take 34.59 seconds. If we include the simulation of the spatial random effects then the total time is 42.82 seconds. In addition, to estimate the parameters of the model with all regressors, INLA and SPM take 8.55 seconds and 3.74 seconds, respectively.

For model selection, INLA provides the Widely Applicable Information Criterion (WAIC) (Watanabe, 2010) and the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). In addition, Celeux et al. (2006) proposed several different definitions of the DIC for models with latent variables which in the case considered here are the spatial random effects. Here, we use SPM to compute the type 2 DIC proposed by Celeux et al. (2006). The type 2 DIC is based on the integrated likelihood with the spatial random

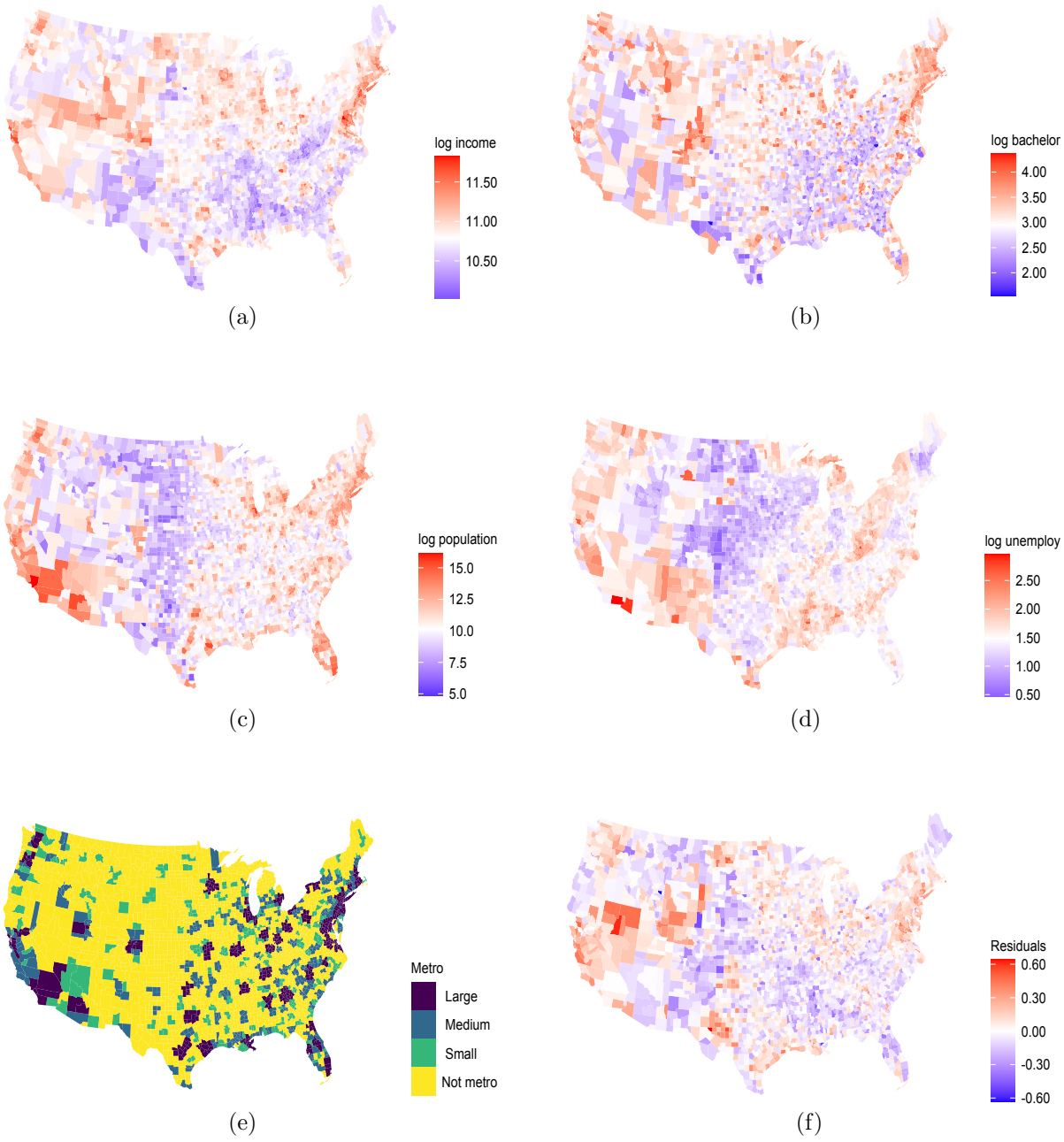


Figure 5: United States socioeconomic data by county: (a) logarithm of median household income; (b) logarithm of percent of adults with a bachelor's degree or higher; (c) logarithm of population; (d) logarithm of unemployment; (e) metro areas; (f) residuals of linear model without spatial random effects.

effects integrated out and thus it is preferable to the DIC that is computed with estimated spatial random effects Celeux et al. (2006). The WAIC computed by INLA selects a model with two regressors: logarithm of percent adults with less than high school education and the indicator variable that the county belongs to a medium sized metro area. Meanwhile, the DIC computed by INLA selects a model with only one regressor: the indicator variable that the county belongs to a medium sized metro area. In contrast, the DIC type 2 computed by SPM chooses the full model with all 8 regressors. In terms of computational time, a full search of the model space with $2^8 = 256$ possible models with INLA takes 2157.7 seconds (35.96 minutes). In contrast, for the same task SPM takes a total of 6.74 seconds, with 4.1 seconds spent in the computation of the spectral decomposition of $\mathbf{H}$ and transformation of the data to the spectral domain, and 2.64 seconds spent in the estimation and computation of model selection criteria for all $2^8 = 256$ possible models. Therefore, in this application SPM is 320 times faster than INLA.

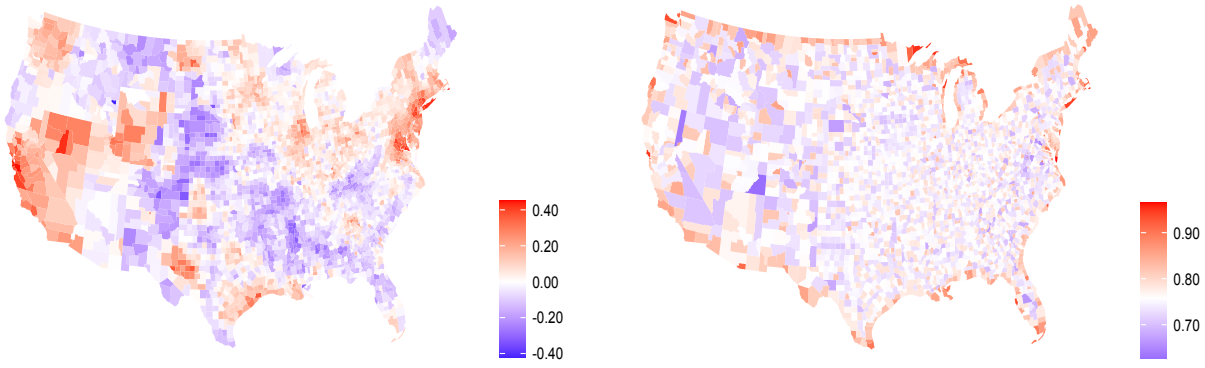
Table 2 presents posterior summaries for the parameters of the full hierarchical model computed both with SGS and SPM. Results for SGS are based on two chains, started from different starting values, with 15,000 iterations each and burnin of 1,000 iterations; the Gelman-Rubin convergence diagnostic (Gelman and Rubin, 1992) implemented in the R package Coda (Plummer et al., 2006) indicated convergence with estimated potential scale reduction factor equal to one for τ, σ^2, $\beta_0, \ldots, \beta_8$. Results from INLA largely agree with results from SGS and SPM, thus INLA results are not presented in the table. For most of the parameters, estimates and credible intervals computed with SGS and SPM are very close. The only somewhat noticeable difference is for τ, with estimates (95% credible intervals) of 0.1393 (0.0966,0.1902) and 0.1382 (0.0979,0.1952) from SGS and SPM, respectively. From a practical point of view, these numerical differences between the SGS and SPM results seem to be negligible.

With respect to this specific application, the larger the county population size, the larger tends to be the median household income. In addition, the educational level of the county's population is clearly important. Further, the higher the unemployment rate, the lower tends to be the median household income. Furthermore, median household incomes tend to be higher in metro areas when compared to non-metro areas, and larger metro areas tend to have higher median household incomes. Finally, to interpret the results for the spatial dependence parameter τ we need to recall from Keefe et al. (2019) that smaller values of τ indicate stronger spatial dependence, whereas larger values of τ indicate weaker spatial dependence, and values of τ larger than 10 indicate practically independent random effects. In the current case study, the estimate of the spatial dependence parameter τ is 0.1412 indicating strong spatial dependence.

Finally, Figure 6 presents maps of the United States with the posterior mean and the posterior standard deviation for the spatial random effects per county. The map of posterior mean reflects the strong spatial dependence in the spatial random effects. This map is particularly interesting because it indicates the regions for which the median household income is lower or higher than predicted by the regressors. For example, median household income is lower than predicted by the regressors in most of New Mexico. On another hand, median household income is higher than predicted by the regressors in most of California and in the

Table 2: Logarithm of median household income case study. Posterior summaries from SGS and SPM for parameters of hierarchical model with ICAR spatial random effects. Baseline education is some college or associate degree. Baseline metro status is non-metro area. Regressors: logarithm of population; logarithm of percent of adults with less than high school; logarithm of percent of adults with high school; logarithm of percent of adults with a bachelor's degree or higher; logarithm of unemployment; indicator of large metro area; indicator of medium metro area; and indicator of small metro area.

Parameter	**Posterior estimate**		**95% Credible Interval**	
	SGS	SPM	SGS	SPM
β_0 (intercept)	11.74	11.74	(11.52, 11.97)	(11.52, 11.97)
β_1 (log population)	0.0099	0.0099	(0.0051, 0.0147)	(0.0051, 0.0147)
β_2 (log less high school)	-0.1506	-0.1506	(-0.1682, -0.1329)	(-0.1683, -0.1329)
β_3 (log high school)	-0.1220	-0.1223	(-0.1609, -0.0837)	(-0.1606, -0.0840)
β_4 (log bachelor)	0.0233	0.0231	(-0.0043, 0.0505)	(-0.0044, 0.0506)
β_5 (log unemployment)	-0.2235	-0.2236	(-0.2448, -0.2020)	(-0.2449, -0.2022)
β_6 (metro large)	0.1508	0.1509	(0.1332, 0.1681)	(0.1333, 0.1684)
β_7 (metro medium)	0.0676	0.0676	(0.0527, 0.0826)	(0.0528, 0.0825)
β_8 (metro small)	0.0238	0.0238	(0.0098, 0.0376)	(0.0099, 0.0377)
τ	0.1393	0.1406	(0.0966, 0.1902)	(0.0999, 0.1978)
σ^2	0.0048	0.0049	(0.0038, 0.0058)	(0.0040, 0.0059)



(a) (b)

Figure 6: Spatial random effects: (a) posterior mean; (b) posterior standard deviation.

Atlantic Coast from the DC metropolitan area going north to the Boston metropolitan area. Finally, even though beyond the scope of this article, a closer examination of the spatial random effects may suggest other regressors to include in the analysis.

9. Discussion

Our novel contributions in this paper are two-fold. First, we prove the equivalence result between the use of an improper CAR prior with centering on the fly and the use of the sum-zero constrained ICAR prior proposed by Keefe et al. (2018, 2019). Second, we develop the SGS and SPM algorithms for fast Bayesian posterior computation for Gaussian hierarchical models with ICAR spatial random effects.

Our results provide fundamental insights about ICAR priors. Specifically, the current prevalent view is that ICAR priors are improper and centering on the fly is used to ensure posterior propriety. Our results indicate that the ICAR specification with centering on the fly is actually equivalent to using a singular Gaussian distribution as the prior. Hence, the prior is not only proper, but is defined on a lower dimensional space than the space on which the spatial random effects are defined. This insight has fundamental practical implications that are both methodological and computational.

One methodological implication is the ability to specify reference priors for the hyperparameters of hierarchical models with ICAR random effects (Keefe et al., 2019). Another methodological implication, which is the subject of current research, is the ability to use Bayes factors for Bayesian model selection. Computational implications include the ability to use our fast and scalable SGS and SPM algorithms. Our results show that computations for Bayesian model selection are greatly accelerated by our SPM algorithm.

We have performed a simulation study to compare the frequentist properties of SGS and SPM methods with those of INLA. SGS uses the reference prior proposed by Keefe et al. (2019) and SPM uses an approximate reference prior that we propose here. Our results show that the default settings in INLA do not quantify uncertainty correctly; in contrast SGS and SPM provide adequate quantification of uncertainty. Specifically, the mean square error of INLA estimates of variances may be much larger than that of SGS and SPM. In addition, while frequentist coverage of 95% INLA credible intervals for the variances may be much lower than nominal level, the frequentist coverage of 95% SGS and SPM credible intervals are much closer to the nominal 95% level. Further, the frequentist coverage of 95% SGS credible intervals is close to or above nominal for all parameters and all sample sizes considered. Therefore, for the analysis of Gaussian hierarchical models with ICAR random effects, we recommend the use of either the reference prior or the approximate reference prior.

There are many other avenues for future research. One such avenue is the extension of the new methods proposed here to problems where the variance of the measurement errors varies spatially. Specifically, that situation would correspond to assume in Equation (6) that the vector of errors had distribution $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{D})$ for a known diagonal matrix $\mathbf{D}$. We envision that extensions of the methods proposed here could be applied to appropriately modified and practically useful Gaussian hierarchical models with ICAR priors. Another promising avenue for future research is the extension of the ideas and methods presented here to count data. Finally, our current ongoing work includes research on model selection for Gaussian hierarchical models with ICAR spatial random effects (Porter et al., 2021).

Appendix:. Proofs of main results

Proof of Proposition 3.1.. By Bayes' Theorem the full conditional distribution of $\boldsymbol{\phi}$ is

$$\begin{aligned}
p(\boldsymbol{\phi}|\mathbf{y},\alpha,\boldsymbol{\beta},\sigma^2,\tau) &\propto p(\mathbf{y}|\boldsymbol{\phi},\alpha,\boldsymbol{\beta},\sigma^2,\tau)p(\boldsymbol{\phi}|\sigma^2,\tau)\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})'(\mathbf{y}-\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\phi})\right\}\\
&\quad \exp\left\{-\frac{\tau}{2\sigma^2}\boldsymbol{\phi}'\mathbf{H}\boldsymbol{\phi}\right\}\mathbb{1}(\mathbf{1}'\boldsymbol{\phi}=0)\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\phi}'\boldsymbol{\phi}-2\boldsymbol{\phi}'(\mathbf{y}-\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta})+\tau\boldsymbol{\phi}'\mathbf{Q}\mathbf{S}\mathbf{Q}'\boldsymbol{\phi}\right]\right\}\mathbb{1}(\mathbf{1}'\boldsymbol{\phi}=0)\\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\phi}'(\mathbf{I}+\tau\mathbf{Q}\mathbf{S}\mathbf{Q}')\boldsymbol{\phi}-2\boldsymbol{\phi}'(\mathbf{y}-\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta})\right]\right\}\mathbb{1}(\mathbf{1}'\boldsymbol{\phi}=0)\\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\phi}'\mathbf{Q}(\mathbf{I}+\tau\mathbf{S})\mathbf{Q}'\boldsymbol{\phi}-2\boldsymbol{\phi}'\mathbf{Q}\mathbf{Q}'(\mathbf{y}-\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta})\right]\right\}\mathbb{1}(\mathbf{1}'\boldsymbol{\phi}=0),
\end{aligned}$$

where the last step uses the fact that $\mathbf{Q}$ is orthogonal.

Now let $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)'=\mathbf{Q}'\boldsymbol{\phi}$ be a vector of spectral random effects. Thus, $\boldsymbol{\phi}=\mathbf{Q}\boldsymbol{\xi}$ and the Jacobian of the transformation is $d\boldsymbol{\phi}/d\boldsymbol{\xi}=\mathbf{Q}$. Further, note that $\mathbb{1}(\mathbf{1}'\boldsymbol{\phi}=0)=\mathbb{1}(\xi_n=0)$. Hence,

$$\begin{aligned}
p(\boldsymbol{\xi}|\mathbf{y},\alpha,\boldsymbol{\beta},\sigma^2,\tau) &\propto |\mathbf{Q}|\exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\xi}'(\mathbf{I}+\tau\mathbf{S})\boldsymbol{\xi}-2\boldsymbol{\xi}'\mathbf{Q}'(\mathbf{y}-\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta})\right]\right\}\mathbb{1}(\xi_n=0)\\
&\propto \prod_{i=1}^{n}\exp\left\{-\frac{1}{2\sigma^2}\left[\xi_i^2(1+\tau s_i)-2\xi_i\mathbf{q}_i'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-\alpha\xi_i\mathbf{q}_i'\mathbf{1})\right]\right\}\mathbb{1}(\xi_n=0)\\
&= \prod_{i=1}^{n-1}\exp\left\{-\frac{1}{2\sigma^2}\left[\xi_i^2(1+\tau s_i)-2\xi_i\mathbf{q}_i'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right]\right\}\mathbb{1}(\xi_n=0),
\end{aligned}$$

where the last equality uses the facts that $\xi_n=0$ and $\mathbf{q}_i'\mathbf{1}=0$ for $i=1,\ldots,n-1$. Thus, $\xi_1,\ldots,\xi_{n-1}$ are conditionally independent given $\mathbf{y},\alpha,\boldsymbol{\beta},\sigma^2,\tau$ and have full conditional distributions $\xi_i|\mathbf{y},\alpha,\boldsymbol{\beta},\sigma^2,\tau\sim N\left(\mathbf{q}_i'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/(1+\tau s_i),\sigma^2/(1+\tau s_i)\right), i=1,\ldots,n-1$. Let $\mathbf{Q}^*=(\mathbf{q}_1,\ldots,\mathbf{q}_{n-1})$. Then, these full conditional distributions may be written in matrix form as $\boldsymbol{\xi}_{1:(n-1)}|\mathbf{y},\alpha,\boldsymbol{\beta},\sigma^2,\tau\sim N(\mathbf{m},\sigma^2\mathbf{S}^*)$, where $\mathbf{S}^*=\text{diag}((1+\tau s_1)^{-1},\ldots,(1+\tau s_{n-1})^{-1})$ and $\mathbf{m}=\mathbf{S}^*\mathbf{Q}^*(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$. Therefore, the full conditional distribution of $\boldsymbol{\phi}=\mathbf{Q}\boldsymbol{\xi}=\mathbf{Q}^*\boldsymbol{\xi}_{1:(n-1)}$ is $\boldsymbol{\phi}|\mathbf{y},\alpha,\boldsymbol{\beta},\sigma^2,\tau\sim N(\mathbf{Q}^*\mathbf{m},\sigma^2\mathbf{Q}^*\mathbf{S}^*\mathbf{Q}^{*\prime})$. □

Proof of Proposition 3.2.. To obtain the full conditional distribution for $\boldsymbol{\omega}$, substitute $\boldsymbol{\phi}$ with $\boldsymbol{\omega}$ in the likelihood function. Then, by Bayes' Theorem the full conditional density for ϕ^* is

$$\begin{aligned}
p(\boldsymbol{\omega}|\mathbf{y},\alpha,\boldsymbol{\beta},\sigma^2,\tau) &\propto p(\mathbf{y}|\boldsymbol{\omega},\alpha,\boldsymbol{\beta},\sigma^2,\tau)p(\boldsymbol{\omega}|\sigma^2,\tau)\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\omega})'(\mathbf{y}-\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\omega})\right\}\\
&\quad \exp\left\{-\frac{\tau}{2\sigma^2}\boldsymbol{\omega}'\mathbf{H}\boldsymbol{\omega}\right\}\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\omega}'(\mathbf{I}+\tau\mathbf{H})\boldsymbol{\omega}-2\boldsymbol{\omega}(\mathbf{y}-\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta})\right\}.
\end{aligned}$$

Hence, the full conditional distribution for $\boldsymbol{\omega}$ is $\boldsymbol{\omega}|\mathbf{y},\alpha,\boldsymbol{\beta},\sigma^2,\tau\sim N(\mathbf{k},\mathbf{C})$, where $\mathbf{C}=\sigma^2(\mathbf{I}+\tau\mathbf{H})^{-1}$ and $\mathbf{k}=\mathbf{C}\sigma^{-2}(\mathbf{y}-\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta})=(\mathbf{I}+\tau\mathbf{H})^{-1}(\mathbf{y}-\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta})$.

Thus, simulating an intermediate $\boldsymbol{\omega}$ from this working intermediate full conditional distribution and centering $\boldsymbol{\omega}$ using $\boldsymbol{\phi} = (\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}')\boldsymbol{\omega}$ obtains a simulated value for $\boldsymbol{\phi}$ with implied full conditional distribution $N(\mathbf{Rk}, \mathbf{RCR}')$. □

The following lemma is useful for proving the equivalence between the sum-zero constrained ICAR prior and the improper ICAR prior with centering on the fly.

Lemma A1 (Keefe et al., 2018).. $\mathbf{RQ} = (\mathbf{q}_1, \ldots, \mathbf{q}_{n-1}, \mathbf{0})$.

Proof of Theorem 3.1.. We now show that the full conditional distribution of $\boldsymbol{\phi}$ for the improper CAR model centered on the fly and the full conditional based on the sum-zero constrained ICAR prior by Keefe et al. (2018, 2019) are the same. Since both distributions are multivariate Gaussian distributions, we need to show that both mean vectors and covariance matrices coincide.

First, consider the covariance matrices. Note that

$$\begin{aligned}
\mathbf{C} &= \sigma^2(\mathbf{I} + \tau\mathbf{H})^{-1} = \sigma^2(\mathbf{I} + \tau\mathbf{Q}\mathbf{S}\mathbf{Q}')^{-1} \\
&= \sigma^2\mathbf{Q}\text{diag}((1+\tau s_1)^{-1}, \ldots, (1+\tau s_n)^{-1})\mathbf{Q}'.
\end{aligned}$$

Thus, applying Lemma A1 we get

$$\begin{aligned}
\mathbf{RCR}' &= \sigma^2\mathbf{RQ}\text{diag}((1+\tau s_1)^{-1}, \ldots, (1+\tau s_n)^{-1})\mathbf{Q}'\mathbf{R}' \\
&= \sigma^2(\mathbf{q}_1, \ldots, \mathbf{q}_{n-1}, \mathbf{0})\text{diag}((1+\tau s_1)^{-1}, \ldots, (1+\tau s_n)^{-1})(\mathbf{q}_1, \ldots, \mathbf{q}_{n-1}, \mathbf{0})' \\
&= \sigma^2(\mathbf{q}_1, \ldots, \mathbf{q}_{n-1})\text{diag}((1+\tau s_1)^{-1}, \ldots, (1+\tau s_{n-1})^{-1})(\mathbf{q}_1, \ldots, \mathbf{q}_{n-1})' \\
&= \sigma^2\mathbf{Q}^*\mathbf{S}^*\mathbf{Q}^{*\prime}.
\end{aligned}$$

Therefore, the covariance matrices coincide.

Second, consider the mean vectors. Applying Lemma A1 we get

$$\begin{aligned}
\mathbf{Rk} &= \mathbf{RC}\sigma^{-2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \alpha\mathbf{1}) \\
&= \mathbf{RQ}\text{diag}((1+\tau s_1)^{-1}, \ldots, (1+\tau s_n)^{-1})\mathbf{Q}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \alpha\mathbf{1}) \\
&= (\mathbf{q}_1, \ldots, \mathbf{q}_{n-1}, \mathbf{0})\text{diag}((1+\tau s_1)^{-1}, \ldots, (1+\tau s_n)^{-1})\mathbf{Q}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \alpha\mathbf{1}) \\
&= ((1+\tau s_1)^{-1}\mathbf{q}_1, \ldots, (1+\tau s_n)^{-1}\mathbf{q}_{n-1}, \mathbf{0}) \begin{pmatrix} \mathbf{q}_1'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \vdots \\ \mathbf{q}_{n-1}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \mathbf{q}_n'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \alpha\sqrt{n} \end{pmatrix} \\
&= ((1+\tau s_1)^{-1}\mathbf{q}_1, \ldots, (1+\tau s_n)^{-1}\mathbf{q}_{n-1}) \begin{pmatrix} \mathbf{q}_1'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \vdots \\ \mathbf{q}_{n-1}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{pmatrix} \\
&= \mathbf{Q}^*\mathbf{S}^*\mathbf{Q}^{*\prime}\mathbf{y} = \mathbf{Q}^*\mathbf{m}.
\end{aligned}$$

Therefore, the mean vectors also coincide. □

Acknowledgments. The work of Ferreira was supported in part by National Science Foundation Award 1853549. The work of Porter, Franck, and Ferreira was supported in part by a grant from the Virginia Tech College of Science Dean's Discovery Fund. We are grateful for the comments and suggestions of the Associate Editor and two anonymous reviewers that have led to a greatly improved manuscript.

References

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2014. Hierarchical Modeling and Analysis for Spatial Data. CRC Press, Boca Raton, FL. 2nd edition.

Berger, J., 2006. The case for objective Bayesian analysis. Bayesian analysis 1, 385–402.

Berger, J.O., de Oliveira, V., Sansó, B., 2001. Objective Bayesian analysis of spatially correlated data. Journal of the American Statistical Association 96, 1361–1374.

Bernardinelli, L., Clayton, D., Montomoli, C., 1995. Bayesian estimates of disease maps: how important are priors? Statistics in Medicine 14, 2411–2431.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society – Series B , 192–236.

Besag, J., Green, P., Higdon, D., Mengersen, K., 1995. Bayesian computation and stochastic systems (with discussion). Statistical Science 10, 3–66.

Besag, J., Kooperberg, C., 1995. On conditional and intrinsic autoregressions. Biometrika 82, 733–746.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics 43, 1–20.

Best, N.G., Arnold, R.A., Thomas, A., Waller, L.A., Conlon, E.M., 1999. Bayesian models for spatially correlated disease and exposure data, in: Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting, Oxford University Press. p. 131.

Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V., Pebesma, E.J., 2013. Applied spatial data analysis with R. Springer, New York. 2nd edition.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88, 9–25.

Celeux, G., Forbes, F., Robert, C.P., Titterington, D.M., 2006. Deviance information criteria for missing data models. Bayesian Analysis 1, 651–673.

Clayton, D., Kaldor, J., 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics , 671–681.

Crainiceanu, C.M., Ruppert, D., 2004. Likelihood ratio tests in linear mixed models with one variance component. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66, 165–185.

De Oliveira, V., Ferreira, M.A.R., 2011. Maximum likelihood and restricted maximum likelihood estimation for a class of Gaussian Markov random fields. Metrika 74, 167–183.

Efron, B., 2015. Frequentist accuracy of Bayesian estimates. Journal of the Royal Statistical Society – Series B 77, 617.

Ferreira, M.A.R., 2019. The limiting distribution of the Gibbs sampler for the intrinsic conditional autoregressive model. Brazilian Journal of Probability and Statistics 33, 734–744.

Ferreira, M.A.R., Bertolde, A.I., Holan, S., 2010. Analysis of economic data with multi-scale spatio-temporal models, in: O'Hagan, A., West, M. (Eds.), Handbook of Applied Bayesian Analysis. Oxford University Press, Oxford, pp. 295–318.

Ferreira, M.A.R., De Oliveira, V., 2007. Bayesian reference analysis for Gaussian Markov random fields. Journal of Multivariate Analysis 98, 789–812.

Ferreira, M.A.R., Holan, S.H., Bertolde, A.I., 2011. Dynamic multiscale spatio-temporal models for Gaussian areal data. Journal of the Royal Statistical Society – Series B 73, 663–688.

Ferreira, M.A.R., Salazar, E., 2014. Bayesian reference analysis for exponential power regression models. Journal of Statistical Distributions and Applications 1, 1–12.

Ferreira, M.A.R., Suchard, M.A., 2008. Bayesian analysis of elapsed times in continuous-time Markov chains. Canadian Journal of Statistics 36, 355–368.

Fonseca, T.C.O., Ferreira, M.A.R., Migon, H.S., 2008. Objective Bayesian analysis for the Student-t regression model. Biometrika 95, 325–333.

Fonseca, T.C.O., Migon, H.S., Ferreira, M.A.R., 2012. Bayesian analysis based on the jeffreys prior for the hyperbolic distribution. Brazilian Journal of Probability and Statistics 26, 327–343.

Freni-Sterrantino, A., Ventrucci, M., Rue, H., 2018. A note on intrinsic conditional autoregressive models for disconnected graphs. Spatial and Spatio-temporal Epidemiology 26, 25–34.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7, 457–511.

He, D., Sun, D., He, L., 2021. Objective bayesian analysis for the Student-t linear regression. Bayesian Analysis 16, 129–145.

Hodges, J.S., Carlin, B.P., Fan, Q., 2003. On the precision of the conditionally autoregressive prior in spatial models. Biometrics 59, 317–322.

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., Eskin, E., 2008. Efficient control of population structure in model organism association mapping. Genetics 178, 1709–1723.

Keefe, M.J., Ferreira, M.A.R., Franck, C.T., 2018. On the formal specification of sum-zero constrained intrinsic conditional autoregressive models. Spatial Statistics 24, 54–65.

Keefe, M.J., Ferreira, M.A.R., Franck, C.T., 2019. Objective Bayesian analysis for Gaussian hierarchical models with intrinsic conditional autoregressive priors. Bayesian Analysis 14, 181–209.

Lavine, M.L., Hodges, J.S., 2012. On rigorous specification of ICAR models. The American Statistician 66, 42–49.

Lee, D., 2013. CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. Journal of Statistical Software 55, 1–24.

Magnus, J.R., Neudecker, H., 1999. Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, Chichester. revised edition.

Martins, T.G., Simpson, D., Lindgren, F., Rue, H., 2013. Bayesian computing with INLA: new features. Computational Statistics & Data Analysis 67, 68–83.

Plummer, M., Best, N., Cowles, K., Vines, K., 2006. Coda: Convergence diagnosis and output analysis for mcmc. R News 6, 7–11.

Porter, E.M., Franck, C.T., Ferreira, M.A.R., 2021. Objective Bayesian model selection for spatial hierarchical models with intrinsic conditional autoregressive priors. Submitted .

Porter, E.M., Keefe, M.J., Franck, C.T., Ferreira, M.A.R., 2019. ref.ICAR: Objective Bayes Intrinsic Conditional Autoregressive Model for Areal Data. R package version 1.0.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Rue, H., 2021. Personal communication.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. CRC Press.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71, 319–392.

Salazar, E., Ferreira, M.A.R., Migon, H.S., 2012. Objective bayesian analysis for exponential power regression models. Sankhya - Series B 74, 107–125.

Sørbye, S.H., Rue, H., 2014. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. Spatial Statistics 8, 39–51.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 583–639.

Sun, D., Berger, J.O., 2007. Objective bayesian analysis for the multivariate normal model, in: Bernardo, J.M., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., West, M. (Eds.), Bayesian Statistics 8. Oxford University Press, Oxford, pp. 525–547.

Thomas, A., Best, N., Lunn, D., Arnold, R., Spiegelhalter, D., 2004. GeoBugs user manual. Cambridge: Medical Research Council Biostatistics Unit .

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research 11, 3571–3594.