Bayesian Model Selection for Spatial Data and Cost-constrained Applications

Erica M. Porter

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Christopher T. Franck, Co-chair
Marco A.R. Ferreira, Co-chair
Stephen C. Adams
Leanna L. House

June 13, 2023
Blacksburg, Virginia

Keywords: Spatial statistics, Bayesian model selection

Bayesian Model Selection for Spatial Data and Cost-constrained Applications

Erica M. Porter

ABSTRACT

Bayesian model selection is a useful tool for identifying an appropriate model class, dependence structure, and valuable predictors for a wide variety of applications. In this work we consider objective Bayesian model selection where no subjective information is available to inform priors on model parameters *a priori*, specifically in the case of hierarchical models for spatial data, which can have complex dependence structures. We develop an approach using trained priors via fractional Bayes factors where standard Bayesian model selection methods fail to produce valid probabilities under improper reference priors. This enables researchers to concurrently determine whether spatial dependence between observations is apparent and identify important predictors for modeling the response. In addition to model selection with objective priors on model parameters, we also consider the case where the priors on the model space are used to penalize individual predictors *a priori* based on their costs. We propose a flexible approach that introduces a tuning parameter to cost-penalizing model priors that allows researchers to control the level of cost penalization to meet budget constraints and accommodate increasing sample sizes.

Bayesian Model Selection for Spatial Data and Cost-constrained Applications

Erica M. Porter

GENERAL AUDIENCE ABSTRACT

Spatial data, such as data collected over a geographic region, is relevant in many fields. Spatial data can require complex models to study, but use of these models can impose unnecessary computations and increased difficulty for interpretation when spatial dependence is weak or not present. We develop a method to simultaneously determine whether a spatial model is necessary to understand the data and choose important variables associated with the outcome of interest. Within a class of simpler, linear models, we propose a technique to identify important variables associated with an outcome when there exists a budget or general desire to minimize the cost of collecting the variables.

Acknowledgments

I am grateful to my co-advisors, Chris Franck and Marco Ferreira for all of the guidance and knowledge over the past few years. Chris, I have appreciated the patience you showed me, the countless writing edits, and the time you made to help me work through problems and decisions. Marco, you have been an invaluable source of research knowledge, publication expertise, and more. Thank you for all of your support and for helping me navigate research and graduate school.

Leanna, thank you for all of your advice and encouragement over the years, and for letting me hang out with Trixie sometimes. You are a very thoughtful professor and person, and Virginia Tech is lucky to have you. Stephen, I appreciated your enthusiasm during our research meetings; thank you for your perspective, optimism, and insights into machine learning, research, and writing.

To Steve Walsh, Shane Bookhultz, and Adam Edwards, thank you for being great friends to me through everything. The laughs, smiles, conversations, and games you provided sustained me throughout graduate school, and I am so glad to know you.

To my family, thank you for being a forever constant and the greatest source of love in my life. Thank you for bearing with me through the stress and tears, rescuing me when my car broke down, and providing much-needed holiday fun and memories. And to my dog, Dottie, thank you for being my best friend these last 1.5 years. I wouldn't have made it to the end of my Ph.D. without your company; thanks for all of your love and energy, and for making sure I got outside every day.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Model Selection for Areal Data

Spatial data is prevalent across many disciplines such as environmental applications, disease mapping, and neuroimaging. Areal data, consisting of subregions that spatially partition a finite region, is one type of spatial data that addresses numerous relationships impacted by geographical proximity. Areal data is of particular interest in fields including ecology (Ver Hoef et al., 2018), sociology (Goodchild and Janelle, 2004), and epidemiology (Lee, 2011). Areal data can be especially useful in cases where observations are naturally recorded and maintained at a specific geographic level, e.g. by neighborhood, counties, or states. For example, within sociology areal data can be used to record and analyze demographics and identify potential income disparities between neighboring counties. Areal data can also be used to analyze public policy according to states or voting districts and to make census data available in an aggregated format without divulging information about individual households (Hogan and Tchernis, 2004). In some cases, particularly in health applications, individual-level data is proprietary, but aggregated data for sub-populations is available to researchers, making areal data models an essential tool for analysis (Goodchild and Janelle, 2004). As another example, ecology studies often analyze satellite images and data collected on a grid (Zhu et al., 2010) or irregular subregions formed by natural or managed habitats (Ver Hoef et al., 2018).

Bayesian hierarchical models, including conditional autoregressive (CAR) models, are often used to model areal data because they are flexible enough to accommodate both fixed regression effects which are usually of great interest to subject matter experts, and also spatial dependence parameters which are essential in the analysis of spatial data. The hierarchical model includes elements for fixed regression effects, natural variation, and a vector of spatial random effects corresponding to each subregion in the set of areal data. We focus here on the class of intrinsic CAR models (ICAR), where the vector of spatial random effects is assigned its own prior.

When using hierarchical models to analyze areal data, researchers must select both predictor variables and spatial dependence structure. Performing analysis on geographically proximate and correlated data assuming they are independent can be hazardous and lead to incorrect or misleading conclusions about patterns in the data, which may result in misinformed decisions for policy, health, etc. Conversely, using complex hierarchical models for data that are geographically proximate but exhibit no true correlation with each other can result in unnecessary computation and additional time to interpret extraneous effects. Additionally, identifying which predictors are valuable and merit inclusion or further study is an important task. Historically, model selection has often been performed separately for these two model elements, by first testing for spatial dependence among the observations and then using standard criteria or other selection approaches to select a subset of predictors, or vice versa. This can obscure the true model structure, as fixed and spatial random effects can influence inference for the other (Hodges and Reich, 2010). To address the need for more cohesive model selection methods for areal data, we develop simultaneous Bayesian model selection for fixed effects and spatial model structure in ICAR models. We assign the model parameters a reference prior that obviates the need to specify hyperparameters to which the analysis might be sensitive and which has been shown to have favorable performance with respect

to frequentist coverage rate, average interval length, and mean squared error. Since the reference prior is improper, Bayesian model using traditional Bayes factors does not produce valid probabilities. Therefore we approach simultaneous selection of fixed effects and spatial model structure using fractional Bayes factors (O'Hagan, 1995), which use a fraction of the likelihood to train the prior and provide valid, accurate selection. Chapter 2 presents our approach, which provides simultaneous, objective Bayesian model selection and demonstrates its accuracy via simulation and real data applications.

1.2 Cost-Penalized Model Selection

While Chapter 2 emphasizes the importance of objective Bayesian model selection approaches, sometimes a researcher may want to influence the model selection with a certain goal in mind. Since Bayesian model selection requires priors on both model parameters and the model space, it provides a convenient way to incorporate information about or penalties on particular models *a priori*. Specifically, we study the use of model priors to penalize models that contain costly predictors. Medical applications often use existing data to select predictors to be collected, many of which have to conform to budgets. Several machine learning methods (e.g. Bolón-Canedo et al. (2014); Kong et al. (2016); Ling et al. (2004)) and decision-theoretic approaches (e.g. Fouskakis and Draper (2008); Miyawaki and MacEachern (2022)) have been proposed with the purpose of penalizing costly predictors when selecting a subset of predictors. Bayesian model selection is valuable because it provides posterior probabilities for both models and individual predictors that a researcher wants to make decisions about; these probabilities are easily interpretable and provide a useful way for uncertainty quantification. Fouskakis et al. (2009a) proposed a model prior to penalize more costly predictors relative to the others' costs. The prior reduces prior probability on expensive

covariates, but the cost penalty greatly diminishes at larger sample sizes. Thus, Bayesian model selection using Fouskakis et al. (2009a)'s cost-penalizing model prior selects the same model as Bayesian model selection with uniform model priors as the sample size grows. To address this paradox, we introduce a tuning parameter to the model prior so that researchers can adjust the cost penalization to fit their budget and/or desired model performance. We develop an inclusion path based on this tuning parameter, which plots the posterior inclusion probabilities for each candidate predictor across several values of the tuning parameter to visualize the changing impact of the cost penalty on the posterior and selection results. Chapter 3 presents our tuning-parameter based approach and inclusion path, along with results from simulations and an application to a classification problem for diagnosing heart disease.

1.3 Dissertation Outline

The remainder of this dissertation is organized as follows. Chapter 2 provides a literature review, motivation, development, and evaluation of our fractional Bayes factor approach for simultaneous selection of fixed effects and spatial model structure in ICAR models. Chapter 3 contains a literature review, proposed tuning-parameter based method, and results for our cost-penalized Bayesian model selection approach. Chapter 4 describes version 2.0 of the R package, ref.ICAR, that implements objective ICAR model inference and the model selection approach discussed in Chapter 2, with code for users to recreate examples using the open source code. Chapter 5 summarizes the ideas and contributions discussed in this dissertation and outlines some potential topics for future related work. Finally, Appendix A provides necessary facts for understanding the content in Chapter 2, proofs for all propositions and theorems stated, and additional simulation results associated with the ICAR model selection

presented in Chapter 2.

Chapter 2

Objective Bayesian Model Selection for Spatial Hierarchical Models with Intrinsic Conditional Autoregressive Priors

Erica M. Porter[1], Christopher T. Franck[1], Marco A.R. Ferreira[1]

[1]Department of Statistics, Virginia Tech, Blacksburg, Virginia, 24061, U.S.A.

Abstract

We develop Bayesian model selection via fractional Bayes factors to simultaneously assess spatial dependence and select regressors in Gaussian hierarchical models with intrinsic conditional autoregressive (ICAR) spatial random effects. Selection of covariates and spatial model structure is difficult, as spatial confounding creates a tension between fixed and spatial random effects. Researchers have commonly performed selection separately for fixed and random effects in spatial hierarchical models. Simultaneous selection methods relieve the researcher from arbitrarily fixing one of these types of effects while selecting the other. Notably, Bayesian approaches to simultaneously select covariates and spatial effects are limited. Our use of fractional Bayes factors allows for selection of fixed effects and spatial model structure under automatic reference priors for model parameters, which obviates the need to specify hyperparameters for priors. We also show the equivalence between two ICAR specifications and derive the minimal training size for the fractional Bayes factor applied to the ICAR model under the reference prior. We perform a simulation study to assess the performance of our approach and we compare results to the Deviance Information Criterion and Widely Applicable Information Criterion. We demonstrate that our fractional Bayes factor approach assigns low posterior model probability to spatial models when data is truly independent and reliably selects the correct covariate structure with highest probability within the model space. Finally, we demonstrate our Bayesian model selection approach with applications to county-level median household income in the contiguous United States and residential crime rates in the neighborhoods of Columbus, Ohio.

Keywords: Bayesian model selection, spatial statistics, areal data, ICAR random effects, fractional Bayes factor

2.1 Introduction

Bayesian hierarchical models are often used to model spatial data because they are flexible enough to accommodate both fixed regression effects and spatial random effects. In particular, hierarchical models with conditional autoregressive (CAR) structure (Besag, 1974) and intrinsic conditional autoregressive (ICAR) structure (Besag et al., 1991) for spatial random effects have been used for inference and prediction in fields such as ecology (Ver Hoef et al., 2018), neuroscience (Liu et al., 2016), disease mapping (Reich et al., 2006; Lee, 2011; Jin et al., 2005; White et al., 2017), and public policy (Logan et al., 2020). While modeling and estimation methods for models with CAR and ICAR structures have seen a variety of methodological developments and applications, simultaneous selection of spatial model structure and covariates in hierarchical models with ICAR spatial random effects has seen limited development. Thus, when choosing which specific covariates to include and whether spatial dependence persists in the presence of these covariates, researchers often resort to two-stage procedures. We focus our attention on Bayesian selection methods in this work. For example, researchers have adapted selection techniques to compare various proposed CAR models (Lee, 2011; Song and De Oliveira, 2012; Best et al., 2005) and to determine the impact of covariates when a spatial correlation structure is assumed (Best et al., 1999). Such approaches require researchers to either fix the spatial model structure and select covariates or fix the covariates and assess the need for spatial model structure. The order of selection is arbitrary and has been seen in case studies to potentially provide conflicting results. For example, Lee and Mitchell (2013) studied two covariates in a disease mapping application. They first fit a Bayesian model with both covariates and no spatial random effects. In the non-spatial model, 95% credible intervals revealed both covariates were non-null, however a Moran's I test indicated spatial correlation in the residuals. Upon finding spatial correlation in the residuals, the authors fit spatial models including the Besag-York-Mollié (BYM, Besag

et al., 1991), Leroux (Leroux et al., 1999), and locally adaptive spatial models. In some of these spatial models, one of the covariates was then found to be plausibly null. This sort of recursive approach can lead to uncertainty about covariates in the model since it does not directly assess spatial effects and covariates simultaneously. While Lee and Mitchell (2013) successfully develop novel methods for accommodating localized spatial dependence, this multi–stage approach that selects the mean and covariance structures in separate stages indicates a need for more cohesive and simultaneous selection methods for areal data. Bayesian methods for simultaneous selection of spatial model structure and covariates have seen less development. Since simultaneous selection of spatial model structure and regressors would provide a framework for researchers to make concurrent probabilistic decisions in spatial contexts, we propose a Bayesian approach for simultaneous selection of fixed effects and spatial model structure in Gaussian hierarchical models with ICAR priors.

Current literature on model selection for hierarchical models with ICAR priors has suffered from the crucial limitation that, until recently, these models did not have a fully specified expression for the likelihood function with integrated out random effects. Without such an expression, the development of formal Bayesian model selection was not possible. Fortunately, Keefe et al. (2018) recently proposed a formal specification of sum-zero constrained ICAR models that fully specifies the constant of proportionality in these models. We explore these recent results to develop formal Bayesian model selection for hierarchical models with ICAR random effects.

To the best of our knowledge, the closest published work related to our proposed method is by Song and De Oliveira (2012). Specifically, Song and De Oliveira (2012) proposed the use of posterior model probabilities to choose between classes of Gaussian CAR models and simultaneous autoregressive (SAR) models with default priors for model parameters. However, their proposed methodology differs from ours in two important ways. First, Song and

De Oliveira (2012) considered CAR and SAR models as direct models for the observations, whereas we consider the more usual framework in the statistical literature of using CAR priors for spatial random effects in a hierarchical model. Second, their approach assumes that all competing models have the same mean structure, and it is restricted to model selection of covariance structure. Hence, their approach cannot perform covariate selection. In contrast, we provide an approach to perform joint selection of covariates (fixed effects) and spatial model structure.

There are several published criteria other than posterior probabilities for model selection in hierarchical models with CAR and ICAR spatial random effects. Currently, the most frequently used model selection criterion for such hierarchical models is the Deviance Information Criterion (DIC, Spiegelhalter et al. (2002)). For example, the DIC has been used for model selection in the contexts of generalized multivariate CAR models (Jin et al., 2005), coregionalized models for areal data (Jin et al., 2007), locally adaptive spatial CAR models (Lee and Mitchell, 2013), and disease mapping (Martinez-Beneito et al., 2017). Another model comparison tool, proposed by White et al. (2017) for hierarchical models with CAR random effects, is cross-validation, where a part of the sample is used for model fitting and the other observations are held out for model evaluation. In such cross-validation setting, White et al. (2017) proposed as model comparison criteria predictive interval coverage, predictive mean square error, and predictive mean absolute error. Although ingenious and practically useful, these published model selection criteria do not provide the natural and straightforward quantification of model uncertainty provided by posterior model probabilities.

To enable simultaneous selection of fixed effects and spatial model structure, we develop a Bayesian model selection method for hierarchical models with an ICAR component. In particular, we examine a sum-zero constrained ICAR prior for spatial random effects in a Bayesian hierarchical model (Keefe et al., 2018). We devise a fractional Bayes factor

(O'Hagan, 1995) approach for model selection via posterior model probabilities. Fractional Bayes factors use a portion of the likelihood to update priors on parameters, which enables our automatic Bayesian model selection with an improper reference prior on model parameters. Model selection consistency, which refers to the method's ability to select the true model as sample size increases if the true model is in the candidate set, is a well-known result when using fractional Bayes factors (O'Hagan, 1995). Thus our approach provides consistent simultaneous selection of fixed effects and spatial model structure in Bayesian hierarchical models and allows for direct probabilistic statements about inclusion of covariates and spatial model structure.

We describe our formal Bayesian model selection approach in the following sections. In Section 2.2 we introduce the hierarchical model with a sum-zero constrained ICAR prior, prove an equivalence result for two ICAR specifications, and provide the reference prior we consider for the parameters of the ICAR component. In Section 2.3 we present the motivation and implementation of our proposed method which uses fractional Bayes factors for simultaneous selection of fixed effects and spatial model structure in Gaussian hierarchical models with ICAR priors. In Section 2.4 we study the performance of our proposed method with a simulation study that includes varying levels of spatial dependence. Section 2.5 demonstrates two applications of our method, including median income and socioeconomic data at the county-level for the contiguous US in 2017 and residential crime rates in the neighborhoods of Columbus, Ohio in 1980. Finally, in Section 2.6 we discuss the practical impact of our fractional Bayes factor approach and future avenues for research. Proofs of theoretical results and additional model selection background and simulation results are provided in the Supplementary Material.

2.2 Hierarchical Model Specification

We consider a hierarchical model for areal data measured over a contiguous region that is partitioned into n disjoint subregions indexed by $1, \ldots, n$. Consider the following hierarchical model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\theta} + \boldsymbol{\phi}, \tag{2.1}$$

where $\boldsymbol{Y}$ is a $n \times 1$ response vector, X is a $n \times p$ matrix of covariates, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients corresponding to fixed effects. Additionally, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ is a $n \times 1$ vector of independent unstructured random effects with distribution $N(\mathbf{0}, \sigma^2 I_n)$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$ is a $n \times 1$ vector of spatial random effects. The first column of the matrix X is assumed to be a vector of ones and, thus, the first element of $\boldsymbol{\beta}$ is an intercept. The vectors of random effects, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, are assumed to be independent *a priori*.

The spatial random effects $\boldsymbol{\phi}$ are assigned a sum-zero constrained intrinsic conditional autoregressive (ICAR) prior (Keefe et al., 2018, 2019). We consider a signal-to-noise ratio parameterization where τ/σ^2 represents the precision for the vector of spatial random effects, where the parameter $\tau \in (0, \infty)$ controls the strength of spatial dependence and σ^2 denotes the variance of the unstructured random effects. Small values of τ indicate strong spatial dependence while values tending towards infinity correspond to independent data. The sum-zero constraint $\sum_{i=1}^{n} \phi_i = 0$ appears explicitly in the density for $\boldsymbol{\phi}$ as follows.

$$p(\boldsymbol{\phi}|\sigma^2, \tau) = (2\pi)^{-(n-1)/2} \left(\frac{\tau}{\sigma^2}\right)^{(n-1)/2} \left(\prod_{i=1}^{n-1} d_i\right)^{1/2} \exp\left\{-\frac{\tau}{2\sigma^2}\boldsymbol{\phi}^T H \boldsymbol{\phi}\right\} \mathbb{1}(\mathbf{1}_n^T \boldsymbol{\phi} = 0), \tag{2.2}$$

where $\mathbf{1}_n$ is a vector of ones and H is a positive semi-definite precision matrix defined as

$$(H)_{ij} = \begin{cases} h_i, & \text{if } i = j, \\ -g_{ij}, & \text{if } i \in N_j, \\ 0, & \text{otherwise,} \end{cases} \tag{2.3}$$

and $d_1 \geq d_2 \geq \ldots \geq d_{n-1} > d_n = 0$ are the ordered eigenvalues of H. The matrix H is fixed and is chosen by the researcher to specify the neighborhood structure of the study region. For example, a common choice for H classifies two subregions as neighbors if they share a border. In that case, $\{N_j; j = 1 \ldots n\}$ denotes the set of regions that are neighbors to region j, h_i indicates the total number of neighbors of region i, and $g_{ij} = 1$ if regions i and j are neighbors and $g_{ij} = 0$ if regions i and j are not neighbors. We assume that there are no islands in the region of interest, that is, all of the subregions are connected. As a consequence, H has rank $n - 1$ and one null eigenvalue (e.g., see Ferreira and De Oliveira, 2007; De Oliveira and Ferreira, 2011). Let the spectral decomposition of H be $H = QDQ^T$, where $Q = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n)$ is a $n \times n$ matrix whose columns are the normalized eigenvectors of H and $D = \text{diag}(d_1, d_2, \ldots, d_n)$ is a diagonal matrix with the ordered eigenvalues of H along the diagonal. Let $\tilde{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_{n-1})$. Then the distribution of the spatial random effects $\boldsymbol{\phi}$ can be written as the following singular Gaussian distribution (Keefe et al., 2018, 2019):

$$\boldsymbol{\phi} \sim N\Big(\mathbf{0}, \frac{\sigma^2}{\tau}\Sigma_\phi\Big), \tag{2.4}$$

where $\Sigma_\phi = \tilde{Q}\text{diag}(d_1^{-1}, \ldots, d_{n-1}^{-1})\tilde{Q}^T$ is the Moore-Penrose pseudoinverse (Penrose, 1955) of the precision matrix H. See Auxiliary Fact A3 in the Supplementary Material for details concerning how the sum-zero constraint is represented in Equations (2.2) and (2.4).

2.2.1 Equivalence Between ICAR Specifications

There are three main ways to impose the sum-zero constraint on the ICAR prior. The first way is through centering on the fly, where the spatial random effects are centered to sum to zero at each iteration of the MCMC algorithm. The second way is through obtaining the full conditional distribution of the spatial random effects, which as we explain below is a proper multivariate Gaussian distribution, and then use standard multivariate Gaussian results to obtain the full conditional distribution of the spatial random effects conditional on their sum being equal to zero. Finally, the third way is to use the sum-zero constrained ICAR model proposed by Keefe et al. (2018, 2019). Ferreira (2019) and Ferreira et al. (2021) have shown that the first and the third ways are equivalent for Gaussian hierarchical models. In this section, we show that the second and third ways are equivalent.

The following propositions and theorem present results about the distribution of spatial random effects $\boldsymbol{\phi}$ and show the equivalence between sampling from the improper ICAR prior conditional on a sum-zero constraint and sampling from the formal sum-zero constrained prior.

Proposition 2.1 (Ferreira et al. (2021))**.** *Assume the hierarchical model given by Equations (2.1) and (2.2). Partition the design matrix as* $X = [\mathbf{1}_n, F]$ *and, similarly, partition the vector of regression coefficients as* $\boldsymbol{\beta} = (\alpha, \boldsymbol{\nu}^T)^T$ *where* α *is an intercept. Then, the full conditional distribution of* $\boldsymbol{\phi}$ *is*

$$\boldsymbol{\phi}|\tau, \sigma^2, \boldsymbol{Y}, \boldsymbol{\beta} \sim N(\tilde{Q}\mathbf{s}, \sigma^2 \tilde{Q} D^* \tilde{Q}^T), \tag{2.5}$$

where $D^* = diag((1+\tau d_1)^{-1}, \ldots, (1+\tau d_{n-1})^{-1})$*, and* $\mathbf{s} = D^* \tilde{Q}^T (\boldsymbol{Y} - F\boldsymbol{\nu})$.

Next, let $\boldsymbol{\omega}$ be a vector of spatial random effects that *a priori* follows the improper ICAR

prior (Besag et al., 1991). Then, the prior density for $\boldsymbol{\omega}$ is defined up to a constant of proportionality and, with the signal-to-noise ratio parameterization, is given by

$$p(\boldsymbol{\omega}|\tau,\sigma^2) \propto \exp\Big\{-\frac{\tau}{2\sigma^2}\boldsymbol{\omega}^T H\boldsymbol{\omega}\Big\}. \tag{2.6}$$

If we substitute $\boldsymbol{\phi}$ by $\boldsymbol{\omega}$ in Equation (2.1) and use the prior for $\boldsymbol{\omega}$ given in Equation (2.6), straightforward application of Bayes' Theorem yields the full conditional distribution (e.g., see Ferreira et al., 2021)

$$\boldsymbol{\omega}|\tau,\sigma^2,\mathbf{Y},\boldsymbol{\beta} \sim N(g,V), \tag{2.7}$$

where $V=\sigma^2(I_n+\tau H)^{-1}=\sigma^2 Q(I_n+\tau D)^{-1}Q^T$ and $g=(I_n+\tau H)^{-1}(\mathbf{Y}-X\boldsymbol{\beta})=Q(I_n+\tau D)^{-1}Q^T(\mathbf{Y}-X\boldsymbol{\beta})$. Note that $(I_n+\tau H)$ is diagonally dominant and, therefore, non-singular. Hence, the matrix V is well defined. However, assigning a flat prior for the intercept in the model and the improper CAR prior given in Equation (2.6) leads to an improper posterior distribution for ω.

Alternatively to sampling the sum-zero-constrained spatial random effects vector $\boldsymbol{\phi}$ from its full conditional distribution (2.5), we may consider using the Besag spatial random effects vector $\boldsymbol{\omega}$ sampled from its full conditional distribution (2.7) conditional on the constraint $\mathbf{1}_n^T\boldsymbol{\omega}=0$. The details of this distribution are given in Proposition 2.2.

Proposition 2.2. *Assume the model given by Equation (2.1) but substituting $\boldsymbol{\phi}$ by $\boldsymbol{\omega}$. In addition, assume for $\boldsymbol{\omega}$ the prior given by Equation (2.6). Then, the full conditional distribution of $\boldsymbol{\omega}$ conditional on the constraint $\boldsymbol{1}_n^T\boldsymbol{\omega}=0$ is*

$$\boldsymbol{\omega}|\,\boldsymbol{1}_n^T\boldsymbol{\omega}=0,\tau,\sigma^2,\,\boldsymbol{Y},\boldsymbol{\beta} \sim N(\mu^*,\Sigma^*), \tag{2.8}$$

where $\mu^*=g-V\boldsymbol{1}_n(\boldsymbol{1}_n^TV\boldsymbol{1}_n)^{-1}(\boldsymbol{1}_n^Tg-0)$ and $\Sigma^*=V-V\boldsymbol{1}_n(\boldsymbol{1}_n^TV\boldsymbol{1}_n)^{-1}\boldsymbol{1}_n^TV$.

Proof. See Proofs of Main Results in the Supplementary Material. □

The following theorem shows the equivalence between sampling from the full conditional distribution of $\boldsymbol{\phi}$ implied by the sum-zero constrained ICAR prior of Keefe et al. (2018, 2019) and sampling from the full conditional distribution of the spatial random effects $\boldsymbol{\omega}$ implied by the improper ICAR prior of Besag et al. (1991) with respect to the sum-zero constraint, that is, sampling $\boldsymbol{\omega}$ conditional on the spatial random effects summing to zero.

Theorem 2.3. *Assume that all subregions are connected. Then, sampling from the full conditional distribution given in Equation (2.8) of the spatial random effects $\boldsymbol{\omega}$ implied by the improper ICAR prior (2.6) conditional on the sum-zero constraint $\mathbf{1}_n^T\boldsymbol{\omega} = 0$ is equivalent to sampling from the full conditional distribution given in Equation (2.5) implied by the sum-zero constrained ICAR prior given in Equation (2.2).*

Proof. See Proofs of Main Results in the Supplementary Material. □

The equivalence result given in Theorem 2.3 is of fundamental importance because it implies that the model selection approach we propose can be applied to Gaussian hierarchical models with Besag ICAR spatial random effects such as implemented in the widely used R package R-INLA (Rue et al., 2009).

Next, we denote the vector of unknown model parameters for the hierarchical spatial model by $\boldsymbol{\eta} = (\boldsymbol{\beta}, \sigma^2, \tau)$. Following the approach of Keefe et al. (2018), we impose the formal sum-zero constraint that implies as prior for the spatial random effects $\boldsymbol{\phi}$ the singular Gaussian distribution given in Equation (2.4). After that, we integrate out the vector of spatial random effects $\boldsymbol{\phi}$ to obtain for $\mathbf{Y}|\boldsymbol{\eta}$ the Gaussian distribution

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \tau \sim N(X\boldsymbol{\beta}, \sigma^2(I_n + \tau^{-1}\Sigma_\phi)). \tag{2.9}$$

We are interested in selecting which covariates to include in the model by choosing among competing $X\boldsymbol{\beta}$, and whether to include spatial dependence. We use the ordinary linear model (OLM) as the independent data model, which does not accommodate spatial correlation. In the case of the OLM, the unknown parameters are $\boldsymbol{\beta}$ and σ^2, and the response $\mathbf{Y}$ follows the Gaussian distribution

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2 \sim N(X\boldsymbol{\beta}, \sigma^2 I_n). \tag{2.10}$$

2.2.2 Priors on Model Parameters

We adopt a Bayesian approach and specify priors for $\boldsymbol{\eta}$. We consider the recently proposed reference prior for the parameters $\boldsymbol{\beta}$, σ^2, and τ of the hierarchical model given in Equations (2.1) and (2.4) (Keefe et al., 2019), which serves as an automatic prior with favorable properties for inference in Gaussian hierarchical models with ICAR priors. The joint reference prior for $\boldsymbol{\eta}$ in the hierarchical spatial model is given by

$$\pi(\boldsymbol{\beta}, \sigma^2, \tau) \propto \frac{1}{\sigma^2}\frac{1}{\tau}\left[\sum_{j=1}^{n-p}\left(\frac{\xi_j}{\tau+\xi_j}\right)^2 - \frac{1}{n-p}\left\{\sum_{j=1}^{n-p}\left(\frac{\xi_j}{\tau+\xi_j}\right)\right\}^2\right]^{\frac{1}{2}}, \tag{2.11}$$

where $\xi_1, \ldots, \xi_{n-p}$ are the ordered eigenvalues of $Q^{*T}\Sigma_\phi Q^*$ such that the columns of Q^* are normalized eigenvectors corresponding to the non-zero eigenvalues of the projection matrix $G = I_n - X(X^TX)^{-1}X^T$. The prior on σ^2 is $\pi(\sigma^2) \propto 1/\sigma^2$, and the conditional reference prior on the vector of regression coefficients is $\pi(\boldsymbol{\beta}|\sigma^2, \tau) \propto 1$. Thus $\pi(\tau)$ takes the form of (2.11) excluding $1/\sigma^2$. Note that improper priors must be treated carefully when used for model selection, which we address further in Section 2.3.

Subjective information for setting hyperparameters is not always available and expert elicitation is challenging, as evidenced by the absence of such approaches in the spatial statistics

literature. The reference prior obviates the need to choose hyperparameters for priors in hierarchical models for areal data and has been shown to perform well for estimation in spatial ICAR models. Keefe et al. (2019, Section 5 and supplementary material) show that inference procedures based on the reference prior have favorable performance in terms of frequentist coverage rate, average interval length, and mean squared error (MSE) for $\boldsymbol{\beta}$, σ^2, and τ. Thus, the reference prior in (2.11) can be used to reliably estimate all model parameters in $\boldsymbol{\eta}$ and to identify appropriate subsets of covariates. Finally, for the OLM, the joint reference prior for $\boldsymbol{\eta}$ is $\pi(\boldsymbol{\eta}) \propto 1/\sigma^2$. In Section 2.3 we describe how the reference prior in (2.11) can be used with fractional Bayes factors to simultaneously select spatial model structure and covariates.

2.3 Bayesian Model Selection via Fractional Bayes Factors

We next describe simultaneous Bayesian model selection for spatial dependence and covariates based on models (2.9) and (2.10). Bayesian model selection relies on integrated likelihoods of the form

$$p(\mathbf{Y}|M_c) = \int p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\pi(\boldsymbol{\eta}_c|M_c)d\boldsymbol{\eta}_c, \tag{2.12}$$

where model M_c has corresponding parameter vector $\boldsymbol{\eta}_c$ and $c = 1, \ldots, C$.

To compare two models M_1 and M_2, we may use the Bayes factor BF_{12} that is defined as a

ratio of the two models' integrated likelihoods

$$BF_{12} = \frac{p(\mathbf{Y}|M_1)}{p(\mathbf{Y}|M_2)}. \tag{2.13}$$

To alleviate numerical underflow when forming all posterior model probabilities, we form all Bayes factors with respect to a single baseline model, M_l, which has the largest integrated likelihood of all models in the model space $\mathcal{M} = \{M_c,\ c = 1, \ldots, C\}$, where C is the total number of candidate models. From the set of Bayes factors $\{BF_{1l}, \ldots, BF_{Cl}\}$ formed with respect to baseline M_l, the posterior probability of a single model M_r in the model space can then be found using Bayes' Rule:

$$P(M_r|\mathbf{Y}) = \frac{p(\mathbf{Y}|M_r)P(M_r)}{\sum_{c=1}^{C} p(\mathbf{Y}|M_c)P(M_c)} = \left(\sum_{c=1}^{C} BF_{cl}P(M_c) \right)^{-1} \times BF_{rl} \times P(M_r). \tag{2.14}$$

Formulation of posterior model probabilities in Equation (3.3) requires prior probabilities to be assigned to the competing models. For a moment, consider covariate subset selection for K total candidate predictors in the class of OLMs, where $K = p - 1$. We follow the recommendations of Scott and Berger (2010) by first assigning a uniform prior on all groups of models with a fixed number of covariates k, then evenly splitting the share of probability among models in that set. For example, if $K = 2$, then the candidate models all include either zero, one, or two covariates. The model with zero covariates receives 1/3 of the prior probability, as does the model with two covariates. Each model with a single covariate receives 1/6 of the prior model probability. This approach imparts Bayesian multiplicity correction to the selection procedure. We modify the subset selection approach slightly to accommodate both OLMs and spatial models. In this work, we set prior probability for independence at 1/2, and also give 1/2 prior probability to spatial dependence. Thus, we further divide prior probabilities suggested by Scott and Berger (2010) for the independence models

in half, and incorporate all possible subset models with the inclusion of spatial dependence, so that the full candidate model set contains all possible combinations of candidate predictors in both spatial dependence and independence settings. Then the prior probability for a model M_c with k_c covariates is

$$P(M_c) = \frac{1}{2(K+1)} \binom{K}{k_c}^{-1}. \tag{2.15}$$

We develop Bayesian model selection via fractional Bayes factors with the following motivation in mind. If improper priors are assigned to parameters in competing models, the full Bayes factor is defined only up to an undefined constant and thus cannot be used for valid model comparison. In particular, the conditional reference prior on $\boldsymbol{\beta}$ is improper and cannot be used in the full Bayes factor when we consider selection of covariates. For example, consider a comparison between two candidate models, M_1 and M_2, where M_1 represents an intercept-only spatial model and M_2 represents a spatial model with $k \geq 1$ covariates. Then $\pi(\boldsymbol{\eta}) = a_1 \cdot \frac{1}{\sigma^2}\pi(\tau)$ for M_1 and $\pi(\boldsymbol{\eta}) = a_2 \cdot \frac{1}{\sigma^2}\pi(\tau)$ for M_2, where $a_1 \neq a_2$ due to differing sizes for $\boldsymbol{\beta}$. Then the Bayes factor, BF_{12}, becomes

$$BF_{12} = \frac{a_1 \cdot \int p(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \tau, M_1)\pi(\boldsymbol{\beta})\pi(\sigma^2)\pi(\tau)d\boldsymbol{\beta}d\sigma^2 d\tau}{a_2 \cdot \int p(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \tau, M_2)\pi(\boldsymbol{\beta})\pi(\sigma^2)\pi(\tau)d\boldsymbol{\beta}d\sigma^2 d\tau}, \tag{2.16}$$

where a_1/a_2 is an undefined constant and thus BF_{12} is not well-defined. In addition, while σ^2 appears in every model in $\mathcal{M}$, use of the full Bayes factor with $\pi(\sigma^2) \propto 1/\sigma^2$ tacitly assumes the normalizing constant for $\pi(\sigma^2)$ is the same across all models, where σ^2 may include variation from important regressors that are missing in some models. Thus, the same improper prior on σ^2 for models with different specifications of $\boldsymbol{\beta}$ may not be reasonable when performing model selection via full Bayes factors. Assigning proper priors to all model parameters elicits a proper Bayes factor as defined in Equation (3.1). However, specification

of sensible priors for spatial dependence models is difficult. The sensitivity of Bayes factors and the resulting model selection to hyperparameter specification is well established (Kass and Raftery, 1995; Berger and Pericchi, 1996; Chipman et al., 2001; Franck and Gramacy, 2020).

Rather than approaching Bayesian model selection with proper priors, approaches that use training samples to calibrate reference improper priors, including partial Bayes factors and fractional Bayes factors (FBF), have been proposed (O'Hagan, 1995; Berger and Pericchi, 1996). The partial Bayes factor separates out a subset of the data as a training sample, which is then used to update the priors on parameters. The partial Bayes factor uses for training the joint distribution of the specific observations selected for training. Selecting training observations for correlated data is difficult, as a subset of randomly selected points may not contain much information about the dependence structure and the τ parameter. The underlying Markovian structure may be lost by splitting the likelihood based on spatially correlated observations between training and selection, and training based on observations that do not properly reflect the overall dependence structure may result in poor model selection. The intrinsic Bayes factor averages over partial Bayes factors obtained from some or all possible training samples (Berger and Pericchi, 1996). However, this process is computationally expensive and it is not clear if all possible minimal training sets would have the same size necessary to capture the dependence structure. To overcome this difficulty, we develop a FBF approach, which uses a fraction of the likelihood rather than reserving specific observations for training. We thus use FBF methodology to approximate partial Bayes factors. The FBF updates the prior on model parameters $\pi(\boldsymbol{\eta}_c)$ using a fraction $b = m/p$ of the likelihood, obtaining the updated prior

$$\pi^*(\boldsymbol{\eta}_c) = \frac{\pi(\boldsymbol{\eta}_c)\{p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\}^b}{\int \pi(\boldsymbol{\eta}_c)\{p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\}^b d\boldsymbol{\eta}_c}. \tag{2.17}$$

Then the fractional integrated likelihood, $q_c(b, \mathbf{Y})$, for a single model M_c using the updated prior is

$$q_c(b, \mathbf{Y}) = \int \pi^*(\boldsymbol{\eta}_c)\{p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\}^{1-b} d\boldsymbol{\eta}_c, \tag{2.18}$$

where $p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\}^{1-b}$ is the likelihood remaining to calculate the fractional integrated likelihood. Finally, note that we can rewrite the fractional integrated likelihood as

$$\begin{aligned} q_c(b, \mathbf{Y}) &= \int \frac{\pi(\boldsymbol{\eta}_c)p^b(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\{p^{1-b}(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\}d\boldsymbol{\eta}_c}{\int p^b(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\pi(\boldsymbol{\eta}_c)d\boldsymbol{\eta}_c} \\ &= \frac{\int p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\pi(\boldsymbol{\eta}_c)d\boldsymbol{\eta}_c}{\int p^b(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\pi(\boldsymbol{\eta}_c)d\boldsymbol{\eta}_c}. \end{aligned} \tag{2.19}$$

Since the fractional integrated likelihood uses a ratio of two integrals each containing the same $\pi(\boldsymbol{\eta}_c)$, all undefined normalizing constants discussed in Equation (2.16) cancel out in the computation of the fractional integrated likelihood. The FBF, which we denote by BF_{12}^b, for two models M_1 and M_2 is defined as the ratio of two fractional integrated likelihoods. That is,

$$BF_{12}^b = \frac{q_1(b, \mathbf{Y})}{q_2(b, \mathbf{Y})}. \tag{2.20}$$

We use this FBF approach to form posterior model probabilities that are model selection consistent (O'Hagan, 1995), which we demonstrate with a simulation study in Section 2.4.

2.3.1 Parameter estimation and model selection under the FBF

An advantage of a fractional Bayes framework is that the trained prior, denoted $\pi^*(\boldsymbol{\eta}_c)$ for parameter vector $\boldsymbol{\eta}_c$, multiplied by the likelihood after training, is proportional to the same

posterior distribution as if $\pi(\boldsymbol{\eta}_c)$ is used with the full likelihood. That is,

$$\begin{aligned}\pi^*(\boldsymbol{\eta}_c)\{p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\}^{1-b} &= \frac{\pi(\boldsymbol{\eta}_c)\{p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\}^b}{\int \pi(\boldsymbol{\eta}_c)\{p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\}^b d\boldsymbol{\eta}_c} \times \{p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\}^{1-b} \\ &\propto \pi(\boldsymbol{\eta}_c)\{p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\}^b\{p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\}^{1-b} \\ &\propto \pi(\boldsymbol{\eta}_c)p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c) \\ &\propto p(\boldsymbol{\eta}_c|\mathbf{Y}, M_c).\end{aligned} \tag{2.21}$$

Thus, point and interval estimation for parameters $\boldsymbol{\eta}_c$ is unchanged when using a FBF approach for model selection, and the trained prior that results from using the FBF resolves a tension for use of priors for either estimation or model selection. Therefore, researchers may select prior $\pi(\boldsymbol{\eta}_c)$ for the purpose of parameter estimation and benefit from a FBF approach with $\pi^*(\boldsymbol{\eta}_c)$ for model selection without conflict.

2.3.2 Integrated Likelihood Methods

To form the fractional integrated likelihood $q_c(b, \mathbf{Y})$ of a single model M_c using a FBF approach, we need both $\int p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\pi(\boldsymbol{\eta}_c)d\boldsymbol{\eta}_c$ and $\int p^b(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\pi(\boldsymbol{\eta}_c)d\boldsymbol{\eta}_c$. Under the Gaussian hierarchical model with reference prior presented in Section 2.2.2, parameters $\boldsymbol{\beta}$ and σ^2 can be analytically integrated out of the integrated likelihood, but τ must be integrated out via approximation methods. Therefore, the integrated likelihoods for the independent model have tractable expressions. Note that, since we use a prior for τ that depends on a projection of the matrix of covariates for a given model, the fractional integrated likelihood for model M_c depends on the model-specific matrix of covariates. That is, $\boldsymbol{\beta}_c$ is the vector of regression coefficients for covariates contained in M_c, X_c is the matrix of covariates corresponding to $\boldsymbol{\beta}_c$, p_c is the length of $\boldsymbol{\beta}_c$, and $\xi_{c1}, \ldots, \xi_{c,n-p}$ are the ordered eigenvalues of $Q_c^{*T}\Sigma_\phi Q_c^*$ such that the columns of Q_c^* are normalized eigenvectors corresponding to the

non-zero eigenvalues of the projection matrix $I_n - X_c(X_c^T X_c)^{-1}X_c^T$. Then the denominator of the fractional integrated likelihood in (2.19) for a spatial model under the reference prior reduces to the one-dimensional integral

$$\int p^b(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\pi(\boldsymbol{\eta}_c)d\boldsymbol{\eta}_c \propto \int_0^\infty |\Omega|^{-\frac{b}{2}}|X_c^T\Omega^{-1}X_c|^{-\frac{1}{2}}(\tau)^{-1}\left[\frac{b}{2}S_c^2\right]^{\frac{p_c-nb}{2}} \times \left[\sum_{j=1}^{n-p_c}\left(\frac{\xi_{cj}}{\tau+\xi_{cj}}\right)^2 - \frac{1}{n-p_c}\left\{\sum_{j=1}^{n-p_c}\left(\frac{\xi_{cj}}{\tau+\xi_{cj}}\right)\right\}^2\right]^{\frac{1}{2}} d\tau, \tag{2.22}$$

where $\Omega = I_n + \tau^{-1}\Sigma_\phi$ and $S_c^2 = \mathbf{Y}^T(\Omega^{-1} - \Omega^{-1}X_c(X_c^T\Omega^{-1}X_c)^{-1}X_c^T\Omega^{-1})\mathbf{Y}$.

We use an adaptive quadrature approach to approximate integrals $\int p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\pi(\boldsymbol{\eta}_c)d\boldsymbol{\eta}_c$ and $\int p^b(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\pi(\boldsymbol{\eta}_c)d\boldsymbol{\eta}_c$ over τ for the FBF. See the Supplementary Material for further details and complete integrated likelihood expressions for the OLM and spatial ICAR model. Next, Section 2.3.3 discusses the choice of training fraction, specifically the minimal training fraction that will make the integrals in the FBF finite so that adaptive quadrature can be applied.

2.3.3 FBF Training Fraction

The training fraction for the FBF should be chosen to be small while still ensuring propriety of the fractional integrated likelihood. Consider a training fraction of the form $b = m/n$, where m is the corresponding training size. The minimal training size, which we use in this work, is the smallest integer value for m such that $\int p^b(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\pi(\boldsymbol{\eta}_c)d\boldsymbol{\eta}_c$ is finite for all models considered in $\mathcal{M} = \{M_c,\ c = 1, \ldots, C\}$. In particular, if m is chosen to be too small, the integral $\int p^b(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\pi(\boldsymbol{\eta}_c)d\boldsymbol{\eta}_c$ will diverge for one or more models in $\mathcal{M}$ and the corresponding FBF cannot be formed for all models. Additionally, as m increases beyond the minimal training size, the posterior model probabilities more closely resemble the prior

model probabilities. Over–training the prior at the cost of the likelihood forfeits statistical power and reduces the ability of the FBF to detect signal. Thus, the minimal training size should be used when known for a class of models. To understand the behavior of the denominator of the fractional integrated likelihood for the ICAR model, we first consider expressions leading to the integral over τ as in (2.22) by using the eigenvalue decomposition of functions of X and Σ_ϕ.

The following propositions outline the fractional integrated likelihood results which lead to the minimal training size for the FBF with the reference prior for spatial ICAR models. The above text has demonstrated that it is essential for the FBF-trained prior to be proper. The purpose of the upcoming Propositions 2.4 and 2.5 is to find the smallest value of the training size m that will yield a proper FBF-trained prior, overcome the arbitrary constant issue in Equation (2.16), and lead to valid Bayesian model selection. Proposition 2.4 addresses propriety with respect to σ^2 in the denominator of the fractional integrated likelihood and in the updated prior (see Equations (2.19) and (2.17)). For notational convenience, let $p^{(b)}(\mathbf{Y}|\sigma^2,\tau,M) = \int p^b(\mathbf{Y}|\boldsymbol{\beta},\sigma^2,\tau,M)\pi(\beta)d\boldsymbol{\beta}$ and $p^{(b)}(\mathbf{Y}|\tau,M) = \int\int p^b(\mathbf{Y}|\boldsymbol{\beta},\sigma^2,\tau,M)\pi(\boldsymbol{\beta})\pi(\sigma^2)d\boldsymbol{\beta}d\sigma^2$.

Proposition 2.4. *To ensure* $\int p^b(\boldsymbol{Y}|\boldsymbol{\eta},M)\pi(\boldsymbol{\eta})d\boldsymbol{\eta} < \infty$, *consider the tail behavior of* $p^{(b)}(\boldsymbol{Y}|\sigma^2,\tau,M)\pi(\sigma^2)$ *over* σ^2. *For a given value of* τ,

(i) $p^{(b)}(\boldsymbol{Y}|\sigma^2,\tau,M)\pi(\sigma^2) = O\big((\sigma^2)^{\frac{p-nb}{2}-1}\big)$ *as* $\sigma^2 \to \infty$.

(ii) $p^{(b)}(\boldsymbol{Y}|\sigma^2,\tau,M)\pi(\sigma^2) = O\big(\exp\{\frac{-b}{2\sigma^2}S^2\}\big)$ *as* $\sigma^2 \to 0$.

Proof. See Proofs of Main Results in Supplementary Material. □

Next, Proposition 2.5 addresses propriety of the trained prior and $p^{(b)}(\mathbf{Y}|\tau,M)$ with respect to τ after both $\boldsymbol{\beta}$ and σ^2 have been integrated out analytically.

Proposition 2.5. *Assume* $\frac{nb-p}{2} > 0$. *To ensure* $\int p^b(\boldsymbol{Y}|\boldsymbol{\eta}, M)\pi(\boldsymbol{\eta})d\boldsymbol{\eta} < \infty$, *consider the behavior of* $p^{(b)}(\boldsymbol{Y}|\tau, M)$ *over* τ. $p^{(b)}(\boldsymbol{Y}|\tau, M)$ *is a continuous function on* $\tau \in (0, \infty)$ *and*

(i) $p^{(b)}(\boldsymbol{Y}|\tau, M) = O(\tau^{\frac{1-b}{2}+1})$ *as* $\tau \to 0$.

(ii) $p^{(b)}(\boldsymbol{Y}|\tau, M) = O(1)$ *as* $\tau \to \infty$.

Proof. See Proofs of Main Results in Supplementary Material. □

Proposition 2.6 demonstrates that the reference prior for τ is proper.

Proposition 2.6. *(Keefe et al., 2019) The marginal reference prior for* τ *is a continuous function on* $(0, \infty)$ *where*

(i) $\pi(\tau) = O(1)$ *as* $\tau \to 0$

(ii) $\pi(\tau) = O(\tau^{-2})$ *as* $\tau \to \infty$

Theorem 2.7 provides the minimum training size for the application of our FBF approach.

Theorem 2.7. *Consider model* (2.9) *and the reference prior in* (2.11). *The minimal training size for the FBF is* $m = p + 1$.

Proof. See Proofs of Main Results in Supplementary Material. □

Finally, it is straightforward to show that results similar to that of Theorem 2.7 also hold for the OLM with a reference prior as well as for the SAR model for spatial areal data with the independence Jeffreys prior developed by De Oliveira and Song (2008). For both of these models, the minimal training size for the FBF is also $m = p + 1$.

2.4 Simulation Study

To investigate the utility of our FBF approach to simultaneously select covariates and presence of spatial dependence, we perform Monte Carlo simulations for 100 square grid regions of size $n = 10^2, 20^2, 30^2$. To study selection performance for a variety of spatial settings, we examine varying levels of the signal-to-noise ratio. In particular, we fix $\sigma^2 = 1$ for simulations and consider $\tau = 0.01, 0.1, 1, 10, \infty$ for the response, where small values of τ correspond to strong spatial dependence, and infinite τ corresponds to independence. Further, we consider $k = 5$ covariates with $\boldsymbol{\beta} = (5, 2, 1, 0, 0, 0)^T$. Since many spatial applications also contain spatially correlated covariates, covariates are generated from a model of the form (2.9) with mean 0, $\sigma^2 = 1$, and strong spatial correlation where $\tau = 0.1$ for each covariate. We obtain model selection results based on 100 simulated data sets for each combination of these levels of sample size and spatial dependence.

For each simulated data set our method computes posterior model probabilities via FBFs for each of the $2^6 = 64$ models, including 32 OLMs and 32 spatial models with all possible combinations of the $k = 5$ covariates. Additionally, since we use Bayesian model selection, posterior inclusion probabilities for individual fixed effects are easy to calculate. Figure 2.1 displays boxplots of the posterior inclusion probabilities for each of the 5 covariates alongside the probability of selecting a spatial model at each sample size. Red diamonds correspond to the mean probability across the 100 data sets represented in each boxplot. Overall our method correctly assigns high posterior model probability to the correct model, both in terms of spatial structure and identification of null and non-null covariates. For small τ the true model contains spatial dependence, and as sample size increases the probability of selecting a spatial model quickly approaches 1 for $\tau \in \{0.01, 0.1, 1\}$ and moves toward 0 for τ at ∞, which corresponds to the OLM. Note that only τ at ∞ corresponds to true

independence, reducing the model in (2.9) to (2.10). However, practically, most real spatial data sets have small τ less than 10. Thus, for larger finite values of τ, e.g. $10 \leq \tau < \infty$, the true dependence structure is spatial, but there is very little spatial signal to detect. As demonstrated in Figures 2.1b, 2.1d, and 2.1f, the decision to select a spatial model or OLM at $\tau = 10$ is much more equivocal than for small values of τ, due to the diminishing strength of spatial correlation between the observations. Figure 2.1f indicates that for $n = 900$, our method has correctly identified all spatial data sets with small $\tau \in \{0.01, 0.1, 1\}$ as requiring spatial random effects. The most difficult case of non-null covariate selection occurs in Figure 2.1a under $n = 100$ and the strongest level of spatial dependence with $\tau = 0.01$, where we assigned x_2 a low signal-to-noise ratio. Results for all three sample sizes indicate accurate identification of covariates, as the posterior inclusion probabilities for the three null covariates x_3, x_4, and x_5 move towards 0 from the prior, and posterior inclusion probabilities for the non-null covariates x_1 and x_2 quickly approach 1. For any single covariate x, if you sum over the model priors corresponding to each model that contains x, that probability is 1/2. Additionally, recall from Section 2.3 that the prior probability of spatial dependence is 1/2. The horizontal bar in Figure 2.1 corresponds to this prior probability at 1/2, where all posterior probabilities can be seen moving off of this prior.

These results indicate that our method tends to select the correct model in terms of both spatial and fixed effects as n increases. When data is truly independent corresponding to $\tau = \infty$, our method correctly selects the simpler OLM without spatial random effects the majority of the time, it does not attribute high probability to null covariates, and it assigns high probability to non-null covariates with relatively small signal-to-noise ratios. Performance quickly improves as the sample size increases, indicating that our FBF approach provides consistent model selection for fixed and spatial effects simultaneously.

In addition to our FBF model selection approach, we also considered the Deviance Infor-

mation Criterion (DIC, Spiegelhalter et al. (2002)) and the Widely Applicable Information Criterion (WAIC, Watanabe (2010)). The DIC and WAIC are most often computed using the likelihood that includes $\boldsymbol{\phi}$ with estimated values for the spatial random effects plugged into the criteria calculations. The likelihood for $\mathbf{Y}|\boldsymbol{\phi}, \boldsymbol{\beta}, \sigma^2 \sim N(X\boldsymbol{\beta} + \boldsymbol{\phi}, \sigma^2 I_n)$ follows as:

$$p(\mathbf{Y}|\boldsymbol{\phi}, \boldsymbol{\beta}, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\Big\{ -\frac{1}{2\sigma^2}(\mathbf{Y} - X\boldsymbol{\beta} - \boldsymbol{\phi})^T(\mathbf{Y} - X\boldsymbol{\beta} - \boldsymbol{\phi})\Big\}. \quad (2.23)$$

We also consider a version of the DIC that uses the likelihood with the spatial random effects integrated out:

$$p(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \tau) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}|\Omega|^{-1/2}\exp\Big\{ -\frac{1}{2\sigma^2}(\mathbf{Y} - X\boldsymbol{\beta})^T\Omega^{-1}(\mathbf{Y} - X\boldsymbol{\beta})\Big\}, \quad (2.24)$$

where $\Omega = I_n + \tau^{-1}\Sigma_\phi$. The likelihood in (2.24) corresponds to the distribution in Equation (2.9) used in our FBF approach. This version of DIC, called type 2 DIC by Celeux et al. (2006) in the context of missing data models, was studied in Ferreira et al. (2021) for the spatial hierarchical models considered here.

We use an MCMC algorithm to compute the DIC, WAIC, and type 2 DIC for all 64 models for each data set in the simulation study as described above (Gelman et al., 2014). We use the same reference priors when computing DIC, WAIC, and type 2 DIC as for the FBF, so parameters in the ICAR models are assigned the prior in Equation (2.11) and parameters in the OLMs are sampled with prior $\pi(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$. For each ICAR model, we sample parameters in $\boldsymbol{\eta}$ using a Metropolis-within-Gibbs algorithm with a Gibbs step for $\boldsymbol{\beta}$ and a joint Metropolis-Hastings step for τ and σ^2 (Keefe et al., 2019). To obtain samples from likelihood (2.23), we simulate the spatial random effects $\boldsymbol{\phi}$ using composite sampling. The full conditional distribution for $\boldsymbol{\phi}$ and the complete algorithm for sampling from the parameters of the ICAR model are listed in the Supplementary Material. For each OLM, we sample σ^2 from its marginal posterior $p(\sigma^2|\mathbf{Y})$ and use a Gibbs sampler to sample $\boldsymbol{\beta}$ from its conditional posterior $p(\boldsymbol{\beta}|\sigma^2, \mathbf{Y})$. For each model, 30,000 MCMC iterations are obtained

with the first 10,000 iterations discarded as burn-in.

The computations for all results reported here were performed using a $2 \times$ E5 $-$ 2683v42.1GHz (Broadwell) CPU supercomputer from Advanced Research Computing at Virginia Tech. Our FBF selection approach takes 9.28, 65.89, and 365.73 seconds for a single data set with sample size equal to 100, 400, and 900, respectively. The DIC selection approach takes 953.39; 2,491.87; and 6,315.48 seconds for all models for a single data set with sample size equal to 100, 400, and 900, respectively. The WAIC takes 1,041.75; 2,254.34; and 6,156.46 seconds for a single data set with sample size equal to 100, 400, and 900. Finally, the type 2 DIC selection approach takes 1,059.91; 2,350.69; and 6,835.21 seconds for a single data set with sample size equal to 100, 400, and 900, respectively.

For each simulated data set, we calculated the DIC, WAIC, and type 2 DIC for all 64 candidate models using MCMC and identified the model with lowest DIC, WAIC and type 2 DIC values, and the model with highest posterior model probability according to our FBF approach. Figure 2.2 plots the proportion of data sets for which DIC, WAIC, type 2 DIC (abbreviated as DIC-2), and our FBF approach correctly identify the correct covariate structure containing only x_1 and x_2 and the correct spatial model structure, where the true model for τ at ∞ is the OLM with no spatial random effects. As discussed above and seen in Figure 2.1, the true dependence structure when $\tau = 10$ is spatial, but spatial correlation is weak in this setting, making the need for spatial random effects in the model ambiguous. Therefore we compare selection results only at values of $\tau \in \{0.01, 0.1, 1, \infty\}$ in Figure 2.2.

Each panel of Figure 2.2 demonstrates that the FBF approach performs better than DIC, WAIC, and type 2 DIC for selection in all data settings considered here. Our FBF approach also successfully identifies the correct model with respect to spatial random effects more than 80% of the time for $n = 100$ at all levels of spatial dependence, and correctly identifies the spatial model structure for every data set for $n = 400$ and $n = 900$. Additional simulation

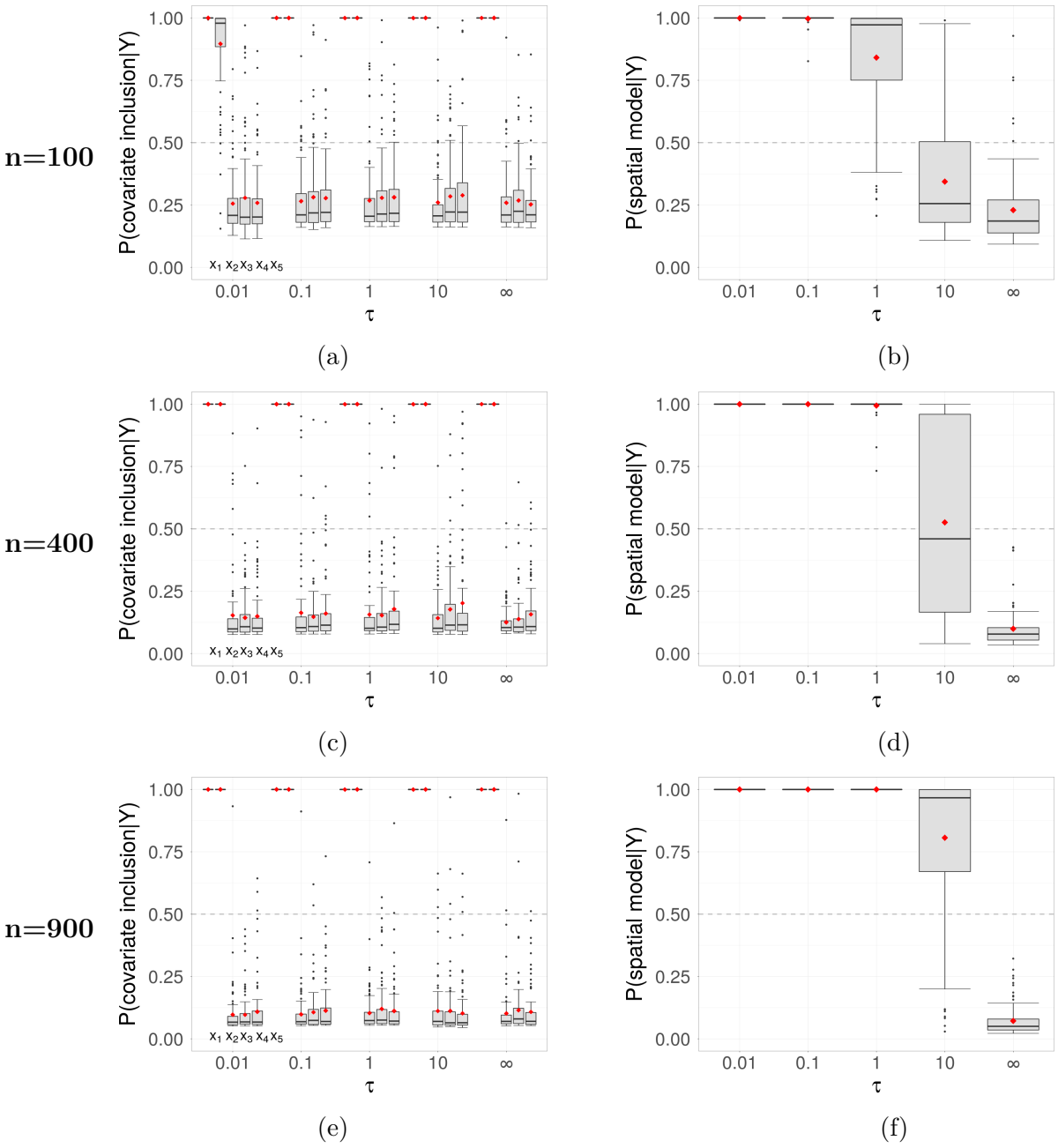


Figure 2.1: Covariate posterior inclusion probabilities (left) and probability of selecting a spatial model (right) for $\tau \in \{0.01, 0.1, 1, 10, \infty\}$ for n = 100 (top row), n = 400 (middle row), and n = 900 (bottom row). Each boxplot represents probabilities from 100 simulated data sets. The reference FBF selection method correctly assigns high probability to non-null covariates, and to spatial models for small τ. The posterior inclusion probabilities for non-null covariates x_1 and x_2 are exactly 1 for all simulated data sets where $n = 400$ and $n = 900$, and for $n = 100$ with $\tau > 0.01$. Thus, the boxplots for these covariates appear as lines at 1.

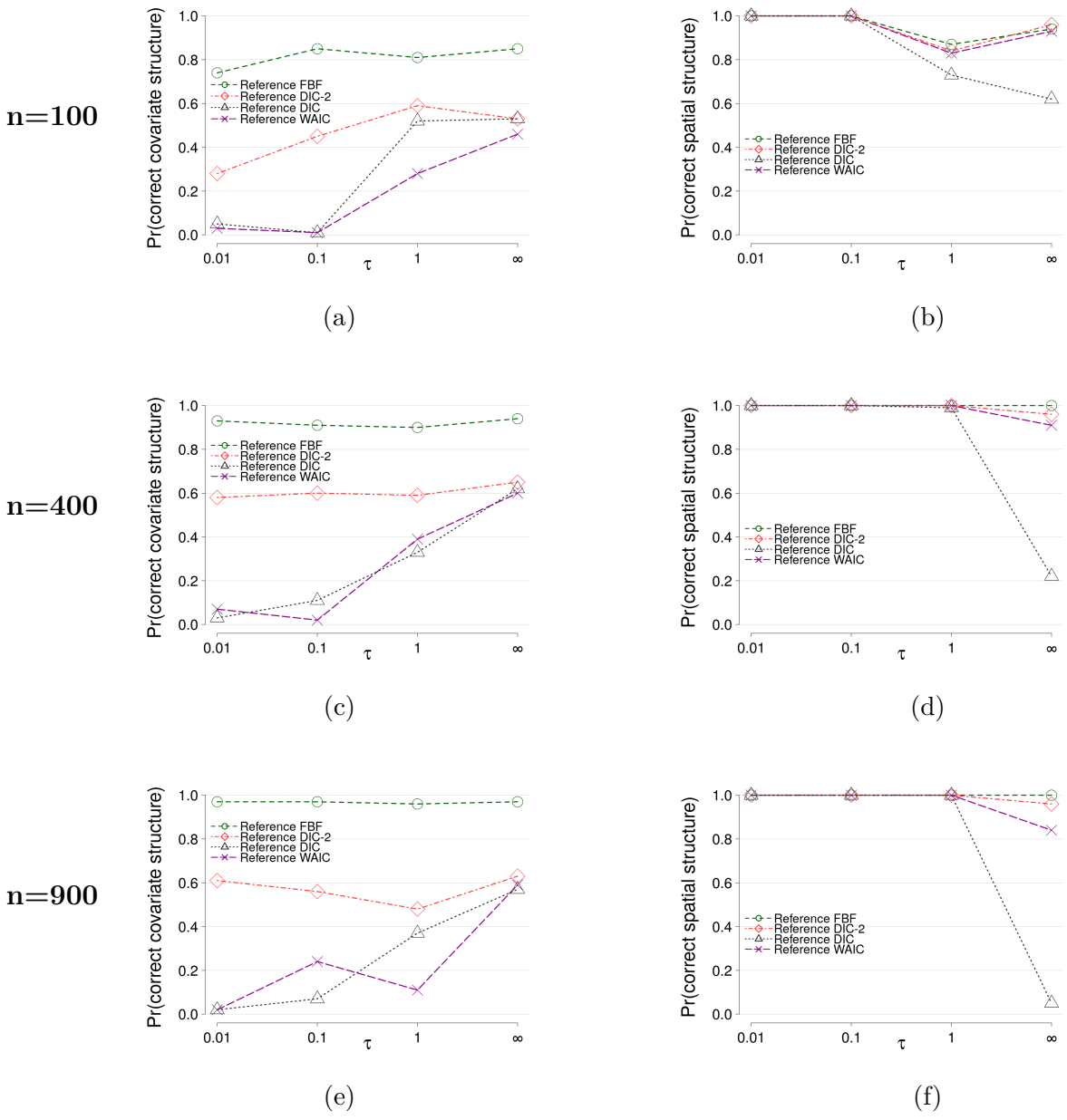


Figure 2.2: Proportion of times out of 100 simulated data sets that the reference FBF, reference DIC-2, reference DIC, and reference WAIC methods select the correct covariate and spatial dependence structure for $\tau \in \{0.01, 0.1, 1, \infty\}$ for n = 100 (top row), n = 400 (middle row), and n = 900 (bottom row). The reference FBF selection method reliably selects covariates and spatial dependence for all values of τ and performs better than DIC-2, DIC, and WAIC for selection in all data settings.

results generated with covariates with no spatial dependence for the sample sizes and coefficient vector described above appear in the Supplementary Material. These results exhibit similar patterns to those in Figure 2.2, as the performance of the FBF is superior to that of the DIC, WAIC, and type 2 DIC for all data settings. Thus, including results provided in the Supplementary Material, for all n and τ considered here, the FBF performs better in each setting in this simulation study. This simulation study demonstrates the reliability of our fully automatic FBF approach to accurately and simultaneously select both spatial model structure and covariate structure.

2.5 Case Studies

To illustrate the practical application of our FBF approach to simultaneously select covariates and spatial random effects in spatial areal datasets of varying sizes, we perform selection for two existing datasets, whose responses include county-level median household income in the contiguous United States and residential crime rates in the neighborhoods of Columbus, Ohio.

To demonstrate the breadth of our method, we show that our FBF approach can also be used to select between different types of spatial random effects. In particular, for the two case studies that follow, we also considered selection with the class of simultaneous autoregressive (SAR) models in the model space $\mathcal{M}$. We adopt the model form and independence Jeffreys prior for the SAR model from De Oliveira and Song (2008). The SAR model for response $\mathbf{Y}$ is given by the following autoregression.

$$\mathbf{Y} = X\boldsymbol{\beta} + (I_n - B)^{-1}\boldsymbol{\epsilon}, \tag{2.25}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ and $B = \gamma W$, with unknown spatial parameter γ and $W = (w_{ij})_{n\times n}$ is a known, symmetric weight matrix with all $w_{ij} \geq 0$ and $w_{ij} > 0$ if $i \in N_j$. As with the ICAR model, we treat adjacent subregions as neighbors with the SAR model. The spatial parameter $\gamma \in (\lambda_n^{-1}, \lambda_1^{-1})$, where $\lambda_1 \geq \lambda_2 \ldots, \geq \lambda_n$ are the ordered eigenvalues of W, and $\gamma = 0$ corresponds to the OLM with distribution $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n)$. Then SAR response $\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \gamma$ has the following Gaussian distribution.

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \gamma \sim N(X\boldsymbol{\beta}, (I_n - B)^{-1} M (I_n - B^T)^{-1}), \tag{2.26}$$

where $M = \sigma^2 I_n$. We consider the independence Jeffreys prior $\pi^J(\boldsymbol{\beta}, \sigma^2, \gamma)$ for SAR model parameters $\boldsymbol{\beta}$, σ^2, and γ (De Oliveira and Song, 2008).

$$\pi^J(\boldsymbol{\beta}, \sigma^2, \gamma) \propto \frac{1}{\sigma^2} \left\{ \sum_{i=1}^{n} \left(\frac{\lambda_i}{1 - \gamma\lambda_i} \right)^2 - \frac{1}{n} \left[\sum_{i=1}^{n} \frac{\lambda_i}{1 - \gamma\lambda_i} \right]^2 \right\}^{\frac{1}{2}}. \tag{2.27}$$

We use the same training size, $m = p + 1$, for the SAR model with independence Jeffreys prior, as the prior induces similar behavior in the integrated likelihood to that produced by the reference prior for the ICAR model. Derivation of the fractional integrated likelihood $q_c(b, \mathbf{Y})$ for the SAR model with independence Jeffreys prior is detailed in the Fractional Integrated Likelihood Calculations section of the Supplementary Material.

Upon including SAR models in the candidate set, we adjust the model priors from Section 2.3. This results in a 50/25/25 split between prior probability for OLMs, ICAR models, and SAR models, with the remaining probability within each class attributed by model size, as described in Section 3. Thus, the prior probability for an OLM M_c with k_c covariates is

$$P(M_c) = \frac{1}{2(K+1)} \binom{K}{k_c}^{-1}, \tag{2.28}$$

and the prior probability for an ICAR or SAR model M_c with k_c covariates is

$$P(M_c) = \frac{1}{4(K+1)} \binom{K}{k_c}^{-1}. \tag{2.29}$$

The following case studies perform selection using the FBF with minimal training size $m = p + 1$ where OLM, ICAR, and SAR models are included in the model space.

2.5.1 Case Study: US Socioeconomic Application

To demonstrate our formal Bayesian model selection approach for areal data, we first consider an application to median household income by county in the contiguous United States in 2017. We consider the logarithm of median household income as the response variable and we select among five candidate predictors: logarithm of the county population in 2017; logarithm of the unemployment rate in 2017; and three indicator variables for whether the county belongs to a large metropolitan area, a medium metropolitan area, or a small metropolitan area. The baseline covariate level corresponds to a non-metropolitan county. Figure 2.3 plots the response variable and all the candidate covariates over a map of the contiguous US counties.

Following the approach to simultaneous selection of spatial model structure and fixed effects presented in Section 2.3 and adapted as described above to include SAR models in the candidate set, we form posterior model probabilities for all 96 models that include either an ICAR, SAR, or independent model structure as in Equations (2.9), (2.26), and (2.10) and every combination of the five covariates. We use a training size of $m = 7$, according to the minimal training fraction found in Section 2.3.3. Model selection using the FBF approach selects with probability 1 the ICAR model of form (2.9) with all five candidate predictors.

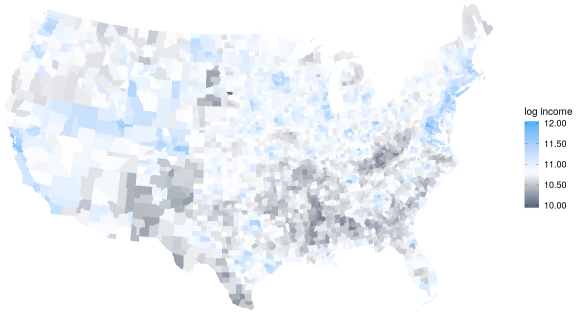
To assess the impact of priors on the model space, we also performed selection among the 96 models using uniform priors on every model. Note that between the ICAR and SAR sets this causes 2/3 prior probability of selecting a spatial model.

The prior probability of the ICAR model with all covariates is equal to 0.0417 under Equation (2.15) and decreases to 0.0104 under the use of equal prior probability for every model. But even with the latter prior specification, the posterior probability of the ICAR model with all five covariates is also equal to 1. Therefore, the model selection result in this case study does not appear to be highly sensitive to reasonable specifications of model prior probabilities.

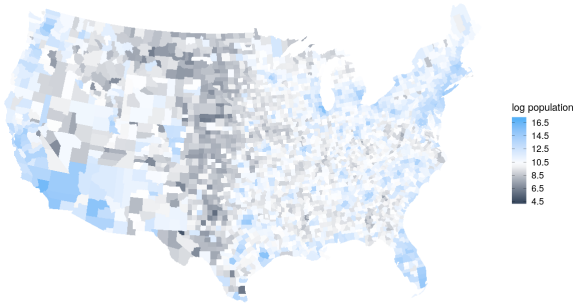
In keeping with the results seen in Figure 2.2, the type 2 DIC also selects the ICAR model with all covariates, with corresponding value equal to -3926.387. In contrast, both the DIC and WAIC select the OLM with all five candidate predictors, with criteria values equal to -2,306.965 and -2,306.213, respectively. The estimated τ value for this data set is 0.1575, which indicates strong spatial dependence. Finally, the results from the simulation study presented in Section 2.4 indicate that the reference FBF selection method is more reliable. These results demonstrate performance of our reference FBF selection method when applied to large spatial data sets.

2.5.2 Case Study: Columbus, OH Crime Rates

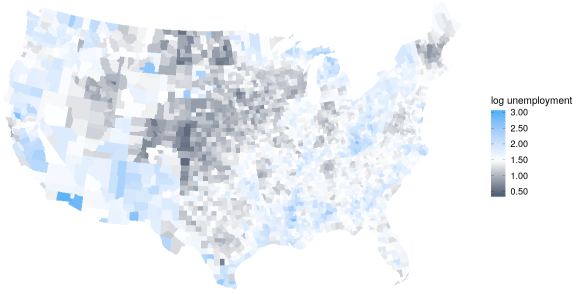
Next we consider a data set containing crime rates in the 49 neighborhoods of Columbus, OH in 1980. This data set has been previously analyzed by Anselin (1988) and Banerjee et al. (2015) and can be obtained from the `spData` package in R (Bivand et al., 2019). The response variable is residential burglaries and vehicle thefts per thousand households in each of the $n = 49$ neighborhoods of Columbus, OH. We consider five available candidate predictors: housing value, household income, open space in the neighborhood, percentage of housing



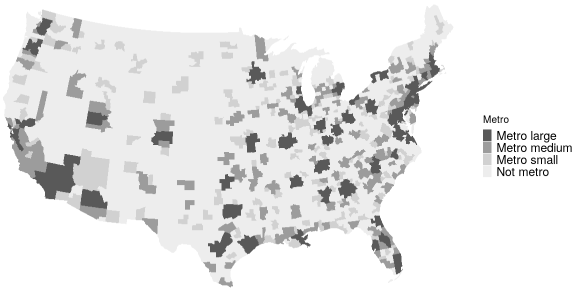
(a) Log median income by county



(b) Log population by county



(c) Log unemployment rate by county



(d) Metro area classification by county

Figure 2.3: Map of United States socioeconomic variables by county in 2017: (a) logarithm of median household income; (b) logarithm of population; (c) logarithm of unemployment rate; (d) metro area classification.

units without plumbing, and distance to the Columbus business district. Using the minimal training size $m = 7$ to select between the 96 candidate models, our FBF approach selects with probability 0.1422 the OLM with three covariates: housing value, household income, and distance to the Columbus business district. Table 2.1 lists the candidate predictors and their corresponding posterior inclusion probabilities; the selected model contains the three covariates with the largest posterior inclusion probabilities. Figure 2.4 plots the response variable and the three selected covariates over a map of the 49 neighborhoods in Columbus, OH. In contrast to the previous case study, $n = 49$ is a small sample size and thus posterior probabilities do not move as far off the model priors. In particular, the prior probability for an OLM with three covariates was 0.0083. The total posterior model probability of selecting an OLM was 0.6770, indicating that the decision about spatial structure for this application has moved only slightly off the 1/2 prior probability of selecting an OLM. Table 2.2 lists the covariate structure, dependence structure, posterior probability, and DIC-2, DIC, and WAIC values for the top models indicated by the FBF approach. The top five models are OLMs and the model with sixth highest probability is the ICAR model containing the same covariates as the selected model. Covariates household income and distance to the Columbus business district have the two largest posterior inclusion probabilities and are included in all of the top 6 models.

We also performed selection for this data set assigning uniform priors on the model space. This setup selected with posterior probability 0.1458 the OLM with the covariates housing value, household income, and distance to the Columbus business district. The prior probability for an OLM with three covariates increased to 0.0104, so the model selected by our initial FBF setup received even more prior mass from uniform priors. The total posterior model probability of selecting an OLM was 0.5089, which is an increase from the prior probability of 0.3333 of selecting an OLM. The posterior inclusion probabilities for the candidate

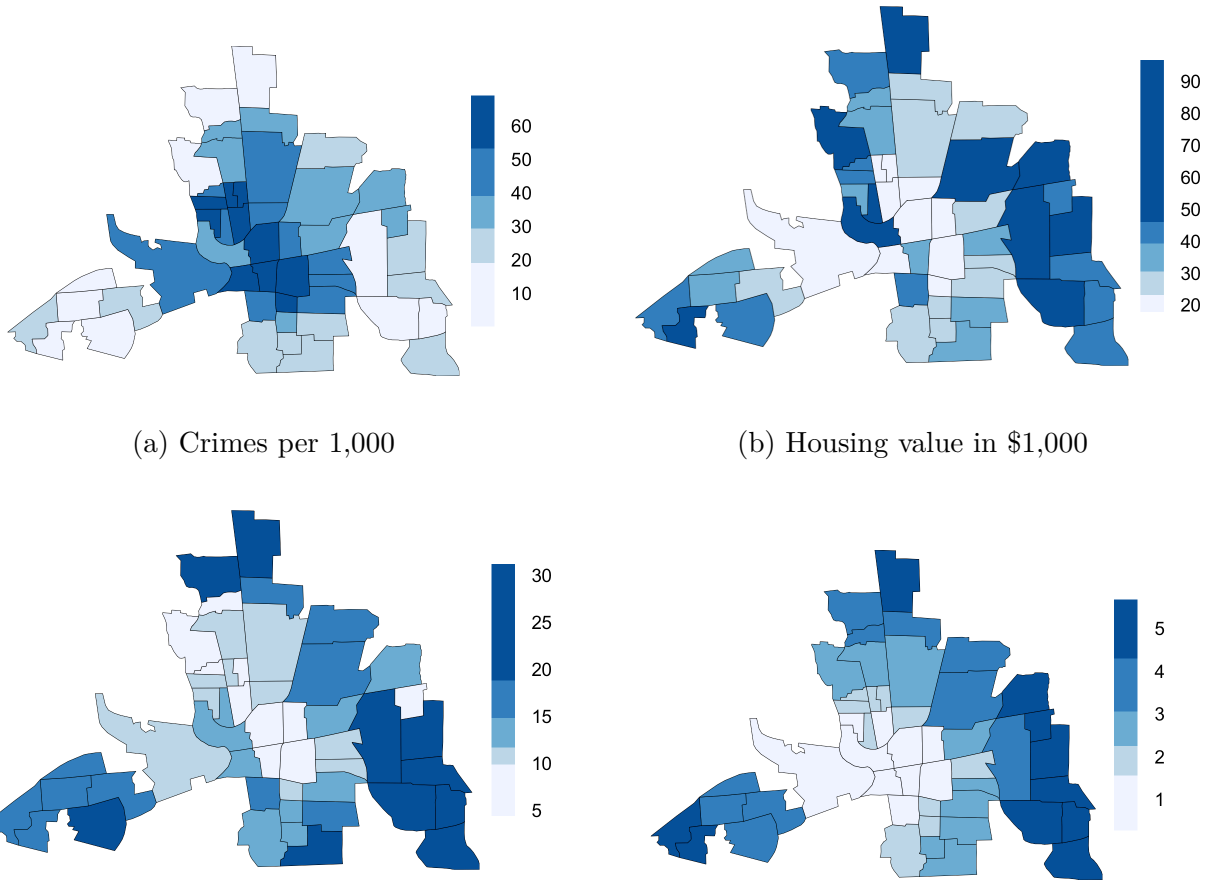
predictors are 0.6827, 0.9033, 0.1816, 0.3002, and 0.8830 when performing selection with uniform model priors. Among OLM and ICAR models, the DIC selects the ICAR model with the covariate housing value and WAIC selects the ICAR model with covariates housing value and open space in the neighborhood. The DIC-2 selects the same OLM as the FBF approach. The DIC-2, DIC and WAIC values for their chosen models are 369.736, 265.814 and 288.481, respectively. This coincides with the simulation study in Section 2.4, which indicates that the DIC and WAIC criteria in particular tend to select spatial models over OLMs more often than the FBF approach does. The estimated value of τ for this data set is 1.9794, which does not indicate strong spatial dependence among the observations. Despite the low sample size, this application highlights the ability of our FBF method to select both independent and spatial data models in real spatial applications.

Covariate description	**Posterior inclusion probability**
housing value	0.733
household income	0.931
open space in the neighborhood	0.302
percentage of housing units without plumbing	0.432
distance to Columbus business district	0.918

Table 2.1: Description and posterior inclusion probability for each of the 5 candidate covariates available for predicting theft and burglary rates in the neighborhoods of Columbus, OH.

2.6 Discussion

We have presented a FBF approach that enables automatic, objective Bayesian model selection for hierarchical models with ICAR spatial random effects. We have derived integrated



(a) Crimes per 1,000

(b) Housing value in $1,000

(c) Household income in $1,000

(d) Distance to business district

Figure 2.4: Map of Columbus, OH variables by neighborhood in 1980: (a) crimes per 1,000; (b) housing value; (c) household income; (d) distance to Columbus business district.

likelihood expressions and the resulting FBFs under the reference prior for areal data, which acts as an automatic prior. We found the minimal training size for the FBF for the hierarchical model with an ICAR prior when the reference prior is assigned to all model parameters, and showed through simulation that our approach provides consistent simultaneous selection of fixed effects and spatial model structure. Notably, our FBF approach provides superior results, in terms of both detection of covariates and spatial dependence, to the widely used

Covariate					Model type	Posterior probability	DIC-2	DIC	WAIC
value	income	open	plumb	distance					
✓	✓			✓	OLM	0.142	369.7	369.8	370.7
✓	✓		✓	✓	OLM	0.126	369.9	369.8	370.7
	✓			✓	OLM	0.123	371.3	371.3	372.6
✓	✓	✓	✓	✓	OLM	0.081	372	372	372.2
✓	✓	✓		✓	OLM	0.061	371.7	371.6	372.1
✓	✓			✓	ICAR	0.060	370.7	330.2	343.8

Table 2.2: Top 6 models for the Columbus, OH crime data according to the reference FBF approach. The first set of 5 columns indicates which of the covariates are in the model with the following abbreviations: value (housing value), income (household income), open (open space in the neighborhood), plumb (percentage of housing units without plumbing), and distance (distance to Columbus business district). Columns 6-10 provide the corresponding model type, posterior model probability, DIC-2, DIC, and WAIC values. The top 5 models are OLMs and the model with 6_{th} highest posterior model probability is the ICAR model with the same covariate structure as the model selected by the FBF approach.

model selection criteria DIC and WAIC. When compared to the type 2 DIC, which is calculated using a likelihood with the spatial random effects integrated out, the performance from our FBF approach is superior and more reliable in simulations. However, the type 2 DIC selects the same model as the FBF approach in each of the two case studies presented in Section 2.5. We have demonstrated in Section 2.4 that the FBF approach implemented with the reference prior performs well for selection in spatial ICAR models, and Keefe et al. (2019) established that the reference prior to have favorable properties for estimation. Thus, the FBF approach provides the ability to use a single prior for both estimation and model selection for spatially dependent areal data. Finally, we showed that our FBF selection approach can be applied to spatial areal data sets of many sizes, and can be generalized to select between different types of spatial random effects (e.g. ICAR versus SAR). As is the case for other variable subset selection approaches, the model space grows exponentially with the number of candidate predictors and, as is well known, exhaustive search becomes

computationally burdensome for large p. In this work we examine problems where the entire model space can be enumerated and assigned posterior model probabilities. When the model space is too large for exhaustive search, our FBF-based model selection approach can still be used in conjunction with a stochastic search algorithm to explore the model space such as the genetic algorithm used by Wu et al. (2020).

There are many possible avenues for future research. First, we note that our reference FBF approach provides posterior model probabilities, which could be used in future research to provide Bayesian model averaging for prediction. Other future work may include developing model selection for data in the exponential family. In particular, ongoing work addresses ICAR effects for Poisson and Binomial response data, which commonly occurs in disease-mapping and health data. In principle, one could extend the proposed methodology to the Poisson case by developing automatic and/or objective priors for ICAR random effects in the Poisson context, deriving the minimal training size, and applying FBF methodology to produce Bayesian model selection for count data models.

Chapter 3

Flexible cost-penalized Bayesian model selection: developing inclusion paths with an application to diagnosis of heart disease

Erica M. Porter[1], Christopher T. Franck[1], Stephen Adams[2]

[1]Department of Statistics, Virginia Tech, Blacksburg, Virginia, 24061, U.S.A.
[2]National Security Institute, Virginia Tech, Arlington, Virginia, 22203, U.S.A.

Abstract

We propose a Bayesian model selection approach that allows medical practitioners to select among predictor variables while taking their respective costs into account. Medical procedures almost always incur costs in time and/or money. These costs might exceed their usefulness for modeling the outcome of interest. We develop Bayesian model selection that uses flexible model priors to penalize costly predictors *a priori* and select a subset of predictors useful relative to their costs. Our approach (i) gives the practitioner control over the magnitude of cost penalization, (ii) enables the prior to scale well with sample size, and (iii) enables the creation of our proposed inclusion path visualization, which can be used to make decisions about individual candidate predictors using both probabilistic and visual tools. We demonstrate the effectiveness of our inclusion path approach and the importance of being able to adjust the magnitude of the prior's cost penalization through a dataset pertaining to heart disease diagnosis in patients at the Cleveland Clinic Foundation, where several candidate predictors with various costs were recorded for patients, and through simulated data.

Keywords: Bayesian model selection, cost penalty, cost-effective.

3.1 Introduction

Medical studies are typically expensive to conduct, with costs measured by time, money, or required expertise. Varying costs for predictor variables arise in settings such as medical diagnoses (Detrano et al., 1989), risk calculators (Struck et al., 2020; Lloyd-Jones et al., 2019; Bang et al., 2009; Ridker et al., 2007), and healthcare quality assessments. When collecting or analyzing data to determine which predictors are most useful, their costs should be taken into account to accommodate available budgets. For example, accurate medical diagnoses are crucial to ensuring that patients receive information and begin treatment promptly, if necessary. Tests and metrics available for diagnosing medical conditions, such as heart disease, can range from relatively inexpensive background questionnaires to highly sophisticated, cutting-edge diagnostic tests. Similarly, risk calculators such as those for chronic diseases take information like easily-obtained family medical history and time-consuming updated tests and imaging to estimate the chances of disease onset. While costs are ubiquitous in gathering medical information and data, few statistical variable selection methods address the cost of individual predictor variables to help medical practitioners decide which to obtain. Perhaps surprisingly, medical data are often reported without their associated costs (Bolón-Canedo et al., 2014). Some methods exist to identify a subset of predictors with lower costs. However, to the best of our knowledge, none of the existing methods can alone provide practitioners easy, considerable control to change the impact cost has on selection results, output readily interpretable probabilities and model parameters, and create a convenient visual to compare many different cost-adjusted analyses at once. We propose a Bayesian model selection approach that introduces a tuning parameter to a cost-penalizing prior on predictors and produces an inclusion path for the practitioner to visually examine the predictive power of predictors relative to their costs as the cost penalization is increased or decreased.

The idea of model selection that accounts for cost has been studied in a few areas of the statistical literature. Most important for our proposed method, when candidate predictors have different costs required to collect them, Fouskakis, Ntzoufras, and Draper (2009a) proposed a model prior that penalizes individual candidate predictors *a priori* based on their costs, with an application to quality of healthcare assessment. We refer to the prior developed by Fouskakis et al. (2009a) as the FND prior, for the three authors of the prior. Bayesian model selection using the FND prior leads to selection of a less costly subset of predictors when compared to Bayesian model selection with no regard to cost. However, we have found that cost penalization from the FND prior does not always scale appropriately with sample size. Namely, as sample size increases, the cost penalization provided by the FND prior is overpowered and may not impact selection.

In Section 3.2 we propose an extension of the FND prior that introduces a tuning parameter to adjust the level of cost penalization for candidate predictors, providing necessary flexibility for medical practitioners to specify the cost penalization for their problem and the decision at hand. The tuning parameter gives the practitioner the ability to directly control the amount of cost penalization and create an inclusion path that visualizes the impact of different cost penalizations on the selection of candidate predictors. For illustration, Figure 3.1 shows the idea: the practitioner controls the cost penalization via tuning parameter on the horizontal axis. The y-axis indicates the value of a chosen inclusion metric for each candidate predictor at different levels of cost penalization the practitioner wishes to study. As the practitioner increases the magnitude of cost penalization, the inclusion metric will tend to decrease for predictors whose cost is high relative to their effect size (i.e. predictor 2 and predictor 3 in Figure 3.1). For our method, we choose to use posterior inclusion probabilities for each predictor as the inclusion metric. Our method can be applied to any binary outcome (e.g. diagnosis) where medical practitioners need to make a decision or prediction based on

predictors with quantifiable costs. Our method can accommodate costs recorded in terms of money, time, equipment, computations, or other measures depending on the medical application, each with the consequence of increasing the burden on overall resources.

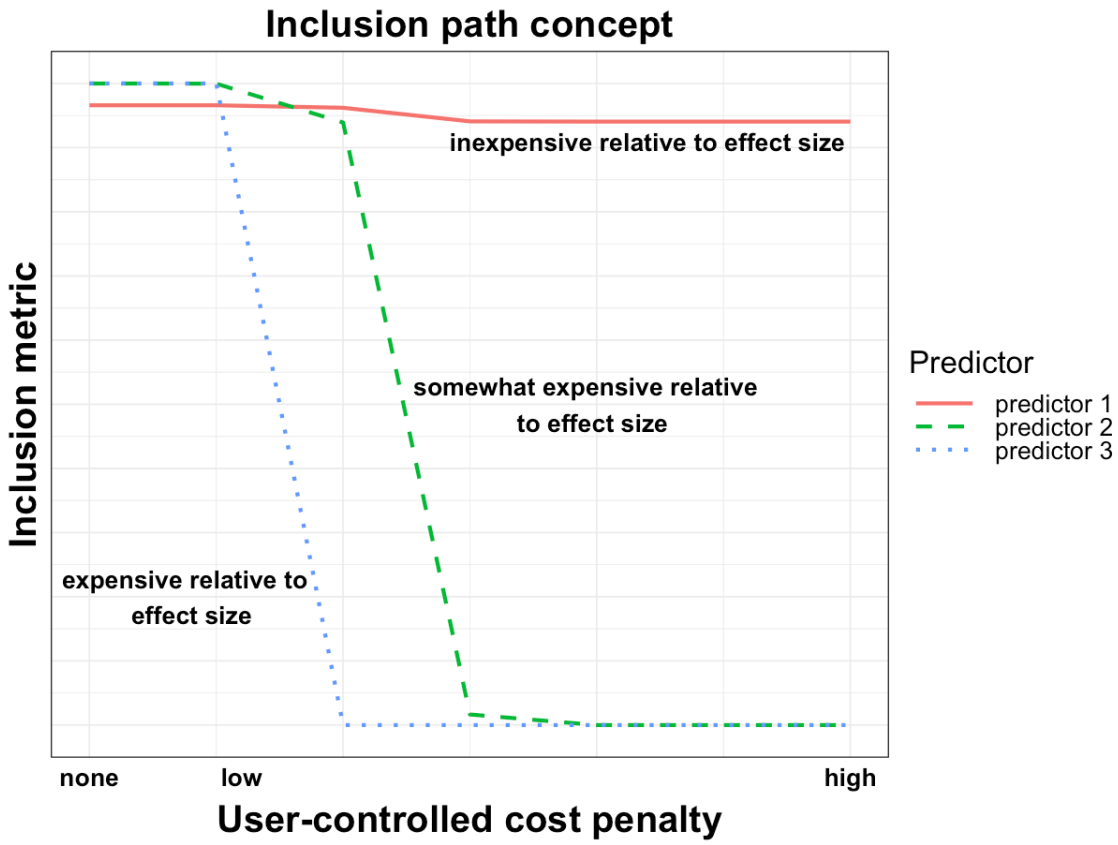


Figure 3.1: Diagram prefacing our proposed inclusion path. The practitioner controls cost penalization to accommodate a budget and uses the inclusion metric value to study how each predictor's importance for modeling the outcome changes as cost is penalized differently.

Machine learning is another area that has seen some development of cost-penalized methods that can be adapted for medical applications. There are three common types of machine learning costs. Missclassification cost, or the cost of incorrectly predicting the label/classification of an observation of the response, has been addressed by adjusting decision points and boosting algorithms (Elkan, 2001; Fan et al., 1999) and active learning addresses labeling cost, i.e. the cost of collecting a value for the response variable (Cohn et al., 1996; Settles,

2009). We focus here on developing methods for penalizing candidate predictor costs. To account for candidate predictor costs, Bolón-Canedo et al. (2014) used a filter model framework, where the evaluation function is a version of Pearson's correlation that accounts for a candidate predictor subset's correlation with the labeled outcome and for multicollinearity between the candidate predictors in the subset. They penalized based on cost by subtracting a weighted average of the cost for each subset of candidate predictors from its corresponding correlation, where the weight is a tuning parameter indicating whether to prioritize the correlation term or the cost-penalizing term. Kong et al. (2016) altered the backward greedy search strategy from selective Bayesian classifiers by iteratively proposing to delete costly candidate predictors while a high level of classification accuracy is maintained. Ling et al. (2004) proposed decision trees that choose a locally optimal predictor at each step, where the splitting criterion is the minimal total cost of training data, instead of the usual minimal entropy. Zhou et al. (2016) further generalized the decision tree approach by using random forests with a probability vector used in the tree construction process. Adams et al. (2016) developed joint parameter estimation and cost-penalized predictor selection specifically for hidden Markov models by using an expectation-maximization algorithm with additional parameters whose prior includes a weight to penalize individual candidate predictor costs. While useful for identifying a subset of predictors with lower costs, none of these individual methods seem to conveniently and simultaneously provide the user-controlled tuning parameter, readily interpretable posterior inclusion probabilities for individual candidate predictors, interpretable model parameters, and useful visualization over a range of cost penalizations that our method provides. Our method allows the practitioner to control the magnitude of cost penalization on candidate predictors so they can meet budgetary requirements and weigh cost and performance for their data using probabilities.

Several methods based on decision theory have also been proposed to penalize for predic-

tor costs. For example, Brown et al. (1999) developed a decision-theoretic approach for multivariate linear regression which added to a quadratic loss function a terminal cost function representing the cost of keeping a particular subset of candidate predictors. Fouskakis and Draper (2008) proposed a decision-theoretic approach for binary outcome generalized linear models that appended a data collection utility component based on marginal predictor costs to the expected utility function, and they applied this utility function to several stochastic optimization algorithms. Recently, Miyawaki and MacEachern (2022) added a cost function to the traditional predictive loss (Lindley, 1968) and applied Bayesian model averaging (BMA) first over purchased predictors and then by marginalizing over potential unpurchased predictors via MCMC. They found that the latter approach performs better than standard BMA but introduces additional sensitivity in prior specification and requires further subjective prior information and assumptions, such as the joint distribution of unobserved predictors. In another MCMC-based approach, Fouskakis et al. (2009b) used a reversible jump MCMC to search the model space constrained to models whose total predictor costs fall below a threshold. In contrast, our method provides a single user-controlled tuning parameter to adjust the magnitude of cost penalization on candidate predictors to produce multiple cost-penalized analyses and produce probabilities for all candidate models and predictors.

The FND prior, which our proposed method extends, penalizes costly predictors relative to a minimum (baseline) cost, and Fouskakis et al. (2009a) used the prior to develop a cost-adjusted selection approach which results in a generalized version of BIC. Fouskakis et al. (2009a) developed cost-adjusted BIC to select among sickness indicators for predicting death within 30 days due to pneumonia. Fouskakis et al. (2009a) compared their selection results to those in which a uniform prior is used for all predictors and models. The latter approach, which Fouskakis et al. (2009a) call a benefit-only analysis, as it ignores costs, selects a more

costly model when applied to a set of $n = 2,532$ pneumonia patients.

When applying the FND prior to other data sets, we have found that the FND prior can lead to selection of less costly models when predictor costs differ, but the cost penalization is not appropriate or sufficient for all sample sizes. In fact, at large sample sizes, the cost penalization imposed by the FND prior greatly diminishes, often causing the resulting cost-penalized model selection to closely resemble that of a standard benefit-only analysis. To establish this phenomenon, we calculated the Kullback-Leibler (KL) divergence between the sets of posterior model probabilities produced by cost-penalized analysis with the FND prior and the benefit-only selection approach as sample size increases for several simulated data sets. Figure 3.2 plots the KL divergence between the posterior model probabilities produced by the two approaches as sample size increases for 10 data sets of initial size $n = 150$. See Section 3.2.5 for more details regarding data generation.

Figure 3.2 shows that the KL divergence value between the cost-penalized and benefit-only methods approaches 0 as the sample size increases. Thus, larger sample sizes lead cost-penalized Bayesian model selection using the FND prior to select a model with structure and cost similar to that of a benefit-only approach. This introduces a paradox since the user adopted the FND prior to control/reduce cost from the selected predictors. But when the sample size of the available data is large, the cost-penalizing ability of the FND prior diminishes and the approach recommends the costly benefit-only model the user was hoping to avoid. Thus, collecting more observations, which often inherently increases medical study costs, can dilute or cancel out the penalization for costly predictors. This phenomenon may lead practitioners and medical researchers to plan for studies that are more expensive than necessary. In this paper we extend the FND prior by proposing simple functions for adjusting the cost penalization according to a tuning parameter. These functions extend the useful FND prior and maintain the property of invariance to cost conversions, making them

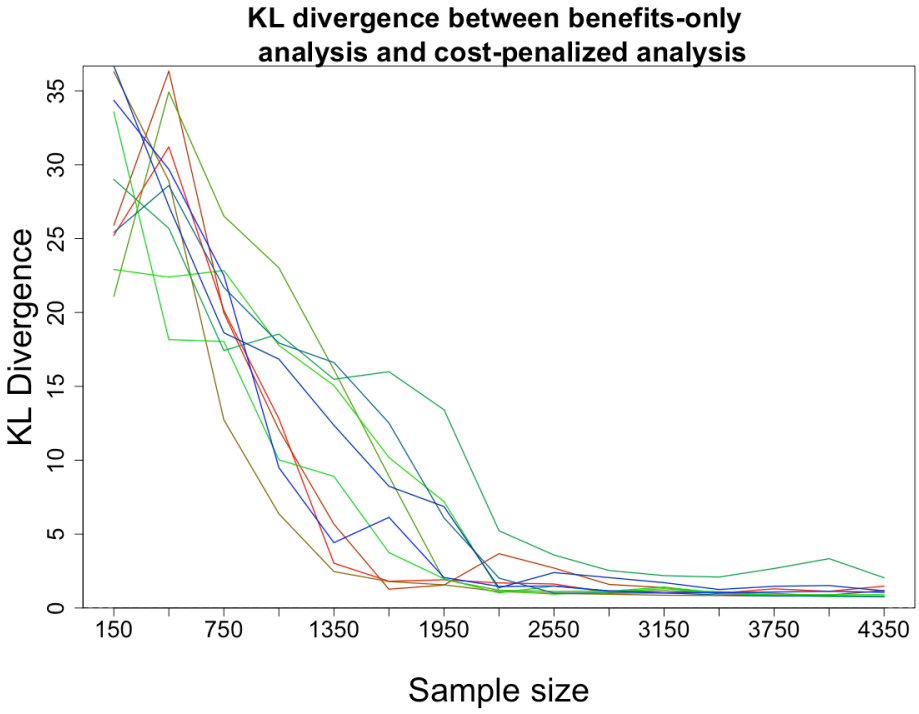


Figure 3.2: KL divergence between posterior model probabilities produced using the FND prior versus benefit-only model selection for 10 data sets across varying sizes. These data were generated from the linear logistic regression setting considered by Fouskakis et al. (2009a). See Section 3.2.5 for more details. The KL divergence between the posterior model probabilities produced by the two selection approaches decreases as the sample size grows for all 10 lines, with 7 of the 10 near 0 by $n = 2500$, indicating that as sample size increases, the impact of the cost penalty on the posterior is reduced.

widely applicable. Our inclusion path approach highlights the change in posterior inclusion probabilities for individual candidate predictors at different levels of the cost penalization according to the functions we suggest. The inclusion path weighs the modeling ability against the cost of the predictors. We demonstrate the utility of our adjusted cost penalization and inclusion path approach first with simulated data and then a data set collected at a medical clinic on Cleveland, Ohio, where the response is the presence of heart disease. There are 13 candidate predictors relevant to diagnosing heart disease available for each patient, with widely varying costs (Detrano et al., 1989). A benefit-only approach selects many costly

predictors that do not appreciably improve modeling of heart disease. We show that our adjusted cost penalization can be used to select models with reduced costs per patient, which can help to meet hospital or insurance budgets, while still retaining cost-effective predictors for physicians to diagnosis of a critical condition like heart disease. Our functions that adjust the cost penalization extend the utility of the FND prior. The resulting inclusion path plots the changing impact of candidate predictors where the practitioner now has input over the magnitude of cost penalization.

The remainder of this paper is organized as follows. Section 3.2 introduces cost-penalized Bayesian model selection, describes the FND prior, details our functions and properties for adjusting the cost penalization on predictors, outlines our inclusion path approach, and explains the setup for our simulated data. Section 3.3 demonstrates the utility of the FND prior and presents our adjusted cost penalization and inclusion path approach, first using simulated data and then when applied to the Cleveland heart disease data. All selection results for our method are compared to results produced by the FND prior. Section 3.4 summarizes the impact of our findings and outlines avenues for potential future research and applications in cost-penalized model selection.

3.2 Data and Methods

3.2.1 Bayesian model selection

We consider Bayesian model selection with the goal of penalizing candidate predictors based on their costs through a flexible class of model priors. Each candidate predictor has an associated cost. We use Bayesian model selection to obtain posterior model probabilities for each possible combination of predictors, with the goal being to select a subset of predictors

that accurately classify the outcome while penalizing on the basis of the costs of those predictors. For p candidate predictors, the corresponding model space is $\mathcal{M} = \{0,1\}^p = \{M_1, M_2 \ldots, M_K\}$, where $K = |\mathcal{M}| = 2^p$ is the total number of candidate models. To compare two models, say M_1 and M_2, we may use the Bayes factor, BF_{12}, that is defined as the ratio of the two models' integrated likelihoods:

$$BF_{12} = \frac{p(\mathbf{Y}|M_1)}{p(\mathbf{Y}|M_2)}, \tag{3.1}$$

or the posterior model odds:

$$PO_{12} = \frac{p(M_1|\mathbf{Y})}{p(M_2|\mathbf{Y})}. \tag{3.2}$$

Then the posterior model probability of a single model M_ℓ in the model space can be found using Bayes' Rule:

$$P(M_\ell|\mathbf{Y}) = \frac{p(\mathbf{Y}|M_\ell)P(M_\ell)}{\sum_{k=1}^{K} p(\mathbf{Y}|M_k)P(M_k)} = \left(\sum_{k=1}^{K} BF_{k\ell} \times \frac{P(M_k)}{P(M_\ell)}\right)^{-1} = \left(\sum_{k=1}^{K} PO_{k\ell}\right)^{-1}, \tag{3.3}$$

where $P(M_k)$ is the prior model probability for model M_k. We now describe the FND prior in Section 3.2.2.

3.2.2 Cost-penalizing model selection

Fouskakis et al. (2009a) developed the FND prior for variable selection for the linear logistic regression setting, with the goal of optimizing selection of sickness indicators for improving quality of health care assessments while penalizing expensive candidate predictors. Their motivating data set from the RAND Corporation was used to select from a list of 83 sickness

indicators to model patient deaths within 30 days of admission due to pneumonia, where costs were measured as the time required to observe/record predictors for each patient (Keeler et al., 1990). In particular, both the benefit-only analysis and cost-penalized analysis using the FND prior selected 13 predictors from 83 total candidate predictors, but the total cost of the two sets of selected predictors were 22.5 and 9.5 (in minutes), respectively, resulting in a cost reduction of more than 50%. Their approach is invariant to cost conversions, devised a penalty related to BIC relative to a baseline cost, and can be used to reproduce the traditional BIC when all predictor costs are equal. Further, by using expressions based on posterior model odds, Fouskakis et al. (2009a) were able to produce posterior model probabilities from a generalized cost-adjusted version of the BIC, which shares many similarities with the approximations and behaviors of the traditional BIC.

Following Fouskakis et al. (2009a), we consider the linear logistic regression setting in this paper, later applying it to the binary diagnosis of heart disease. For convenience, we use the notation of Fouskakis et al. (2009a) for indicator functions and predictor costs. We use a linear logistic regression model where $Y_i \in \{0, 1\}$ and X_{ij} denotes the j_{th} predictor for observation i, where $i = 1, \ldots, n$ and $j = 0, \ldots, p$. Let γ_j be an indicator that is equal to 1 if predictor X_j is included in the model and 0 if it is not. Then $\boldsymbol{\gamma}$ is a length p vector of 0's and 1's indicating whether each of the p predictors are in the model or not. Note that the intercept is included in every model so $X_{i0} = 1$ and $\gamma_0 = 1$ for all models. Then the Bayesian modeling framework is

$$\begin{aligned}
(Y_i|\boldsymbol{\gamma}) &\overset{\text{indep}}{\sim} \text{Bernoulli}[p_i(\boldsymbol{\gamma})], \\
\eta_i(\boldsymbol{\gamma}) &= \log\left[\frac{p_i(\boldsymbol{\gamma})}{1 - p_i(\boldsymbol{\gamma})}\right] = \sum_{j=0}^{p} \beta_j \gamma_j X_{ij}, \\
\eta(\boldsymbol{\gamma}) &= \boldsymbol{X}\text{diag}(\boldsymbol{\gamma})\beta = \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}.
\end{aligned} \tag{3.4}$$

For the vector of regression coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, we assign a Gaussian prior distribution $\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ with form $N(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \Sigma_{\boldsymbol{\gamma}})$. Assuming prior ignorance for the mean and prior event probability $p_i(\gamma) = 1/2$ for all observations, then *a prior* $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is distributed as:

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma} \sim N(\mathbf{0}, 4n(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}). \tag{3.5}$$

The cost-penalized prior by Fouskakis et al. (2009a) for a single γ_j is proportional to:

$$P(\gamma_j) \propto \exp\Big[-\frac{\gamma_j}{2}\Big(\frac{c_j}{c_0} - 1\Big) \log n\Big], \tag{3.6}$$

where c_j is the marginal cost per observation for predictor X_j and $c_0 = \min\{c_j, j = 1, \ldots, p\}$ is defined as the baseline cost, corresponding to the cheapest (least expensive) candidate predictor. Then the FND prior for a particular model corresponding to $\boldsymbol{\gamma}$ follows as:

$$P(\boldsymbol{\gamma}) = \exp\left\{ -\frac{1}{2} \log n \sum_{j=1}^{p} \gamma_j \left(\frac{c_j}{c_0} - 1\right) - \sum_{j=1}^{p} \log\left[n^{-\frac{1}{2}(\frac{c_j}{c_0}-1)} + 1\right]\right\}. \tag{3.7}$$

A natural comparator to (3.7) is a uniform prior on the model space:

$$P(\boldsymbol{\gamma}) = \frac{1}{2^p}, \tag{3.8}$$

which produces the benefit-only selection when used with Equation (3.3). We use a Laplace approximation to approximate the integrated likelihoods with distribution (3.4) and prior (3.5), and we obtain posterior model probabilities for all models in $\mathcal{M}$ as in Equation (3.3).

We develop our inclusion path approach, presented in Section 3.2.3, using posterior inclusion probabilities for the candidate predictors. The posterior inclusion probability for predictor

X_j is obtained by adding up the posterior model probabilities from each model containing X_j, i.e. where $\gamma_j = 1$:

$$P(\gamma_j = 1|\mathbf{Y}) = \sum_{\boldsymbol{\gamma}_k \in \mathcal{M}|\gamma_j=1} P(\boldsymbol{\gamma}_k|\mathbf{Y}). \tag{3.9}$$

Note that (3.7) uses the marginal cost $\{c_j, j = 1, \ldots p\}$ of each candidate predictor, which is the cost of collecting that predictor for a single observation. When adhering to an overall budget, the practitioner may also want to consider the total cost per observation or the total cost of the model. The total cost per observation is equal to the sum of the marginal cost of all predictors in the model, i.e. $\sum_{j=1}^{p} \gamma_j$, and the total cost of the model for the given observations is $n \times \sum_{j=1}^{p} \gamma_j$.

3.2.3 Adjusted cost-penalizing functions

The cost penalization from the FND prior in Equation (3.7) increases with sample size, i.e., through the $\log(n)$ term. However, as demonstrated in Figure 3.2, the cost penalization does not tend to grow quickly enough with n, so the FND prior may not be sufficient for all data sets or budgets, particularly in the case of large sample sizes. We propose an extension of the FND prior that gives the practitioner more flexibility when penalizing predictors based on their costs *a priori*. One way to improve the flexibility of the FND prior is the ability to adjust the cost penalization rather than relying on a fixed penalization for every application with cost. If the cost of the model selected based on existing data exceeds a current or future budget, it would be crucial to be able to increase the cost penalization to find an effective but affordable model. For other data, it might be the case that the model selected using the FND prior does not provide the desired performance and then it may be crucial to decrease the cost penalization to allow for selection of additional predictors that would improve the

model's overall performance while still penalizing for high costs, just to a lesser extent. To give practitioners more control over the FND prior so that it may scale to their problem, we propose two functions of the cost ratio c_j/c_0 according to a tuning parameter b. These functions change the magnitude of cost penalization in the FND prior.

We propose functions $g(c_j/c_0)$ to adjust the cost ratio according to tuning parameter b that satisfy the following properties:

(a) $b = 0$ implies $g(c_j/c_0) = 1$ for all $j = 1, \ldots, p$, reducing the prior to the uniform model prior (3.8) and resulting in a benefit-only model selection.

(b) $b = 1$ makes $g(c_j/c_0)$ equal to c_j/c_0 and reproduces the FND prior as seen in Equation (3.7).

(c) When $b > 1$, $g(c_j/c_0)$ penalizes predictors with costs $c_j > c_0$ more highly than the FND prior in (3.7). When $0 < b < 1$, $g(c_j/c_0)$ penalizes predictors with costs $c_j > c_0$ less than the FND prior but more than a benefit-only analysis. Higher values of b increase $g(c_j/c_0)$ and the resulting cost penalization, leading to lower prior inclusion probabilities for predictors with costs above the baseline.

(d) When $c_j = c_0$, the j_{th} candidate predictor cost is the same as the baseline cost. Changing b does not introduce/increase penalization on any of the predictors with baseline cost.

Two sensible choices for functions of the cost ratio are linear and exponential functions, described below.

Linear function of the cost ratio

First we propose a linear function of the cost ratio for each $c_j > c_0$ according to a slope equal to the difference between c_j and c_0. For a given value of the tuning parameter b, the linear function of the cost ratio for predictor X_j is

$$g\left(\frac{c_j}{c_0}\right) = \frac{(c_j - c_0) \cdot b + c_0}{c_0}. \tag{3.10}$$

Note that (3.10) satisfies properties (a)-(d). We call the prior that uses the cost ratio function in (3.10) the linear cost prior (LCP). That is, the LCP is the model prior with form (3.7) where the cost ratio c_j/c_0 is replaced with the cost ratio function $g(c_j/c_0) = ((c_j - c_0) \cdot b + c_0)/c_0$.

Exponential function of the cost ratio

Next we propose an exponential function of the cost ratio c_j/c_0. For a given value of the tuning parameter b, the exponential function of the cost ratio for predictor X_j is

$$g\left(\frac{c_j}{c_0}\right) = \left(\frac{c_j}{c_0}\right)^b, \tag{3.11}$$

which also satisfies properties (a)-(d). We call the prior that uses the exponential cost ratio function in (3.11) the exponential cost prior (ECP). That is, the ECP is the model prior with form (3.7) where the cost ratio c_j/c_0 is replaced with the cost ratio function $g(c_j/c_0) = (c_j/c_0)^b$.

3.2.4 Inclusion paths

Using the LCP and ECP and varying b, we create an inclusion path for each candidate predictor. To do this, we compute and plot posterior inclusion probabilities for each of the candidate predictors across multiple values of b. The inclusion path visualizes a probabilistic way to learn the order in which the candidate predictors would be chosen or discarded, e.g. according to some threshold, if a different magnitude of cost penalization is used. The costs $\{c_j, j = 1, \dots p\}$ remain unchanged; $g(c_j/c_0)$ changes how heavily candidate predictors with larger costs are penalized *a priori*. Our inclusion path technique is inspired by the interpretable path diagrams like LASSO and ridge (Tibshirani, 1996; Hoerl and Kennard, 1970). From the plot of inclusion paths, the researcher can all at once study candidate predictors' posterior inclusion probabilities for the benefit-only analysis, the FND prior, and a wide range of either the LCP or ECP with differing b values. This is because we formulated the LCP and ECP to satisfy properties (a)-(d) so that the established uniform priors and the FND prior can be easily recreated from either the LCP or ECP.

Next we outline the simulation settings that we will use to illustrate our inclusion path approach.

3.2.5 Simulation study settings

To study the model selection behavior resulting from the FND prior, LCP, and ECP, consider a simulated data set of size $n = 150$ with form (3.4) and vector of regression coefficients

$$\boldsymbol{\beta} = (1, 0, 0, 0, 0.5, 0.5, 0.5, 0.8, 0.8, 0.8)^T, \tag{3.12}$$

and corresponding cost vector

$$\boldsymbol{c} = (1, 3, 9, 1, 3, 9, 1, 3, 9), \tag{3.13}$$

representing predictors with all combinations of null, smaller, and larger effect sizes and baseline, cheap, and expensive costs. The predictors $X_1, \ldots, X_9$ are generated independently from a $N(0, 1)$ distribution. The vector of regression coefficients (3.12) and predictor costs (3.13) were also used to generate the 10 series of data sets increasing in size used to calculate KL divergence values between the benefit-only analysis and selection using the FND prior in Section 3.1.

In practice, medical costs can be measured and recorded according to money, time, labor, and many other quantities. The FND prior and our LCP and ECP extensions are all invariant to cost units/conversions since the cost penalizations are expressed relative to the baseline cost. Since monetary cost resonates with many populations, we choose to consider the costs for our simulated data in dollar amounts. Thus, our simulation setting has 512 candidate models with costs ranging from \$0.00 (corresponding to the intercept-only) and \$39.00 per observation. Section 3.3 presents selection results for data generated as above using Bayesian model selection as described in Section 3.2 with uniform priors, the FND prior, the LCP, and the ECP applied to the model space.

3.3 Results

3.3.1 Simulation results with existing cost-penalizing methods

Consider a data set of size $n = 150$ generated as described in Section 3.2.5. Table 3.1 highlights the difference between the selection results from a benefit-only approach and selection using the FND prior. We can see from Table 3.1 that the benefit-only model selection ap-

proach selects the model with the following predictors: the cheap and expensive predictors with smaller effect size and all predictors that have the larger effect size. This model has a total cost of \$25.00 per observation. The posterior model probability for this benefit-only model is 0.332, moving off of the prior probability of 0.002, and the corresponding C-statistic is 0.84. Meanwhile, model selection using the FND prior on the model space selects the model with the following predictors: X_7 (baseline predictor with larger effect size) and X_8 (cheap predictor with larger effect size), for a total cost of \$4.00 per observation, a 84% reduction in cost from the benefit-only model. The posterior model probability for this model is 0.491, moving off of its prior probability of 0.0008, and the corresponding C-statistic is 0.782. We can see that the FND prior leads to a remarkable reduction in the cost per observation and successfully prioritizes predictors with larger effect sizes while it considers their costs, while incurring modest loss in accuracy as measured by the C-statistic.

However, as is often the case with Bayesian methods, the influence of the prior is reduced as the sample size increases. In this case, the fixed cost penalization from the FND prior diminishes as the sample size increases. As a result, for large sample sizes, the FND prior's cost-penalized selection more closely resembles the benefit-only model selection. To view the changing impact of the cost penalization on selection as sample size increases, consider a similar data set of size $n = 600$. To mimic collection of additional data, we added 450 new data points simulated from the settings in (3.12) and (3.13) to the set of $n = 150$ studied in Table 3.1, resulting in a data set of size $n = 600$. For this larger data set we again performed model selection first using a benefit-only analysis with uniform priors (3.8) and then using cost-penalized Bayesian model selection with the FND prior on the model space. Both selection approaches choose the same model with all 6 predictors with non-null effect sizes, with a total cost of \$26.00 per observation. Table 3.2 highlights the identical selection results from these two analyses for the size $n = 600$ data set. The C-statistic for the

Benefit-only selection

$n = 150$

Predictor	**Effect size**	**Cost level**	**Cost**
X_1	null	baseline	1
X_2	null	cheap	3
X_3	null	expensive	9
X_4	smaller	baseline	1
X_5	smaller	cheap	3
X_6	smaller	expensive	9
X_7	larger	baseline	1
X_8	larger	cheap	3
X_9	larger	expensive	9
			$25/obs

FND selection

$n = 150$

predictor	**Effect size**	**Cost level**	**Cost**
X_1	null	baseline	1
X_2	null	cheap	3
X_3	null	expensive	9
X_4	smaller	baseline	1
X_5	smaller	cheap	3
X_6	smaller	expensive	9
X_7	larger	baseline	1
X_8	larger	cheap	3
X_9	larger	expensive	9
			$4/obs

Table 3.1: Candidate predictors and their corresponding effect sizes and costs. Rows colored in orange ■ are the predictors selected by (left) the benefit-only analysis and (right) selection using the FND prior for a data set of size $n = 150$. At this small sample size, the benefit-only model costs more than 4 times the model selected using the FND prior.

model containing every non-null predictor is 0.865, with posterior model probability equal to 0.594 from the benefit-only analysis (model prior 0.002) and 0.927 using the FND prior for selection (model prior 2e-29).

With the two selection results being identical, we can see that the cost penalization provided by the FND prior was ineffective after the addition of only 450 more observations. Thus, the reward for incurring the expense of additional data collection is to recommend a more expensive model, the same as when no cost penalty is used. If looking to decide which predictors to collect, for example, in a future study, a practitioner may use the FND prior and come to very different decisions based on the size of their existing data, by only a few hundred observations. This points to a need to be able to adjust the cost penalization

to provide cost-effective model selection for each medical application at hand. Section 3.3.2 demonstrates how the LCP and ECP can be used with our inclusion path approach to adjust the cost penalization so that it can be maintained at different sample sizes.

Benefit-only selection

$n = 600$

Predictor	**Effect size**	**Cost level**	**Cost**
X_1	null	baseline	1
X_2	null	cheap	3
X_3	null	expensive	9
X_4	smaller	baseline	1
X_5	smaller	cheap	3
X_6	smaller	expensive	9
X_7	larger	baseline	1
X_8	larger	cheap	3
X_9	larger	expensive	9
			$26/obs

FND selection

$n = 600$

predictor	**Effect size**	**Cost level**	**Cost**
X_1	null	baseline	1
X_2	null	cheap	3
X_3	null	expensive	9
X_4	smaller	baseline	1
X_5	smaller	cheap	3
X_6	smaller	expensive	9
X_7	larger	baseline	1
X_8	larger	cheap	3
X_9	larger	expensive	9
			$26/obs

Table 3.2: Candidate predictors and their corresponding effect sizes and costs. Rows colored in orange ■ are the predictors selected by (left) the benefit-only analysis and (right) cost-penalized selection using the FND prior for a data set of size $n = 600$. After the addition of only 450 more observations to the data set studied in Table 3.1, the benefit-only selection and Bayesian model selection using the FND prior choose the model with the same predictors and cost per observation.

3.3.2 Inclusion paths using adjusted cost penalization

Visualizing the effect of the tuning parameter b enables the practitioner to see the change in all the predictors' importance in the posterior for a range of cost penalizations. This forms

the basis for our proposed inclusion path. Here, we demonstrate the proposed inclusion path approach with the LCP and ECP assigned to the model space. Consider a data set of size $n = 450$, where the impact of the FND prior begins to diminish, simulated as described in Section 3.2.5. We first apply the LCP with cost ratio function (3.10) and a range of values for the tuning parameter b. We apply each of the resulting LCP model priors to the model space and obtain posterior inclusion probabilities for each of the 9 candidate predictors. Then we create an inclusion path by plotting the posterior inclusion probabilities as a function of tuning parameter b. Figure 3.3 plots the inclusion path for each combination of candidate predictor cost and effect size when the LCP is used to adjust the amount of cost penalization using a linear function of the cost ratio.

From Figure 3.3, the posterior inclusion probabilities for 8 of the 9 predictors decrease as the cost penalization is increased via larger values of b. The predictor with the larger effect size and baseline cost maintains a posterior inclusion probability at 1 across all values of b, as it is not penalized *a priori* by the FND, LCP, or ECP priors, according to property (d) from Section 3.2.3. The posterior inclusion probabilities for the cheap and expensive null predictors are 0.09 and 0.25, respectively, in the benefit-only analysis corresponding to $b = 0$. The values of $b = 0.25$ and 0.5 introduce cost penalization that is less severe than that from the FND prior but higher than the (nonexistent) cost penalization from the benefit-only analysis. The posterior inclusion probabilities for the cheap and expensive null predictors drop to 0.02 and 0.0007 when $b = 0.25$ is used with the LCP, and their corresponding posterior inclusion probabilities equal 0 for $b > 0.5$ and $b > 0.25$, respectively. Thus, even slightly penalizing these null predictors for having costs above the baseline helps to move their inclusion probabilities to 0. The posterior inclusion probability of the null predictor with baseline cost is 0.12 in the benefit-only analysis and equals 0.05 for $b \geq 3.5$, as this predictor has a null effect but is not penalized by any variation of the LCP.

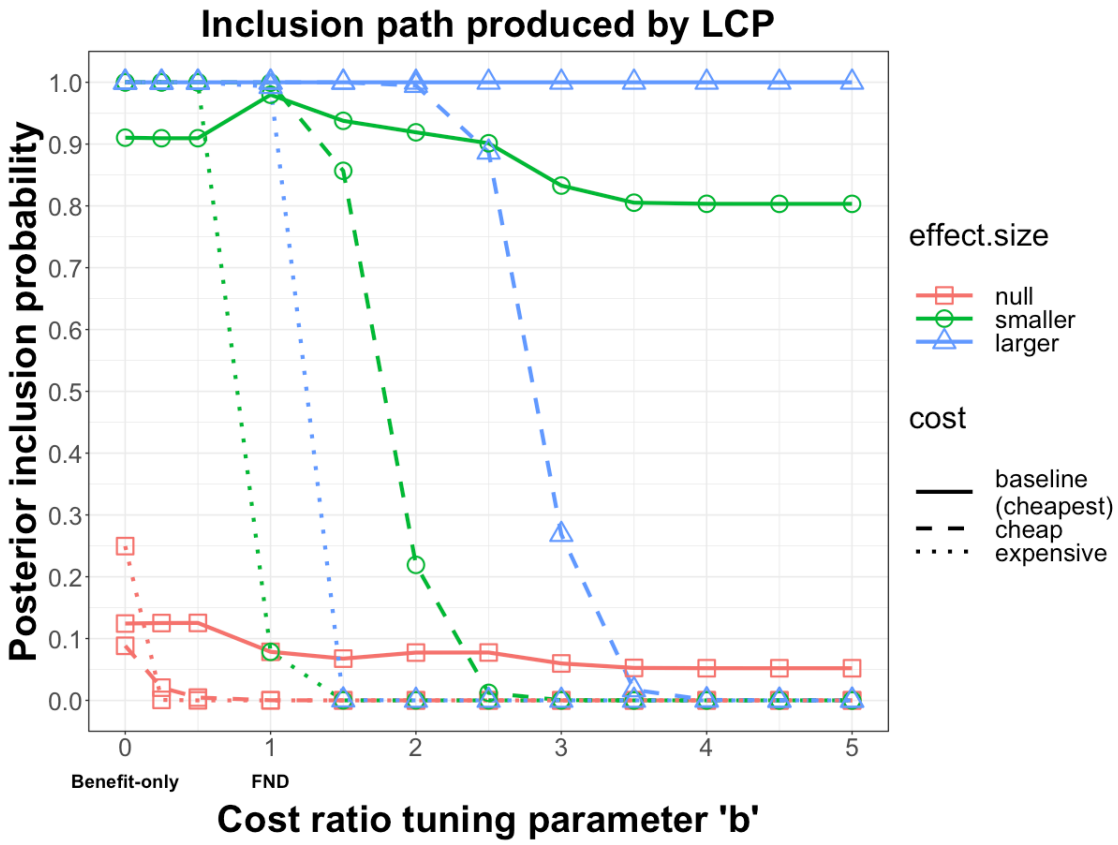


Figure 3.3: Posterior inclusion probabilities for each of the 9 predictors with baseline, cheap, and expensive costs and null, smaller, and larger effect sizes for the $n = 450$ data set. The x-axis represents the tuning parameter used to linearly adjust cost penalization through the cost ratio. Model selection was performed using the LCP on the model space with tuning parameter values $b = 0$, 0.25, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5. Larger values of b represent an increase in the cost penalization for all predictors with cost above the baseline, which is the cost of the cheapest (least expensive/costly) candidate predictor.

When determining which predictors are selected with each variation of the LCP, we can consider comparing the posterior inclusion probabilities for the remaining non-null predictors to a threshold, e.g. 0.5 as in the median probability model (Barbieri and Berger, 2004). The inclusion path allows us to see which predictors' inclusion probabilities meet the desired threshold for the different versions of cost-penalized selection at one time. For example, in the benefit-only analysis, which corresponds to the LCP with $b = 0$, every one of the 6 non-null predictors have posterior inclusion probabilities beyond the 0.5 threshold under

consideration. Using the FND prior, which corresponds to the LCP with $b = 1$, the posterior inclusion probabilities for all non-null predictors except for the expensive one with smaller effect size exceed the 0.5 threshold. Then, by increasing b to 3 and using the LCP for Bayesian model selection, only the non-null predictors with baseline cost and smaller and larger effect sizes have posterior inclusion probabilities above 0.5. Thus, our inclusion path makes it clear that the models selected by each of these analyses would contain 6, 5, and 2 total predictors, respectively. The practitioner may choose a different threshold for the posterior inclusion probabilities to suit their application. The goal of the inclusion path is to provide a way to select the best model that conforms to the practitioner's budget by deciding which predictors to invest in. We use 0.5 for illustration here.

Next, Figure 3.4 plots the inclusion path for each of the 9 candidate predictors when the ECP, which adjusts the amount of cost penalization using an exponential function (3.11) of the cost ratio, is used on the model space with tuning parameter values b between 0 and 5. For this data set, the order in which the posterior inclusion probabilities for individual predictors decrease and cross the 0.5 probability threshold is the same as it was for the LCP in Figure 3.3. Specifically, the posterior inclusion probabilities decrease with b first according to cost and then by effect size. Using the ECP, the posterior inclusion probabilities are much closer for the expensive predictor with larger effect size and the cheap predictor with smaller effect size, as all costs above the baseline are penalized at a faster rate for $b > 1$ than when linear adjustment is applied. Now imagine that the practitioner is planning a study and wishes to refrain from collecting some of the predictors, either to lower the overall cost or to be able to record more total observations. If the cost per observation for all 9 predictors studied here is too high, the practitioner might consider dropping/excluding predictors in an order first according to cost and then according to effect size, as indicated by the paths in Figures 3.3 and 3.4. A visual such as those in Figures 3.3 and 3.4 is important because it

enables the practitioner to see how the different levels of cost penalization affect the posterior inclusion probabilities of individual predictors while also accounting for their effect sizes.

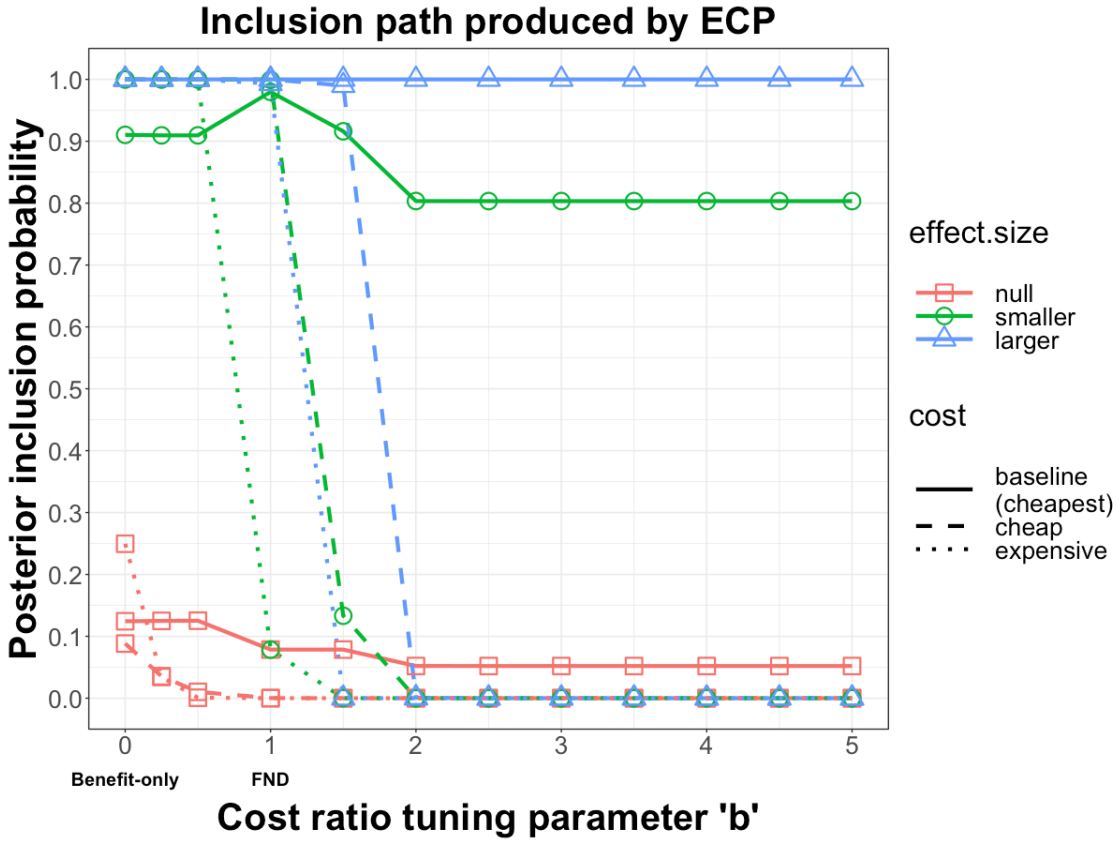


Figure 3.4: Posterior inclusion probabilities for each of the 9 predictors with baseline, cheap, and expensive costs and null, smaller, and larger effect sizes for the $n = 450$ data set. The x-axis represents the tuning parameter used to exponentially adjust cost penalization through the cost ratio. Model selection was performed using the ECP with cost ratio (3.11) with tuning parameter values $b = 0$, 0.25, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5.

Note that the choice of which prior, LCP or ECP, to use for the inclusion path should depend on the data set and overall budget. Note that for b between 0 and 1 the LCP decreases prior inclusion probability and the resulting posterior inclusion probabilities for non-baseline predictors at a faster rate than the ECP does, much as a straight line is steeper than an exponential curve close to the origin. But for $b > 1$ the ECP decreases the prior inclusion probabilities more quickly, as the exponential rate surpasses the slope of the linear

function. If a degree of cost penalization less than that provided by the FND prior is desired, the ECP may produce a more useful inclusion path, as the posterior inclusion probabilities will change more gradually for $0 < b < 1$, creating a more distinct ordering for the individual inclusion paths.

3.3.3 Case study: selecting cost-effective predictors to model diagnosis of heart disease

We apply our adjusted cost-penalizing model selection to a data set of clinical test results originally from Detrano et al. (1989). The data consists of the medical records of $n = 297$ patients collected at the Cleveland Clinic Foundation in 1988. We use these data for illustration because diagnosis of medical conditions such as heart disease is an important classification problem and the data are available along with the financial cost of each candidate predictor. We obtained these data from the UCI Machine Learning Repository. The response is a binary variable that indicates the presence of heart disease in each patient; $Y_i = 1$ if the patient has heart disease and $Y_i = 0$ if the patient does not have heart disease, where $i = 1, \ldots, n$. There are 13 candidate predictors. A summary of numeric candidate predictors appears in Table 3.3 and categorical candidate predictors are summarized in Table 3.4. Thus, there is a total of 8,192 models in the model space $\mathcal{M}$. Each predictor has an associated cost per patient, listed in the third column of Tables 3.3 and 3.4. The costs of the individual predictors range from \$1.00 to \$102.90, so the cost per observation ranges from \$0.00 (intercept-only) to \$600.57 for all 13 predictors per patient. Costs listed in Tables 3.3 and 3.4 are per patient and are specified in Canadian dollars, based on information from the Ontario Health Insurance Program.

Predictor name	**Description**	**Cost (in $)**	**Mean**	**St. dev.**	**Odds ratio***
age	Patient age (years)	1	54.54	9.05	1.61
resting BP	Resting blood pressure (mm Hg)	1	131.69	17.76	1.37
cholesterol	Serum cholestoral (mg/dl)	7.27	247.35	51.99	1.18
heart rate	Maximum heart rate achieved	102.90	149.60	22.94	0.36
ST depression	ST depression induced by exercise relative to rest	87.30	1.06	1.17	2.91

Table 3.3: Numeric candidate predictors for the Cleveland heart disease data. *The odds ratios are expressed in terms of an increase of one standard deviation in the predictor.

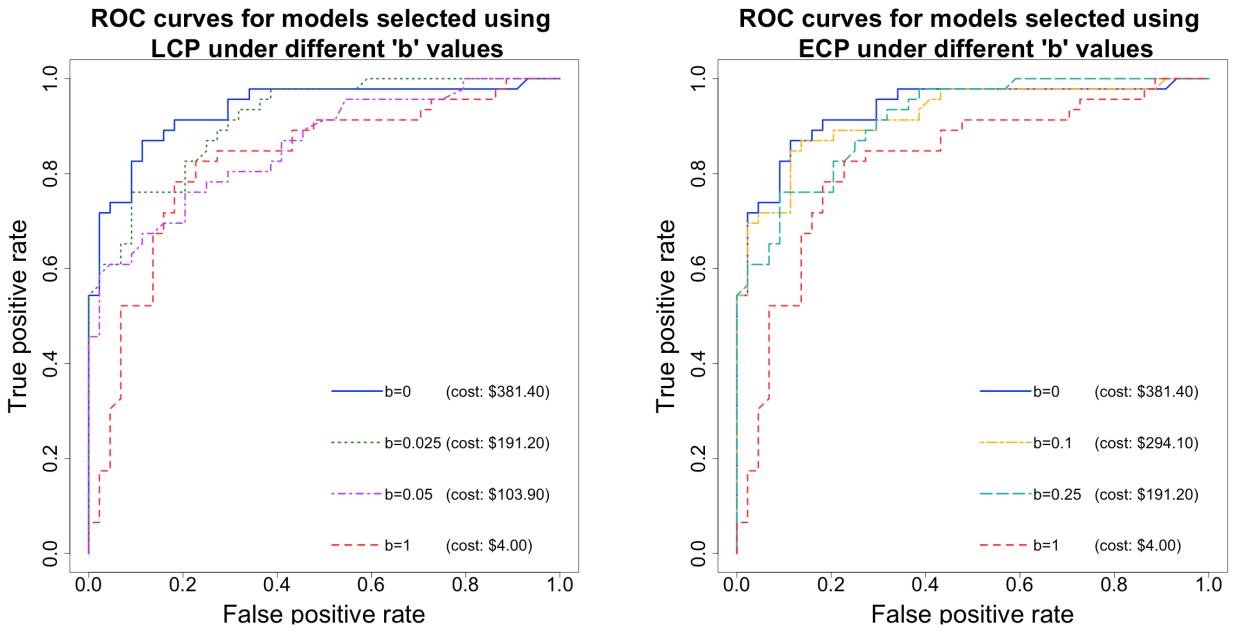
We performed Bayesian model selection using the heart disease data, applying a benefit-only approach and cost-penalized model selection as described in Section 3.2 with the FND, LCP, and ECP priors assigned to the model space. Benefit-only model selection with uniform priors (3.8) on the model space selects the model with the following predictors: sex, chest pain, resting BP, ST depression, peak ST segment, major vessels, and defect type with corresponding posterior model probability 0.072, up from a prior model probability equal to 0.000122 and with a total cost of $381.40 per patient. Model selection using the FND prior on the model space selects the model with the four baseline predictors (sex, age, chest pain, and resting BP), with corresponding posterior model probability equal to 0.5. This model that contains only baseline predictors has prior model probability equal to 0.063 using the FND prior, and a cost equal to $4.00 per patient. While cost-penalized selection with the FND prior leads to a less costly model that is only 2.3% of the cost of the benefit-only model, this model performs worse than more costly models in terms of classification. Figure 3.5 displays ROC curves and corresponding costs for models selected using different values of b with the LCP and ECP. Detection of a health condition such as heart disease is critical

Predictor name	**Description**	**Cost (in $)**	**Values**	**Percent observed**	**Odds ratio****
sex	Patient sex	1	0-Female	32.3%	-
			1-Male	67.7%	3.57
blood sugar	Fasting blood sugar	5.20	0-False	85.5%	-
			1-True	14.5%	1.02
exercise angina	Exercise-induced angina	87.30	0-No	67.3%	-
			1-Yes	32.7%	7.00
chest pain	Chest pain type	1	1- Typical angina	7.7%	-
			2-Atypical angina	16.5%	0.51
			3-Non-anginal pain	27.9%	0.63
			4-Asymptomatic	47.8%	6.04
EKG	Resting electrocardiogram results	15.50	0-Normal	49.5%	-
			1-ST-T wave abnormality	1.3%	5.02
			2-Probable/definite left	49.2%	1.97
peak ST segment	Slope of the peak exercise ST segment	87.30	1-Upsloping	46.8%	-
			2-Flat	46.1%	5.31
			3-Downsloping	7.1%	3.81
major vessels	Major vessels colored by flourosopy	100.90	0	58.6%	-
			1	21.9%	6.01
			2	12.8%	12.70
			3	6.7%	16.24
defect type	Type of heart defect	102.90	3-Normal defect	55.2%	-
			6-Fixed defect	6.1%	6.87
			7-Reversible defect	38.7%	11.19

Table 3.4: Categorical candidate predictors for the Cleveland heart disease data. **Odds ratios measure the shift in multiplicative odds from the reference category (denoted by -).

for doctors to be able to provide effective treatment and advice to affected patients. The cost $4.00 model might not have acceptable false positive rates for the hospital to trust the

diagnosis, and thus it is desirable to be able to adjust the cost penalization, particularly between $0 < b < 1$, for these data. Our method is key for striking a balance between the benefit-only approach and the FND approach, by furnishing a spectrum of cost penalization choices. The C-statistics for the models considered in Figure 3.5a are 0.935, 0.935, 0.934, and 0.857 for the models selected using the LCP with $b = 0$, 0.025, 0.05, and 1, respectively. We can see that penalizing based on predictor costs (e.g. with $b = 0.025$) can lead to great cost reduction without loss of performance compared to the benefit-only model. Perhaps surprisingly, the model with cost equal to only \$4.00 per patient has a C-statistic of 0.857, which indicates that it might have some utility in a triage scenario. However, our method provides the ability for practitioners to see which predictors can supplement the model to improve diagnosis beyond easily-obtained predictors like age and sex and a subjective pain rating from the patient.



(a) (b)

Figure 3.5: ROC curves for models selected for the Cleveland heart disease data using Bayesian model selection with the (a) LCP and (b) ECP model priors and different values of tuning parameter b.

To study the individual candidate predictors for diagnosing heart disease, we applied our inclusion path approach with adjusted cost penalization to the heart disease data from Detrano et al. (1989). Figure 3.6 contains the inclusion path using the LCP with tuning parameter values of b between 0 and 0.25. The lines corresponding to each predictor's inclusion path are color-coded such that lines corresponding to predictors with the same cost are the same color; darker shades of red correspond to more costly predictors, and then line types differ between distinct predictors that have the same cost. Values of b above 0.25 are not shown here, as the posterior inclusion probabilities do not change for $b > 0.25$ for these data, and the cost \$4.00 model containing only the baseline predictors is selected with highest posterior model probability for all $b > 0.1$.

From Figure 3.6 we can see that the posterior inclusion probabilities for the baseline predictors sex and chest pain are high for the benefit-only analysis, and remain equal to 1 for all values of $b \geq 0.05$. The posterior inclusion probability for baseline predictor age is only 0.09 in the benefit-only analysis, increases to 0.47 at $b = 0.05$ and then reaches and remains at 0.96 for $b \geq 0.05$. The final baseline predictor, resting BP, has posterior model probability equal to 0.59 in the benefit-only analysis. Its posterior inclusion probability is 0.86 at $b = 0.025$ and equals 0.53 for $b \geq 0.1$, always remaining above a 0.5 probability threshold. The posterior inclusion probabilities for the remaining 9 predictors with costs above the baseline decrease towards 0 for $b > 0$. Predictors heart rate, exercise angina, cholesterol, EKG, and blood sugar have posterior inclusion probabilities below 0.5 in the benefit-only analysis which all move towards 0 as the cost penalty increases, indicating that these predictors would not be chosen based on a benefit-only or a cost-penalized analysis. The posterior inclusion probabilities for ST depression and defect type are equal to 0.75 and 0.96, respectively in the benefit-only analysis. However, both have posterior inclusion probabilities below 0.5 for all versions of the LCP that penalize for cost with $b > 0$, indicating

that their costs may outweigh their effect sizes. Finally, the peak ST segment predictor has high posterior inclusion probability only for $b = 0$ and 0.025 and the probability for the major vessels predictor exceeds 0.5 for $b = 0, 0.025$, and 0.05.

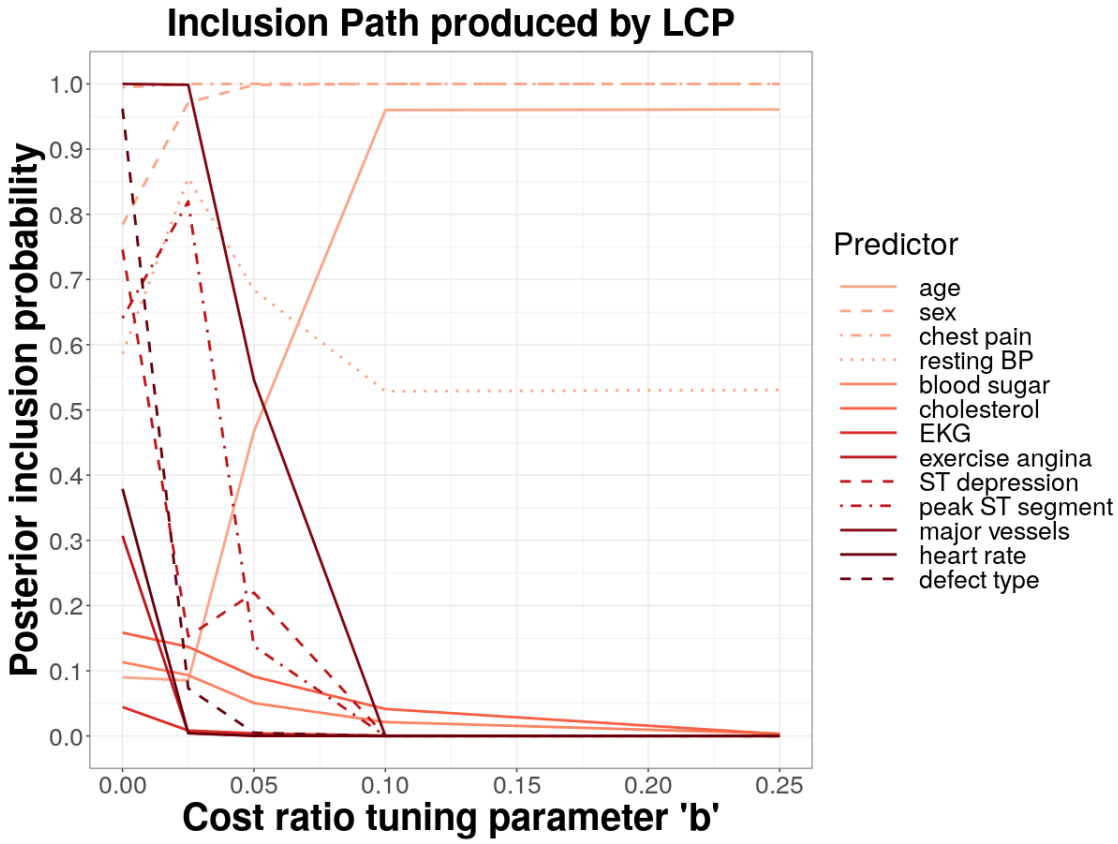


Figure 3.6: Posterior inclusion probabilities for each of the 13 predictors from the heart disease data. The x-axis represents the tuning parameter used to linearly adjust cost penalization by using a function of the cost ratio. Model selection was performed using the LCP with cost ratio function (3.10) and tuning parameter values $b = 0$, 0.025, 0.05, 0.1, and 0.25. Values of b above 0.25 are not shown here because there is no further change in the posterior inclusion probabilities.

Figure 3.7 contains the inclusion path using the ECP with exponential function of the cost ratio from Equation (3.11); lines corresponding to each predictor are again color-coded according to cost and distinguished by line type from other predictors with the same cost. The order of the inclusion paths is the same as in Figure 3.6, i.e. the non-baseline predictors whose posterior inclusion probabilities decreased towards zero for smaller values of b with the

LCP also decrease towards zero with smaller values of b here. However, the ECP provides a more gradual change in prior and thus posterior inclusion probabilities for $0 < b < 1$. Thus, Figure 3.7 is appealing to make decisions about which predictors to use in addition to the four baseline predictors to improve the ability of the model to accurately diagnose heart disease. Since we saw from Figure 3.5 that the cursory, baseline predictors included at $b = 1$ may not be ideal by themselves, we might think of sliding b below 1 until reaching a set of predictors that provide a cost-penalized model whose performance is adequate and whose overall cost satisfies a hospital or patient budget. Based on a 0.5 probability threshold, a hospital might consider, in addition to the four baseline predictors that are easy to obtain, also including predictors such as major vessels, peak ST segment, defect type, and ST depression. Suppose, for example, that a particular hospital has a maximum budget of $200.00 per patient for the purpose of diagnosing heart disease. The hospital could recommend its providers record the four baseline predictors for each patient, as well as the major vessels and peak ST segment predictors, for a total cost of $192.20 per patient. The inclusion paths make it clear which predictors would be included/excluded from selection at each level of cost penalization, and can be used to visualize inclusion for several predictors in any medical setting with quantifiable costs over a wide range of cost penalizations.

3.4 Discussion

We have presented an approach to adjust cost-penalizing model priors for cost-adjusted Bayesian model selection. Our approach extends the well-established FND prior (Fouskakis et al., 2009a) by giving the practitioner the ability to adjust the level of cost penalization on candidate predictors and visualize the resulting inclusion probabilities. We proposed linear and exponential functions, according to a tuning parameter b, of the ratios of marginal

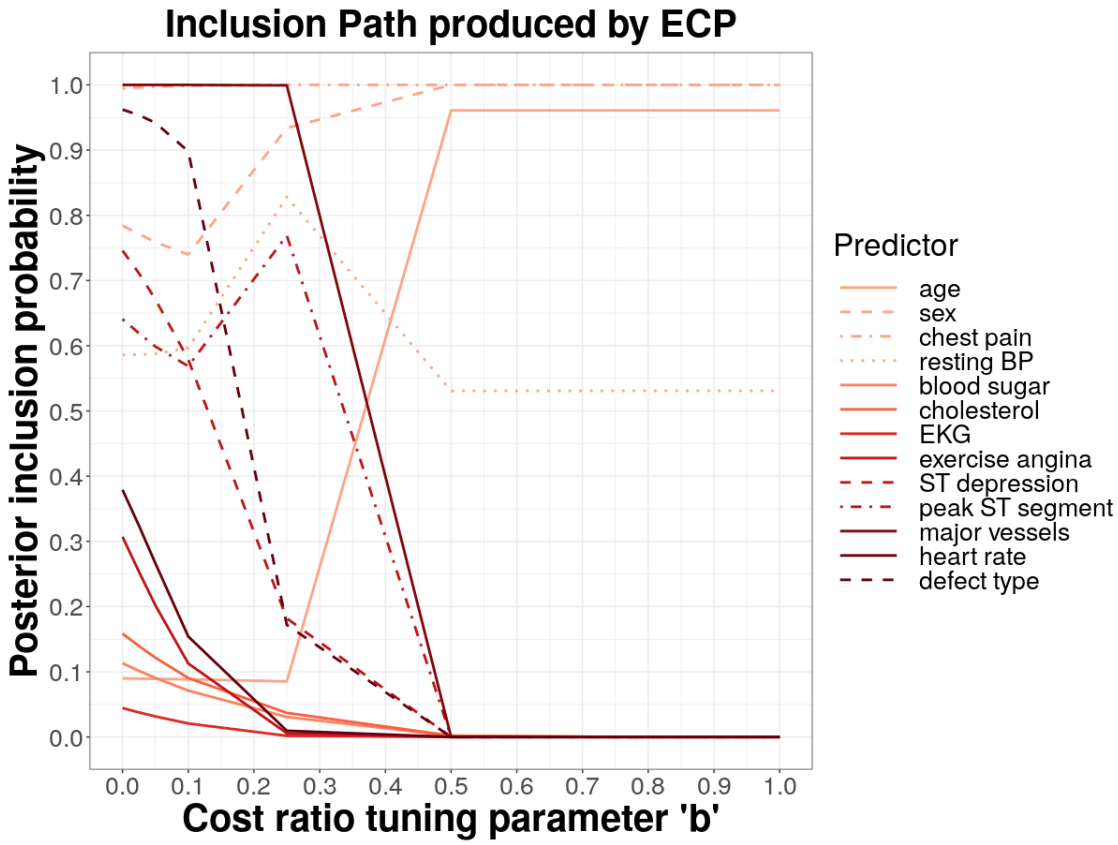


Figure 3.7: Posterior inclusion probabilities for each of the 13 predictors from the heart disease data. The x-axis represents the tuning parameter used to exponentially adjust cost penalization by using a function of the cost ratio. Model selection was performed using the ECP with tuning parameter values $b = 0$, 0.025, 0.05, 0.1, 0.25, 0.5, 0.75, and 1.

predictor costs relative to a baseline cost. The resulting LCP and ECP priors using these functions of the cost ratios provide adjusted levels of cost penalization controlled by the practitioner via the tuning parameter. The properties that our cost ratio functions adhere to ensure that our adjusted priors can easily reproduce a benefit-only analysis and allow for unit conversion without changing model selection results, making it useful for costs measured in a variety of ways. We have shown, through simulation, that adjusting the cost penalty according to either of our proposed functions helps to maintain the cost penalization for larger sample sizes. Our inclusion path approach, which plots the change in posterior inclusion probabilities together across a range of adjusted cost penalties, provides a visual

tool to learn the relative importance of predictors when accounting for their costs and to make decisions about individual predictors to meet an overall budget.

This work extends the utility of the FND prior by adjusting the penalization on costly predictors in model selection. For example, suppose that the model selected using the FND prior has a total predictor cost that exceeds the practitioner's or hospital's designated budget. Then our proposed inclusion path can easily be used to see which predictor(s) have probabilities that fall below the desired threshold as the practitioner slides b towards higher values. Similarly, if the practitioner seeks a higher-performing model that still penalizes candidate predictor costs to some extent, they can slide b down, closer to 0, to learn which additional predictor(s) will improve model fit without causing an undue increase in the cost per observation. We presented a case study where it may be useful to adjust the amount of cost penalization to be less than that in the FND prior by using values of $0 < b < 1$. Specifically, we applied our approach to a data set of 297 heart disease patients and found that decreasing the cost penalization from the FND prior helps to identify predictors that can improve the diagnosis of heart disease while still appropriately penalizing the most costly predictors. The LCP and ECP both provide useful options for adjusting the magnitude of cost penalization. Notably, between 0 and 1 the LCP increases the penalty on costly predictors at a faster rate than the ECP does, while for $b > 1$ the ECP penalizes predictor costs more quickly.

Avenues for future research include developing cost-penalized model selection for a sequential decision-making framework, for example, in medical diagnoses where practitioners may order additional tests depending on a patient's initial results. Another extension may include determining whether collecting additional observations or including more predictors would maximize performance or precision of a model, when subjected to a budget.

Together, our tuning parameter-based functions of the cost ratios and our inclusion path

proposed in this manuscript give the practitioner considerable flexibility to weigh each predictor's cost with its modeling ability, with probabilities providing a concrete measure of inclusion, especially relative to other predictors.

Chapter 4

Objective Bayesian Analysis for Areal Data Using R package ref.ICAR

4.1 Introduction

The `ref.ICAR` package version 2.0 provides functions to conduct objective Bayesian analysis for areal data using the reference prior proposed by Keefe et al. (2019) for all model parameters. This model provides an approach for modeling spatially correlated areal data, using an intrinsic conditional autoregressive (ICAR) component on a vector of spatial random effects with a reference prior for all model parameters. Ferreira et al. (2021) developed faster MCMC sampling for the ICAR model with reference prior, and Porter et al. (2023b) developed objective Bayesian model selection based on fractional Bayes factors for the same model. By providing functions for sampling, inference, and model selection based on the automatic reference prior, `ref.ICAR` enables users to perform a full Bayesian analysis without the need to expertly specify priors or hyperparameters. Porter (2019) described version 1.0 of `ref.ICAR`.

4.2 ref.ICAR Functions

`ref.ICAR` can be used to analyze areal data corresponding to a contiguous region, provided a shapefile or neighborhood matrix and data. The functions implemented by `ref.ICAR` are summarized below.

- `shape.H()` Takes a file path to a shapefile and forms the appropriate neighborhood matrix for analysis with the ICAR reference prior model.
- `ref.MCMC()` Implements the posterior sampling algorithm proposed by Keefe et al. (2019). Generates MCMC chains for parameter and regional inferences.
- `ref.summary()` Provides posterior inferences for the model parameters, τ, $\boldsymbol{\beta}$, and σ^2.

Includes posterior medians, highest posterior density (HPD) intervals, and acceptance rates for each parameter.

- `reg.summary()` Provides fitted posterior values and summaries for each subregion in the areal data set. Includes posterior medians and highest posterior density (HPD) intervals by region.
- `ref.plot()` Outputs trace plots for the model parameters, τ, $\boldsymbol{\beta}$, and σ^2.
- `ref.analysis()` Performs analysis by sequentially implementing each of the functions above. This function produces plots and a list containing MCMC chains, parameter estimates, regional estimates, and sampling acceptance rates.
- `probs.icar()` Performs simultaneous, objective Bayesian model selection for covariates and spatial model structure for areal data. This function provides posterior model probabilities for all candidate ICAR models and OLMs.

4.3 ICAR Model Setup

The model implemented by `ref.ICAR` is summarized below.

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\theta} + \boldsymbol{\phi} \tag{4.1}$$

where

- $\mathbf{Y}$ is an $n \times 1$ data vector for the response variable, where n corresponds to the number of regions in the shapefile. The current version of the package does not allow for missing data.
- X is a matrix of covariates. This can include a vector $\mathbf{1}_n$ for an intercept, and additional columns corresponding to quantitative predictors.

- $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effect regression coefficients, where p corresponds to the number of columns in X.
- $\boldsymbol{\theta}$ is an $n \times 1$ vector of independent and normally distributed unstructured random effects defined with mean 0 and variance σ^2.
- $\boldsymbol{\phi}$ is an $n \times 1$ vector of spatial random effects that is assigned an intrinsic CAR prior with the sum-zero constraint $\sum_{i=1}^{n} \phi_i = 0$ Keefe et al. (2018).

The model assumes a signal-to-noise ratio parameterization for the variance components of the random components of the model, so σ^2 and τ are used as below.

$$\boldsymbol{\phi} \sim \left(\mathbf{0}, \frac{\sigma^2}{\tau}\Sigma_\phi\right)$$

The parameter τ controls the strength of spatial dependence, and given the neighborhood structure, Σ_ϕ is a fixed matrix. Specifically, Σ_ϕ is the Moore-Penrose inverse of H, where the neighborhood matrix H is an $n \times n$ symmetric matrix constructed as follows.

$$(H)_{ij} = \begin{cases} h_i & \text{if } i = j \\ -g_{ij} & \text{if } i \in N_j \\ 0 & \text{otherwise,} \end{cases} \tag{4.2}$$

where $g_{ij} = 1$ if subregions i and j are neighbors, $g_{ij} = 0$ if subregions i and j are not neighbors,and h_i is the number of neighbors of subregion i. Therefore, the neighborhood matrix H is an $n \times n$ symmetric matrix where the diagonal elements correspond to the number of neighbors for each subregion in the data, and each off-diagonal element equals -1 if the corresponding subregions are neighbors.

Provided a path to a shapefile, the `shape.H()` function in `ref.ICAR` constructs H as specified above, and checks for symmetry and contiguous regions (i.e. no islands) prior to analysis.

The functions `shape.H()` and `ref.analysis()` accept a file path to a shapefile. If a user wants to analyze areal data without a corresponding shapefile (e.g. neuroimaging), they will need to construct H as above and use this H in `ref.MCMC()`. `ref.plot()`, `ref.summary()`, and `reg.summary()` can then be used with the MCMC chains obtained from `ref.MCMC()`.

4.4 Example: Objective ICAR Inference

Consider an example of areal data over the contiguous United States. Figure 4.1 represents the average SAT scores reported in 1999 for each of the contiguous United States and Washington D.C. This example will explore these data and use the `ref.ICAR` package to fit a model to the response, Verbal SAT scores, considering spatial dependence and a single covariate, percent of eligible students that took the SAT in each state in 1999. This data was analyzed in *Hierarchical Modeling and Analysis for Spatial Data* Banerjee et al. (2015). The data are available online at https://www.counterpointstat.com/hierarchical-modeling-and-analysis-for-spatial-data.html. We include the data in the `ref.ICAR` package with permission from the authors. The shapefile is found from http://www.arcgis.com/home/item.html?id=f7f805eb65eb4ab787a0a3e1116ca7e5.

These data and the accompanying shapefile are attached to the `ref.ICAR` package. The files can be loaded into R as shown below. The `st_read()` function from package `sf` is used to read the shapefile.

```
> library(sf)
> system.path <- system.file("extdata", "us.shape48.shp", package = "ref.ICAR",
                             mustWork = TRUE)
> shp.layer <- gsub('.shp','',basename(system.path))
```

```
> shp.path <- dirname(system.path)
> us.shape48 <- st_read(dsn = path.expand(shp.path), layer = shp.layer,
                        quiet = TRUE)
```

The SAT data can then be loaded into R from `ref.ICAR` using `read.table()`.

```
> library(utils)
> data.path <- system.file("extdata", "states-sats48.txt", package = "ref.ICAR",
                           mustWork = TRUE)
> sats48 <- read.table(data.path, header = T)
> us.shape48$verbal <- sats48$VERBAL
> us.shape48$percent <- sats48$PERCENT
```

Now that the shapefile and data are loaded, the observed data can be plotted as a choropleth map (Figure 4.1). This map illustrates the spatial dependence to be analyzed by the model. The Midwestern states and Utah exhibit the highest average SAT scores, and overall, neighboring states have similar average scores.

```
> library(ggplot2)
> library(classInt)
> library(dplyr)

> breaks_qt <- classIntervals(c(min(us.shape48$verbal) - .00001, us.shape48$verbal),
                              n = 7, style = "quantile")
> us.shape48_sf <- mutate(us.shape48, score_cat = cut(verbal, breaks_qt$brks))
> ggplot(us.shape48_sf) +
       geom_sf(aes(fill=score_cat)) +
```

```
scale_fill_brewer(palette = "OrRd") +
labs(title="Plot of observed \n verbal SAT scores") +
theme_bw() +
theme(axis.ticks.x=element_blank(),
    axis.text.x=element_blank(),
    axis.ticks.y=element_blank(),
    axis.text.y=element_blank(),
    axis.title=element_text(size=25,face="bold"),
    plot.title = element_text(face="bold", size=25, hjust=0.5)) +
guides(fill=guide_legend("Verbal score"))
```

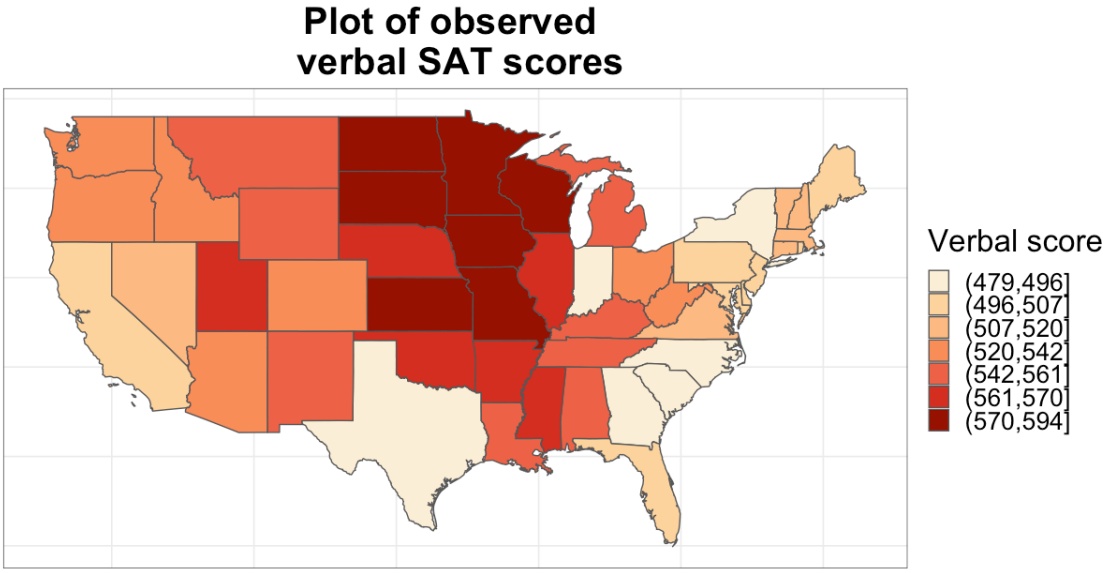


Figure 4.1: Plot of observed verbal SAT scores to be used for inference with the ICAR model.

Similarly, the covariate, percent of eligible students taking the SAT, can be plotted over the contiguous United States.

```
> breaks_qt <- classIntervals(c(min(us.shape48$percent) - .00001,
                                us.shape48$percent), n = 7, style = "quantile")
> us.shape48_sf <- mutate(us.shape48, pct_cat = cut(percent,
                          breaks_qt$brks))
> ggplot(us.shape48_sf) +
        geom_sf(aes(fill=pct_cat)) +
        scale_fill_brewer(palette = "OrRd") +
        labs(title="Plot of observed \n percent SAT takers") +
        theme_bw() +
        theme(axis.ticks.x=element_blank(),
            axis.text.x=element_blank(),
            axis.ticks.y=element_blank(),
            axis.text.y=element_blank(),
            axis.title=element_text(size=25,face="bold"),
            plot.title = element_text(face="bold", size=25, hjust=0.5)) +
        guides(fill=guide_legend("Percent taking"))
```

These data exhibit a seemingly inverse relationship to the SAT scores; lower percentages of students take the SAT in the Midwest.

Now employing the functions in `ref.ICAR`, the `shape.H()` function first takes the path to the shape file (obtained above), and returns a list of two objects. This list contains the neighborhood matrix, H and a `SpatialPolygonsDataFrame` object corresponding to the shapefile, to be used by the remaining functions.

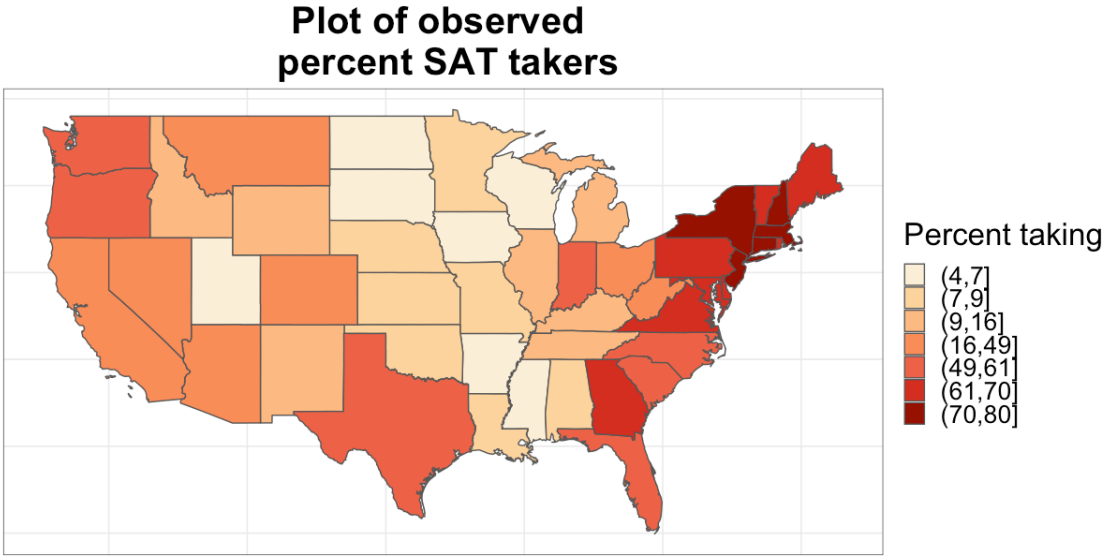


Figure 4.2: Plot of observed percentage of eligible students taking the SAT exam to be used as a covariate for inference with the ICAR model.

```
> library(ref.ICAR)
> library(spdep)


> shp.data <- shape.H(system.path)
> H <- shp.data$H
> class(shp.data$map)
  [1] "sf"         "data.frame"
> length(shp.data$map)
  [1] 7
```

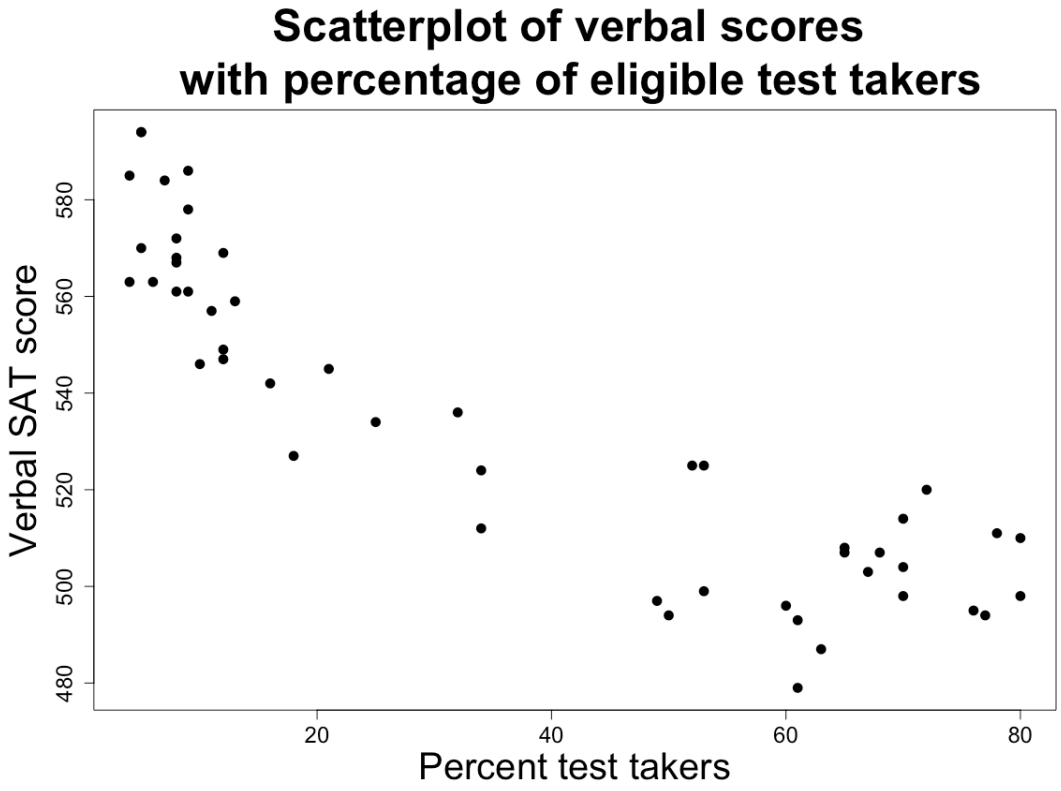


Figure 4.3: Plot of observed percentage of eligible students taking the SAT exam versus the observed verbal SAT scores.

The response and covariates, Y and X must be defined before fitting the model. The response, Y, is Verbal SAT scores. X has two columns corresponding to an intercept and the predictor, percent of eligible students taking the SAT in 1999.

```
> Y <- sats48$VERBAL
> x <- sats48$PERCENT
> X <- cbind(1,x)
```

Then sampling can be performed using `ref.MCMC()`. The default starting values are used below, with MCMC iterations and burn-in larger than the default. The sampling for 'ref.MCMC()' is based on developments by Ferreira et al. (2021), who express the spa-

tial hierarchical model in the spectral domain to obtain the faster Spectral Gibbs Sampler (SGS). Previous versions of `ref.ICAR` implemented the Spectral Decomposition of the Precision (SDP) algorithm proposed by Keefe et al. (2019). See Ferreira et al. (2021) for an outline of the algorithm and computational comparisons.

```
> ref.SAT <- ref.MCMC(y=Y,X=X,H=H,iters=15000,burnin=5000,
                      verbose=FALSE)
> names(ref.SAT)
  [1] "MCMCchain"     "tauc.MCMC"     "sigma2.MCMC"   "beta.MCMC"
  [5] "phi.MCMC"      "accept.phi"    "accept.sigma2" "accept.tauc"
```

The object ref.SAT contains MCMC chains for each of the parameters in the model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\theta} + \boldsymbol{\phi}$, using a signal-to-noise ratio parameterization. From these, the function `ref.plot()` creates trace plots for each parameter to visually confirm convergence.

```
> par(mfrow=c(2,2))
> ref.plot(ref.SAT$MCMCchain,X,burnin=5000,num.reg=length(Y))
```

The remaining components for the analysis are the functions for parameter and regional inferences. The function `ref.summary()` provides posterior medians and intervals for the model parameters $\boldsymbol{\beta}$, τ, and σ^2. The function `ref.summary()` provides medians and Highest Posterior Density intervals for the fitted y values for each subregion in the data.

```
> library(coda)
```

```
> summary.params <- ref.summary(MCMCchain=ref.SAT$MCMCchain,
                                tauc.MCMC=ref.SAT$tauc.MCMC,
```

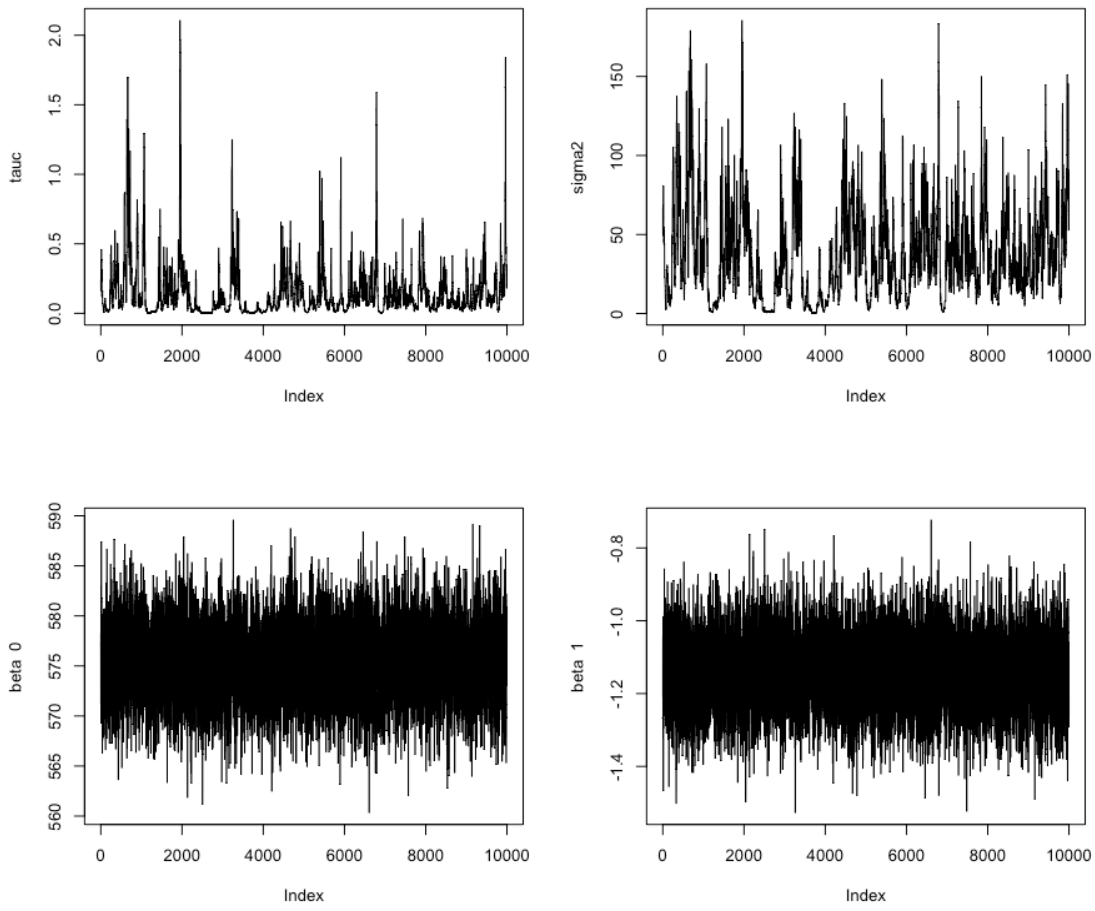


Figure 4.4: Trace plots for parameters of the ICAR model. These indicate that the MCMC chains for each parameter have converged after the specified iterations.

```
sigma2.MCMC=ref.SAT$sigma2.MCMC,
beta.MCMC=ref.SAT$beta.MCMC,
phi.MCMC=ref.SAT$phi.MCMC,
accept.phi=ref.SAT$accept.phi,
accept.sigma2=ref.SAT$accept.sigma2,
accept.tauc=ref.SAT$accept.tauc,
iters=15000,burnin=5000)
```

```
> names(summary.params)
  [1] "beta.median"   "beta.hpd"    "tauc.median"   "tauc.hpd"
  [5] "sigma2.median" "sigma2.hpd"  "tauc.accept"   "sigma2.accept"
> summary.params
  $beta.median
  [1] 575.461382  -1.142777

  $beta.hpd
        lower       upper
  var1 568.028129 582.5163300
  var2  -1.334464  -0.9487258

  $tauc.median
  [1] 0.08581395

  $tauc.hpd
        lower        upper
  0.0003402439 0.5212723440

  $sigma2.median
  [1] 31.2014

  $sigma2.hpd
      lower      upper
  0.1604447 96.0084162
```

```
$tauc.accept
1] 0.3436667


$sigma2.accept
[1] 0.3436667
```

The posterior medians for β_0 and β_1 are 575.46 and -1.14, respectively. Additionally, the HPD interval for β_1 does not include 0, which indicates that as the percent of eligible students taking the SAT increases, average Verbal SAT score tends to decrease. The τ median is 0.086, with HPD interval between 0.0003 and 0.521. Now we can use the `reg.summary()` function to obtain posterior estimates for each subregion, taking into account the percentage of eligible SAT takers covariate.

```
> summary.region <- reg.summary(ref.SAT$MCMCchain,X,Y,burnin=5000)
> us.shape48$verbalfits <- summary.region$reg.medians
> breaks_qt <- classIntervals(c(min(us.shape48$verbalfits) - .00001,
                us.shape48$verbalfits), n = 7, style = "quantile")


> us.shape48_sf <- mutate(us.shape48, reg_cat = cut(verbalfits, breaks_qt$brks))
> ggplot(us.shape48_sf) +
       geom_sf(aes(fill=reg_cat)) +
       scale_fill_brewer(palette = "OrRd") +
       labs(title="Plot of fitted \n verbal SAT scores") +
       theme_bw() +
       theme(axis.ticks.x=element_blank(),
           axis.text.x=element_blank(),
```

```
        axis.ticks.y=element_blank(),
        axis.text.y=element_blank(),
        axis.title=element_text(size=25,face="bold"),
        plot.title = element_text(face="bold", size=25, hjust=0.5)) +
    guides(fill=guide_legend("Region medians"))
```

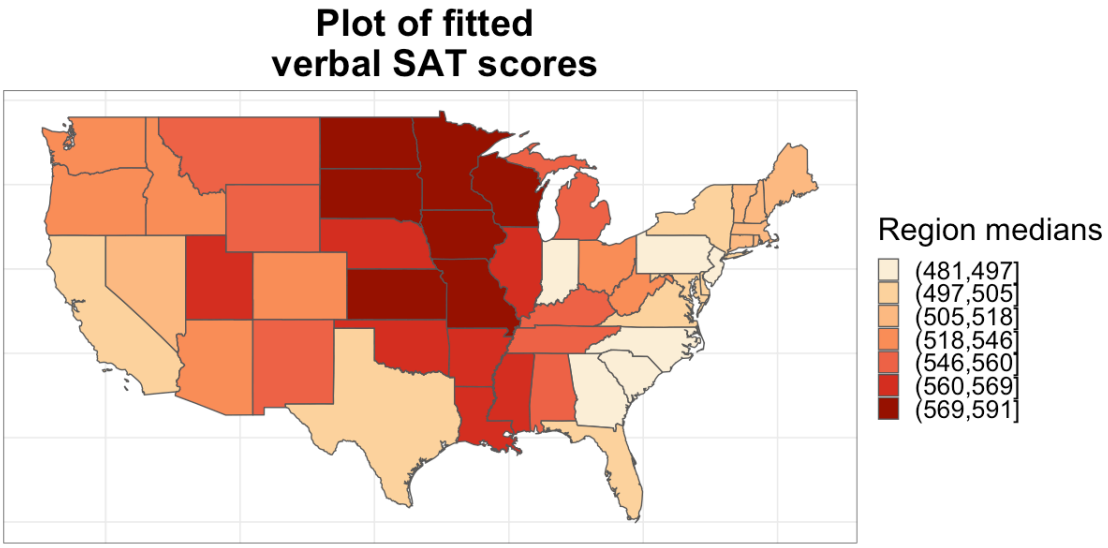


Figure 4.5: Plot of posterior fitted verbal SAT scores after inference with the ICAR model.

4.5 Example: Objective Model Selection for Areal Data

Porter et al. (2023b) developed objective Bayesian model selection for simultaneous selection of covariates and spatial model structure for areal data. Since the joint reference prior on

model parameters is improper (Keefe et al., 2019), fractional Bayes factor methodology is used to approximate Bayes factors and obtain valid posterior model probabilities for all candidate ICAR models and OLMs from the provided candidate covariates. See Porter et al. (2023b) for the method details and simulation results, including the minimal training size for the fractional Bayes factor that is recommended for this approach. The following is a code example that uses open source data analyzed in Porter et al. (2023b) with multiple candidate covariates. The data is available with the `spdep` package, which is imported by `ref.ICAR`. The outcome of interest is the residential crime rate across the 49 neighborhoods of Columbus, Ohio. The five candidate predictors include average housing value, average household income, amount of open space in each neighborhood, the number of housing units without available plumbing, and distance from the Columbus business district.

```
> library(pracma)
> library(hier.part)
> library(stats)
> library(spdep)

> # read in the data as contained in the spdep package
> columbus <- st_read(system.file("shapes/columbus.shp", package="spData")[1],
                      quiet=TRUE)
```

Model selection requires the same form of neighborhood matrix H that is used for inference from Equation (4.3). The code below uses the spectral decomposition of the spatial hierarchical model described in Ferreira et al. (2021) for faster computations. While this data set has a small sample size, the computational advantages of using the spectral decomposition are greater as the sample size increases.

```
> # create neighborhood matrix
> columbus.listw <- poly2nb(columbus)
> summary(columbus.listw)
> W <- nb2mat(columbus.listw, style="B")
> Dmat <- diag(apply(W,1,sum))
> num.reg <- length(columbus$CRIME)

> H <- Dmat - W
> H <- (H+t(H))/2
> rownames(H) <- NULL
> isSymmetric(H)  # check that neighborhood matrix is symmetrix before proceeding

> # spectral quantities for use in model selection
> H.spectral <- eigen(H, symmetric=TRUE)
> Q <- H.spectral$vectors
> eigH <- H.spectral$values
> phimat <- diag(1/sqrt(eigH[1:(num.reg-1)]))
> Sig_phi <- matrix(0,num.reg, num.reg) #initialize
    for(i in 1:(num.reg-1)){
        total <- (1/(eigH[i]))*Q[,i]%*%t(Q[,i])
        Sig_phi <- Sig_phi + total
    }
```

Now, upon defining the response vector and design matrix for the full candidate model, `probs.icar()` calculates and returns posterior inclusion probabilities for each candidate ICAR model and OLM. `probs.icar()` is a wrapper function that uses adaptive quadrature

to integrate the fractional integrated likelihoods over the τ parameter, form the set of fractional Bayes factors, and compute the corresponding posterior model probabilities for the model space.

```
> # define response and design matrix
> Y <- columbus$CRIME
> X <- cbind(1, columbus$HOVAL, columbus$INC, columbus$OPEN, columbus$PLUMB,
            columbus$DISCBD)
> b <- (ncol(X)+1)/num.reg  # specify the minimal training size for this example

> # perform model selection
> columbus.select <- probs.icar(Y=Y,X=X,H=H,
                                H.spectral=H.spectral,
                                Sig_phi=Sig_phi,
                                b=b,verbose=FALSE)
```

From the model selection results contained in `columbus.select`, the highest probability model is found to be the OLM with the first, second, and fifth candidate covariates (average housing value, average household income, and distance from the Columbus business district), indicating that this data set does not require a spatial dependence model.

```
> # print the model with highest posterior model probability
> columbus.select$probs.mat[which.max(columbus.select$probs.mat[,1]),]
    model prob    model type      model form
19  0.1193955     Independent     Y ~ Intercept + X1 + X2 + X5

> # print vector of posterior inclusion probabilities for each covariate
```

```
> post.include.cov <- matrix(NA,nrow = 1, ncol=ncol(X)-1)
> labels <- c(rep(NA, ncol(X)-1))
> for(i in 1:(ncol(X)-1)){labels[i] <- paste("X", i, sep="")}
> colnames(post.include.cov) <- labels
> for(j in 1:ncol(X)-1){
  post.include.cov[,j] <- sum(columbus.select$probs.mat[grep(paste("X",j,sep=""),
                              columbus.select$probs.mat$'model form'), 1])
  }
```

Examining the posterior inclusion probabilities for each of the candidate covariates confirms that the selected model contains the three covariates with highest posterior inclusion probability.

```
> post.include.cov
           X1        X2        X3        X4        X5
  [1,] 0.7454222 0.9238743 0.3009956 0.4312049 0.9273156
```

Chapter 5

Discussion and Future Work

This work has discussed different uses and implementations for Bayesian model selection. We developed objective Bayesian model selection for simultaneous selection of fixed regression effects and spatial model structure in spatial hierarchical models with an ICAR component. By providing accurate methods for simultaneous selection of fixed and random effects, researchers are not required to arbitrarily fix either the mean structure or the spatial model structure in order to select the other. By using an automatic reference prior for model parameters, specification of hyperparameters for vague proper priors is not required when expert or subjective prior information is not available. Our use of fractional Bayes factors for simultaneous selection of fixed effects and spatial model structure enables valid posterior model probabilities for all candidate models and a simulation study demonstrated the accurate selection results for many levels of spatial dependence. Two case studies, one concerning income in all US counties and one studying crime rates at the neighborhood level in Columbus, Ohio show the ubiquity of spatial areal data that require selection methods for ICAR models. Chapter 4 presented code, to be publicly available, for performing a full objective Bayesian analysis for areal data, including inference and model selection capabilities.

We also studied the case where Bayesian model selection is used to penalize predictors for their costs *a priori* in order to select a less expensive subset of predictors than standard variable selection techniques would select. We introduced useful flexibility to a cost-penalizing model prior by using a tuning parameter that adjusts the magnitude of cost penalization

through a function of the predictors' cost ratios. We produced an inclusion path that uses posterior inclusion probabilities to visualize the influence of each predictor as the cost penalization is changed according to the tuning parameter. This enables practitioners to control the level of cost penalization in order to satisfy an overall budget, produce a model with the desired performance, and/or ensure that the cost penalization is maintained in the case of large sample sizes. An application to the diagnosis of heart disease demonstrated the utility of being able to control cost penalization to balance model accuracy and financial burden.

Imminent steps to improve, expand, and disseminate the methods presented in this dissertation include detailed open source software updates, summary with related works in a textbook chapter, and carrying out the review process for manuscripts related to Chapters 2 and 3. Specifically, software updates to the existing R package, ref.ICAR, are forthcoming to reflect all of the capabilities presented in Chapter 4. A corresponding manuscript aimed towards the *R Journal* is in preparation with detailed, user-friendly explanations and demonstrations of methods for objective Bayesian analysis of spatial data with ICAR priors so that the software can be easily understood and used by both statisticians and practitioners. Further, a textbook chapter summarizing the model selection approach from Chapter 2 and related literature for ICAR models has been drafted to appear in a forthcoming textbook *Recent Advances in Spatial and Spatio-Temporal Modeling*, which will provide an avenue for readers to make comparisons with methods for other types of spatial models. Chapter 3 is a full manuscript publicly available (Porter et al., 2023a) and currently under review at *Statistics in Medicine*, and the proposed methods can be extended to allow for flexible decision-making in a sequential manner. Pending peer review and feedback, further extensions and comparisons for the cost-penalized model selection may be developed.

Bibliography

Adams, S., Beling, P. A., and Cogill, R. (2016). "Feature selection for hidden Markov models and hidden semi-Markov models." *IEEE Access*, 4: 1642–1657.

Anselin, L. (1988). *Spatial econometrics: methods and models*. Springer Science & Business Media.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: CRC Press, 2nd edition edition.

Bang, H., Edwards, A. M., Bomback, A. S., Ballantyne, C. M., Brillon, D., Callahan, M. A., Teutsch, S. M., Mushlin, A. I., and Kern, L. M. (2009). "Development and Validation of a Patient Self-assessment Score for Diabetes Risk." *Annals of Internal Medicine*, 151(11): 775–783. PMID: 19949143.
URL https://www.acpjournals.org/doi/abs/10.7326/0003-4819-151-11-200912010-00005

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *The Annals of Statistics*, 32(3): 870 – 897.
URL https://doi.org/10.1214/009053604000000238

Berger, J. O. and Pericchi, L. R. (1996). "The Intrinsic Bayes Factor for Model Selection and Prediction." *Journal of the American Statistical Association*, 91(433): 109–122.
URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476668

Besag, J. (1974). "Spatial Interaction and the Statistical Analysis of Lattice Systems."

Journal of the Royal Statistical Society: Series B (Methodological), 36(2): 192–225.
URL https://doi.org/10.1111/j.2517-6161.1974.tb00999.x

Besag, J., York, J., and Mollié, A. (1991). "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics*, 43(1): 1–20.
URL https://doi.org/10.1007/BF00116466

Best, N., Arnold, R., Thomas, A., Waller, L., and M. Conlon, E. (1999). "Bayesian Models for Spatially Correlated Disease and Exposure Data." *Bayesian Statistics*, 6: 65–82.

Best, N., Richardson, S., and Thomson, A. (2005). "A comparison of Bayesian spatial models for disease mapping." *Statistical Methods in Medical Research*, 14(1): 35–59. PMID: 15690999.
URL https://doi.org/10.1191/0962280205sm388oa

Bivand, R., Nowosad, J., and Lovelace, R. (2019). *spData: Datasets for Spatial Analysis*. R package version 0.3.2.
URL https://CRAN.R-project.org/package=spData

Bolón-Canedo, V., Porto-Díaz, I., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2014). "A framework for cost-based feature selection." *Pattern Recognition*, 47(7): 2481–2489.

Brown, B., Fearn, T., and Vannucci, M. (1999). "The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach." *Biometrika*, 86(3): 635–648.
URL https://doi.org/10.1093/biomet/86.3.635

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). "Deviance information criteria for missing data models." *Bayesian analysis*, 1(4): 651–673.

Chipman, H., George, E. I., and McCulloch, R. E. (2001). *The Practical Implementation of Bayesian Model Selection*, 67–116. Model Selection. Beachwood, OH: Institute of Mathematical Statistics.
URL https://doi.org/10.1214/lnms/1215540964

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). "Active learning with statistical models." *Journal of artificial intelligence research*, 4: 129–145.

De Oliveira, V. (2007). "Objective Bayesian analysis of spatial data with measurement error." *Canadian Journal of Statistics*, 35(2): 283–301.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.5550350206

De Oliveira, V. and Ferreira, M. A. R. (2011). "Maximum Likelihood and Restricted Maximum Likelihood Estimation for a Class of Gaussian Markov Random Fields." *Metrika*, 74: 167–183.

De Oliveira, V. and Song, J. J. (2008). "Bayesian Analysis of Simultaneous Autoregressive Models." *Sankhya: The Indian Journal of Statistics, Series B*, 70(2): 323–350.
URL http://www.jstor.org/stable/41234438

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., and Froelicher, V. (1989). "International application of a new probability algorithm for the diagnosis of coronary artery disease." *The American Journal of Cardiology*, 64(5): 304–310.
URL https://www.sciencedirect.com/science/article/pii/0002914989905249

Dietrich, C. R. (1991). "Modality of the restricted likelihood for spatial Gaussian random fields." *Biometrika*, 78(4): 833–839.
URL https://doi.org/10.1093/biomet/78.4.833

Elkan, C. (2001). "The foundations of cost-sensitive learning." In *International joint conference on artificial intelligence*, volume 17, 973–978. Lawrence Erlbaum Associates Ltd.

Fan, W., Stolfo, S. J., Zhang, J., and Chan, P. K. (1999). "AdaCost: misclassification cost-sensitive boosting." In *Icml*, volume 99, 97–105.

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(4): 734–744.
URL https://doi.org/10.1214/19-BJPS435

Ferreira, M. A. R. and De Oliveira, V. (2007). "Bayesian reference analysis for Gaussian Markov random fields." *Journal of Multivariate Analysis*, 98(4): 789 – 812.
URL http://www.sciencedirect.com/science/article/pii/S0047259X06001138

Ferreira, M. A. R., Porter, E. M., and Franck, C. T. (2021). "Fast and scalable computations for Gaussian hierarchical models with intrinsic conditional autoregressive spatial random effects." *Computational Statistics and Data Analysis*, 162: 107264.
URL https://www.sciencedirect.com/science/article/pii/S0167947321000980

Fouskakis, D. and Draper, D. (2008). "Comparing Stochastic Optimization Methods for Variable Selection in Binary Outcome Prediction, With Application to Health Policy." *Journal of the American Statistical Association*, 103(484): 1367–1381.
URL https://doi.org/10.1198/016214508000001048

Fouskakis, D., Ntzoufras, I., and Draper, D. (2009a). "Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care." *The Annals of Applied Statistics*, 3(2): 663 – 690.
URL https://doi.org/10.1214/08-AOAS207

— (2009b). "Population-based reversible jump Markov chain Monte Carlo methods for Bayesian variable selection and evaluation under cost limit restrictions." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(3): 383–403.
URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2008.00658.x

Franck, C. T. and Gramacy, R. B. (2020). "Assessing Bayes Factor Surfaces Using Interactive Visualization and Computer Surrogate Modeling." *The American Statistician*, 74(4): 359–369.
URL https://doi.org/10.1080/00031305.2019.1671219

Gelman, A., Hwang, J., and Vehtari, A. (2014). "Understanding predictive information criteria for Bayesian models." *Statistics and computing*, 24(6): 997–1016.

Goodchild, M. and Janelle, D. (2004). *Spatially Integrated Social Science*. Spatial Information Systems. Oxford University Press.
URL https://books.google.com/books?id=T66ZP1ieZ5UC

Hodges, J. S. and Reich, B. J. (2010). "Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love." *The American Statistician*, 64(4): 325–334.
URL https://doi.org/10.1198/tast.2010.10052

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, 12(1): 55–67.
URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634

Hogan, J. W. and Tchernis, R. (2004). "Bayesian Factor Analysis for Spatially Correlated Data, With Application to Summarizing Area-Level Material Deprivation From Census Data." *Journal of the American Statistical Association*, 99(466): 314–324.
URL https://doi.org/10.1198/016214504000000296

Jin, X., Banerjee, S., and Carlin, B. P. (2007). "Order-free co-regionalized areal data models with application to multiple-disease mapping." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5): 817–838.
URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00612.x

Jin, X., Carlin, B. P., and Banerjee, S. (2005). "Generalized hierarchical multivariate CAR models for areal data." *Biometrics*, 61(4): 950–61.

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90(430): 773–795.
URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572

Keefe, M. J., Ferreira, M. A. R., and Franck, C. T. (2018). "On the formal specification of sum-zero constrained intrinsic conditional autoregressive models." *Spatial Statistics*, 24: 54 – 65.
URL http://www.sciencedirect.com/science/article/pii/S2211675317301574

— (2019). "Objective Bayesian Analysis for Gaussian Hierarchical Models with Intrinsic Conditional Autoregressive Priors." *Bayesian Analysis*, 14(1): 181–209.
URL https://doi.org/10.1214/18-BA1107

Keeler, E. B., Kahn, K. L., Draper, D., Sherwood, M. J., Rubenstein, L. V., Reinisch, E. J., Kosecoff, J., and Brook, R. H. (1990). "Changes in Sickness at Admission Following the Introduction of the Prospective Payment System." *JAMA*, 264(15): 1962–1968.
URL https://doi.org/10.1001/jama.1990.03450150062032

Kong, G., Jiang, L., and Li, C. (2016). "Beyond accuracy: Learning selective Bayesian classifiers with minimal test cost." *Pattern Recognition Letters*, 80: 165–171.

Kuo, B.-S. (1999). "Asymptotics of ML Estimator for Regression Models with a Stochastic Trend Component." *Econometric Theory*, 15(1): 24–49.
URL http://www.jstor.org/stable/3533141

Lee, D. (2011). "A comparison of conditional autoregressive models used in Bayesian disease mapping." *Spatial and Spatio-temporal Epidemiology*, 2(2): 79 – 89.
URL http://www.sciencedirect.com/science/article/pii/S1877584511000049

Lee, D. and Mitchell, R. (2013). "Locally adaptive spatial smoothing using conditional auto-regressive models." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(4): 593–608.
URL https://doi.org/10.1111/rssc.12009

Leroux, B., Lei, X., and Breslow, N. (1999). *Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence*, 135–178. Statistical Models in Epidemiology, the Environment and Clinical Trials. New York: Springer.

Lindley, D. V. (1968). "The Choice of Variables in Multiple Regression." *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(1): 31–53.
URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1968.tb01505.x

Ling, C. X., Yang, Q., Wang, J., and Zhang, S. (2004). "Decision trees with minimal costs." In *Proceedings of the twenty-first international conference on Machine learning*, 69. ACM.

Liu, Z., Berrocal, V. J., Bartsch, A. J., and Johnson, T. D. (2016). "Pre-surgical fMRI Data Analysis Using a Spatially Adaptive Conditionally Autoregressive Model." *Bayesian Analysis*, 11(2): 599–625.
URL https://doi.org/10.1214/15-BA972

Lloyd-Jones, D. M., Braun, L. T., Ndumele, C. E., Smith, S. C., Sperling, L. S., Virani, S. S., and Blumenthal, R. S. (2019). "Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease: A Special Report From the American Heart Association and American College of Cardiology." *Circulation*, 139(25): e1162–e1177.
URL https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000000638

Logan, J. R., Bauer, C., Ke, J., Xu, H., and Li, F. (2020). "Models for Small Area Estimation for Census Tracts." *Geographical Analysis*, 52(3): 325–350.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gean.12215

Martinez-Beneito, M. A., Botella-Rocamora, P., and Banerjee, S. (2017). "Towards a Multidimensional Approach to Bayesian Disease Mapping." *Bayesian Analysis*, 12(1): 239–259.
URL https://doi.org/10.1214/16-BA995

Miyawaki, K. and MacEachern, S. N. (2022). "Economic variable selection." *Canadian Journal of Statistics*.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.11675

Muirhead, R. J. (2005). *Aspects of Multivariate Statistical Theory*, volume 197. John Wiley & Sons.

O'Hagan, A. (1995). "Fractional Bayes Factors for Model Comparison." *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 99–138.
URL http://www.jstor.org/stable/2346088

Penrose, R. (1955). "A generalized inverse for matrices." *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3): 406–413.

Porter, E. M. (2019). "Applying an Intrinsic Conditional Autoregressive Reference Prior for Areal Data." Ph.D. thesis, Virginia Tech.

Porter, E. M., Franck, C. T., and Adams, S. (2023a). "Flexible cost-penalized Bayesian model selection: developing inclusion paths with an application to diagnosis of heart disease." URL https://arxiv.org/abs/2305.06262

Porter, E. M., Franck, C. T., and Ferreira, M. A. R. (2023b). "Objective Bayesian Model Selection for Spatial Hierarchical Models with Intrinsic Conditional Autoregressive Priors." *Bayesian Analysis*, 1(1): 1–27.
URL https://doi.org/10.1214/23-BA1375

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006). "Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models." *Biometrics*, 62(4): 1197–1206.
URL https://doi.org/10.1111/j.1541-0420.2006.00617.x

Ridker, P. M., Buring, J. E., Rifai, N., and Cook, N. R. (2007). "Development and Validation of Improved Algorithms for the Assessment of Global Cardiovascular Risk in Women: The Reynolds Risk Score." *JAMA*, 297(6): 611–619.
URL https://doi.org/10.1001/jama.297.6.611

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

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2): 319–392.
URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2008.00700.x

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *Annals of Statistics*, 38(5): 2587–2619.
URL https://doi.org/10.1214/10-AOS792

Settles, B. (2009). "Active learning literature survey." Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Song, J. J. and De Oliveira, V. (2012). "Bayesian model selection in spatial lattice models." *Statistical Methodology*, 9(1-2): 228–238.
URL https://doi.org/10.1016/j.stamet.2011.01.003

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 583–639.

Struck, A. F., Tabaeizadeh, M., Schmitt, S. E., Ruiz, A. R., Swisher, C. B., Subramaniam, T., Hernandez, C., Kaleem, S., Haider, H. A., Cissé, A. F., Dhakar, M. B., Hirsch, L. J., Rosenthal, E. S., Zafar, S. F., Gaspard, N., and Westover, M. B. (2020). "Assessment of the Validity of the 2HELPS2B Score for Inpatient Seizure Risk Prediction." *JAMA Neurology*, 77(4): 500–507.
URL https://doi.org/10.1001/jamaneurol.2019.4656

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1): 267–288.
URL http://www.jstor.org/stable/2346178

Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., and Fortin, M.-J. (2018). "Spatial autoregressive models for statistical inference from ecological data." *Ecological Monographs*, 88(1): 36–59.
URL https://doi.org/10.1002/ecm.1283

Verbyla, A. (1990). "A conditional derivation of residual maximum likelihood." *Australian Journal of Statistics*, 32(2): 227–230.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.1990.tb01015.x

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.

White, P., Gelfand, A., and Utlaut, T. (2017). "Prediction and model comparison for areal unit data." *Spatial Statistics*, 22: 89–106.
URL https://doi.org/10.1016/j.spasta.2017.09.002

Wu, H.-H., Ferreira, M. A., Elkhouly, M., and Ji, T. (2020). "Hyper nonlocal priors for variable selection in generalized linear models." *Sankhya A*, 82(1): 147–185.

Zhou, Q., Zhou, H., and Li, T. (2016). "Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features." *Knowledge-based systems*, 95: 1–11.

Zhu, J., Huang, H.-C., and Reyes, P. E. (2010). "On selection of spatial linear models for lattice data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3): 389–402.
URL https://doi.org/10.1111/j.1467-9868.2010.00739.x

Appendices

Appendix A

A.1 Auxiliary Facts

Auxiliary Fact A1

Consider a random vector $\boldsymbol{x}$ with $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Now partition $\boldsymbol{x}$ into two sub-vectors such that $\boldsymbol{x}^T = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)$. Likewise, partition the mean vector and covariance matrix such that $\boldsymbol{\mu}^T = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{pmatrix}$. Then, the conditional distribution of $\boldsymbol{x}_1$ given $\boldsymbol{x}_2$ is (Muirhead, 2005)

$$\boldsymbol{x}_1 | \boldsymbol{x}_2 \sim N\big(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T\big)$$

Auxiliary Fact A2

Consider a random vector $\boldsymbol{x}$ with $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, a $k \times n$ matrix $\boldsymbol{A}$ with rank k and $0 < k < n$. In addition, consider a fixed vector $\boldsymbol{u}$ of length k. Then, application of Auxiliary Fact A1 implies that the distribution of $\boldsymbol{x} | \boldsymbol{A}\boldsymbol{x} = \boldsymbol{u}$ is $N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ with (Rue and Held, 2005)

$$\boldsymbol{\mu}^* = \boldsymbol{\mu} - \boldsymbol{\Sigma}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)^{-1}(\boldsymbol{A}\boldsymbol{\mu} - \boldsymbol{u}),$$

and

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{\Sigma}.$$

Auxiliary Fact A3

To demonstrate the equivalence between Equations (2.2) and (2.4), consider $\mathbf{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a singular covariance matrix with dimension n and rank $n-k$. Let $\boldsymbol{\Sigma} = PDP^T$ be the spectral decomposition of $\boldsymbol{\Sigma}$ where $P = (p_1, \ldots, p_n)$ is a matrix with columns $p_1, \ldots, p_n$ equal to the normalized eigenvectors of $\boldsymbol{\Sigma}$, and $D = \text{diag}(d_1, \ldots, d_n)$ with $d_1 \geq \ldots \geq d_{n-k} > d_{n-k+1} = \ldots = d_n = 0$ the ordered eigenvalues of $\boldsymbol{\Sigma}$. Then the following results hold (see Ferreira et al. (2021)):

(i) For index n of the eigenvector corresponding to the null eigenvalue of $\boldsymbol{\Sigma}$, $E(p_i^T\mathbf{Z}) = p_i^T\boldsymbol{\mu}$ and $Var(p_i^T\mathbf{Z}) = p_i^T\boldsymbol{\Sigma}p^T = 0$ for index $i = n = k+1, \ldots, n$ corresponding to a null eigenvector of $\boldsymbol{\Sigma}$.

(ii) $P(p_i^T\mathbf{Z} = p_i^T\boldsymbol{\mu}) = 1$. Thus, the singular Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ implicitly encodes k linear constraints $p_i^T\boldsymbol{Z} = p_i^T\boldsymbol{\mu}$. Then the density of the singular Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is:

$$p(\mathbf{Z}) = (2\pi)^{-(n-k)/2}\left(\prod_{i=1}^{n-k} d_i\right)^{1/2} \exp\left\{-\frac{1}{2}(\mathbf{Z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^+(\mathbf{Z}-\boldsymbol{\mu})\right\} \prod_{i=n-k+1}^{n} \mathbb{1}(p_i^T\mathbf{Z} = p_i^T\boldsymbol{\mu}),$$

Auxiliary Fact A4

Consider the following lemma (Verbyla, 1990; Dietrich, 1991; Kuo, 1999; De Oliveira, 2007).

Lemma A.1. *Suppose $X_{n\times p}$ is full rank with $n > p$, and Σ is an $n \times n$ symmetric positive definite matrix. Then there exists a full rank matrix L of dimension $n \times (n-p)$ with the following properties:*

(i) $L^T X = 0$;

(i) $L^T L = I_{n-p}$;

(iii) $\Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1} = L(L^T\Sigma L)^{-1}L^T$;

(iv) $\log(|\Sigma|) + \log(|X^T\Sigma^{-1}X|) = \log(|L^T\Sigma L|) + c$, *where* c *depends on* X *but not on* Σ;

(v) $L^T\Sigma L$ *is an* $(n-p)\times(n-p)$ *diagonal matrix with positive diagonal elements.*

Recall that Q^* has columns equal to the normalized eigenvectors corresponding to the non-zero eigenvalues of the projection matrix G. We obtain a matrix with the above properties by setting $L = Q^*U$, where $U\Psi U^T$ is the spectral decomposition of $Q^{*T}\Sigma_\phi Q^*$. Then the following results hold (De Oliveira, 2007).

$$\log(|\Omega|) + \log(|X^T\Omega^{-1}X|) = \sum_{j=1}^{n-p}\log(1+\frac{1}{\tau}\xi_j) + c, \tag{L.1}$$

$$S^2 = \sigma_*^2\sum_{j=1}^{n-p}\left(\frac{1+\tau_*^{-1}\xi_j}{1+\tau^{-1}\xi_j}\right)Z_j^2, \tag{L.2}$$

where σ_*^2 and τ_* are the true values of σ^2 and τ, respectively, and $\{Z_j^2\} \overset{iid}{\sim} \chi_1^2$.

A.2 Proofs of Main Results

A.2.1 Equivalence between ICAR Specifications

Proof of Proposition 2.1. This proof is reproduced from Ferreira et al. (2021) for convenience. The notation corresponds to that used in the manuscript.

Recall from the manuscript that $X = [\mathbf{1}_n, F]$ and $\boldsymbol{\beta} = (\alpha, \boldsymbol{\nu}^T)^T$, where α is an intercept. By Bayes' Theorem the full conditional distribution of $\boldsymbol{\phi}$ is

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

where the last step uses the fact that Q is orthogonal.

Now let $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)' = Q'\boldsymbol{\phi}$ be a vector of spectral random effects. Thus, $\boldsymbol{\phi} = Q\boldsymbol{\zeta}$ and the Jacobian of the transformation is $d\boldsymbol{\phi}/d\boldsymbol{\zeta} = Q$. Further, note that $\mathbb{1}(\mathbf{1}_n'\boldsymbol{\phi} = 0) = \mathbb{1}(\zeta_n = 0)$. Hence,

$$\begin{aligned}
p(\boldsymbol{\zeta}|\mathbf{y}, \alpha, \boldsymbol{\nu}, \sigma^2, \tau) &\propto |Q| \exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\zeta}'(I + \tau D)\boldsymbol{\zeta} - 2\boldsymbol{\zeta}' Q'(\mathbf{y} - \alpha\mathbf{1}_n - F\boldsymbol{\nu})\right]\right\} \mathbb{1}(\zeta_n = 0) \\
&\propto \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}\left[\zeta_i^2(1 + \tau d_i) - 2\zeta_i \mathbf{q}_i'(\mathbf{y} - F\boldsymbol{\nu}) - \alpha\zeta_i \mathbf{q}_i'\mathbf{1}_n)\right]\right\} \mathbb{1}(\zeta_n = 0) \\
&= \prod_{i=1}^{n-1} \exp\left\{-\frac{1}{2\sigma^2}\left[\zeta_i^2(1 + \tau d_i) - 2\zeta_i \mathbf{q}_i'(\mathbf{y} - F\boldsymbol{\nu})\right]\right\} \mathbb{1}(\zeta_n = 0),
\end{aligned}$$

where the last equality uses the facts that $\zeta_n = 0$ and $\mathbf{q}_i'\mathbf{1}_n = 0$ for $i = 1, \ldots, n-1$. Thus, $\zeta_1, \ldots, \zeta_{n-1}$ are conditionally independent given $\mathbf{y}, \alpha, \boldsymbol{\nu}, \sigma^2, \tau$ and have full conditional distributions $\zeta_i|\mathbf{y}, \alpha, \boldsymbol{\nu}, \sigma^2, \tau \sim N\left(\mathbf{q}_i'(\mathbf{y} - F\boldsymbol{\nu})/(1 + \tau d_i), \sigma^2/(1 + \tau d_i)\right), i = 1, \ldots, n-1$. Let $\tilde{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_{n-1})$. Then, these full conditional distributions may be written in matrix form as $\boldsymbol{\zeta}_{1:(n-1)}|\mathbf{y}, \alpha, \boldsymbol{\nu}, \sigma^2, \tau \sim N(\boldsymbol{s}, \sigma^2 D^*)$, where $D^* = \text{diag}((1 + \tau d_1)^{-1}, \ldots, (1 + \tau d_{n-1})^{-1})$ and $\boldsymbol{s} = D^*\tilde{Q}(\mathbf{y} - F\boldsymbol{\nu})$. Therefore, the full conditional distribution of $\boldsymbol{\phi} = Q\boldsymbol{\zeta} = \tilde{Q}\boldsymbol{\zeta}_{1:(n-1)}$ is $\boldsymbol{\phi}|\mathbf{y}, \alpha, \boldsymbol{\nu}, \sigma^2, \tau \sim N(\tilde{Q}\boldsymbol{s}, \sigma^2 \tilde{Q} D^* \tilde{Q}')$. □

Proof of Proposition 2.2. Proposition 2.2 follows directly from Equation (2.7) upon direct application of standard results given in Auxiliary Fact A2 for the multivariate Normal distribution. □

Proof of Theorem 2.3. Let $\mathbf{1}_n$ be the vector of ones. Note that the matrix H has as eigenvector $\mathbf{q}_n = n^{-1/2}\mathbf{1}_n$ with corresponding eigenvalue $d_n = 0$. Further, because all subregions are connected the other $n-1$ eigenvalues are positive.

We now proceed to show that simulating from the full conditional distribution conditional on the constraint $\mathbf{1}_n^T\boldsymbol{\omega} = 0$ given in Equation (2.8) is equivalent to simulating from the full conditional distribution (2.5).

Note that because of orthogonality of eigenvectors, $\mathbf{1}_n^T\mathbf{q}_i = 0$ for $i = 1, \ldots, n-1$. In addition, $\mathbf{1}_n^T\mathbf{q}_n = n^{1/2}$ and recall that $d_n = 0$. Hence,

$$\begin{aligned}
\mathbf{1}_n^T V \mathbf{1}_n &= \sigma^2 \mathbf{1}_n^T Q(I_n + \tau D)^{-1} Q^T \mathbf{1}_n \\
&= \sigma^2 (0, \ldots, 0, n^{1/2}) diag((1+\tau d_1)^{-1}, \ldots, (1+\tau d_n)^{-1})(0, \ldots, 0, n^{1/2})^T \\
&= \sigma^2 \frac{n}{1+\tau d_n} \\
&= n\sigma^2.
\end{aligned}$$

In addition,

$$\begin{aligned}
\mathbf{1}_n^T V &= \sigma^2 \mathbf{1}_n^T Q(I_n + \tau D)^{-1} Q^T \\
&= \sigma^2 (0, \ldots, 0, n^{1/2}) diag((1+\tau d_1)^{-1}, \ldots, (1+\tau d_n)^{-1}) Q^T \\
&= \sigma^2 (0, \ldots, 0, n^{1/2}) Q^T.
\end{aligned}$$

Thus, the covariance matrix Σ^* is

$$\begin{aligned}
\Sigma^* &= V - V\mathbf{1}_n(\mathbf{1}_n^T V \mathbf{1}_n)^{-1}\mathbf{1}_n^T V \\
&= \sigma^2 Q(I_n + \tau D)^{-1}Q^T - \sigma^2 Q(0,\ldots,0,n^{1/2})^T \frac{1}{n\sigma^2}\sigma^2(0,\ldots,0,n^{1/2})Q^T \\
&= \sigma^2 Q\left\{(I_n + \tau D)^{-1} - (0,\ldots,0,1)^T(0,\ldots,0,1)\right\}Q^T \\
&= \sigma^2 \tilde{Q} D^* \tilde{Q}^T.
\end{aligned}$$

Thus, the covariance matrices of the distributions given by Equations (2.5) and (2.8) coincide.

Now let us consider the mean vectors. Let $R = I_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^T$. Note that $RQ = (I_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^T)(\mathbf{q}_1,\ldots,\mathbf{q}_n) = (\mathbf{q}_1,\ldots,\mathbf{q}_{n-1},\mathbf{0})$. Then, the mean vector of the distribution given by Equation (2.8) is

$$\begin{aligned}
\mu^* &= g - V\mathbf{1}_n(\mathbf{1}_n^T V\mathbf{1}_n)^{-1}(\mathbf{1}_n^T g - 0) \\
&= g - \sigma^2 Q(0,\ldots,0,n^{1/2})^T \frac{1}{n\sigma^2}\mathbf{1}_n^T g \\
&= g - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T g \\
&= Rg \\
&= RQ(I_n + \tau D)^{-1}Q^T(\mathbf{Y} - X\boldsymbol{\beta}). \\
&= (\mathbf{q}_1,\ldots,\mathbf{q}_{n-1},\mathbf{0})(I_n + \tau D)^{-1}Q^T(\mathbf{Y} - \alpha\mathbf{1}_n - F\boldsymbol{\nu}) \\
&= ((1+\tau d_1)^{-1}\mathbf{q}_1,\ldots,(1+\tau d_{n-1})^{-1}\mathbf{q}_{n-1},\mathbf{0}) \begin{pmatrix} \mathbf{q}_1^T(\mathbf{Y} - F\boldsymbol{\nu}) \\ \vdots \\ \mathbf{q}_{n-1}^T(\mathbf{Y} - F\boldsymbol{\nu}) \\ \mathbf{q}_n^T(\mathbf{Y} - F\boldsymbol{\nu}) - \alpha\sqrt{n} \end{pmatrix}
\end{aligned}$$

$$
\begin{aligned}
&= ((1+\tau d_1)^{-1}\mathbf{q}_1, \ldots, (1+\tau d_{n-1})^{-1}\mathbf{q}_{n-1}) \begin{pmatrix} \mathbf{q}_1^T(\mathbf{Y} - F\boldsymbol{\nu}) \\ \vdots \\ \mathbf{q}_{n-1}^T(\mathbf{Y} - F\boldsymbol{\nu}) \end{pmatrix} \\
&= \tilde{Q}D^*\tilde{Q}^T(\mathbf{Y} - F\boldsymbol{\nu}) = \tilde{Q}\mathbf{s}.
\end{aligned}
$$

Thus, the mean vector of the distributions given by Equations (2.5) and (2.8) also coincide. Therefore, because Gaussian distributions are completely characterized by their first two moments, the distributions given by Equations (2.5) and (2.8) are equivalent. □

A.2.2 FBF Minimal Training Size

Proof of Proposition 2.4. Consider all terms in $\int p^b(\mathbf{Y}|\boldsymbol{\eta}, M)\pi(\boldsymbol{\eta})d\boldsymbol{\eta}$ that involve σ^2 after $\boldsymbol{\beta}$ is integrated out.

$$
\begin{aligned}
p^{(b)}(\mathbf{Y}|\sigma^2, \tau, M)\pi(\sigma^2) =& (2\pi)^{\frac{p-nb}{2}}(\sigma^2)^{\frac{p-nb}{2}-1}|\Omega|^{-\frac{b}{2}}b^{-\frac{p}{2}}|X^T\Omega^{-1}X|^{-\frac{1}{2}}\times \\
& \exp\Big\{-\frac{b}{2\sigma^2}S^2\Big\}
\end{aligned} \tag{A.1}
$$

Consider a given value of τ. Then, as $\sigma^2 \to \infty$, $\exp\{\frac{-b}{2\sigma^2}S^2\} \to 1$ and the expression in (A.1) behaves as $O\big((\sigma^2)^{\frac{p-nb}{2}-1}\big)$. As $\sigma^2 \to 0$, (A.1) is dominated by $\exp\{\frac{-b}{2\sigma^2}S^2\}$. □

Proof of Proposition 2.5. From (A.1) recognize an Inverse-Gamma(α,β) kernel with respect to σ^2, where $\alpha = \frac{nb-p}{2}$ and $\beta = \frac{bS^2}{2}$. From Proposition 2.4, when $\frac{nb-p}{2} \leq 0$, the integral of expression (A.1) is divergent and thus $\int p^b(\mathbf{Y}|\boldsymbol{\eta}, M)\pi(\boldsymbol{\eta})d\boldsymbol{\eta}$ also diverges. If $\frac{nb-p}{2} > 0$, the integration with respect to σ^2 can be completed using Inverse-Gamma properties. In that case, integrating (A.1) over σ^2 results in

$$p^{(b)}(\mathbf{Y}|\tau, M) \propto |\Omega|^{-\frac{b}{2}}|X^T\Omega^{-1}X|^{-\frac{1}{2}}\Big\{\frac{b}{2}S^2\Big\}^{\frac{p-nb}{2}} \tag{A.2}$$

Consider the expression $(|\Omega|^b|X^T\Omega^{-1}X|)^{-\frac{1}{2}}$. We can rewrite this as

$$\begin{aligned}(|\Omega|^b|X^T\Omega^{-1}X|)^{-\frac{1}{2}} &= (|\Omega|^{b-1}|\Omega||X^T\Omega^{-1}X|)^{-\frac{1}{2}} \\ &= (|\Omega|)^{\frac{1-b}{2}}(|\Omega||X^T\Omega^{-1}X|)^{-\frac{1}{2}}\end{aligned}$$

Consider Ω. Note the spectral decomposition of H is $H = QDQ^T$, where $Q = (\mathbf{q}_1, \mathbf{q}_1, \dots \mathbf{q}_n)$ is a $n\times n$ matrix whose columns are normalized eigenvectors of H and $D = \text{diag}(d_1, d_2, \dots, d_n)$ where $d_1 \geq d_2 \geq \dots \geq d_{n-1} > d_n = 0$ are the ordered eigenvalues of H. Then the spectral decomposition of Σ_ϕ is $\Sigma_\phi = QD^+Q^T$, where D^+ is the Moore-Penrose inverse of D (Penrose, 1955). Then

$$\Omega = I_n + \tau^{-1}\Sigma_\phi = Q\text{diag}\Big(1 + \frac{\tau^{-1}}{d_1}, \dots, 1 + \frac{\tau^{-1}}{d_{n-1}}, 1\Big)Q^T$$

$$\implies |\Omega| = \prod_{i=1}^{n-1}\Big(1 + \frac{\tau^{-1}}{d_i}\Big) = O(\tau^{1-n}) \text{ as } \tau \to 0$$

Now consider $|\Omega||X^T\Omega^{-1}X|$. From Lemma A.1 and (L.1),

$$|\Omega||X^T\Omega^{-1}X| \propto \prod_{i=1}^{n}\Big(1 + \frac{\xi_i}{\tau}\Big) = O(\tau^{p-n}) \text{ as } \tau \to 0$$

Therefore,

$$\begin{aligned}(|\Omega|^b|X^T\Omega^{-1}X|)^{-\frac{1}{2}} &= (|\Omega|)^{\frac{1-b}{2}}(|\Omega||X^T\Omega^{-1}X|)^{-\frac{1}{2}} \\ &= \Big(O(\tau^{1-n})\Big)^{\frac{1-b}{2}}\Big(O(\tau^{p-n})\Big)^{-\frac{1}{2}} \\ &= O\big(\tau^{\frac{1-b+nb-p}{2}}\big) \text{ as } \tau \to 0\end{aligned}$$

Now consider S^2. By Lemma A.1 and (L.2), $S^2 = O(\tau)$ as $\tau \to 0$. Therefore,

$$\begin{aligned} p^{(b)}(\mathbf{Y}|\tau, M) &= O\big(\tau^{\frac{1-b+nb-p}{2}}\big)O(\tau^{\frac{p-nb}{2}}) \\ &= O(\tau^{\frac{1-b}{2}}) \text{ as } \tau \to 0 \end{aligned}$$

Now consider the behavior of (A.2) as $\tau \to \infty$.

$$|\Omega| = \prod_{i=1}^{n-1} \left(1 + \frac{\tau^{-1}}{d_i}\right) = O(1) \text{ as } \tau \to \infty$$

$$|\Omega||X^T\Omega^{-1}X| \propto \prod_{i=1}^{n} \left(1 + \frac{\xi_i}{\tau}\right) = O(1) \text{ as } \tau \to \infty$$

So, $(|\Omega|^b|X^T\Omega^{-1}X|)^{-\frac{1}{2}} = O(1)$ as $\tau \to \infty$. Similarly, $S^2 = \sigma_*^2 \sum_{j=1}^{n-p} \left(\frac{1+\tau_{c*}^{-1}\xi_j}{1+\tau_c^{-1}\xi_j}\right) Z_j^2 = O(1)$ as $\tau \to \infty$. Therefore, $p^{(b)}(\mathbf{Y}|\tau, M) = O(1)$ as $\tau \to \infty$. □

Proof of Theorem 2.7. Proposition 2.4 shows that $\frac{nb-p}{2} > 0$ is a necessary condition to integrate over $p^b(\mathbf{Y}|\boldsymbol{\eta}, M)\pi(\boldsymbol{\eta})$ with respect to σ^2. The reference prior $\pi(\tau)$ is proper as a result of Proposition 2.6. Since $\pi(\tau)$ is proper and the tail behavior of $p^{(b)}(\mathbf{Y}|\tau, M)$ is well-behaved according to Proposition 2.5 $\int p^b(\mathbf{Y}|\boldsymbol{\eta}, M)\pi(\boldsymbol{\eta})d\boldsymbol{\eta} < \infty$ if $\frac{nb-p}{2} > 0 \implies b > \frac{p}{n}$. Thus, $m = p + 1$ is the smallest integer for which $\int p^b(\mathbf{Y}|\boldsymbol{\eta}, M)\pi(\boldsymbol{\eta})d\boldsymbol{\eta} < \infty$ and the fractional integrated likelihood converges. □

A.3 Fractional Integrated Likelihood Calculations

To form the FBF, we derive the fractional integrated likelihoods for models (2.9) and (2.10). The usual integrated likelihood $\int p(\mathbf{Y}|\boldsymbol{\eta}_c, M_c)\pi(\boldsymbol{\eta}_c)d\boldsymbol{\eta}_c$ for a single model $M_c \in \mathcal{M}$ can then

be found by setting the training fraction to $b = 1$ in the expression for the denominator of the fractional integrated likelihood $q(b, \mathbf{Y})$.

The likelihood function of the OLM raised to training fraction b follows as:

$$p^b(\mathbf{Y}|X, \boldsymbol{\beta}, \sigma^2) = (2\pi)^{-\frac{nb}{2}}(\sigma^2)^{-\frac{nb}{2}}\exp\left\{ -\frac{b}{2\sigma^2}(\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta})\right\}$$

Under the signal-to-noise ratio parameterization for the Gaussian hierarchical model, the likelihood of the spatial model with form (2.9) raised to fraction b follows as:

$$p^b(\mathbf{Y}|X, \boldsymbol{\beta}, \sigma^2, \tau) = (2\pi)^{-\frac{nb}{2}}(\sigma^2)^{-\frac{nb}{2}}|\Omega|^{-\frac{b}{2}}\exp\left\{ -\frac{b}{2\sigma^2}(\mathbf{Y} - X\boldsymbol{\beta})^T\Omega^{-1}(\mathbf{Y} - X\boldsymbol{\beta})\right\},$$

where $\Omega = I_n + \tau^{-1}\Sigma_\phi$.

Since we consider $\pi(\boldsymbol{\beta}) \propto 1$, only the likelihood contains $\boldsymbol{\beta}$. Thus, we can first integrate the likelihoods raised to b with respect to $\boldsymbol{\beta}$. Then, using notation from Section 2.3.3,

$$p^{(b)}(\mathbf{Y}|\sigma^2, M) = (2\pi)^{\frac{p-nb}{2}}(\sigma^2)^{\frac{p-nb}{2}}b^{-\frac{p}{2}}|X^TX|^{-\frac{1}{2}}\exp\left\{ -\frac{b}{2\sigma^2}\mathbf{Y}^T(I - X(X^TX)^{-1}X^T)\mathbf{Y}\right\}$$

and

$$\begin{aligned} p^{(b)}(\mathbf{Y}|\sigma^2, \tau, M) =& (2\pi)^{\frac{p-nb}{2}}(\sigma^2)^{\frac{p-nb}{2}}|\Omega|^{-\frac{b}{2}}b^{-\frac{p}{2}}|X^T\Omega^{-1}X|^{-\frac{1}{2}}\times \\ & \exp\left\{ -\frac{b}{2\sigma^2}\mathbf{Y}^T(\Omega^{-1} - \Omega^{-1}X(X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1})\mathbf{Y}\right\}, \end{aligned}$$

for the OLM and spatial models, respectively. The following pages contain the integration $\int p^b(\mathbf{Y}|\boldsymbol{\eta}, M)\pi(\boldsymbol{\eta})d\boldsymbol{\eta}$ for the OLM and for the spatial model under the reference prior described in Section 2.2.2.

A.3.1 OLM fractional integrated likelihood

The reference prior in the case of the OLM is $\pi(\boldsymbol{\eta}) \propto 1/\sigma^2$. Then the denominator of the fractional integrated likelihood for the OLM under the reference prior follows as

$$\begin{aligned}
\int p^b(\mathbf{Y}|\boldsymbol{\eta}, M)\pi(\boldsymbol{\eta})d\boldsymbol{\eta} &= \int_H \Big[p(\mathbf{Y}|X, \boldsymbol{\beta}, \sigma^2)\Big]^b \pi(\boldsymbol{\eta})d\boldsymbol{\eta} \\
&= \int_0^\infty \int_{-\infty}^\infty (2\pi)^{-\frac{nb}{2}}(\sigma^2)^{-\frac{nb}{2}} \exp\Big\{ -\frac{b}{2\sigma^2}(\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta})\Big\}\frac{1}{\sigma^2}d\boldsymbol{\beta}d\sigma^2 \\
&= \int_0^\infty (2\pi)^{\frac{p-nb}{2}}(\sigma^2)^{\frac{p-nb}{2}} b^{-\frac{p}{2}}|X^TX|^{-\frac{1}{2}} \times \\
&\qquad\qquad \exp\Big\{ -\frac{b}{2\sigma^2}\mathbf{Y}(I - X(X^TX)^{-1}X^T)\mathbf{Y}\Big\}\frac{1}{\sigma^2}d\sigma^2 \\
&= (2\pi)^{\frac{p-nb}{2}} b^{-\frac{p}{2}}|X^TX|^{-\frac{1}{2}}\Gamma\Big(\frac{nb-p}{2}\Big)\Big(\frac{1}{2}\Big[b\cdot\mathbf{Y}^T(I - X(X^TX)^{-1}X^T)\mathbf{Y}\Big]\Big)^{\frac{p-nb}{2}}
\end{aligned}$$

A.3.2 ICAR fractional integrated likelihood

The reference prior in the case of the spatial hierarchical model (2.1) is given in Equation (2.11). Then the denominator of the fractional integrated likelihood of an ICAR model, under the reference prior for τ, σ^2, and $\boldsymbol{\beta}$ follows as

$$
\begin{aligned}
\int p^b(\mathbf{Y}|\boldsymbol{\eta}, M)\pi(\boldsymbol{\eta})d\boldsymbol{\eta} &\propto \int_H \Big[p(\mathbf{Y}|X, \boldsymbol{\beta}, \sigma^2, \tau)\Big]^b \pi(\boldsymbol{\eta})d\boldsymbol{\eta} \\
&\propto \int_0^\infty \int_0^\infty \int_{-\infty}^\infty (2\pi)^{-\frac{nb}{2}}(\sigma^2)^{-\frac{nb}{2}}|\Omega|^{-\frac{b}{2}}\exp\Big\{-\frac{b}{2\sigma^2}(\mathbf{Y}-X\boldsymbol{\beta})^T\Omega^{-1}(\mathbf{Y}-X\boldsymbol{\beta})\Big\}\times \\
&\qquad \frac{1}{\sigma^2}\frac{1}{\tau}\left[\sum_{j=1}^{n-p}\Big(\frac{\xi_j}{\tau+\xi_j}\Big)^2 - \frac{1}{n-p}\left\{\sum_{j=1}^{n-p}\Big(\frac{\xi_j}{\tau+\xi_j}\Big)\right\}^2\right]^{\frac{1}{2}} d\boldsymbol{\beta}d\sigma^2 d\tau \\
&\propto \int_0^\infty \int_0^\infty (2\pi)^{\frac{p-nb}{2}}(\sigma^2)^{\frac{p-nb}{2}}|\Omega|^{-\frac{b}{2}}b^{-\frac{p}{2}}|X^T\Omega^{-1}X|^{-\frac{1}{2}}\times \\
&\qquad \exp\Big\{-\frac{b}{2\sigma^2}\mathbf{Y}^T(\Omega^{-1}-\Omega^{-1}X(X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1})\mathbf{Y}\Big\}\times \\
&\qquad \frac{1}{\sigma^2}\frac{1}{\tau}\left[\sum_{j=1}^{n-p}\Big(\frac{\xi_j}{\tau+\xi_j}\Big)^2 - \frac{1}{n-p}\left\{\sum_{j=1}^{n-p}\Big(\frac{\xi_j}{\tau+\xi_j}\Big)\right\}^2\right]^{\frac{1}{2}} d\sigma^2 d\tau \\
&\propto \int_0^\infty \int_0^\infty (2\pi)^{\frac{p-nb}{2}}(\sigma^2)^{\frac{p-nb}{2}-1}|\Omega|^{-\frac{b}{2}}b^{-\frac{p}{2}}|X^T\Omega^{-1}X|^{-\frac{1}{2}}(\tau)^{-1}\times \\
&\qquad \exp\Big\{-\frac{b}{2\sigma^2}\mathbf{Y}^T(\Omega^{-1}-\Omega^{-1}X(X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1})\mathbf{Y}\Big\}\times \\
&\qquad \left[\sum_{j=1}^{n-p}\Big(\frac{\xi_j}{\tau+\xi_j}\Big)^2 - \frac{1}{n-p}\left\{\sum_{j=1}^{n-p}\Big(\frac{\xi_j}{\tau+\xi_j}\Big)\right\}^2\right]^{\frac{1}{2}} d\sigma^2 d\tau \\
&\propto \int_0^\infty (2\pi)^{\frac{p-nb}{2}}|\Omega|^{-\frac{b}{2}}b^{-\frac{p}{2}}|X^T\Omega^{-1}X|^{-\frac{1}{2}}(\tau)^{-1}\times \\
&\qquad \Gamma\Big(\frac{nb-p}{2}\Big)\Big[\frac{b}{2}\mathbf{Y}^T(\Omega^{-1}-\Omega^{-1}X(X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1})\mathbf{Y}\Big]^{\frac{p-nb}{2}}\times \\
&\qquad \left[\sum_{j=1}^{n-p}\Big(\frac{\xi_j}{\tau+\xi_j}\Big)^2 - \frac{1}{n-p}\left\{\sum_{j=1}^{n-p}\Big(\frac{\xi_j}{\tau+\xi_j}\Big)\right\}^2\right]^{\frac{1}{2}} d\tau.
\end{aligned}
$$

A.3.3 SAR fractional integrated likelihood

Finally, the independence Jeffreys prior for the SAR model given in Equation (2.26) is given in Equation (2.27). Then the denominator of the fractional integrated likelihood of a SAR model, under the independence Jeffreys prior for γ, σ^2, and $\boldsymbol{\beta}$ follows as

$$
\begin{aligned}
\int p^b(\mathbf{Y}|\boldsymbol{\eta}, M)\pi(\boldsymbol{\eta})d\boldsymbol{\eta} &\propto \int_H \Big[p(\mathbf{Y}|X, \boldsymbol{\beta}, \sigma^2, \tau)\Big]^b \pi(\boldsymbol{\eta})d\boldsymbol{\eta} \\
&= \int_{\lambda_n^{-1}}^{\lambda_1^{-1}} \int_0^{\infty} \int_{-\infty}^{\infty} (2\pi)^{-\frac{nb}{2}} (\sigma^2)^{-\frac{nb}{2}} |\Sigma_\gamma^{-1}|^{\frac{b}{2}} \exp\Big\{ -\frac{b}{2\sigma^2}(\mathbf{Y} - X\boldsymbol{\beta})^T \Sigma_\gamma^{-1}(\mathbf{Y} - X\boldsymbol{\beta})\Big\} \\
&\qquad \times \frac{1}{\sigma^2}\left\{ \sum_{i=1}^n \Big(\frac{\lambda_i}{1-\gamma\lambda_i}\Big)^2 - \frac{1}{n}\Big[\sum_{i=1}^n \frac{\lambda_i}{1-\gamma\lambda_i}\Big]^2 \right\}^{\frac{1}{2}} d\boldsymbol{\beta} d\sigma^2 d\gamma \\
&\propto \int_{\lambda_n^{-1}}^{\lambda_1^{-1}} \int_0^{\infty} (2\pi)^{-\frac{nb}{2}} (\sigma^2)^{\frac{p-nb}{2}-1} |\Sigma_\gamma^{-1}|^{\frac{b}{2}} |X^T\Sigma_\gamma^{-1}X|^{-\frac{1}{2}} (b)^{-\frac{p}{2}} \\
&\qquad \times \exp\Big\{ -\frac{b}{2\sigma^2}\mathbf{Y}^T(\Sigma_\gamma^{-1} - \Sigma_\gamma^{-1}X(X^T\Sigma_\gamma^{-1}X)^{-1}X^T\Sigma_\gamma^{-1})\mathbf{Y}\Big\} \\
&\qquad \times \left\{ \sum_{i=1}^n \Big(\frac{\lambda_i}{1-\gamma\lambda_i}\Big)^2 - \frac{1}{n}\Big[\sum_{i=1}^n \frac{\lambda_i}{1-\gamma\lambda_i}\Big]^2 \right\}^{\frac{1}{2}} d\sigma^2 d\gamma \\
&\propto \int_{\lambda_n^{-1}}^{\lambda_1^{-1}} (2\pi)^{-\frac{nb}{2}} |\Sigma_\gamma^{-1}|^{\frac{b}{2}} |X^T\Sigma_\gamma^{-1}X|^{-\frac{1}{2}} (b)^{-\frac{p}{2}} \Gamma\Big(\frac{nb-p}{2}\Big) \\
&\qquad \times \Big\{ \frac{b}{2}\mathbf{Y}^T(\Sigma_\gamma^{-1} - \Sigma_\gamma^{-1}X(X^T\Sigma_\gamma^{-1}X)^{-1}X^T\Sigma_\gamma^{-1})\mathbf{Y}\Big\}^{\frac{p-nb}{2}} \\
&\qquad \times \left\{ \sum_{i=1}^n \Big(\frac{\lambda_i}{1-\gamma\lambda_i}\Big)^2 - \frac{1}{n}\Big[\sum_{i=1}^n \frac{\lambda_i}{1-\gamma\lambda_i}\Big]^2 \right\}^{\frac{1}{2}} d\gamma
\end{aligned}
$$

A.4 MCMC algorithm for ICAR and OLM parameters

Below is the MCMC algorithm we use to obtain parameter samples to calculate DIC and WAIC. For ICAR models, we sample parameters in $\boldsymbol{\eta}$ using a Metropolis-within-Gibbs algorithm with a Gibbs step for $\boldsymbol{\beta}$ and a joint Metropolis-Hastings step for τ and σ^2 (Keefe et al., 2019). Then we simulate the spatial random effects $\boldsymbol{\phi}$ using composite sampling. For OLMs, we sample σ^2 from its marginal posterior $p(\sigma^2|\mathbf{Y})$ and use a Gibbs sampler to sample $\boldsymbol{\beta}$ from its conditional posterior $p(\boldsymbol{\beta}|\sigma^2, \mathbf{Y})$.

Algorithm 1 MCMC Algorithm for ICAR and OLM parameters

1. Initialize $\eta^{(0)} = (\boldsymbol{\beta}^{(0)}, \sigma^{2(0)}, \tau^{(0)}, \boldsymbol{\phi}^{(0)})$

For i in 1 to K
{
2. Generate $\log(\sigma^{2*}) \sim N(\sigma^{2(i-1)}, \delta_s)$ and $\log(\tau^*) \sim N(\tau^{(i-1)}, \delta_t)$.
3. Compute joint acceptance probability for σ^{2*} and τ^*:

$$\alpha = \min\left[1, \frac{P(\boldsymbol{\eta}^*|\text{Data})}{P(\boldsymbol{\eta}^{(i)}|\text{Data})}\frac{q(\boldsymbol{\eta}|\boldsymbol{\eta}^*)}{q(\boldsymbol{\eta}^*|\boldsymbol{\eta})}\right]$$

4. Generate $(\boldsymbol{\beta}^*|\mathbf{Y}, \sigma^2, \tau, X) \sim N_p(\mu^*, \Sigma^*)$, where

$$\mu^* = \left(X^T(I_n + \tau^{-1(i-1)}\Sigma_\phi)^{-1}X\right)^{-1} X^T(I_n + \tau^{-1(i-1)}\Sigma_\phi)^{-1}\mathbf{Y},$$
$$\Sigma^* = \sigma^{2(i-1)}\left(X^T(I_n + \tau^{-1(i-1)}\Sigma_\phi)^{-1}X\right)^{-1}.$$

5. Use composite sampling to generate $\boldsymbol{\phi}^{(i)}$ $(i = 1, \ldots, K)$ from its full conditional distribution (see Supplementary Material).
}
6. Obtain parameter samples for the corresponding OLM, sampling $\boldsymbol{\beta}$ and σ^2 from the following distributions:

$$\sigma^2|\mathbf{Y} \sim \text{InverseGamma}\Big(\frac{n-k}{2}, \frac{\mathbf{Y}^T(I_n - X(X^TX)^{-1}X^T)\mathbf{Y}}{2}\Big),$$
$$\boldsymbol{\beta}|\sigma^2, \mathbf{Y} \sim N_p\Big((X^TX)^{-1}X^T\mathbf{Y}, \sigma^2(X^TX)^{-1}\Big).$$

A.5 Sampling from ϕ

We use composite sampling to obtain samples for the spatial random effects $\boldsymbol{\phi}$. We can easily recognize the full conditional distribution for $\boldsymbol{\phi}$ from Auxiliary Fact A1. Consider a random vector $\boldsymbol{x}$ with $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ such that $\boldsymbol{x}^T = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)$, $\boldsymbol{\mu}^T = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)$, and $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{pmatrix}$.

Then, from Auxiliary Fact A1, the conditional distribution of $\boldsymbol{x}_1$ given $\boldsymbol{x}_2$ is $N(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, where $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T$

Using this result we can let $\boldsymbol{x}_1 = \boldsymbol{\phi}$ and $\boldsymbol{x}_2 = \mathbf{Y}$.

Then, we have

$$\begin{aligned} \boldsymbol{\mu}_1 &= \mathbf{0} \\ \boldsymbol{\mu}_2 &= X\boldsymbol{\beta} \\ \boldsymbol{\Sigma}_{22} &= \sigma^2 \left(I_n + \frac{1}{\tau}\Sigma_\phi \right) \\ \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{12}^T &= \boldsymbol{\Sigma}_{11} = \frac{\sigma^2}{\tau}\Sigma_\phi. \end{aligned}$$

Therefore,

$$(\boldsymbol{\phi}|\mathbf{Y}, \sigma^2, \tau, X, \boldsymbol{\beta}) \equiv \boldsymbol{x}_1|\boldsymbol{x}_2 \sim N(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}).$$

A.6 Simulation Results

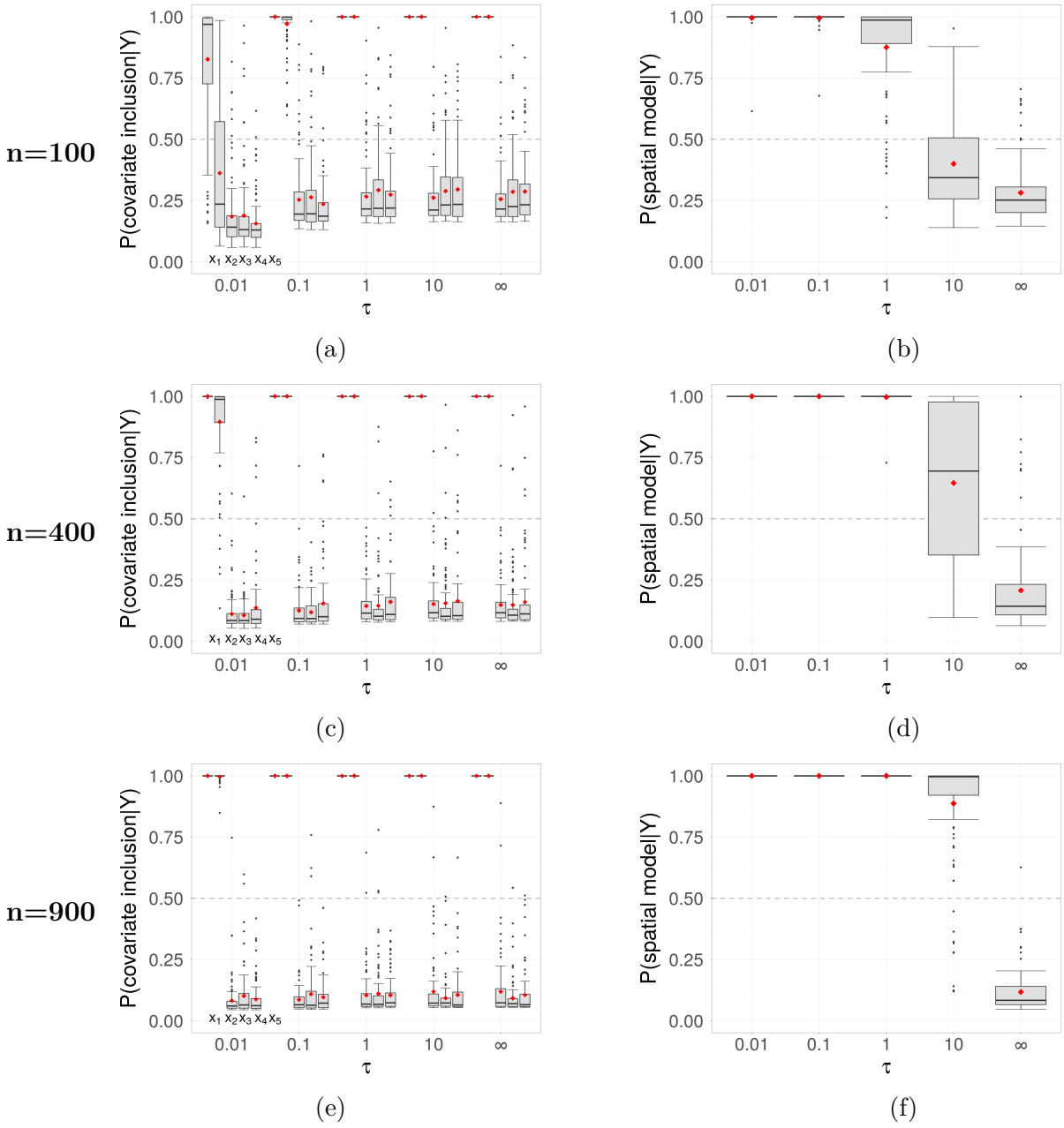


Figure S.1: Covariate posterior inclusion probabilities and probability of selecting a spatial model for $\tau \in \{0.01, 0.1, 1, 10, \infty\}$ for n = 100 (top row), n = 400 (middle row), and n = 900 (bottom row). The reference FBF selection method assigns high probability to non-null covariates for all values of τ and to spatial models for small τ. In contrast to results presented in the manuscript, covariates were generated independently here, i.e. with no spatial correlation.

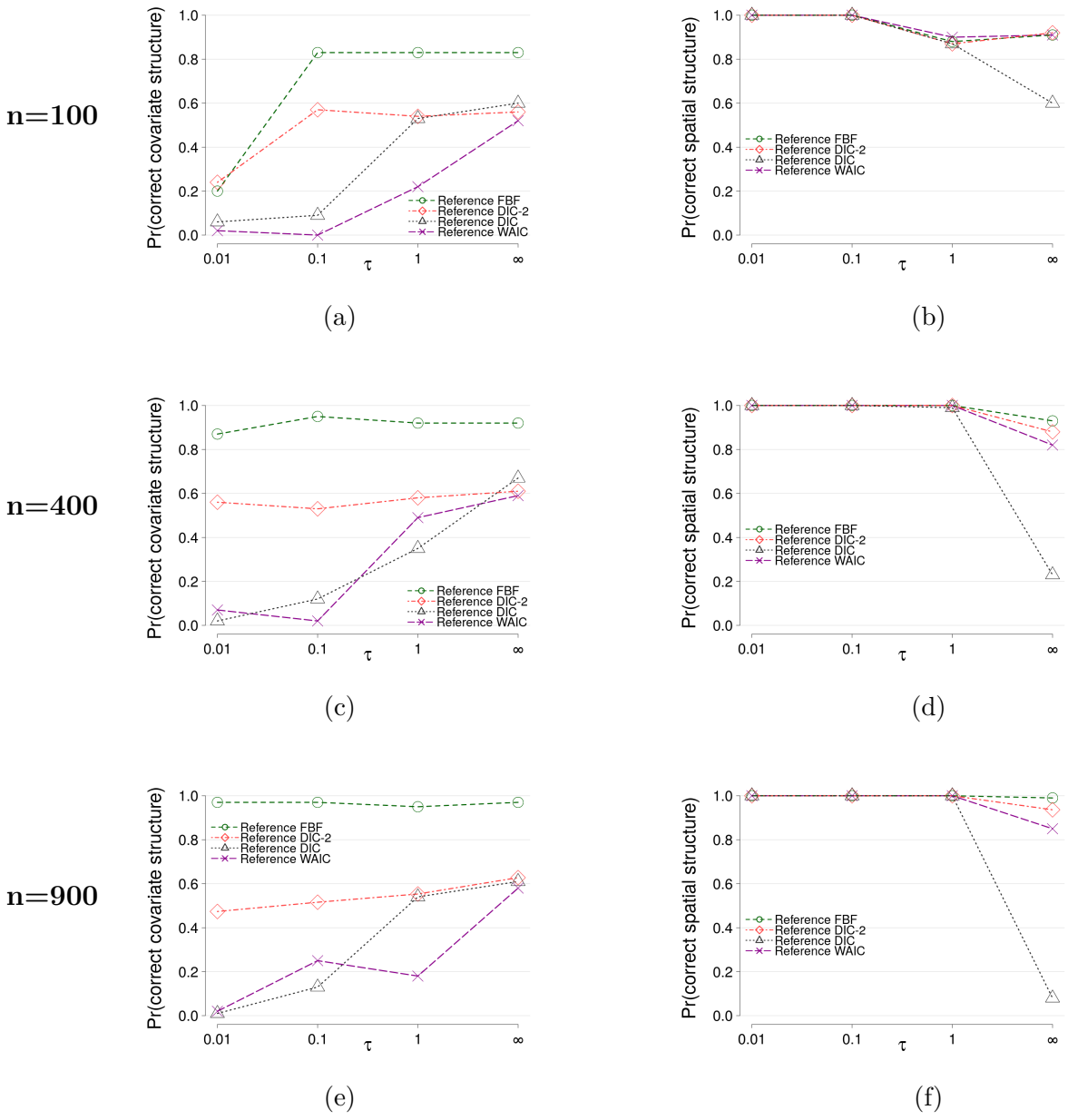


Figure S.2: Proportion of times the reference FBF, DIC-2, DIC, and WAIC methods select the correct covariate and spatial dependence structure for $\tau \in \{0.01, 0.1, 1, \infty\}$ for n = 100 (top row), n = 400 (middle row), and n = 900 (bottom row). The reference FBF selection method reliably selects covariates and spatial dependence for all values of τ and performs better than each of the information criteria. In contrast to results presented in the manuscript, covariates were generated independently here, i.e. with no spatial correlation.