Environmental Modelling & Software 24 (2009) 1027–1035

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

The EM algorithm in a distributed computing environment for modelling environmental space–time data

Alessandro Fassò*, Michela Cameletti

University of Bergamo, Department of Information Technology and Mathematical Methods, Viale Marconi n. 5, 24044 Dalmine (BG), Bergamo, Italy

ARTICLE INFO

Article history:
Received 20 May 2008
Received in revised form
16 February 2009
Accepted 18 February 2009
Available online 21 March 2009

Keywords:
Separability
Spatial covariance
Kalman filter
Environmental statistics
Distributed computing

ABSTRACT

Statistical models for spatio-temporal data are increasingly used in environmetrics, climate change, epidemiology, remote sensing and dynamical risk mapping. Due to the complexity of the relationships among the involved variables and dimensionality of the parameter set to be estimated, techniques for model definition and estimation which can be worked out stepwise are welcome. In this context, hierarchical models are a suitable solution since they make it possible to define the joint dynamics and the full likelihood starting from simpler conditional submodels. Moreover, for a large class of hierarchical models, the maximum likelihood estimation procedure can be simplified using the Expectation–Maximization (EM) algorithm.

In this paper, we define the EM algorithm for a rather general three-stage spatio-temporal hierarchical model, which includes also spatio-temporal covariates. In particular, we show that most of the parameters are updated using closed forms and this guarantees stability of the algorithm unlike the classical optimization techniques of the Newton–Raphson type for maximizing the full likelihood function. Moreover, we illustrate how the EM algorithm can be combined with a spatio-temporal parametric bootstrap for evaluating the parameter accuracy through standard errors and non-Gaussian confidence intervals.

To do this a new software library in form of a standard R package has been developed. Moreover, realistic simulations on a distributed computing environment allow us to discuss the algorithm properties and performance also in terms of convergence iterations and computing times.

Software availability

Name: R package Stem
Developer: Michela Cameletti
E-mail: michela.cameletti@unibg.it
Software required: R
Availability: downloadable from: http://cran.r-project.org/web/packages/Stem/index.html

1. Introduction

Statistical modelling of spatio-temporal data has to take into account various sources of variability and correlation arising from time at various frequencies, from space at various scales, their interaction and other covariates which may be purely spatial quantities or pure time-series without a spatial dimension, or even dynamical fields on space and time. Hierarchical models for spatio-temporal process can cope with this complexity in a straightforward and flexible way. For this reason they are receiving more and more attention from both the Bayesian and frequentist point of view (see, for example, Wikle et al. (1998), Wikle (2003) and Clark and Gelfand (2006)), the latter being the approach adopted in this paper.

A hierarchical model can be constructed by putting together conditional submodels which are defined hierarchically at different levels. At the first level the observation variability is modelled by the so-called measurement equation, which is essentially given by a signal plus an error. In the classical approach the true signal or trend is a deterministic function; here, for the sake of flexibility, the trend is a stochastic process which is defined at the subsequent levels of the hierarchy, where the inherent complex dynamics is split into sub-dynamics which, in turn, are modelled hierarchically.

In addition to flexibility, a second advantage of this approach is that we can apportion the total uncertainty to the various components or levels. Moreover, from the likelihood point of view, this corresponds to taking a conditional viewpoint for which the joint probability distribution of a spatio-temporal process can be expressed as the product of some simpler conditional distributions defined at each hierarchical stage.

* Corresponding author. Tel.: +39 035 2052323; fax: +39 035 2052310.
E-mail address: alessandro.fasso@unibg.it (A. Fassò).

1364-8152/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2009.02.009

When the spatio-temporal covariance function satisfies the so-called separability property, these models can be easily represented in state-space form. Hence Kalman filtering and smoothing techniques can be used for reconstructing the temporal component of the unobserved trend (Wikle and Cressie, 1999). For example in environmental statistics, Brown et al. (2001) consider the calibration of radar rainfall data by means of a ground-truth monitoring network and Fassò et al. (2007b) study airborne particulate matter and the calibration of a heterogeneous monitoring network.

Moreover, a separable hierarchical model easily provides a spatial estimator of the Kriging type (Cressie, 1993, Chapter 3) so that a spatio-temporal process, together with its uncertainty, can be mapped in time. For example, Stroud et al. (2001), Sahu et al. (2007), Fassò et al. (2007a), Fassò and Cameletti (in press) propose mapping methods for spatio-temporal data, such as rainfall, tropospheric ozone or airborne particulate matters, which are continuous in space and measured by a monitoring network irregularly distributed in the considered areas.

The Expectation–Maximization (EM) algorithm has been originally proposed for maximum likelihood estimation in presence of structural missing data, see e.g. McLachlan and Krishnan (1997). In spatio-temporal modelling, the EM has been recently used by Xu and Wikle (2007) for estimating certain parameterizations and by Amisigo and Van De Giesen (2005) for the concurrent estimation of model parameters and missing data in river runoff series.

In this paper we propose EM estimation and bootstrap uncertainty assessment for a separable hierarchical spatio-temporal model which generalizes Xu and Wikle (2007) and Amisigo and Van De Giesen (2005) as it covers the case of spatio-temporal covariates. This model class is used for air quality applications in Fassò et al. (2007a) and Fassò and Cameletti (in press), which consider also dynamical mapping and introduce some sensitivity analysis techniques for assessing the mapping performance and understanding the model components. In this framework, using the state-space representation, it is easily seen that temporal prediction is an immediate consequence of Kalman filtering for this model, see e.g. Durbin and Koopman (2001).

The rest of the paper is organized as follows. In Section 2, the above separable spatio-temporal model with covariates is formally introduced.

In Section 3, the EM algorithm is discussed extensively. In particular, we show that the maximization step is based on closed-form formulas for all the parameters except for the spatial covariance ones, which are obtained by the Newton–Raphson (NR) algorithm. Hence, we avoid the inversion of the large Hessian matrix which would arise in performing numerical maximization of the full likelihood.

In Section 4, the spatio-temporal parametric bootstrap is introduced for computing standard errors of the parameter estimates and their confidence intervals. This method turns out to be particularly useful for assessing estimate accuracy, especially in our case which is characterized by asymmetric estimate distributions.

Section 5 is devoted to a simulation study that discusses the performances of the EM algorithm in terms of estimate precision and computing time. This is done using realistic data which are generated on the basis of the airborne particulate matter data set discussed by Cameletti (2007) , Fassò et al. (2007a) and Fassò and Cameletti (in press). In particular, Section 5.1 focuses on the implementation issues with special reference to `R` software and the distributed computing environment while the discussion of the results is provided in Sections 5.2 and 5.3.

The conclusions are drawn in Section 6, while the paper ends with Appendixes A and B which contain computational details regarding EM and NR algorithm.

2. The spatio-temporal model

Let $Z(s,t)$ be the observed scalar spatio-temporal process at time t and geographical location s. Let $Z_t = \{Z(s_1,t),\dots,Z(s_n,t)\}$[1] be the network data at time t and at n geographical locations $s_1, \dots, s_n$. Moreover let $Y_t = \{Y_1(t),\dots,Y_p(t)\}$ be a p-dimensional vector for the unobserved temporal process at time t with $p \le n$. The three-stage hierarchical model is defined by the following equations for $t = 1, \dots, T$

$$Z_t = U_t + \varepsilon_t \tag{1}$$

$$U_t = X_t\beta + KY_t + \omega_t \tag{2}$$

$$Y_t = GY_{t-1} + \eta_t \tag{3}$$

In equation (1) the measurement error ε_t is introduced so that U_t can be seen as a smoothed version of the spatio-temporal process Z_t. In the second stage the unobserved spatio-temporal process U_t is defined as the sum of three components: a function of the $(n \times d)$-dimensional matrix X_t of d covariates observed at time t at the n locations, the latent space-constant temporal process Y_t and the model error ω_t. It should be noted that the $(n \times p)$-dimensional matrix K is known and accounts for the weights of the p components of Y_t for each spatial location s_i, $i = 1, \dots, n$. A common choice for K is given by the loadings of a principal component decomposition (see Fassò et al. (2007b) and Wikle and Cressie (1999)). Then in equation (3), the temporal dynamics of Y_t is modelled as a p-dimensional autoregressive process where G is the transition matrix and η_t is the innovation error.

The three error components, namely ε_t, η_t and ω_t, are zero mean and independent over time as well as mutually independent. In particular, the pure measurement error ε_t is a Gaussian white noise process with variance and covariance matrix given by $\sigma_\varepsilon^2 I_n$, where I_n is a n-dimensional identity matrix. The measurement instrument precision σ_ε^2 is supposed constant over space and time as it is the case of a homogeneous network. The case of different instruments belonging to a heterogeneous network is discussed in Fassò et al. (2007b). The innovation η_t of equation (3) is a p-dimensional Gaussian white noise process with variance–covariance matrix Σ_η. Finally, the pure spatial component ω_t of equation (2) is a n-dimensional Gaussian spatial process. It is uncorrelated with ε_t and η_t for each t and its variance–covariance matrix is given by a time-constant spatial covariance function

$$\mathrm{Cov}[\omega(s,t),\omega(s',t)] = \sigma_\omega^2 C_\theta(h)$$

where $h = \|s - s'\|$ is the Euclidean distance between sites s and s'. As the covariance function depends only on h, the spatial process is second-order stationary and isotropic. Moreover, the function $C_\theta(h)$ depends on the parameter θ to be estimated and is continuous at $h = 0$ with $\lim_{h\to 0} C_\theta(h) = 1$. A simple example of covariance function is the exponential which is given by

$$C_\theta(h) = \exp(-\theta h) \tag{4}$$

Other covariance functions defining isotropic second-order stationary spatial processes are discussed, for example, in Banerjee et al. (2004, Chapter 1).

Substitution of equation (2) into equation (1) yields the following two-stage hierarchical model

[1] Here and in the sequel, braces are used for column stacking the vectors involved. Brackets will be used for row stacking instead.