Spatial Dynamic Factor Models with environmental applications

Pasquale Valentini - Luigi Ippoliti
University G. d'Annunzio (Italy), pvalent@unich.it

Dani Gamerman
Universidade Federal do Rio de Janeiro (Brasil)

Abstract: This article is concerned with a dynamic factor model for spatio-temporal environmental variables. The model is proposed in a state-space formulation which, through the Kalman recursions, allows a unified approach to prediction and estimation. Full probabilistic inference for the model parameters is facilitated by adapting standard Markov chain Monte Carlo (MCMC) algorithms for dynamic linear models to our model formulation.

Keywords: Dynamic factor models, Spatio-temporal models

1 Introduction

In recent years, spatio-temporal models have received widespread popularity and have been largely developed through applications in environmental sciences. In fact, the European Environmental Agency and the US Environmental Protection Agency have both devoted significant efforts to developing air quality models for the assessment of air pollution issues and evaluation of feasible solutions. We note nevertheless that there is no single approach which can be considered uniformly as being the most appropriate for a specific problem.

In this paper, we propose a latent regression model which is useful for spatial and temporal predictions of pollutants of interest. The model is developed in a state-space representation which represents a powerful way to provide full probabilistic inference for the model parameters, interpolation and forecast of the variable of interest. To account for spatial interpolation, the spatial dependence is incorporated in the measurement matrix and we describe its construction by discussing a stochastic specification. The possibility of specifying two measurement equations leads to a significant advantage in terms of spatial interpolation and this makes an important difference with respect to other spatio-temporal models proposed in literature. A further important property of the proposed model is that it leads to capture the temporal variation of the multivariate space-time fields.

2 The General Model

Consider the multivariate spatio-temporal processes $\mathbf{X}(\mathbf{s},t) = [X_1(\mathbf{s},t),\ldots,X_{n_x}(\mathbf{s},t)]'$ and $\mathbf{Y}(\mathbf{s},t) = [Y_1(\mathbf{s},t),\ldots,Y_{n_y}(\mathbf{s},t)]'$, where $\mathbf{s} \in S$, with S a some spatial domain and $t \in \{1,2,\ldots\}$ a discrete index of times. For geostatistical data, S is a given subset of $\Re^d$ and s is assumed to vary continuously throughout S . For lattice data, S is assumed to be a given finite or countable collection of points. Lattices may be either regular, as on a grid, or irregular, such as zip codes, census divisions.
It is explicitly assumed that $\mathbf{X}$ is a predictor of $\mathbf{Y}$. Hence, $\mathbf{Y}$ denotes the specific multivariate process of interest to be predicted in time and/or space. Here, the relationship between the multivariate processes is modelled through the structural spatial dynamic factor (SSDF) model.
Let us define the multivariate spatial processes as $\mathbf{Y}(t) = [\mathbf{Y}(\mathbf{s}_1,t)',\ldots,\mathbf{Y}(\mathbf{s}_N,t)']'$, a $\tilde{n}_y \times 1$ vector ($\tilde{n}_y = n_y N$) at N locations for n_y variables, and $\mathbf{X}(t) = [\mathbf{X}(\mathbf{s}_1,t)',\ldots,\mathbf{X}(\mathbf{s}_N,t)']'$, a $\tilde{n}_x \times 1$ vector ($\tilde{n}_x = n_x N$) at N locations for n_x variables.
The measurement equations of the SSDF model are

$$\begin{aligned} \mathbf{X}(t) &= \mathbf{m}_x(t) + \mathbf{H}_x \mathbf{f}(t) + \mathbf{u}_x(t) && (1) \\ \mathbf{Y}(t) &= \mathbf{m}_y(t) + \mathbf{H}_y \mathbf{g}(t) + \mathbf{u}_y(t) && (2) \end{aligned}$$

where $\mathbf{m}_y(t)$ and $\mathbf{m}_x(t)$ are $\tilde{n}_y \times 1$ and $\tilde{n}_x \times 1$ mean components modelling the smooth large-scale temporal variability, $\mathbf{H}_y$ $(\tilde{n}_y \times m)$ and $\mathbf{H}_x$ $(\tilde{n}_x \times l)$ are measurement matrices giving information on the spatial structure of the random fields, and $\mathbf{u}_x(t) \sim N(\mathbf{0}, \mathbf{\Sigma}_{u_x})$ and $\mathbf{u}_y(t) \sim N(\mathbf{0}, \mathbf{\Sigma}_{u_y})$. Throughout the paper it is assumed that $m \ll \tilde{n}_y$ and $l \ll \tilde{n}_x$.
The temporal dynamic of the processes is modelled through the following state equations:

$$\begin{aligned} \mathbf{g}(t) &= \sum_{i=1}^{p} \mathbf{B}_i \mathbf{g}(t-i) + \sum_{j=1}^{q} \mathbf{C}_j \mathbf{f}(t-j) + \boldsymbol{\xi}(t) && (3) \\ \mathbf{f}(t) &= \sum_{k=1}^{s} \mathbf{R}_k \mathbf{f}(t-k) + \boldsymbol{\eta}(t) && (4) \end{aligned}$$

where $\mathbf{C}_i$ $(m \times m)$, $\mathbf{D}_j$ $(m \times l)$, and $\mathbf{R}_k$ $(l \times l)$ are coefficient matrices modelling the temporal evolution of the latent vectors $\mathbf{g}(t) = [g_1(t),\ldots,g_m(t)]'$ and $\mathbf{f}(t) = [f_1(t),\ldots,f_l(t)]'$, respectively. Finally, $\boldsymbol{\xi}(t)$ and $\boldsymbol{\eta}(t)$ are independent Gaussian error terms for which we assume, $\boldsymbol{\xi}(t) \sim N(\mathbf{0}, \mathbf{\Sigma}_\xi)$ and $\boldsymbol{\eta}(t) \sim N(\mathbf{0}, \mathbf{\Sigma}_\eta)$.
SSDF analysis may be used to identify possible clusters of locations whose temporal behaviour is primarily described by a potentially small set of common dynamic latent factors.

3 The Structural Spatial Dynamic Factor Model

It is customary for dynamic factor models to refer to the unobserved (state) processes as the common factors and to refer to the coefficients that link the factors with the observed series as the factor loadings. Because of their spatial nature, the factor loadings are equivalently defined as spatial patterns (Lopes et al., 2008; Ippoliti et al, 2010). The latent factors, $\mathbf{f}(t)$ and $\mathbf{g}(t)$, are able to capture the temporal variation of the multivariate space-time fields, and the spatial dependence can be modeled by the columns of the matrices $\mathbf{H}_y$ and $\mathbf{H}_x$ through multivariate Gaussian Random Field for geostatistical data, or through multivariate Markov random field (MRF) for lattice data.

4 The State Space Formulation

Given the SSDF model the temporal dynamic is modelled through state equations (3) and (4). The specification of equation (4) is necessary to predict in time the latent process $\mathbf{f}(t)$ and thus to obtain $k-$step ahead forecasts of $\mathbf{g}(t)$ through equation (3). The joint generation process of $\mathbf{g}(t)$ and $\mathbf{f}(t)$ is a VAR(p) process of the type

$$\mathbf{d}(t) \quad = \quad \mathbf{\Phi}_1\mathbf{d}(t-1) + \ldots + \mathbf{\Phi}_p\mathbf{d}(t-p) + \boldsymbol{\epsilon}(t) \tag{5}$$

where

$$\mathbf{d}(t) = \left[\begin{array}{c}\mathbf{g}(t)\\ \mathbf{f}(t)\end{array}\right], \quad \mathbf{\Phi}_i = \left[\begin{array}{cc}\mathbf{C}_i & \mathbf{D}_i\\ \mathbf{0} & \mathbf{R}_i\end{array}\right], \quad \boldsymbol{\epsilon}(t) = \left[\begin{array}{c}\boldsymbol{\xi}(t)\\ \boldsymbol{\eta}(t)\end{array}\right] \text{ and } p \geq max(s,q).$$

The presence of the measurement and the state variables naturally leads to the state-space representation of the SSDF model.

5 Nonstationary case

In the case in which the two spatio-temporal processes $X(\mathbf{s};t)$ and $Y(\mathbf{s};t)$ are not stationary in time, we assume that factors are generated by cointegrated vector autoregressive processes. In this case the factors are represented by the error correction specification of the vector autoregressive process of equation (5):

$$\Delta\mathbf{d}(t) \quad = \quad \tilde{\mathbf{A}}\mathbf{d}(t-1) + \sum_{i=1}^{p-1}\tilde{\mathbf{\Phi}}_i\Delta\mathbf{d}(t-i) + \boldsymbol{\epsilon}(t) \tag{6}$$

where $\tilde{\mathbf{A}} = -\mathbf{I} + \sum_{i=1}^{p}\mathbf{\Phi}_i$, $\tilde{\mathbf{\Phi}}_i = -\sum_{j=i+1}^{p}\mathbf{\Phi}_l$, and Δ is the difference operator (i.e. $\Delta\mathbf{d}(t) = \mathbf{d}(t) - \mathbf{d}(t-1)$).

Let $\mathbf{\Phi}(z)$ denote the characteristic polynomial associated with the process (6). We assume that latent exogenous variables are cointegrated with cointegrating rank r_f

(Cho, 2010) and also $rank(\tilde{\mathbf{A}}) = r$, $r = m + l - c > r_f$ with $m + l > c$ and c are the unit roots of $Det(\mathbf{\Phi}(z))$.
Because of the exogeneity of X, the matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{\Phi}}_i$ are upper block triangular matrices: $\tilde{\mathbf{A}} = \begin{bmatrix} \tilde{\mathbf{A}}_1 & \tilde{\mathbf{A}}_{12} \\ \mathbf{0} & \tilde{\mathbf{A}}_2 \end{bmatrix}$ and $\tilde{\mathbf{\Phi}}_i = \begin{bmatrix} \tilde{\mathbf{\Phi}}_{1i} & \tilde{\mathbf{\Phi}}_{12i} \\ \mathbf{0} & \tilde{\mathbf{\Phi}}_{2i} \end{bmatrix}$.
Then, equation (6) can be rewritten in the following two equations:

$$\begin{aligned} \Delta\mathbf{g}(t) &= \mathbf{A}\mathbf{B}'\mathbf{d}(t-1) + \mathbf{A}_{2f}\mathbf{B}'_f\mathbf{f}(t-1) + \sum_{i=1}^{p-1} \mathbf{K}_i \Delta\mathbf{d}(t-i) + \boldsymbol{\xi}(t) && (7) \\ \Delta\mathbf{f}(t) &= \mathbf{A}_f\mathbf{B}'_f\mathbf{f}(t-1) + \sum_{i=1}^{p-1} \tilde{\mathbf{\Phi}}_{2i}\Delta\mathbf{f}(t-j) + \boldsymbol{\eta}(t) && (8) \end{aligned}$$

where $\mathbf{A}$ is $m \times r_d$, $\mathbf{B}$ is $(m+l) \times r_d$, $\mathbf{A}_f$ and $\mathbf{B}_f$ are $l \times r_f$, $\mathbf{A}_{2f}$ is $m \times r_f$, $\mathbf{K}_i = [\tilde{\mathbf{\Phi}}_{1i} \;\; \tilde{\mathbf{\Phi}}_{12i}]$ and $r_d \leq m + l$.

6 Inference

Full probabilistic inference for the model parameters is carried out by elicitating the independent prior distributions. Posterior inference for the proposed class of spatial dynamic factor models is facilitated by MCMC algorithms. The common factors are jointly sampled by means of the well known forward filtering backward sampling (FFBS) algorithm (Carter and Kohn 1994) which exploits the state space representation of the general model. All other full conditional distributions are "standard" multivariate normal distributions or gamma distributions. An exception is for the spatial parameters and the covariance matrices which are sampled using a Metropolis-Hastings step.

References

Carter C. K., Kohn R. (1994) On Gibbs sampling for state space models. *Biometrika*, 81, 541-553.

Cho S. (2010), Inference of cointegrated model with exogenous variables, *SIRFE Working Paper* 10-A04, Seoul National University.

Ippoliti L., Valentini P., Gamerman D. (2010) Space-time modelling of coupled spatio-temporal environmental variables, *Technical Report* N. 229, DME/IM-UFRJ.

Lopes H. F., Salazar E., Gamerman D. (2008) Spatial dynamic factor analysis. *Bayesian Analysis*, 3, 759-792.