econometrics

MDPI

Article

Bayesian Model Averaging with the Integrated Nested Laplace Approximation

Virgilio Gómez-Rubio [1,*,†,‡], Roger S. Bivand [2,‡] and Håvard Rue [3,‡]

[1] Department of Mathematics, School of Industrial Engineering, Universidad de Castilla-La Mancha, E-02071 Albacete, Spain
[2] Department of Economics, Norwegian School of Economics, 5045 Bergen, Norway; roger.bivand@nhh.no
[3] CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia; haavard.rue@kaust.edu.sa
* Correspondence: Virgilio.Gomez@uclm.es; Tel.: +34-967-59-92-00 (ext. 8291)
† Current address: Department of Mathematics, School of Industrial Engineering, Universidad de Castilla-La Mancha, Avda. España s/n, 02071 Albacete, Spain.
‡ These authors contributed equally to this work.

Received: 25 October 2019; Accepted: 20 May 2020; Published: 1 June 2020

Abstract: The integrated nested Laplace approximation (INLA) for Bayesian inference is an efficient approach to estimate the posterior marginal distributions of the parameters and latent effects of Bayesian hierarchical models that can be expressed as latent Gaussian Markov random fields (GMRF). The representation as a GMRF allows the associated software `R-INLA` to estimate the posterior marginals in a fraction of the time as typical Markov chain Monte Carlo algorithms. INLA can be extended by means of Bayesian model averaging (BMA) to increase the number of models that it can fit to conditional latent GMRF. In this paper, we review the use of BMA with INLA and propose a new example on spatial econometrics models.

Keywords: Bayesian model averaging; INLA; spatial econometrics

1. Introduction

Bayesian model averaging (BMA; see, for example, Hoeting et al. 1999) is a way to combine different Bayesian hierarchical models that can be used to estimate highly parameterized models. By computing an average model, the uncertainty about the model choice is taken into account when estimating the uncertainty of the model parameters.

As BMA often requires fitting a large number of models, this can be time consuming when the time required to fit each of the models is large. The integrated nested Laplace approximation (INLA; Rue et al. 2009) offers an alternative to the computationally intensive Markov chain Monte Carlo (MCMC; Gilks et al. 1996) methods. INLA focuses on obtaining an approximation of the posterior marginal distributions of the models parameters of latent GMRF models (Rue and Held 2005). Hence, BMA with INLA is based on combining the resulting marginals from all the models averaged.

Bivand et al. (2014) and Bivand et al. (2015) used this approach to fit some spatial econometrics models by fitting conditional models on some of the hyperparameters of the original models. The resulting models are then combined using BMA to obtain estimates of the marginals of the hyperparameters of the original model. Gómez-Rubio and Rue (2018) embedded INLA within MCMC so that the joint posterior distribution of a subset of the model parameters is estimated using the Metropolis–Hastings algorithm (Hastings 1970; Metropolis et al. 1953). This requires fitting a model with INLA (conditional on some parameters) at each step, so that the resulting models can also be combined to obtain the posterior marginals of the remainder of the model parameters.

While typical BMA focuses on averaging models with various fixed and random effects, the different models that appear in the methodology of BMA with INLA presented in this paper (Bivand et al. 2014; Bivand et al. 2015) are due to a discretization of the parametric space of some of the hyperparameters. The models are fit conditional on some values of these hyperparameters, so that BMA with INLA accounts for the uncertainty about these parameters, but the underlying model structure is the same for all the models. However, the basic methodology for averaging models is the same as in typical BMA as weights are assigned with the different models, which are in turn used to produce an average model.

Hence, BMA with INLA relies on a weighted sum of the conditional marginals obtained from a family of conditional models on some hyperparameters. Weights are computed by using Bayes' rule, and they depend on the marginal likelihood of the conditional model and the prior distribution of the conditioning hyperparameters. This new approach is described in detail in Section 4. We will illustrate how this works by developing an example on spatial econometrics models described in Section 2.

Compared to Bivand et al. (2014) and Bivand et al. (2015) we extend BMA with INLA to more than one dimension. In addition, we provide details about how to explore the parametric space of the conditioning hyperparameters, as well as how to consider the grid for numerical integration and how to define it using an internal scale to deal with unbounded parameters for computational convenience. The extent of the grid is based on maximum likelihood estimates and their standard errors of the model hyperparameters. The Delta method is used here to estimate the standard error of the transformed parameters in the internal scale. Hence, the numerical integration approach is more thorough and robust than in Bivand et al. (2014) and Bivand et al. (2015).

This paper is organized as follows. Spatial econometrics models are summarized in Section 2. Next, an introduction to INLA is given in Section 3. This is followed by a description of Bayesian model averaging (with INLA) in Section 4. An example is developed in Section 5. Finally, Section 6 gives a summary of the paper and includes a discussion on the main results.

2. Spatial Econometrics Models

Spatial econometrics models (LeSage and Pace 2009) are often employed to account for spatial autocorrelation in the data. Usually, these models include one or more spatial autoregressive terms. Halleck Vega and Elhorst (2015) proposed a model (termed the general nesting spatial model, GNS) that includes an autoregressive term on the response and another one in the error term:

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{u}.$$

Here, $\boldsymbol{y}$ is the response, $\boldsymbol{X}$ a matrix of covariates with coefficients $\boldsymbol{\beta}$, $\boldsymbol{W}$ an adjacency matrix. and $\boldsymbol{W}\boldsymbol{X}$ lagged covariates with coefficients γ. Finally, $\boldsymbol{u}$ is an error term. This error term is often modeled to include spatial autocorrelation:

$$\boldsymbol{u} = \lambda \boldsymbol{W}\boldsymbol{u} + \boldsymbol{e}.$$

Here, $\boldsymbol{e}$ is a vector of independent Gaussian observations with zero mean and precision τ.

The previous model can be rewritten as:

$$\boldsymbol{y} = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\gamma}) + \boldsymbol{u}'$$

with $\boldsymbol{u}'$ an error term with a multivariate Gaussian distribution with zero mean and precision matrix:

$$\tau(\boldsymbol{I} - \rho \boldsymbol{W}^{\top})(\boldsymbol{I} - \lambda \boldsymbol{W}^{\top})(\boldsymbol{I} - \lambda \boldsymbol{W})(\boldsymbol{I} - \rho \boldsymbol{W}).$$

Note that here, the same adjacency matrix $\boldsymbol{W}$ is used for the two autocorrelated terms, but different adjacency matrices could be used. The range of ρ and λ is determined by the eigenvalues of $\boldsymbol{W}$. When $\boldsymbol{W}$ is taken to be row-standardized, the range is the interval $(1/\lambda_{min}, 1))$, where λ_{min} is the minimum

eigenvalue of $\boldsymbol{W}$ (see, for example, Haining 2003). In this case, the lagged covariates $\boldsymbol{WX}$ represent the average value at the neighbors, which is useful when interpreting the results.

In a Bayesian context, a prior needs to be set on every model parameter. For the spatial autocorrelation parameter, a uniform in the interval $(-1, 1)$ will be used. For the coefficients of the covariates, a normal with zero mean and precision 1000 is used and for precision τ, Gamma with parameters 0.01 and 0.01.

When the lagged covariates term $WX\gamma$ is dropped, the resulting model is often referred to as the spatial autoregressive combined (SAC) model (LeSage and Pace 2009). Other important models in spatial econometrics appear when some of the terms in the GNS model are dropped (Halleck Vega and Elhorst 2015). See, for example, Gómez-Rubio et al. (2017) and how these models are fit with INLA.

In spatial econometrics models, how changes in the value of the covariates affect the response in neighboring regions is of interest. These spill-over effects or impacts (LeSage and Pace 2009) are caused by the term $(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}$ that multiplies the covariates, and they are defined as:

$$\frac{\partial y_i}{\partial x_{jr}},\ i, j = 1, \ldots, n;\ r = 1, \ldots, p$$

where n is the number of observations, p the number of covariates, and x_{jr} the value of covariate r in region j.

Hence, for each covariate k, there will be an associated matrix of impacts. The diagonal values are known as direct impacts as they measure the effect of changing a covariate on the same areas. The off-diagonal values are known as indirect impacts as they measure the change of the response at neighboring areas when covariate k changes. Finally, total impacts are the sum of direct and indirect impacts.

Gómez-Rubio et al. (2017) described the impacts for different spatial econometrics models. For the SAC model, the impact matrix for covariate k is:

$$\boldsymbol{S_r(W)} = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{I}\beta_r.$$

In practice, average impacts are reported as a summary of the direct, indirect, and total impacts (for details, see, for example, Gómez-Rubio et al. 2017; LeSage and Pace 2009). In particular, the average direct impact is the trace of $\boldsymbol{S_r(W)}$ divided by n; the average indirect impact is defined as the sum of the off-diagonal elements divided by n; and the average total impact is the sum of all elements of $\boldsymbol{S_r(W)}$ divided by n. The average total impact is also the sum of the average direct and average indirect impacts.

For the SAC model, the average total impact is $\beta_r/(1-\rho)$, and the average direct impact is $n^{-1}tr((\boldsymbol{I} - \rho\boldsymbol{W})^{-1})\beta_r$. The average indirect impact can be computed as the difference between the average total and average direct impacts, $[1/(1-\rho) - n^{-1}tr((\boldsymbol{I} - \rho\boldsymbol{W})^{-1})]\beta_r$, or by computing the sum of the off-diagonal elements of $\boldsymbol{S_r(W)}$ divided by n.

Note that computing the impacts depends on parameters ρ and β_r. For this reason, the joint posterior distribution of both parameters would be required. As explained below, this is not a problem, because it can be rewritten as $\pi(\beta_r, \rho \mid \boldsymbol{y}) = \pi(\beta_r \mid \boldsymbol{y}, \rho)\pi(\rho \mid \boldsymbol{y})$.

3. The Integrated Nested Laplace Approximation

Markov chain Monte Carlo algorithms (see, for example, Brooks et al. 2011) are typically used to estimate the joint posterior distribution of the ensemble of parameters and latent effects of a hierarchical Bayesian model. However, these algorithms are often slow when dealing with high parameterized models.

Rue et al. (2009) proposed a computationally efficient numerical method to estimate the posterior marginals of the hyperparameters and latent effects. This method is called integrated nested Laplace approximation (INLA) because it is based on repeatedly using the Laplace approximation to estimate the posterior marginals. In addition, INLA focuses on models that can be expressed as a latent Gaussian Markov random fields (GMRF; Rue and Held 2005).

In this context, the model is expressed as follows. For each observation y_i in the ensemble of observations y, its likelihood is defined as:

$$y_i|\boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \sim f(y_i|\boldsymbol{\mathcal{X}}, \boldsymbol{\theta}),\ i = 1, \ldots, n$$

where $f(\cdot)$ denotes the density of the likelihood function.

For each observation y_i, its mean μ_i will be conveniently linked to the linear predictor η_i using the appropriate function (which will depend on the likelihood used). The linear predictor will include different additive effects such as fixed effects and random effects.

Next, the vector of latent effects $\boldsymbol{\mathcal{X}}$ is defined as a GMRF with zero mean and precision matrix $\boldsymbol{\Sigma}$ (which may depend on the vector of hyperparameters $\boldsymbol{\theta}$):

$$\boldsymbol{\mathcal{X}} \sim GMRF(0, \boldsymbol{\Sigma}(\boldsymbol{\theta})).$$

Finally, the hyperparameters are assigned a prior distribution. This is often done assuming prior independence. Without loss of generality, this can be represented as:

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}).$$

Note that $\pi(\boldsymbol{\theta})$ is a multivariate distribution, but that in many cases, it can be expressed as the product of several univariate (or small dimension) prior distributions.

For the GNS model, the model fit is such that the likelihood $f(\cdot)$ is a univariate Gaussian distribution with the mean equal to a linear predictor η_i and precision τ_ε, which will be set to a very small value (as explained in Appendix A) to remove the error term. The linear predictor of observation i is:

$$\eta_i = \left[(\boldsymbol{I} - \rho \boldsymbol{W})^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\gamma})\right]_i + u_i^{'},\ i = 1, \ldots, n.$$

The structure of GMRF is defined on the vector of latent effects:

$$\boldsymbol{\mathcal{X}} = \left(\boldsymbol{\eta}, \boldsymbol{\beta}, (\boldsymbol{I} - \rho \boldsymbol{W})^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\gamma}), \boldsymbol{u}^{'}\right)$$

with a vector of hyperparameters $\boldsymbol{\theta} = (\tau_\varepsilon, \tau, \rho, \lambda)$. Precision matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ of $\boldsymbol{\mathcal{X}}$ is a block diagonal matrix defined according to the different latent effects in the model.

Hence, if $\boldsymbol{\theta}$ represents the vector of h hyperparameters and $\boldsymbol{\mathcal{X}}$ the vector of l latent effects, INLA provides the posterior marginals:

$$\{\pi(\theta_i|\mathcal{D})\}_{i=1}^{h}$$

and

$$\{\pi(\mathcal{X}_i|\mathcal{D})\}_{i=1}^{l}.$$

Note that $\mathcal{D}$ represents the observed data, and it includes response $\boldsymbol{y}$, covariates $\mathbf{X}$, and any other known quantities required for model fitting.

In addition to the marginals, INLA can be used to estimate other quantities of interest. See, for example, Rue et al. (2017) for a recent review.

INLA can provide accurate approximations to the marginal likelihood of a model, which are computed as:

$$\tilde{\pi}(\mathcal{D}) = \int \left.\frac{\pi(\boldsymbol{\theta}, \boldsymbol{\mathcal{X}}, \mathcal{D})}{\tilde{\pi}_{\mathrm{G}}(\boldsymbol{\mathcal{X}} \mid \boldsymbol{\theta}, \mathcal{D})}\right|_{\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}^*(\boldsymbol{\theta})} d\boldsymbol{\theta}.$$

Here, $\tilde{\pi}_{\mathrm{G}}(\boldsymbol{\mathcal{X}} \mid \boldsymbol{\theta}, \mathcal{D})$ is a Gaussian approximation to the distribution of $\boldsymbol{\mathcal{X}} \mid \boldsymbol{\theta}, \mathcal{D}$, and $\boldsymbol{\mathcal{X}}^*(\boldsymbol{\theta})$ is the posterior mode of $\boldsymbol{\mathcal{X}}$ for a given value of $\boldsymbol{\theta}$. This approximation seems to be accurate in a wide range of examples (Gómez-Rubio and Rue 2018; Hubin and Storvik 2016).

As described in Section 4, the marginal likelihood plays a crucial role in BMA as it determines the weights, together with the prior distribution of some of the hyperparameters in the model.

As stated above, INLA approximates the posterior marginals of the parameters of latent GMRF models. Hence, an immediate question is whether INLA will work for different models. Gómez-Rubio and Rue (2018) introduced the idea of using INLA to fit conditional latent GMRF models by conditioning on some of the hyperparameters. In this context, the vector of hyperparameters $\boldsymbol{\theta}$ is split into $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_{-c}$, so that models are fit with INLA conditional on $\boldsymbol{\theta}_c$, and the posterior marginals of the elements of $\boldsymbol{\theta}_{-c}$ and latent effects $\boldsymbol{\mathcal{X}}$ are obtained, i.e., $\pi(\theta_{-c,i} \mid \mathcal{D}, \boldsymbol{\theta}_c)$ and $\pi(\mathcal{X}_j \mid \mathcal{D}, \boldsymbol{\theta}_c)$, respectively. Here, $\theta_{-c,i}$ represents any element in $\boldsymbol{\theta}_{-c}$ and $\mathcal{X}_j$ any element in $\boldsymbol{\mathcal{X}}$.

In practice, this involves setting hyperparameters $\boldsymbol{\theta}_c$ to some value, so that the model becomes a latent GMRF that INLA can tackle. Gómez-Rubio and Rue (2018) provided some ideas on how these can be chosen. Gómez-Rubio and Palmí-Perales (2019) proposed setting these values to maximum likelihood estimates (for example) and other options, and provided examples that showed that this may still provide good approximations to the posterior marginal distributions of the remainder of the parameters in the model.

4. Bayesian Model Averaging with INLA

As stated above, fitting conditional models by setting some hyperparameters ($\boldsymbol{\theta}_c$) to fixed values can be a way to use INLA to fit wider classes of models. However, this ignores the uncertainty about these hyperparameters $\boldsymbol{\theta}_c$ and makes inference about them impossible. However, BMA may help to fit the complete model, even if it cannot be fit with INLA initially.

This differs from a typical use of BMA in which models with different fixed or random effects are averaged. For example, in Bivand et al. (2014) and Bivand et al. (2015), the different models appeared as a result of setting the spatial autocorrelation of the model to different values (in a discretization of its original parametric space), but otherwise, the structure of the models was the same. For the SAC model, models are fit conditional on $\boldsymbol{\theta}_c = (\rho, \lambda)$, so that $\boldsymbol{\theta}_{-c} = (\boldsymbol{\beta}, \tau)$, and τ_ε is fixed to remove the Gaussian error term (see Appendix A). Hence, we extend the methodology in Bivand et al. (2014) and Bivand et al. (2015) to the multivariate case.

First of all, it is worth noting that the posterior marginals (of the hyperparameters $\boldsymbol{\theta}_{-c}$ and latent effects $\boldsymbol{\mathcal{X}}$) can be written as:

$$\pi(\cdot \mid \mathcal{D}) = \int \pi(\cdot, \boldsymbol{\theta}_c \mid \mathcal{D}) d\boldsymbol{\theta}_c = \int \pi(\cdot \mid \mathcal{D}, \boldsymbol{\theta}_c) \pi(\boldsymbol{\theta}_c \mid \mathcal{D}) d\boldsymbol{\theta}_c.$$

The first term in the integrand, $\pi(\cdot \mid \mathcal{D}, \boldsymbol{\theta}_c)$, is the conditional posterior marginal given $\boldsymbol{\theta}_c$, while the second term is the joint posterior distribution of $\boldsymbol{\theta}_c$, and it can be expressed as:

$$\pi(\boldsymbol{\theta}_c \mid \mathcal{D}) \propto \pi(\mathcal{D} \mid \boldsymbol{\theta}_c) \pi(\boldsymbol{\theta}_c).$$

The first term is the conditional (on $\boldsymbol{\theta}_c$) marginal likelihood, which can be approximated with INLA. The second term is the prior for $\boldsymbol{\theta}_c$, which is known. Hence, the posterior distribution of $\boldsymbol{\theta}_c$ could be computed by re-scaling the previous expression.

Bivand et al. (2014) showed that when $\boldsymbol{\theta}_c$ is unidimensional, numerical integration can be used to estimate the posterior marginal. This is done by using a regular grid of K values $\{\boldsymbol{\theta}_c^{(k)}\}_{k=1}^K$ of a fixed step.

Hence, the posterior marginals of the remainder of the hyperparameters and latent effects can be estimated as:

$$\pi(\cdot \mid \mathcal{D}) \simeq \sum_{k=1}^{K} \pi(\cdot \mid \mathcal{D}, \boldsymbol{\theta}_c^{(k)}) \, w_k$$

with weights w_k defined as:

$$w_k = \frac{\pi(\mathcal{D} \mid \boldsymbol{\theta}_c^{(k)})\pi(\boldsymbol{\theta}_c^{(k)})}{\sum_{k=1}^{K} \pi(\mathcal{D} \mid \boldsymbol{\theta}_c^{(k)})\pi(\boldsymbol{\theta}_c^{(k)})}.$$

Note how the posterior marginal $\pi(\cdot \mid \mathcal{D})$ is expressed as a BMA using the conditional posterior marginals of all the fit models.

Inference on $\boldsymbol{\theta}_c$ is based on the values $\{\boldsymbol{\theta}_c^{(k)}\}_{k=1}^{K}$ and weights $\{w_k\}_{k=1}^{K}$. For example, the posterior mean of the i^{th} element in $\boldsymbol{\theta}_c$ could be computed as $\sum_{k=1}^{K} \theta_{c,i} w_k$. Other posterior quantities could be computed similarly. Note that this also allows for multivariate posterior inference on the elements of $\boldsymbol{\theta}_c$.

The former approach was used in Bivand et al. (2014) and Bivand et al. (2015) in one dimension, but it can be easily extended to higher dimensions by considering a multidimensional grid of points for $\boldsymbol{\theta}_c$, so that each point is the center of a region of a space of equal volume. In practice, it is not necessary to consider the complete space (as it is not feasible), but the region of posterior high probability of $\boldsymbol{\theta}_c$. This may be obtained by, for example, maximizing $\pi(\mathcal{D} \mid \boldsymbol{\theta}_c)\pi(\boldsymbol{\theta}_c)$, which can be easily computed with INLA or by using a maximum likelihood estimate if available (as discussed, for example, in Gómez-Rubio and Palmí-Perales 2019).

For the SAC model, we will be dealing with dimension two as $\boldsymbol{\theta}_c = (\rho, \lambda)$. A regular grid about the posterior mode of (ρ, λ) will be created for each model using the maximum likelihood estimates. The grid is defined in an internal scale to have unbounded variables using the transformation $\gamma_1 = \log(\frac{1+\rho}{1-\rho})$ and $\gamma_2 = \log(\frac{1+\lambda}{1-\lambda})$. The variance of γ_1 and γ_2 can be derived using the Delta method (see Appendix A for details), and it can be computed using the ML standard error estimate of ρ and λ. In particular, each interval is centered at the transformed ML estimate and a semi-amplitude of three standard errors of the variables in the internal scale.

Regarding the grid, this approach may work for small dimensions of $\boldsymbol{\theta}_c$. For large dimensions, a central composite design (CCD; Box and Draper 2007) may be used, as INLA can do for the models that `R-INLA` implements. This was explored in Gómez-Rubio and Palmí-Perales (2019), and the ML estimates of the model parameters could be used to define the CCD points. Note that the dimension of the grid may depend on the model and that it may be difficult to set it beforehand. Grids that are too thin will provide averaged marginals that are too wobbly, and a smaller step may be required.

5. Example: Turnout in Italy

In order to provide an example of the methodology presented in previous sections, we take the turnout dataset described in Ward and Gleditsch (2008), which is available from the website http://ksgleditsch.com/srm_book.html. This dataset records the turnout in the 2001 Italian elections, as well as GDP per capita (`GDPCAP`) in 1997 (in million Lire). The data are comprised of 477 areas, which represent collegi or single-member districts (SDM). Adjacency is defined so that regions whose centroids are 50 km or less away are neighbors to ensure that all regions have neighbors, contiguous regions are neighbors, and Elba is joined to the closest mainland region. In order to assess the impact of the GDP per capita in the estimation of the spatial autocorrelation parameters, two models with and without the covariate are fit. The covariate is included in the log-scale to compare to the results in Ward and Gleditsch (2008).

Figure 1 shows the spatial patterns of these two variables. As can be seen, there was a clear south-north pattern for both variables. Hence, we were interested in fitting spatial econometrics models on turnout in 2011 using GDP per capita in 1997 (in the log-scale) as a predictor so that residual spatial autocorrelation was captured by the two autoregressive terms in the model.

The SAC model is fit with INLA by conditioning on the values of the two spatial autocorrelation parameters. As described in Section 2, after conditioning, the resulting model is a typical mixed-effects model with a particular matrix of covariates and a known structure for the precision matrix of the random effects. Hence, we consider $\boldsymbol{\theta}_c = (\rho, \lambda)$ and $\boldsymbol{\theta}_{-c} = (\beta, \tau)$.

Both autocorrelation parameters are given values in the range $(-1, 1)$. Note that, because the same adjacency matrix is used, the actual domain for both parameters is $(1/\lambda_{min}, 1)$, with λ_{min} the minimum eigenvalue of the adjacency matrix $\boldsymbol{W}$. In this case, $\lambda_{min} = -0.82$, so taking the range $(-1, 1)$ will be enough.

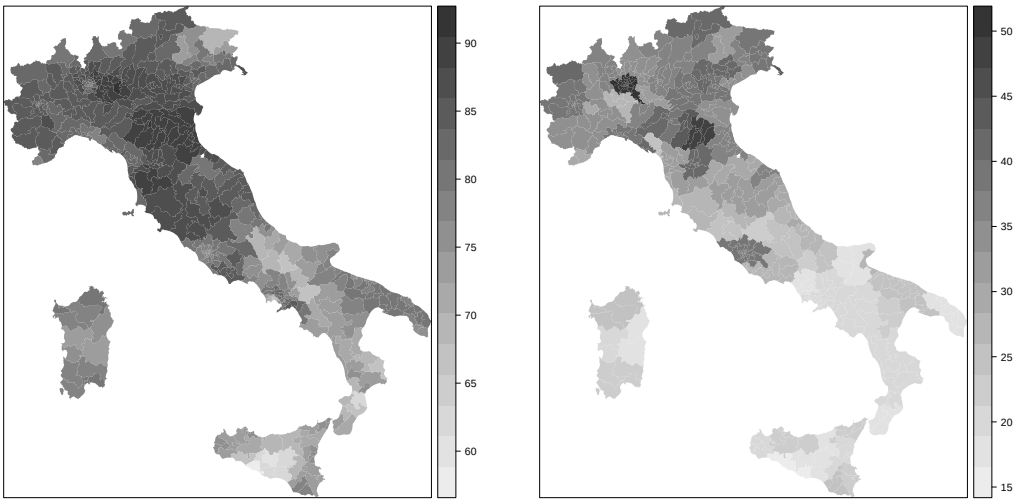

Figure 1. Spatial distribution of the turnout in 2011 (left) and GDP per capita in 1997 (right) in Italy at the collegi level.

Maximum likelihood estimates of the models are obtained with package `spdep` (Bivand et al. 2013), and MCMC estimates are obtained with package `spatialreg` (Bivand and Piras 2015; Bivand et al. 2013). For MCMC, we used 10,000 burn-in iterations, plus another 90,000 iterations for inference, of which only one in ten was kept to reduce autocorrelation. To speed up convergence, the initial values of ρ and λ were set to their maximum likelihood estimates. BMA with INLA estimates was obtained as explained next.

A grid about the posterior mode of (ρ, λ) was created as described in Section 4 using the ML estimates and their standard errors. See Table 1 for the actual values of the ML estimates. Furthermore, a grid of 160×40 points was used to represent the search space and fit the conditional models with `R-INLA` for the model without the covariate and a grid of 40×20 for the model with `log(GDPCAP)`. A different grid was used because the model without the covariate required a thinner grid.

The computation times of the different models were small, and each individual model took about one second to fit on a computer with 32 cores of two Intel Xeon Gold 6130 CPUs at 2.101 GHz with more than 500 Gb of RAM. Although the number of models required to be fit was large, we fit 62 models in parallel (using a single thread each) so that the total time required to fit the model without covariates was about 2 min and less than 1 min for the model with the covariate (because it had a smaller grid). Since all the INLA models were run in parallel, the computational cost was easily distributed across a network of computers to reduce computing time. Merging the different INLA models took a few seconds only.

Table 1 provides a summary of the estimates of the model parameters using different inference methods. First of all are the maximum likelihood (ML) estimates (computed with the function `sacsarlm` in the `spdep` package). Next, ρ and λ were fixed at their ML estimates, and the model was fit with INLA. Next, the posterior marginals of the model parameters using BMA with INLA and MCMC are shown. In general, point estimates obtained with the different methods provided very similar values. MCMC and maximum likelihood also provided very similar estimates of the uncertainty of the point estimates (when available for maximum likelihood). BMA with INLA seemed to provide very similar results to MCMC for both models.

Table 1. Summary statistics of the model parameters. INLA, integrated nested Laplace approximation; BMA, Bayesian model averaging.

Model	Parameter	Max. lik.		INLA-Max. lik.		BMA		MCMC	
		Mean	St.Error	Mean	St. dev.	Mean	St. dev.	Mean	St. Error
No covariates	β_0	5.86	1.54	5.88	0.10	6.39	1.59	6.56	1.95
	ρ	0.93	0.02	0.93	–	0.92	0.02	0.92	0.02
	λ	0.09	0.10	0.09	–	0.12	0.10	0.12	0.10
	τ^{-1}	3.66	–	3.68	0.24	3.71	0.25	3.75	0.28
Covariates	β_0	5.05	2.03	5.04	1.17	5.97	2.17	5.81	2.22
	β_1	1.60	0.51	1.60	0.34	1.81	0.58	1.77	0.59
	ρ	0.87	0.04	0.87	–	0.85	0.04	0.85	0.04
	λ	0.18	0.12	0.18	–	0.23	0.11	0.22	0.11
	τ^{-1}	3.77	–	3.80	0.25	3.85	0.27	3.90	0.30

Note the positive coefficient of log-GDP per capita, which meant a positive association with turnout. Furthermore, the spatial autocorrelation ρ had a higher value than λ, which indicated higher autocorrelation on the response than on the error term.

Figure 2 shows the values of the marginal log-likelihoods of all the conditional models fit, as well as the weights used in BMA. Note that the wide variability of the marginal log-likelihoods made the weights very small. Weights were almost zero for most models with a very few exceptions. Note that this was good because it indicated that the search space for ρ and λ was adequate and that the region of high posterior density was explored because the log-posterior density decayed as the values of (ρ, λ) moved away from the posterior mode. Note that there was an implicit assumption that the posterior of (ρ, λ) was unimodal. Expanding the grid may help to detect other modes if we suspected this were the case. However, in this particular case, we did not believe that the SAC model produced multimodal posteriors as the plots in Figure 1 showed a positive spatial correlation, and negative values of ρ and λ were not likely.

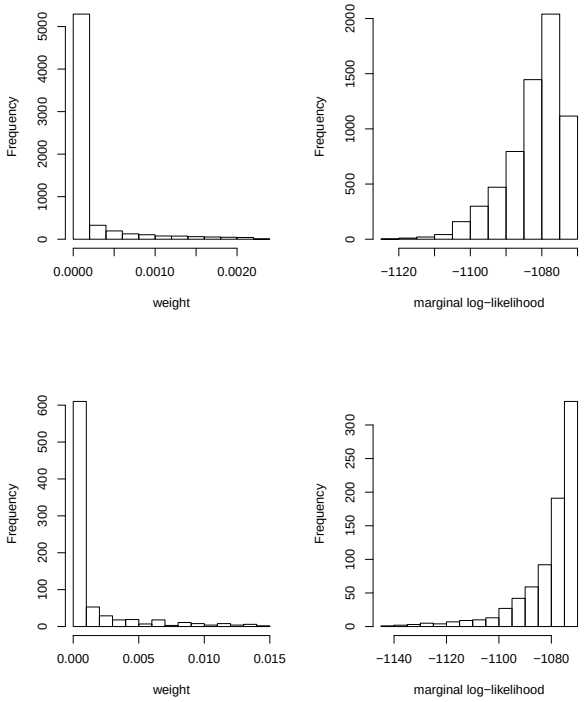

Figure 2. Weights and marginal likelihoods for the model with no covariates (top row) and with covariates (bottom row).

Figure 3 shows the posterior marginals of the spatial autocorrelation parameters using kernel smoothing (Venables and Ripley 2002). For the BMA output, weighted kernel smoothing was used. In general, there was good agreement between BMA with INLA and MCMC. The ML estimates were also very close to the posterior modes.

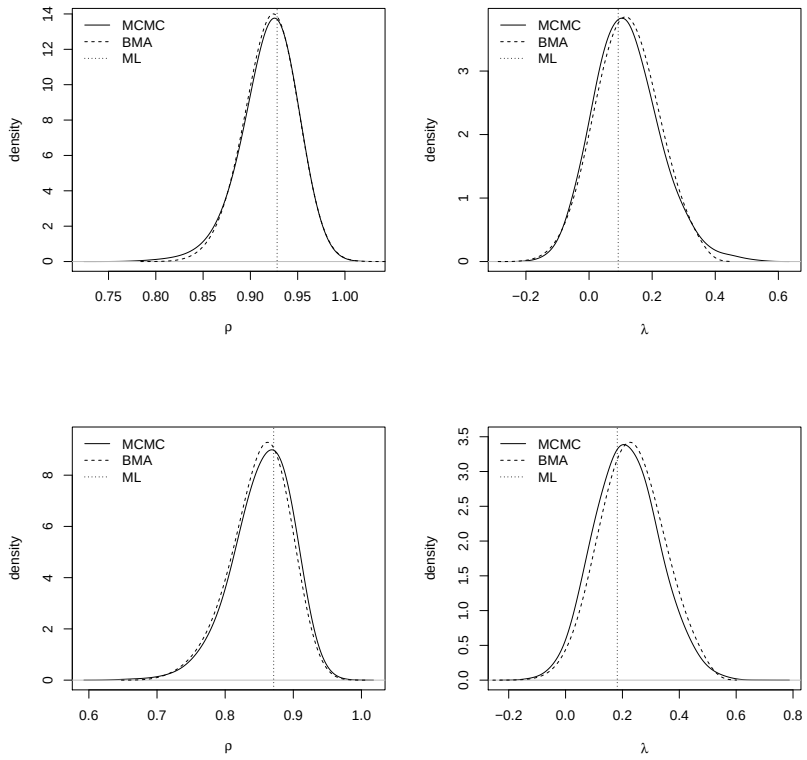

Figure 3. Posterior marginals of the spatial autocorrelation parameters for the model with no covariates (top row) and with covariates (bottom row).

Similarly, Figure 4 shows the joint posterior distribution of the autocorrelation parameters obtained using two-dimensional kernel smoothing. For the BMA output, this was obtained using two-dimensional weighted kernel smoothing with function `kde2d.weighted` in package `ggtern` (Hamilton and Ferry 2018). The ML estimate were also added (as a black dot). As with the posterior marginals, the joint distribution was close between MCMC and BMA with INLA. The posterior mode was also close to the ML estimate. The plots showed a negative correlation between the spatial autocorrelation parameters, which may indicate that they struggled to explain spatial correlation in the data (see also Gómez-Rubio and Palmí-Perales 2019). Furthermore, BMA with INLA was a valid approach to make joint posterior inference on a subset of hyperparameters in the model.

These results showed the validity of relying on BMA with INLA to fit highly parameterized models. This approach also accounted for the uncertainty of all model parameters and should be preferred to other inference methods based on plugging-in or fixing some of the models parameters. In our case, we relied on BMA with INLA so that conditional sub-models were fit and then combined. This had two main benefits. First of all, it allowed for full inference on all model parameters, and secondly, uncertainty about all parameters was taken into account.

Figure 5 shows the posterior marginals of the coefficients and the variance of the models. In general, BMA with INLA and MCMC showed very similar results for all parameters The posterior modes of MCMC and BMA with INLA were very close to the ML estimates. Marginals provided by

INLA with ML estimates were very narrow for the fixed effects, which was probably due to ignoring the uncertainty about the spatial autocorrelation parameters.

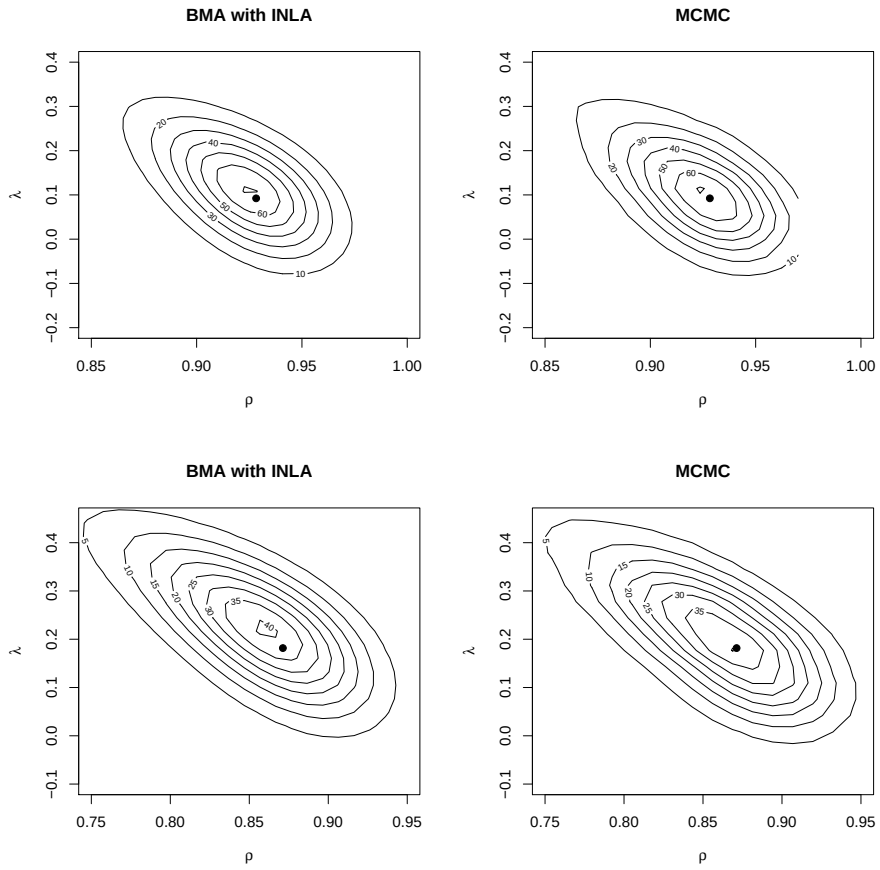

Figure 4. Joint posterior distribution of the spatial autocorrelation parameters for the model with no covariates (top row) and with covariates (bottom row). The black dot represents the maximum likelihood estimate.

Computation of the impacts was important because they measured how changes in the values of the covariates reflected on the changes of the response variable in the current area (direct impacts) and neighboring areas (indirect impacts). Note that impacts were only computed for the model with the covariate included. Table 2 shows the estimates of the different average impacts. In general, all estimation methods provided very similar values of the point estimates, and BMA with INLA and MCMC estimates were very close. Figure 6 displays the posterior marginal distributions of the average impacts. Again, these estimates were very similar among estimation methods.

Table 2. Average impacts estimated with the different methods.

Impact	Max. lik.		INLA-Max. lik.		BMA		MCMC	
	Mean	**St. dev.**	**Mean**	**St. dev.**	**Mean**	**St. dev.**	**Mean**	**St. dev.**
Direct	2.25	–	2.26	0.48	2.45	0.66	2.43	0.71
Indirect	9.97	–	10.18	2.15	9.75	2.25	9.66	2.51
Total	12.22	–	12.43	2.63	12.20	2.64	12.09	2.97

Finally, the two models fit to the data could also be averaged to account for the model uncertainty. Note that this resembled the typical use of BMA as the models were not exactly the same and included different terms (i.e., covariates in this case). The marginals of the common effects could be averaged by

weighting their marginals as seen in Section 4, where the weights now depended on the corresponding marginal likelihoods of the model. For the effects that only appeared in some of the models, the weights could be computed using only the models where these terms appeared. When averaging marginals for individual effects, the functions in package `INLABMA` (Bivand et al. 2015) can be used.

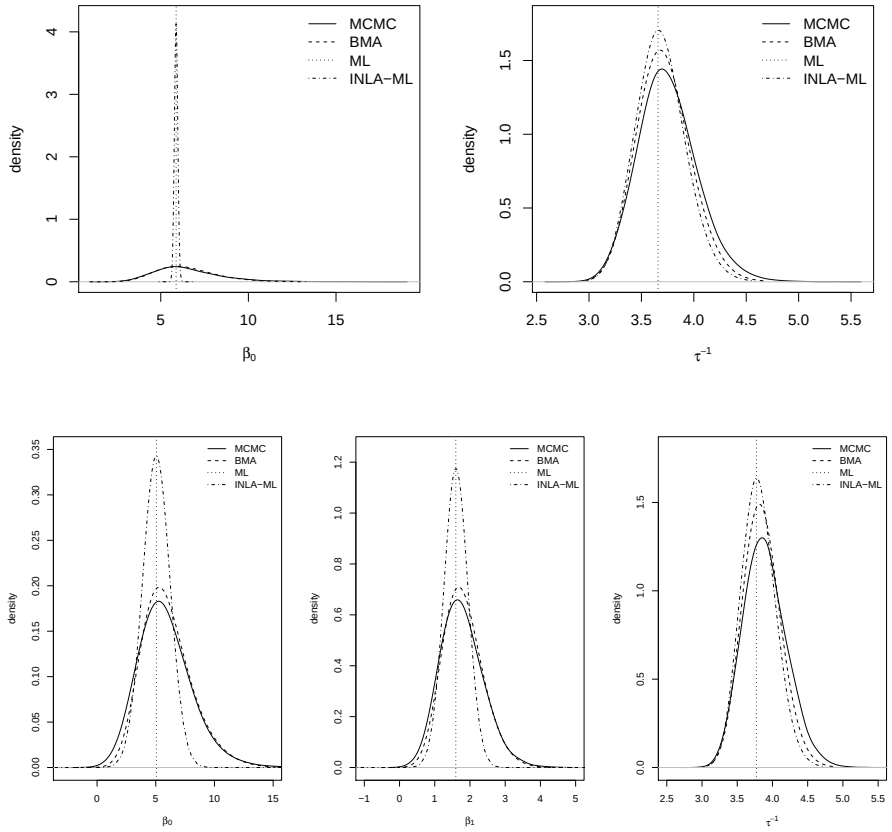

Figure 5. Posterior marginal distributions of the coefficients and variances for the model with no covariates (top row) and with covariates (bottom row).

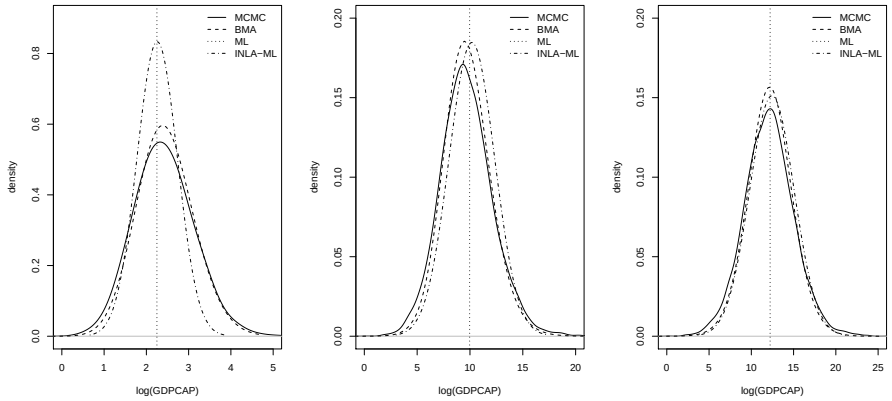

Figure 6. Posterior marginals of the average direct (left), indirect (middle) and total (right) impacts of the log-GDP per capita.

For this particular example, the marginal log-likelihoods for the model with no covariates and with covariates were -1121.80 and -1143.94, respectively. This made the weights be about one and zero for the common effects and one for the effect of the covariate.

6. Discussion

Bayesian model averaging with the integrated nested Laplace approximation was illustrated to make inference about the parameters and latent effects of highly parameterized models. The appeal of this methodology was that it relaxed the constraints on the model, which did not need to be latent GMRF anymore, but conditional latent GMRF. This was laid out with an example based on spatial econometrics models.

The different models that were averaged appeared because of a discretization of a subset of the hyperparameters, which were assumed fixed when the models were fit. Hence, the structure of the model did not change, but the values of some hyperparameters did. By combining the different models using BMA, these hyperparameters were integrated out and their uncertainty accounted for. However, the joint posterior of this subset of the hyperparameters could be obtained, which allowed us to make multivariate posterior inference. This was different from a typical BMA, where models with different latent effects are considered, but this was also possible with INLA once the desired models were fit.

Less expensive alternatives to BMA include setting the values of a subset of the hyperparameters to their posterior modes or maximum likelihood estimates. This could still produce accurate estimates of the posterior marginals of the remainder of the parameters in the model (Gómez-Rubio and Palmí-Perales 2019), but ignored the uncertainty about some parameters in the model, as well as any posterior inference about them.

Although we used numerical integration methods to estimate the joint posterior distribution of the conditioning hyperparameters, this was limited to low dimensions and may not scale well. However, other approaches could be used such as MCMC algorithms (Gómez-Rubio and Rue 2018).

Finally, BMA with INLA made inference about a small subset of hyperparameters in the model possible. This was an interesting feature as INLA focused on marginal inference, and joint posterior analysis required sampling from the internal representation of the latent field (Gómez-Rubio 2020), which could be costly. We also showed how this could be used to compute the posterior marginal distributions of derived quantities (i.e., the impacts) that depended on a small subset of hyperparameters. Our results with BMA with INLA were very close to those obtained with MCMC, which supported the use of BMA with INLA as a feasible alternative for multivariate posterior inference.

Author Contributions: This was a collaborative project. All authors contributed to the paper equally. All authors have read and agreed to the published version of the manuscript.

Funding: Virgilio Gómez-Rubio was funded by Consejería de Educación, Cultura y Deportes (JCCM, Spain) and FEDER, Grant Number SBPLY/17/180501/000491, as well as by Ministerio de Economía y Competitividad (Spain), Grant Number MTM2016-77501-P.

Acknowledgments: We would like to thank Gonzalo García-Donato Layrón for the invitation to contribute to the Special Issue on Bayesian and Frequentist Model Averaging.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Implementation Details

Appendix A.1. Model Fitting

The implementation of the SAC model with INLA was done in two steps. First, the model was fit given ρ and λ. Given these values, the model was a linear mixed-effects model with the design matrix of the fixed effects equal to:

$$(\boldsymbol{I} - \rho \boldsymbol{W})^{-1} \boldsymbol{X}$$

and the vector of random effects has mean zero and precision matrix:

$$\tau\boldsymbol{\Sigma},$$

where τ is a precision parameter to be estimated. Matrix $\boldsymbol{\Sigma}$ is a precision matrix fully determined (given that ρ and λ are known), and it is:

$$\boldsymbol{\Sigma} = (\boldsymbol{I} - \rho\boldsymbol{W}^\top)(\boldsymbol{I} - \lambda\boldsymbol{W}^\top)(\boldsymbol{I} - \lambda\boldsymbol{W})(\boldsymbol{I} - \rho\boldsymbol{W}).$$

This model can be fit with the `R-INLA` package using a Gaussian likelihood and a design matrix for the fixed effects $(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{X}$ and latent random effects with zero mean and precision matrix $\tau\boldsymbol{\Sigma}$. In particular, the latent random effects can be defined using the effect `generic0` in `R-INLA`. Furthermore, the precision of the Gaussian likelihood τ_ε is set to $\exp(15)$ to remove the error term. Otherwise, the model would include another error term in the linear predictor in addition to fixed and random effects. Higher values of the precision led to unstable estimates of the model and the marginal likelihood.

The `generic0` latent effect is a multivariate Gaussian distribution with zero mean and precision matrix $\tau\boldsymbol{\Sigma}$. Because $\boldsymbol{\Sigma}$ is known, its determinant is ignored by `R-INLA` when computing the log-likelihood. For this reason, we added $+\frac{1}{2}\log(|\boldsymbol{\Sigma}|)$ to the computation of the marginal likelihood reported by `R-INLA`, with $|\boldsymbol{\Sigma}|$ the determinant of $\boldsymbol{\Sigma}$.

Once the different models were fit, weights w_k were computed as in Section 4. Models were merged using function `inla.merge`, which took the list of models fit and the vector of weights.

Appendix A.2. Grid Definition

As stated in Section 5, the grid to explore the values of (ρ, λ) was a regular grid defined in an internal scale to make the internal parameters unbounded. The estimate of the variance of the parameters in the internal scale could be obtained by using the Delta method. The transformation used was $\gamma_1 = \log(\frac{1+\rho}{1-\rho})$ and $\gamma_2 = \log(\frac{1+\lambda}{1-\lambda})$. If we considered $g(x) = \log(\frac{1+x}{1-x})$, then $\gamma_1 = g(\rho)$ and $\gamma_2 = g(\lambda)$. Hence, the inverse transformation could be considered if we took function $f(x) = 2\frac{\exp(x)}{1+\exp(x)} - 1$, and the original parameters were defined as $\rho = f(\gamma_1)$ and $\lambda = f(\gamma_2)$.

The Delta method states that if σ^2 is the variance of a variable x with mean μ, then an estimate of the variance of $g(x)$ is $\sigma^2(g^{'}(\mu))^2$, with $g^{'}(x)$ the first derivative of $g(x)$. In our case, we have that:

$$g^{'}(x) = \frac{2}{(1+x)(1-x)}.$$

Using ML estimation, for both ρ and λ, an estimate of the variance is provided by their respective squared standard errors. Hence, the estimate of the variance of γ_1 is:

$$se_\rho^2\left(\frac{2}{(1+\rho)(1-\rho)}\right)^2.$$

Here, se_ρ is the standard error of ρ, and ρ represents the ML estimate. Similarly, an estimate of the variance of γ_2 is developed.

Note that because the regular grid was defined in the internal scale, the prior should be on γ_1 and γ_2. This could be derived (Chapter 5, Gómez-Rubio 2020) by considering a correction due to the transformation. For example, the prior on ρ is a uniform in the interval $(-1, 1)$, which has a constant density in that interval. The prior on γ_1, $\pi(\gamma_1)$, is $\pi(\rho)|\frac{\partial\rho}{\partial\gamma_1}|$, with:

$$\frac{\partial\rho}{\partial\gamma_1} = \frac{\partial(f(\gamma_1))}{\partial\gamma_1} = 2\frac{\exp(\gamma_1)}{(1+\exp(\gamma_1))^2}.$$

Hence, the prior on γ_1 is:

$$\pi(\gamma_1) = \frac{1}{2} 2 \frac{\exp(\gamma_1)}{(1+\exp(\gamma_1))^2} = \frac{\exp(\gamma_1)}{(1+\exp(\gamma_1))^2}.$$

The prior on γ_2 is derived in a similar way, and it is:

$$\pi(\gamma_2) = \frac{\exp(\gamma_2)}{(1+\exp(\gamma_2))^2}.$$

Appendix A.3. Impacts

The computation of the impacts was done by exploiting that models were fit conditional on ρ. For example, to compute the average total impact, the posterior marginal of $\beta_r/(1-\rho)$ was computed given ρ. This was easy as $\pi(\beta_r \mid \mathbf{y}, \rho)$ was estimated by INLA, and the posterior of $\beta_r/(1-\rho)$ could be easily obtained by transforming the posterior marginal of β_r. Then, all the conditional marginals of the average total impacts were combined using BMA and associated weights w_k. Posterior marginals for average direct and indirect impacts could be computed in a similar way.

Appendix A.4. Final Remarks

It is worth stressing that the model fit was done in parallel, which meant that BMA with INLA scaled well when the number of grid points or hyperparameters increased. In our case, we used function `mclapply` to fit all the models required by BMA with INLA.

The `R` code to run all the models shown in this paper with the different methods is available from GitHub at https://github.com/becarioprecario/SAC_INLABMA.

References

Bivand, Roger, and Gianfranco Piras. 2015. Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software* 63: 1–36. [CrossRef]

Bivand, Roger S., Virgilio Gómez-Rubio, and Håvard Rue. 2014. Approximate Bayesian inference for spatial econometrics models. *Spatial Statistics* 9: 146–65. [CrossRef]

Bivand, Roger S., Virgilio Gómez-Rubio, and Håvard Rue. 2015. Spatial data analysis with `R-INLA` with some extensions. *Journal of Statistical Software* 63: 1–31. [CrossRef]

Bivand, Roger S., Edzer Pebesma, and Virgilio Gómez-Rubio. 2013. *Applied Spatial Data Analysis with R*, 2nd ed. New York: Springer.

Box, Geroge E., and Norman R. Draper. 2007. *Response Surfaces, Mixtures, and Ridge Analyses*, 2nd ed. Hoboken: John Wiley & Sons, Inc.

Brooks, Steve, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng. 2011. *Handbook of Markov Chain Monte Carlo*. Boca Raton: Chapman & Hall/CRC Press.

Gilks, Walter R., Sylvia Richardson, and David J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. Boca Raton: Chapman & Hall.

Gómez-Rubio, Virgilio. 2020. *Bayesian Inference with INLA*. Boca Raton: Chapman & Hall/CRC Press.

Gómez-Rubio, Virgilio, and Håvard Rue. 2018. Markov chain monte carlo with the integrated nested laplace approximation. *Statistics and Computing* 28: 1033–51. [CrossRef]

Gómez-Rubio, Virgilio, Roger S. Bivand, and Håvard Rue. 2017. Estimating spatial econometrics models with integrated nested Laplace approximation. *arXiv* arXiv:1703.01273.

Gómez-Rubio, Virgilio, and Francisco Palmí-Perales. 2019. Multivariate posterior inference for spatial models with the integrated nested laplace approximation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68: 199–215. doi:10.1111/rssc.12292. [CrossRef]

Haining, Robert. 2003. *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press.

Halleck Vega, Solmaria, and J. Paul Elhorst. 2015. The SLX model. *Journal of Regional Science* 55: 339–63. doi:10.1111/jors.12188. [CrossRef]

Hamilton, Nicholas E., and Michael Ferry. 2018. ggtern: Ternary diagrams using ggplot2. *Journal of Statistical Software, Code Snippets* 87: 1–17. doi:10.18637/jss.v087.c03. [CrossRef]

Hastings, Wilfred K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97–109. [CrossRef]

Hoeting, Jennifer, David Madigan, Adrian Raftery, and Chris Volin Sky. 1999. Bayesian model averaging: A tutorial. *Statistical Science* 14: 382–401.

Hubin, Aliaksandr, and Geir Storvik. 2016. Estimating the marginal likelihood with Integrated nested Laplace approximation (INLA). *arxiv* arxiv:1611.01450.

LeSage, James, and Robert Kelley Pace. 2009. *Introduction to Spatial Econometrics*. London: Chapman and Hall/CRC.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. Equations of state calculations by fast computing machine. *Journal of Chemical Physics* 21: 1087–91. [CrossRef]

Rue, H., and L. Held. 2005. *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman and Hall/CRC Press.

Rue, Havard, Sara Martino, and Nicolas Chopin. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society, Series B* 71: 319–92. [CrossRef]

Rue, Håvard, Andrea Riebler, Sigrunn H. Sørbye, Janine B. Illian, Daniel P. Simpson, and Finn K. Lindgren. 2017. Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application* 4: 395–421. doi:10.1146/annurev-statistics-060116-054045. [CrossRef]

Venables, William N., and Brian D. Ripley. 2002. *Modern Applied Statistics with S*, 4th ed. New York: Springer. ISBN 0-387-95457-0.

Ward, Michael Don, and Kristian Skrede Gleditsch. 2008. *Spatial Regression Models*. Thousand Oaks: Sage Publications, Inc.