Regression Analysis in a Data Rich Environment

L. Fontanella, L. Ippoliti*, A. Sarra and P. Valentini

Department of Economics, University G. d'Annunzio, Chieti-Pescara, Italy
lfontan@unich.it, ippoliti@unich.it, asarra@unich.it, pvalent@unich.it
**Corresponding author*

Abstract. *This paper discusses a number of conceptual issues pertaining to the study of the relationships existing between two groups of dependent and regressor variables which are supposed to be spatially and temporally correlated. Since it is assumed that this relationships can be studied in a reduced latent space, we propose a factor analytic approach and provide an overview of the motivations for including spatial effects in dynamic factor models. Considerable attention is paid to the inferential framework necessary to carry out estimation and to the different assumptions, constraints and implications embedded in the various model specifications. The discussion combines insights from the traditional (spatial) econometrics literature as well as from geostatistics and image analysis.*

Keywords. *Dynamic factor models; Spatio-temporal processes; Geostatistics; Bayesian inference; Spatial econometrics*

1 Introduction

The idea of borrowing information from different but related sources can be very powerful for statistical analysis. It proved to be very useful in the last decades where complex data structures began to be tackled, as they required sophisticated modeling strategies. In this paper we consider the problem of modeling high-dimensional multivariate, spatially and temporally referenced data. This problem has enjoyed widespread popularity in the last years and requires the definition of general and flexible statistical models where the temporal and cross-sectional dependencies must be accommodated.
Spurred by recent advances in Geo-spatial data acquisition technologies, it is often desirable for these data to examine the relationships existing between one or more dependent variables and some other linked covariates. This can be achieved in a number of ways, though there might be no single approach which can be considered uniformly as being the most appropriate.
Among the different methodologies proposed in the literature, dynamic factor models (DFMs, Molenaar, 1985) have grown significantly in popularity and have been shown to be very useful for exploratory analysis, policy analysis and forecasting in a data rich environment. DFMs have been widely developed in both methodological and practical issues, and have become a standard tool for increasingly high-dimensional modeling of time series. In this work, we devote special attention to the use of a dynamic factor analytic approach in the framework of spatial statistics. It will be shown that this is not only an important area of application but also that this area can receive several benefits from this modeling approach.

Through a fully Bayesian approach, we contribute to the recent literature by melding together dynamic factor models, spatial regression models and geostatistical techniques, in order to explain the multifactorial nature of many spatio-temporal data. We assume that the relationships existing between the groups of dependent and regressor variables can be studied through a temporally dynamic and spatially descriptive model, hereafter referred to as *spatial dynamic structural equation* model (SD-SEM).

2 The spatial dynamic structural equation model

With increased collection of multivariate spatial data, there arises the need for flexible explanatory stochastic models in order to improve estimation precision and to provide simple descriptions of the complex relationships existing among the variables. In the following, we discuss a model formulation which describes the structural relations among the variables in a lower dimensional space.
Assume initially that $\mathbf{Y}$ and $\mathbf{X}$ are two multivariate Gaussian spatio-temporal processes observed at temporal instants $t \in \{1,2,\ldots\}$ and generic locations, $\mathbf{s} \in \mathcal{D}_y$ and $\mathbf{u} \in \mathcal{D}_x$, respectively. For the two different processes, the spatial sites $\mathbf{s}$ and $\mathbf{u}$ can denote the same location but, in general, they need not be the same. Furthermore, both $\mathcal{D}_y$ and $\mathcal{D}_x$ may represent two different spatial domains of interest.
Let n_y be the number of observed variables for $\mathbf{Y}$ and n_x the number of observed variables for $\mathbf{X}$. The most informative case is represented by the isotopic configuration where, for each multivariate process, $\mathbf{Y}$ or $\mathbf{X}$, all variables are measured at all their respective sites. In this case, we can thus write $\mathbf{Y}(\mathbf{s},t) = [Y_1(\mathbf{s},t),\ldots,Y_{n_y}(\mathbf{s},t)]'$ and $\mathbf{X}(\mathbf{u},t) = [X_1(\mathbf{u},t),\ldots,X_{n_x}(\mathbf{u},t)]'$. The opposite case is the completely heterotopic case where not all the variables can be observed at the same site.
Without loss of generality, for the sake of simplicity, we describe here the isotopic case and assume that $\mathbf{Y}$ or $\mathbf{X}$ can be observed at N_y and N_x spatial sites, respectively. Let $\tilde{n}_y = n_y N_y$ and $\tilde{n}_x = n_x N_x$. Then, at a specific time t, the $(\tilde{n}_y \times 1)$ and $(\tilde{n}_x \times 1)$ dimensional spatial processes, $\mathbf{Y}$ and $\mathbf{X}$, are denoted as $\mathbf{Y}(t) = [\mathbf{Y}(\mathbf{s}_1,t)',\ldots,\mathbf{Y}(\mathbf{s}_{N_y},t)']'$ and $\mathbf{X}(t) = [\mathbf{X}(\mathbf{u}_1,t)',\ldots,\mathbf{X}(\mathbf{u}_{N_x},t)']'$. Since it is also assumed that $\mathbf{X}$ is a predictor of $\mathbf{Y}$, which is thus the process of interest, we work within the well known framework of transfer response models covered in many standard time series books.
Our model assumes that each multivariate spatial process, at a specific time t, has the following linear structure

$$\mathbf{X}(t) = \mathbf{m}_x(t) + \mathbf{H}_x\mathbf{f}(t) + \mathbf{u}_x(t) \quad (1)$$

$$\mathbf{Y}(t) = \mathbf{m}_y(t) + \mathbf{H}_y\mathbf{g}(t) + \mathbf{u}_y(t) \quad (2)$$

where $\mathbf{m}_y(t)$ and $\mathbf{m}_x(t)$ are $(\tilde{n}_y \times 1)$ and $(\tilde{n}_x \times 1)$ mean components modeling the smooth large-scale temporal variability, $\mathbf{H}_y$ and $\mathbf{H}_x$ are measurement (factor loadings) matrices of dimensions $(\tilde{n}_y \times m)$ and $(\tilde{n}_x \times l)$, respectively, and $\mathbf{g}(t)$ and $\mathbf{f}(t)$ are m- and l-dimensional vectors of temporal common factors. Also, $\mathbf{u}_y(t)$ and $\mathbf{u}_x(t)$ are Gaussian error terms for which we assume $\mathbf{u}_y(t) \sim N(\mathbf{0}, \mathbf{\Sigma}_{u_y})$ and $\mathbf{u}_x(t) \sim N(\mathbf{0}, \mathbf{\Sigma}_{u_x})$. For simplicity, throughout the paper it is assumed that $\mathbf{\Sigma}_{u_y}$ and $\mathbf{\Sigma}_{u_x}$ are both diagonal matrices and that $m \ll \tilde{n}_y$ and $l \ll \tilde{n}_x$.
In the second level of the hierarchy the temporal dynamics of the common factors are then modeled through the following equations

$$\mathbf{g}(t) = \sum_{i=1}^{p} \mathbf{C}_i\mathbf{g}(t-i) + \sum_{j=0}^{q} \mathbf{D}_j\mathbf{f}(t-j) + \boldsymbol{\xi}(t) \quad (3)$$

$$\mathbf{f}(t) = \sum_{k=1}^{s} \mathbf{R}_k\mathbf{f}(t-k) + \boldsymbol{\eta}(t) \quad (4)$$

where $\mathbf{C}_i$ $(m \times m)$, $\mathbf{D}_j$ $(m \times l)$, and $\mathbf{R}_k$ $(l \times l)$ are coefficient matrices modeling 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 $\boldsymbol{\xi}(t) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\xi)$ and $\boldsymbol{\eta}(t) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$. Equation (3) represents the structural part of the model formulation and appears as a VAR model in which the variables in $\mathbf{g}(t)$, are controlled for the effects of other variables in $\mathbf{f}(t)$. Equations (1-4) thus provide the basic formulation of the SD-SEM.

Model completion requires specific forms for patterns of time-variation in the elements of $\mathbf{m}_y(t)$ and $\mathbf{m}_x(t)$, and for the elements of the factor loadings matrices $\mathbf{H}_y$ and $\mathbf{H}_x$. We note that the pure dynamic factor model is the special case given by equation (2), with $\mathbf{m}_y(t) = \mathbf{0}$, and equation (3) with $\mathbf{D}_j = \mathbf{0}$ for all j, when the typical assumption that the common factors $\mathbf{g}(t)$ are zero-mean and independent over time yields a linear factor representation of the conditional variance matrix of $\mathbf{Y}(t)$. When $\mathbf{m}_y(t) = \mathbf{A}_y \mathbf{z}(t)$, where $\mathbf{A}_y$ is a $(\tilde{n}_y \times r)$ matrix of regression coefficients and $\mathbf{z}(t)$ is a r-vector of observed variables, the model is what has become known as a factor-augmented vector autoregression (FAVAR). Of course, a pure spatial factor model also represents a specific case of the SD-SEM formulation and it is useful for modelling cross-sectional data.

3 Inference

Since the model is developed under the Bayesian paradigm, an important part of the model specification first deals with the choice of prior distributions.

For the factor loadings, let $\mathbf{h}_{y_j} = \left[\mathbf{h}_{y_j}(\mathbf{s}_1)', \mathbf{h}_{y_j}(\mathbf{s}_2)', \ldots, \mathbf{h}_{y_j}(\mathbf{s}_{N_y})'\right]'$ denote the j-th column of $\mathbf{H}_y$ and let $\mathbf{h}_{y_j}(\mathbf{s}_i) = \left[h_{y_j,1}(\mathbf{s}_i), \ldots, h_{y_j,n_y}(\mathbf{s}_i)\right]'$, denote the n_y-dimensional vector of loadings defined for the spatial site $\mathbf{s}_i$. With obvious notation, the same holds for $\mathbf{h}_{x_j}$ but, for simplicity, assume henceforth we only focus on the process of interest $\mathbf{Y}$. In general, taking care of the identifiability constraints (see Lopes and West, 2004), for the factor loadings one can take independent normal priors. However, for spatio-temporal data, one may be more interested in including conditional dependencies within the elements of $\mathbf{Y}(t)$. In fact, because of their spatial nature, the factor loadings are characterized by spatial patterns and consequently, a number of papers have discussed the possibility of inducing flexible correlation structures into the columns of $\mathbf{H}_y$. One possibility is to assume that the columns of the factor loadings matrix follow conditionally independent Gaussian Random fields, that is $\mathbf{h}_{yj} \sim MVN\left(\boldsymbol{\mu}_{h_{y_j}}, \boldsymbol{\Sigma}_{h_{y_j}}\right)$, where $\boldsymbol{\mu}_{h_{y_j}}$ is a mean vector and $\boldsymbol{\Sigma}_{h_{y_j}}$ a parametrized covariance matrix. For $n_y = 1$, a simple geostatistical process ($\mathcal{D}_y$ is spatially continuous) or a conditional autoregressive (CAR) specification ($\mathcal{D}_y$ denotes a regular or an irregular lattice) can be considered for $\mathbf{h}_{yj}$. On the other hand, for $n_y > 1$, the linear model of coregionalization or the multivariate CARs are useful choices for modelling the covariance structure.

By means of an appropriate concatenation of observables $\left[\mathbf{Y}(t)'\, \mathbf{X}(t)'\right]'$ and common factors $\left[\mathbf{g}(t)'\, \mathbf{f}(t)'\right]'$, equations (1-4) can be rewritten in state-space form by appropriately enlarging the state vector according to the order of lagged dependence of the latent factors. In this case, it can be shown (Ippoliti et al.,2012; Valentini et al., 2013) that the state equation assumes the form of a vector autoregressive (VAR) process and all the methods for Bayesian analyis for VARs can be used. Among others, Valentini et al. (2013) also discuss the use of a Stochastic Search Variable Selection (SSVS, George, Sun and Ni, 2008) prior for testing the cointegration structure *within* $\mathbf{g}(t)$ and $\mathbf{f}(t)$ and *between* the two processes.

Posterior inference for the proposed class of spatial dynamic factor models is facilitated by MCMC algorithms. Standard MCMC for dynamic linear models are adapted to our model specification such that posterior and predictive analysis are readily available. Conditional on r and m, the number of common factors, the MCMC scheme described in Lopes and West (2004) can be easily adapted where the common

factors are jointly sampled via the well known forward filtering backward sampling (FFBS) scheme. All other full conditional distributions are "standard" multivariate Gaussian or Gamma distributions. Exceptions refer to the spatial correlation parameters (range parameters and conditional autoregressive parameters under specific priors) of $\mathbf{h}_{y_j}$, or $\mathbf{h}_{x_j}$, which can be sampled using a Metropolis-Hastings step.

4 Discussion

The proposed model has an intuitive appeal and enjoys several advantages. First, our model formulation exploits the spatio-temporal nature of the data and explicitly defines a non-separable spatio-temporal covariance structure of the multivariate process. Second, since the data have a multivariate and multi-dimensional structure, in that several time series can be measured at specific spatial sites, the temporal relationships between dependent and regressor variables is modeled in a latent space. The observed processes are thus described by a potentially small set of common dynamic latent factors with the advantage of overcoming the difficulties of interpreting the relationships under study due to collinearity and low signal-to-noise ratio issues. Temporal forecasts of the variables of interest can also be obtained by only modeling the dynamics of a few common factors. Third, by modeling the spatial variation via spatially structured factor loadings, we entertain the possibility of identifying clusters of spatial sites that share common time series components. Through the spatial modeling of the factor loadings, spatial interpolations of the observed variables are also straightforward. Fourth, several general structures that make use of different covariate information, can be easily accommodated in the different levels of the hierarchy. Fifth, the SD-SEM offers a unified approach suitable to deal with variables and indicators measured at different scales and coming from different spatial sources. Hence, the model provides a simple solution to the misalignment problems which, for example, normally occurs in health care research. Lastly, the model specification is not limited to normally distributed variables, but it can be easily extended to handle more types of variables from an exponential family.
The ideas behind the SD-SEM will be discussed via illustrative examples with real data problems in a forthcoming extended work.

References

[1] George, E., Sun, D., Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics*, **142**, 553–580.

[2] Ippoliti, L., Valentini, P., Gamerman, D. (2012). Space-Time Modelling of Coupled Spatio-Temporal Environmental Variables. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **61**, 175–200.

[3] Lopes, H. F., West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, **14**, 41–67.

[4] Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, **50**, 181–202.

[5] Valentini, P., Ippoliti, L., Fontanella, L. (2013). Modeling US housing prices by spatial dynamic structural equation models. *The Annals of Applied Statistics*, **7**, 763–798.