

Hierarchical Spatial Factor Models for Poisson Count Data

Modelli Gerarchici per Dati Spaziali Multivariati Non Gaussiani

Marco Minozzo¹

Dipartimento di Scienze Statistiche
Università degli Studi di Perugia
minozzo@stat.unipg.it

Riassunto: Nel presente lavoro si affronta l'analisi di dati spaziali multivariati non gaussiani, con particolare riferimento al caso in cui le misure rilevate sono costituite da conteggi. Sull'orma delle attuali proposte presenti in letteratura per la modellazione di dati spaziali univariati, si propone di utilizzare modelli fattoriali gerarchici ("generalized linear mixed models") non Bayesiani. Per la stima dei parametri di questi modelli risultano adattabili alcune delle tradizionali tecniche geostatistiche, mentre mappe dei fattori spaziali latenti non osservabili si possono ottenere utilizzando algoritmi Markov chain Monte Carlo.

Keywords: Factorial kriging analysis, Generalized linear mixed models, Linear factor model, Markov chain Monte Carlo methods, Multivariate geostatistics.

1. Introduction

The present work is concerned with the analysis of non Gaussian multivariate spatial data, and in particular with the case in which the available measurements are counts. Let $y_i(\mathbf{x}_j)$, $i = 1, \dots, N$, $j = 1, \dots, n$, be a set of data resulting from the measurement of N regionalized variables y_i at n spatial locations $\mathbf{x}_j$. Following standard geostatistical practice, we see this set of N regionalized variables as one particular realization of a set of N random functions $Y_i(\mathbf{x})$, $i = 1, \dots, N$, $\mathbf{x} \in D \subseteq \mathbb{R}^2$. When the number N of random functions involved is high, it becomes paramount to explore feasible ways to model the N direct variograms and the $N(N-1)/2$ cross-variograms in terms of a small number of elementary components.

As far as only second order properties are concerned, the pioneering contribution to the modelling of a variogram function matrix is the work by Matheron (1982) which introduced a spatial factor model known as the linear model of coregionalization. More recent discussions of the model, and of the related "factorial kriging analysis", can be found in Goovaerts *et al.* (1993) and Wackernagel (1995). Similar spatial factor models have also been studied by Switzer and Green (1984) and, more recently, by Bailey and Krzanowski (2000).

Although these models, together with the use of factorial cokriging predictions, may be adequate in the case in which the data may be assumed to be (multivariate) Gaussian, in the other cases, particularly when count data are present, as in many ecological data sets, they can lead to very inefficient predictions of the underlying factors. However, whereas in the univariate case the problem has already been tackled by adopting a generalized linear mixed model approach (see, for instance, Diggle *et al.* (1998) and Cressie

¹This work was partially supported by M.U.R.S.T. and University of Perugia (1998) grants.

(2000)), in the multivariate case not much has yet been done. In Section 2 of this work we present a slight generalization of the classical geostatistical proportional covariance model (otherwise known as intrinsic correlation model), whereas in Section 3 we use this generalization to introduce a hierarchical spatial factor model for Poisson data. Finally, in Section 4 we make some discussions.

2. The Spatial Linear Factor Model

Under the hypothesis of joint second order stationarity for the $Y_i(\mathbf{x})$, let $E[Y_i(\mathbf{x})] = \mu_i$, $\text{Var}[Y_i(\mathbf{x})] = \sigma_i^2$ and $E[(Y_i(\mathbf{x}) - \mu_i)(Y_j(\mathbf{x} + \mathbf{h}) - \mu_j)] = C_{ij}(\mathbf{h})$, where $i, j = 1, \dots, N$, and $\mathbf{h}$ is a separation vector of coordinates. For the $Y_i(\mathbf{x})$ we then assume the model

$$Y_i(\mathbf{x}) = \mu_i + Z_i(\mathbf{x}), \quad \text{and} \quad Z_i(\mathbf{x}) = \sum_{p=1}^P a_{ip} F_p(\mathbf{x}) + \xi_i(\mathbf{x}),$$

where $F_p(\mathbf{x})$, $p = 1, \dots, P$, are P non-observable spatial components (common factors) responsible for the cross correlation between the variables $Z_i(\mathbf{x})$, $\xi_i(\mathbf{x})$ are N non-observable spatial components (unique factors) responsible for the residual autocorrelation in the $Z_i(\mathbf{x})$ unexplained by the common factors, and a_{ip} are coefficients. All these non-observable spatial components are assumed to be mean zero second-order stationary with variograms

$$\frac{1}{2} E[(F_p(\mathbf{x} + \mathbf{h}) - F_p(\mathbf{x}))^2] = g(\mathbf{h}), \quad \frac{1}{2} E[(\xi_i(\mathbf{x} + \mathbf{h}) - \xi_i(\mathbf{x}))^2] = \psi_i g(\mathbf{h}),$$

where $g(\mathbf{h})$ is a normalized variogram (conditionally negative definite) function, for which $g(\mathbf{h}) \rightarrow 1$, as $|\mathbf{h}| \rightarrow \infty$, common to all factors, and $\psi_i \geq 0$, $i = 1, \dots, N$. Also, all these regionalized factors are assumed to be mutually uncorrelated, that is, to have all their cross-variograms identically equal to zero.

For the observable random functions $Y_i(\mathbf{x})$, these assumptions imply that the N direct variograms $\gamma_{ii}(\mathbf{h})$ and the $N(N - 1)/2$ cross-variograms $\gamma_{ij}(\mathbf{h})$ are all proportional (intrinsic correlation structure) to the same normalized variogram function $g(\mathbf{h})$, that is,

$$\begin{aligned} \gamma_{ij}(\mathbf{h}) &= \frac{1}{2} E[(Y_i(\mathbf{x} + \mathbf{h}) - Y_i(\mathbf{x}))(Y_j(\mathbf{x} + \mathbf{h}) - Y_j(\mathbf{x}))] \\ &= b_{ij} g(\mathbf{h}) = \left(\sum_{p=1}^P a_{ip} a_{jp} + \psi_i \right) g(\mathbf{h}), \end{aligned}$$

for some real coefficients b_{ij} representing the sills of the variograms $\gamma_{ij}(\mathbf{h})$ (that is, the covariances $C_{ij}(0) = b_{ij}$), between the variables $Y_i(\mathbf{x})$ and $Y_j(\mathbf{x})$.

The specification and estimation of all elements of this spatial linear factor model can be carried out by adapting the steps considered, for instance, in Goovaerts *et al.* (1993) and Wackernagel (1995), whereas optimal (minimum mean square prediction error) linear unbiased predictions for the factors $F_p(\mathbf{x})$, at any spatial location of choice $\mathbf{x}_0$, can be obtained, as in the geostatistical linear coregionalization model (see, e. g., Wackernagel (1995)), by factorial cokriging, that is, by a linear weighted average of the data $y_i(\mathbf{x}_j)$. Now, although traditional geostatistics is almost entirely concerned with second order

properties, in the case in which the joint distribution of all random variables involved is multivariate Gaussian, factorial (simple) cokriging predictions coincide with the conditional expectations $E[F_p(\mathbf{x}_0)|y_i(\mathbf{x}_j) : i = 1, \dots, N; j = 1, \dots, n]$ and so they provide overall optimal unbiased predictions. On the other hand, if the joint distribution of the random variables involved is not (multivariate) Gaussian, then, in general, these conditional expectations are not linear functions of the data $y_i(\mathbf{x}_j)$, and factorial cokriging can provide very inefficient predictions of the factors of interest.

3. The Hierarchical Spatial Factor Model

Let $Y_i(\mathbf{x})$, $i = 1, \dots, N$, be N random functions and $Z_i(\mathbf{x})$, $i = 1, \dots, N$, be N mean zero joint stationary Gaussian processes for which, for any $\mathbf{x}'$ and $\mathbf{x}''$, and for $i \neq j$,

$$Y_i(\mathbf{x}') \perp\!\!\!\perp Y_j(\mathbf{x}'') \mid Z_i(\mathbf{x}'), \quad \text{and} \quad Y_i(\mathbf{x}') \perp\!\!\!\perp Z_j(\mathbf{x}'') \mid Z_i(\mathbf{x}').$$

Moreover, assume that, for any given i and $\mathbf{x}$, conditionally on $Z_i(\mathbf{x})$, the random variables $Y_i(\mathbf{x})$ have conditional distributions $f_i(y; M_i(\mathbf{x}))$, specified by the values of the conditional expectations $M_i(\mathbf{x}) = E[Y_i(\mathbf{x})|Z_i(\mathbf{x})]$, and that $h_i(M_i(\mathbf{x})) = \beta_{i0} + Z_i(\mathbf{x})$, for some parameters β_{i0} and some known link functions $h_i(\cdot)$. In what follows we will assume that the $f_i(y; M_i(\mathbf{x}))$ are all Poisson distributions with means $M_i(\mathbf{x})$ and that $h_i(\cdot) = \ln(\cdot)$.

For the latent part of the model, we will basically adopt the same linear structure considered in Section 2. For the N joint stationary Gaussian processes $Z_i(\mathbf{x})$, let us assume the linear factor model $\mathbf{Z}(\mathbf{x}) = \mathbf{A}\mathbf{F}(\mathbf{x}) + \boldsymbol{\xi}(\mathbf{x})$, where $\mathbf{A}$ is an $N \times P$ matrix of coefficients, and $F_p(\mathbf{x})$, $p = 1, \dots, P$, and $\xi_i(\mathbf{x})$, $i = 1, \dots, N$, are mean zero stationary Gaussian processes with cross-covariances identically equal to zero, and

$$\text{Cov}[F_p(\mathbf{x}), F_p(\mathbf{x} + \mathbf{h})] = \rho(\mathbf{h}), \quad \text{Cov}[\xi_i(\mathbf{x}), \xi_i(\mathbf{x} + \mathbf{h})] = \psi_i \rho(\mathbf{h}),$$

where $\rho(\mathbf{h})$ is a real spatial correlation function for which $\rho(0) = 1$ and $\rho(\mathbf{h}) \rightarrow 0$, as $|\mathbf{h}| \rightarrow \infty$, and ψ_i are non-negative real parameters.

Then it is not difficult to show that the variograms and cross-variograms of the observable $Y_i(\mathbf{x})$, as implied by this model, are given by, for $\mathbf{h} \neq 0$,

$$\begin{aligned} \gamma_{ii}(\mathbf{h}) &= \frac{1}{2} E[(Y_i(\mathbf{x} + \mathbf{h}) - Y_i(\mathbf{x}))^2] \\ &= \tau_i^* + (\tau_i^*)^2 [\exp\{\varsigma_i^2\} - \exp\{\varsigma_i^2 \rho(\mathbf{h})\}], \end{aligned}$$

and, for $i \neq j$,

$$\begin{aligned} \gamma_{ij}(\mathbf{h}) &= \frac{1}{2} E[(Y_i(\mathbf{x} + \mathbf{h}) - Y_i(\mathbf{x}))(Y_j(\mathbf{x} + \mathbf{h}) - Y_j(\mathbf{x}))] \\ &= \tau_i^* \tau_j^* [\exp\{\varsigma_{ij}\} - \exp\{\varsigma_{ij} \rho(\mathbf{h})\}], \end{aligned}$$

where $\tau_i^* = \exp\{\beta_{i0} + \varsigma_i^2/2\}$, $\varsigma_i^2 = \text{Var}[Z_i(\mathbf{x})]$ and $\varsigma_{ij} = \text{Cov}[Z_i(\mathbf{x}), Z_j(\mathbf{x})]$.

Adopting a non-Bayesian inferential framework, and for a specific choice of the function $\rho(\mathbf{h})$, the estimation of the $N + N + N(N - 1)/2$ parameters β_{i0} , ς_i^2 and ς_{ij} of

this hierarchical spatial factor model can be carried out by involving the classical method of moments or by using some iterative algorithm based on some form of least squares. Also, optimal (minimum mean square prediction error) unbiased predictions of the factors $F_p(\mathbf{x})$ at given locations $\mathbf{x}_0$ are provided by the conditional expectations $E[F_p(\mathbf{x}_0)|y_i(\mathbf{x}_j) : i = 1, \dots, N; j = 1, \dots, n]$, which, although not available in close form and difficult to treat analytically, can be numerically evaluated by simulation through Markov chain Monte Carlo methods.

4. Conclusions

Although consistency of the estimation procedures advocated is guaranteed by standard arguments, more efficient estimation algorithms are being explored. However, it must be noted that the simple estimation procedures discussed are particularly appealing in such complex models since they allow a direct link between theoretical elements and corresponding sample quantities.

Though accuracy measures of the factor maps of predictions can be provided by estimates of the error prediction variance, these measures are obtained by assuming that no errors have been made for the estimation of the parameters, and overall measures of goodness of fit of the models should be investigated.

Extensions of the hierarchical spatial factor model here proposed can be considered to include more than one spatial scale on the lines of the classical geostatistical linear model of coregionalization (see, e. g., Matheron (1982) and Wackernagel (1995)).

References

- Bailey T.C. and Krzanowski W.J. (2000) Extensions to spatial factor methods with an illustration in geochemistry, *Mathematical Geology*, 32, 657–682.
- Cressie N. (2000) Geostatistical methods for mapping environmental exposures, in: *Disease and Exposure Mapping*, Elliot P. et al. (Eds.), Oxford University Press, Oxford.
- Diggle P.J., Moyeed R.A. and Tawn J.A. (1998) Model-based geostatistics (with discussion), *Applied Statistics*, 47, 299–350.
- Goovaerts P., Sonnet P. and Navarre A. (1993) Factorial kriging analysis of spring-water contents in the Dyle river basin, Belgium, *Water Resources Research*, 29, 2115–2125.
- Matheron G. (1982) Pour une analyse krigeante des données régionalisées, *Publication N-732*, Centre de Géostatistique, Fontainebleau, France, 22 pp.
- Switzer P. and Green A. (1984) Min/max autocorrelation factor for multivariate spatial imagery, *Technical Report No. 6*, Department of Statistics, Stanford University, Stanford, CA, 23 pp.
- Wackernagel H. (1995) *Multivariate Geostatistics*, Springer, Berlin.