Spatial and Spatio-temporal Epidemiology 4 (2013) 33–49

Contents lists available at SciVerse ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste

Review Article

Spatial and spatio-temporal models with R-INLA

Marta Blangiardo [a,*,1], Michela Cameletti [b,1], Gianluca Baio [c,d], Håvard Rue [e]

[a] MRC-HPA Centre for Environment and Health, Department of Epidemiology and Biostatistics, Imperial College, London, UK
[b] Department of Management, Economics and Quantitative Methods, University of Bergamo, Italy
[c] Department of Statistical Science, University College London, London, UK
[d] Department of Statistics and Quantitative Methods, University of Milano Bicocca, Italy
[e] Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

ARTICLE INFO

Article history:
Received 11 October 2012
Revised 28 November 2012
Accepted 5 December 2012
Available online 2 January 2013

Keywords:
Integrated Nested Laplace Approximation
Stochastic Partial Differential Equation approach
Bayesian approach
Area-level data
Point-level data

ABSTRACT

During the last three decades, Bayesian methods have developed greatly in the field of epidemiology. Their main challenge focusses around computation, but the advent of Markov Chain Monte Carlo methods (MCMC) and in particular of the WinBUGS software has opened the doors of Bayesian modelling to the wide research community. However model complexity and database dimension still remain a constraint.

Recently the use of Gaussian random fields has become increasingly popular in epidemiology as very often epidemiological data are characterised by a spatial and/or temporal structure which needs to be taken into account in the inferential process. The Integrated Nested Laplace Approximation (INLA) approach has been developed as a computationally efficient alternative to MCMC and the availability of an R package (R-INLA) allows researchers to easily apply this method.

In this paper we review the INLA approach and present some applications on spatial and spatio-temporal data.

Contents

1. Introduction

During the last three decades, Bayesian methods have developed greatly and are now widely established in many research areas, from clinical trials (Berry et al., 2011), to health economic assessment (Baio, 2012) to the social

[*] Corresponding author. Tel.: +44 (0)207 594 3309.
E-mail address: m.blangiardo@imperial.ac.uk (M. Blangiardo).
[1] Joint first authors.

1877-5845/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.sste.2012.12.001

sciences (Jackman, 2009), to epidemiology (Greenland, 2006).

The basic idea behind the Bayesian approach is that effectively only one form of uncertainty exists, which is described by suitable probability distributions. Thus, there is no fundamental distinction between observable data or unobservable parameters, which are also considered as random quantities. The uncertainty about the realised value of the parameters given the current state of information (i.e. before observing any new data) is described by a *prior* distribution. The inferential process combines the prior and the (current) data model to derive the *posterior* distribution, which is typically, but not necessarily, the objective of the inference (Bernardo and Smith, 2000; Lindley, 2006).

There are several advantages to the Bayesian approach: for instance the specification of prior distributions allows the formal inclusion of information that can be obtained through previous studies or from expert opinion; the (posterior) probability that a parameter does/does not exceed a certain threshold is easily obtained from the posterior distribution, providing a more intuitive and interpretable quantity than a frequentist *p*-value. In addition, within the Bayesian approach, it is easy to specify a hierarchical structure on the data and/or parameters, which presents the added benefit of making prediction for new observations and missing data imputation relatively straightforward.

Epidemiological data, e.g. in terms of an outcome and one or more risk factors or confounders, are often characterised by a spatial and/or temporal structure which needs to be taken into account in the inferential process. Under these circumstances, the Bayesian approach is generally particularly effective (Dunson, 2001) and has been applied in several epidemiological applications, from ecology (Clark, 2005) to environmental studies (Wikle, 2003; Clark and Gelfand, 2006), to infectious disease (Jewell et al., 2009). For example, if the data consist of aggregated counts of outcomes and covariates, typically disease mapping and/or ecological regression can be specified (Lawson, 2009). Alternatively, if the outcome or risk factors data are observed at point locations, then geostatistical models are considered as suitable representations of the problem (Diggle and Ribeiro, 2007).

Both models can be specified in a Bayesian framework by simply extending the concept of hierarchical structure, allowing to account for similarities based on the neighbourhood or on the distance, for area-level or point-reference data, respectively. However, particularly in these cases, the main challenge in Bayesian statistics resides in the computational aspects. Markov Chain Monte Carlo (MCMC) methods (Brooks et al., 2011; Robert and Casella, 2004), are normally used for Bayesian computation, arguably thanks to the wide popularity of the `BUGS` software (Lunn et al., 2009, 2012). While extremely flexible and able to deal with virtually any type of data and model, in all but trivial cases MCMC methods involve computationally- and time-intensive simulations to obtain the posterior distribution for the parameters. Consequently, the complexity of the model and the database dimension often remain fundamental issues.

The Integrated Nested Laplace Approximation (INLA; Rue et al., 2009) approach has been recently developed as a computationally efficient alternative to MCMC. INLA is designed for *latent Gaussian models*, a very wide and flexible class of models ranging from (generalized) linear mixed to spatial and spatio-temporal models. For this reason, INLA can be successfully used in a great variety of applications (e.g. Li et al., 2012; Riebler et al., 2012; Ruiz-Cárdenas et al., 2012; Martino et al., 2011; Roos and Held, 2011; Schrödle and Held, 2011a,b; Schrödle et al., 2011; Paul et al., 2010), also thanks to the availability of an `R` package named `R-INLA` (Martino and Rue, 2010). Furthermore, INLA can be combined with the Stochastic Partial Differential Equation (SPDE) approach proposed by Lindgren et al. (2011) in order to implement spatial and spatio-temporal models for point-reference data.

The objective of this paper is to present the basic features of the INLA approach as applied to spatial and spatio-temporal data. The paper is structured as follows: first in Section 2 we review the main characteristics of spatial and spatio-temporal data defined at the point and area level; then we provide an overview of the theory behind INLA in Section 3 and present two practical applications on area level data in Sections 3.2 and 3.3. After this in Section 4 we review the SPDE approach to deal with geostatistical data, and then present two practical applications on spatial and spatio-temporal point level data (Sections 4.1 and 4.2). Finally Section 5 discusses some of the issues and provides some conclusions.

2. Spatial and spatio-temporal data

Spatial data are defined as realisations of a stochastic process indexed by space

$$Y(s) \equiv \{y(s), s \in \mathcal{D}\}$$

where $\mathcal{D}$ is a (fixed) subset of $\mathbb{R}^d$ (here we consider $d = 2$). The actual data can be then represented by a collection of observations $\boldsymbol{y} = \{y(s_1), \ldots, y(s_n)\}$, where the set $(s_1, \ldots, s_n)$ indicates the spatial units at which the measurements are taken. Depending on $\mathcal{D}$ being a continuous surface or a countable collection of d-dimensional spatial units, the problem can be specified as a spatially continuous or discrete random process, respectively (Gelfand et al., 2010).

For example, we can consider a collection of air pollutant measurements obtained by monitors located in the set $(s_1, \ldots, s_n)$ of *n points*. In this case, $\boldsymbol{y}$ is a realisation of the air pollution process that changes continuously in space and we usually refer to it as geostatistical or point-reference data. Alternatively, we may be interested in studying the spatial pattern of a certain health condition observed in a set $(s_1, \ldots, s_n)$ of *n areas* (rather than points), defined for example by census tracts or counties; in this case, $\boldsymbol{y}$ may represent a suitable summary, e.g. the number of cases observed in each area.

The first step in defining a spatial model within the Bayesian framework is to identify a probability distribution for the observed data. Usually we select a distribution