Approximate Bayesian inference for spatial econometrics models

V. Gómez-Rubio

Department of Mathematics
U. of Castilla-La Mancha, Albacete, Spain

Universidad de las Palmas de Gran Canaria
12th March 2021

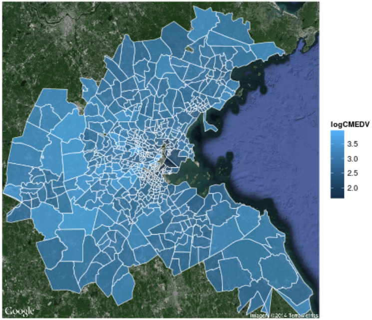
joint work with Roger S. Bivand and Havard Rue

Talk Outline

- Spatial Models for Lattice Data
- Spatial Econometrics Models
- Introduction to the Integrated Nested Laplace Approximation (INLA)
- R-INLA package
- Extending INLA and R-INLA
- Application to (spatial) GLMMs

Spatial Models for Lattice Data

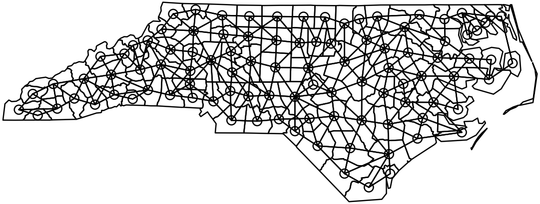
- Lattice data involves data measured at different areas, e.g., neighbourhoods, cities, provinces, states, etc.
- Spatial dependence appears because neighbour areas will show similar values of the variable of interest



Models for lattice data

- We have observations $y = \{y_i\}_{i=1}^{n}$ from the n areas
- y is assigned a multivariate distribution that *accounts for spatial dependence*
- A common way of describing spatial proximity in lattice data is by means of an *adjacency matrix W*
- $W[i, j]$ is non-zero if areas i and j are neighbours
- Usually, two areas are neighbours if the share a common boundary
- There are other definitions of neighbourhood

Adjacency matrix



Regression models

- It is often the case that, in addition to y_i, we have a number of covariates x_i
- Hence, we may want to regress y_i on x_i
- In addition to the covariates we may want to account for the spatial structure of the data
- Different types of regression models can be used to model lattice data:
  - Generalized Linear Models (with spatial random effects)
  - Spatial econometrics models
- Generalized Linear Mixed Models are often used

Linear Mixed Models

- A common approach (for Gaussian data) is to use a linear regression with random effects

$$Y = X\beta + Zu + \varepsilon$$

- The vector random effects u is modelled as a MVN:

$$u \sim N(0, \sigma_u^2 \Sigma)$$

- Σ is defined such as it induces higher correlation with adjacent areas
- Z is a design matrix for the random effects
- $\varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$: error term
- Similar for Generalised Linear Mixed Models

Spatial Econometrics Models

- Slightly different approach to spatial modelling
- Instead of using latent effects, spatial dependence is modelled explicitly
- Autoregressive models are used to make the response variable to depend on the values at its neighbours

Simultaneous Autoregresive Model (SEM)

- This model includes covariates
- Autoregressive on the error term

$$y = X\beta + u; u = \rho W u + e; e \sim N(0, \sigma^2)$$

$$y = X\beta + \varepsilon; \varepsilon \sim N(0, \sigma^2 (I - \rho W)^{-1} (I - \rho W')^{-1})$$

Spatial Lag Model (SLM)

- This model includes covariates
- Autoregressive on the response

$$y = \rho W y + X\beta + e; e \sim N(0, \sigma^2)$$

$$y = (I - \rho W)^{-1} X\beta + \varepsilon;\ \varepsilon \sim N(0, \sigma^2 (I - \rho W)^{-1} (I - \rho W')^{-1})$$

Spatial Durbin Model (SDM)

- This model includes covariates
- Autoregressive on the response
- In addion, we include the lagged-covariates WX as another extra term in the regression

$$y = \rho W y + X\beta + WX\gamma + e = [X, WX][\beta, \gamma] + e;\ ; e \sim N(0, \sigma^2)$$

$$y = \rho W y + XWX\mathcal{B} + e;\ XWX = [X, WX];\ \mathcal{B} = [\beta, \gamma]$$

$$y = (I - \rho W)^{-1} XWX\ \mathcal{B} + \varepsilon;\ \varepsilon \sim N(0, \sigma^2 (I - \rho W)^{-1} (I - \rho W')^{-1})$$

Structure of spatial random effects

There are **many** different ways of including spatial dependence in Σ:

- Simultaneous autoregressive (SAR):

$$\Sigma = [(I - \rho W)'(I - \rho W)]^{-1}$$

- Conditional autoregressive (CAR):

$$\Sigma = (I - \rho W)^{-1}$$

- $\Sigma_{i,j}$ depends on a function of $d(i,j)$. For example:

$$\Sigma_{i,j} = \exp\{-d(i,j)/\phi\}$$

- 'Mixture' of matrices (Leroux et al.'s model):

$$\Sigma = [(1-\lambda)I_n + \lambda M]^{-1};\ \lambda \in (0,1)$$

M precision of instrinsic CAR specification

Bayesian Inference

- Bayesian inference is based on Bayes' rule to compute the probability of the parameters in the model (θ) given the observed data (y):

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)}$$

- $\pi(y|\theta)$ is the likelihood of the model
- $\pi(\theta)$ is the prior distribution of the parameters in the model
- $\pi(y)$ is a normalising constant that is often ignored
- Vague priors are often used for most parameters in the model

Model fitting and computational issues

- Fitting a Bayesian model means computing $\pi(\theta|y)$
- θ contains all parameters in the model and, possibly, other derived quantities
- For example, we could compute posterior probabilities of linear predictors, random effects, sums of random effects, etc.
- Depending on the likelihood and the prior distribution computing $\pi(\theta|y)$ can be very difficult
- In the last 20-30 years some computational approaches have been proposed to estimate $\pi(\theta|y)$ with Monte Carlo methods

Inference with MCMC

- MCMC provides simulations from the ensemble of model parameters, i.e., a multivariate distribution
- This will allow us to estimate the joint posterior distribution
- However, we may be interested in a single parameter or a subset of the parameters
- Inference for this subset of parameters can be done by simply ignoring the samples for the other parameters
- Using the samples it is possible to compute the posterior distribution of any function on the model parameters
- MCMC may require lots of simulations to make valid inference
- Also, we must check that the burn-in period has ended, i.e., we have reached the posterior distribution

Integrated Nested Laplace Approximation

- Sometimes we only need marginal inference on some parameters, i.e., we need $\pi(\theta_i|y)$
- Rue et al. (2009) propose a way of approximating the marginal distributions
- Now we are dealing with (many) univariate distributions
- This is computationally faster because numerical integration techniques are used instead of Monte Carlo sampling

Integrated Nested Laplace Approximation

- We assume that observations **y** are independent given **x** (latent effects) and $\theta = (\theta_1, \theta_2)$ (two sets of hyperparameters)
- The model likelihood can be written down as

$$\pi(\mathbf{y}|\mathbf{x}, \theta) = \prod_{i \in \mathcal{I}} \pi(y_i | x_i, \theta)$$

- x_i is the latent linear predictor η_i and other latent effects

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i \qquad (1)$$

- $\mathcal{I}$ represents the indices of the observations (missing observations are not include here, for example)
- $\theta = (\theta_1, \theta_2)$ is a vector of hyperparameters for the likelihood and the distribution of the latent effects

Integrated Nested Laplace Approximation

- **$\mathbf{x}$** is assumed to be distributed as a Gaussian Markov Random Field with precission matrix $Q(\theta_2)$
- The posterior distribution of the model parameters and hyperparameters is:

$$\pi(\mathbf{x}, \theta|\mathbf{y}) \propto \pi(\theta)\pi(\mathbf{x}|\theta) \prod_{i \in \mathcal{I}} \pi(y_i|x_i, \theta) \propto$$

$$\pi(\theta)|\mathbf{Q}(\theta)|^{n/2} \exp\{-\frac{1}{2}\mathbf{x}^T \mathbf{Q}(\theta)\mathbf{x} + \sum_{i \in \mathcal{I}} \log(\pi(y_i|x_i, \theta)\}$$

Integrated Nested Laplace Approximation

The marginal distributions for the latent effects and hyper-parameters can be written as

$$\pi(x_i|\mathbf{y}) = \int \pi(x_i|\theta, \mathbf{y})\pi(\theta|\mathbf{y})d\theta$$

and

$$\pi(\theta_j|\mathbf{y}) = \int \pi(\theta|\mathbf{y})d\theta_{-j}$$

Integrated Nested Laplace Approximation

Rue et al. (2009) provide a simple approximation to $\pi(\theta|\mathbf{y})$, denoted by $\tilde{\pi}(\theta|\mathbf{y})$, which is then used to compute the approximate marginal distribution of a latent parameter x_i:

$$\tilde{\pi}(x_i|\mathbf{y}) = \sum_k \tilde{\pi}(x_i|\theta_k, \mathbf{y}) \times \tilde{\pi}(\theta_k|\mathbf{y}) \times \Delta_k$$

Δ_k are the weights of a particular vector of values θ_k in a grid for the ensemble of hyperparameters .

R-INLA package

- Available from `http://www.r-inla.org`
- Implementation of INLA as an R package
- `inla()`-function similar to `glm()`
- Model is defined in a `formula`
- Flexible way of defining:
  - Likelihood
  - Prior
  - Latent effects
- Provides marginals of:
  - Model parameters
  - Linear predictor
  - Linear combinations of model parameters
- Tools to manipulate $\pi(\cdot|y)$ to compute $\pi(f(\cdot)|y)$
- Model assessment/choice: Marginal likelihood, DIC, CPO, ...

Summary of implemented latent effects

Name in `f()`	Model
`besag`	Intrinsic CAR
`besagproper`	Proper CAR
`bym`	Convolution model
`rw2d`	2-D random walk
`matern2d`	Matérn correlation
`generic0`	$\Sigma = \frac{1}{\tau} Q^{-1}$
`generic1`	$\Sigma = \frac{1}{\tau}(I_n - \frac{\rho}{\lambda_{max}} C)^{-1}$
`seasonal`	Seasonal variation
`ar1`	Autoreg. order 1
`ar`	Autoreg. order p
`iid?d`	Correlated effects with Wishart prior
`mec`	Classical mearurement error
`meb`	Berkson mearurement error

Full list at `http://www.r-inla.org/models/latent-models`

INLA & Spatial econometrics models

- In principle, INLA can handle a large number of models
- The R-INLA package for the R software implements a number of likelihoods and latent effects
- Several spatial models are implemented (Gómez-Rubio et al., 2014)
- SEM and SLM were not implemented at the time
- The SAR specification was not implemented as a random effect then
- Linear predictors are multiplied by $(I - \rho W)^{-1}$, and this is not implemented either
- What to do then? (Bivand et al., 2014, 2015)

A possible approach

- Conditioning on ρ, SEM and SLM become models that R-INLA can fit
- We can fit different models conditioning on different values of ρ. This will provide

$$\pi(\theta_i | y, \rho = \rho_k),\ \ k = 1, 2, \ldots$$

- The values of ρ can be chosen equally spaced in (-1,1)
- For each fitted model, we can compute the marginal likelihood, i.e., the likelihood of that model: $\pi(y|\rho = \rho_k)$
- Our inference can be based on the model with the largest likelihood
- However, we cannot obtain a marginal distribution for ρ and cannot compute summary statistics

Bayesian Model Averaging

- A better aproach is to combine the different fitted models in some way
- Bayesian Model Averaging provides a way of combining all these models (Bivand et al., 2014, 2015)
- For each fitted model (conditioned on a value of ρ) we have $\pi(\theta_i|y, \rho = \rho_k)$ and $\pi(y|\rho = \rho_k)$
- We can choose a prior distribution for ρ: $\pi(\rho)$
- It should be noted that the marginal distribution of ρ is

$$\pi(\rho|y) = \frac{\pi(y|\rho)\pi(\rho)}{\pi(y)} \propto \pi(y|\rho)\pi(\rho)$$

- The marginal distribution for ρ can be computed by fitting a curve to the values

$$[\rho_k, \pi(y|\rho = \rho_k)\pi(\rho = \rho_k)]$$

Marginal distribution of ρ

- The marginal distribution of a parameter can be written as

$$\pi(\theta_i|y) = \int \pi(\theta_i, \rho|y) d\rho = \int \pi(\theta_i|y, \rho)\pi(\rho|y) d\rho$$

- The previous integral can be aproximated as follows:

$$\sum_k \pi(\theta_i|y, \rho = \rho_k) \frac{\pi(y|\rho = \rho_k)\pi(\rho_k)}{\sum_k \pi(y|\rho = \rho_k)\pi(\rho_k)} = \sum_k w_k \pi(\theta_i|y, \rho = \rho_k)$$

- Finally, a spline can be fitted to the resulting function so that it can be used to compute other quantities, such as the mean, mode, quantiles, etc.

Example: Boston housing data

- We will re-analyse the Boston housing data (Harrison and Rubinfeld, 1978)
- Median of owner-occupied houses using relevant covariates and the spatial structure of the data (Pace and Gilley, 1997)
- We have fitted the Leroux et al.'s model using the previous approach and MCMC to compare the estimates of the model parameters (Bivand et al., 2015)
- In the linear predictor:
  - Fixed effects (i.e., covariates)
  - Spatial effect (Leroux et al.'s model)
  - Error term

Fitting Leroux et al.'s model

- Complex variance-covariance matrix:

$$\Sigma = [(1-\lambda)I_n + \lambda M]^{-1};\ \lambda \in (0,1)$$

- M structure of precision of instrinsic CAR (very sparse matrix!)
- Mixture of i.i.d. Gaussian effect and CAR spatial effect
- Fit models with **R-INLA** conditioning on λ, to obtain:
  - $\pi(\theta|y,\lambda)$, with function `leroux.inla()`
  - $\pi(y|\lambda)$
- λ takes values on a fine grid
- Combine models using the **INLABMA** package
  - BMA of models fitted with R-INLA
  - Takes a list of fitted models AND prior on λ
  - Returns a model in a similar format as `inla()`
- We will be comparing our results with MCMC (**CARBayes** package)

Fitting Leroux et al.'s model

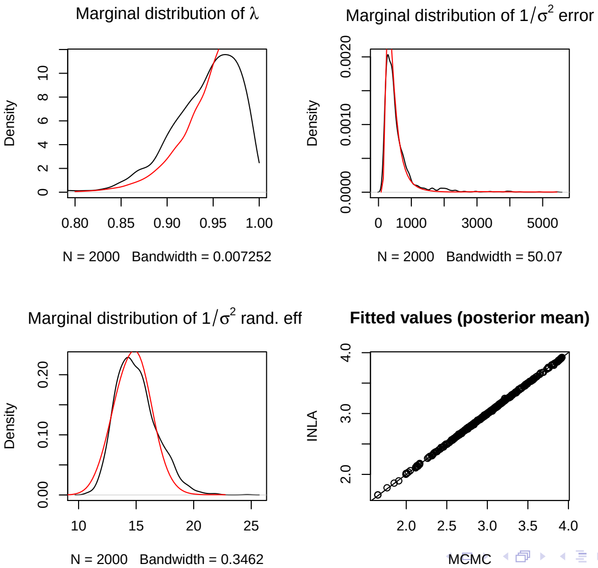
```
#Define parameters for model fitting
rlambda <- seq(0.8, 0.99, length.out = 20)

#Fixed effects in the model
form2 <- log(CMEDV) ~ CRIM + ZN + INDUS + CHAS

#Fit conditioned models (in parallel!!)
lerouxmodels <- mclapply(rlambda,
   function(lambda) {
      leroux.inla(form2, d = as.data.frame(boston),
      W = bmspB, lambda = lambda,
      ...
      )
   })

#BMA with the previous models
bmaleroux <- INLABMA(lerouxmodels, rlambda, 0)
```

Fitting Leroux et al.'s model



New `slm` Latent Class for Spatial Econometrics Models

- Spatial Econometrics models have been added to **R-INLA**
- **R-INLA** includes now a new latent effect:

$$\mathbf{x} = (I_n - \rho W)^{-1}(X\beta + e)$$

- W is a row-standardised adjacency matrix
- ρ is a spatial autocorrelation parameter
- X is a matrix of covariates, with coefficients β
- e are Gaussian i.i.d. errors with variance σ^2
- **SEM**

$$y = X\beta + (I - \rho W)^{-1}(0 + e);\ \ e \sim N(0, \sigma^2 I)$$

- **SLM**

$$y = (I - \rho W)^{-1}(X\beta + e);\ \ e \sim N(0, \sigma^2 I)$$