Estimation by stochastic EM
in a class of spatial factor models

L’uso dell’algoritmo EM stocastico per modelli spaziali fattoriali

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

Riassunto: La stima dei parametri nei modelli spaziali viene tradizionalmente affrontata ricorrendo all’utilizzo di tecniche “ad hoc” spesso prive di proprietà di ottimalità e sono rare in quest’ambito analisi inferenziali basate sullo studio della verosimiglianza. Questo stato di cose è in particolare diffuso nel caso dei modelli spaziali multivariati di tipo geostatistico. In questo lavoro, per modelli geostatistici fattoriali gerarchici (“generalized linear mixed models”) non gaussiani, si propone, per la fase di stima finale dei parametri, un’analisi di verosimiglianza basata sull’utilizzo dell’algoritmo EM stocastico in cui si richiede l’uso di algoritmi Markov chain Monte Carlo.

Keywords: generalized linear mixed models, Monte Carlo EM, Markov chain Monte Carlo, multivariate geostatistics, spatial factor models, stochastic EM algorithm.

1. Hierarchical spatial factor models

From the work of Matheron (1982) till nowadays, multivariate geostatistics has been dominated by the linear model of cokriging (with its simpler counterpart, the proportional covariance model, otherwise known as intrinsic correlation model) and by the related ‘factorial kriging analysis’ (see, for more recent accounts, Wackernagel (1995) and Chilès and Delfiner (1999)). It is fair to say that the widespread use of this model is mainly due to the convenient characterization of the multivariate spatial autocorrelation structure, as well as to the availability of a widely accepted estimation procedure based on the (eigenvalue) principal component decomposition and on the algorithm of Goulard and Voltz (1992). Nevertheless, being this model based only on the form taken by direct and cross variograms, no hypothesis is made about the distribution of the data, and the implied estimation and prediction procedures may be considered adequate only in the case in which the data can be assumed to be multivariate Gaussian. In other cases, particularly in presence of count data, as in many environmental and epidemiological situations, the use of this model can lead to misleading predictions and to erroneous conclusions about the underlying factors. To cope with such situations, in Minozzo (2002) and Minozzo and Fruttini (2003), following the proposals put forward in the univariate case (see, for instance, Diggle et al. (1998) and Cressie (2000)), a hierarchical multivariate spatial model has been proposed by building upon a generalization of the classical geostatistical proportional covariance model.

Let us briefly recall the model. Let $n_i(x_k), i = 1, \ldots, m, k = 1, \ldots, K$, be a set of geo-referenced frequency counts relative to $m$ regionalized variables, gathered at $K$
spatial locations $x_k$. These $m$ regionalized counts are seen as one partial realization of a set of $m$ random functions $N_i(x)$, $i = 1, \ldots, m$, $x \in D \subseteq \mathbb{R}^2$. For these functions we assume, for any $x$, and for $i \neq j$, $N_i(x) \perp N_j(x)$ | $Z_i(x)$ and $N_i(x) \perp Z_j(x)$ | $Z_i(x)$, and, for $x' \neq x''$, and $j = 1, \ldots, m$, $N_i(x') \perp N_j(x'')$ | $Z_i(x')$ and $N_i(x') \perp Z_j(x'')$ | $Z_i(x')$, where $Z_i(x)$, $i = 1, \ldots, m$, are mean zero joint stationary Gaussian processes. We also assume that, for any given $i$ and $x$, conditionally on $Z_i(x)$, the random variables $N_i(x)$ are Poisson distributed with means given by the conditional expectations $M_i(x) = E[N_i(x)|Z_i(x)]$, and that $\ln(M_i(x)) = \beta_i + Z_i(x)$, for some parameters $\beta_i$.

For processes $Z_i(x)$, let us assume the linear factor model $Z_i(x) = \sum_{p=1}^{P} a_{ip} F_p(x) + \xi_i(x)$, where $a_{ip}$ are coefficients, $F_p(x)$ are $P \leq m$ non-observable spatial components (common factors) responsible for the cross-correlation between the variables $Z_i(x)$, and $\xi_i(x)$ are non-observable spatial components (unique factors) responsible for the residual autocorrelation in the $Z_i(x)$ unexplained by the common factors. We assume that $F_p(x)$ and $\xi_i(x)$ are stationary Gaussian processes with $E[F_p(x)] = 0$ and $E[\xi_i(x)] = 0$, and with covariance functions $\text{Cov}[F_p(x), F_q(x+h)] = \rho(h)$ and $\text{Cov}[\xi_i(x), \xi_j(x+h)] = \psi(h)$, where $\rho(h)$ is a real spatial correlation function common to all factors such that $\rho(0) = 1$ and $\rho(h) \to 0$, as $|h| \to \infty$, and $\psi$ are non-negative real parameters. We also assume that the processes $F_p(x)$ and $\xi_i(x)$ have all cross-covariances identically equal to zero.

2. Estimation with the stochastic EM algorithm

Adopting a non-Bayesian inferential framework, estimation of the parameters of the above hierarchical spatial factor model can be carried out by employing the method of moments, adapting some of the standard procedures used for the linear model of coregionalization. Although these estimation procedures may easily be implemented, they lack of sufficient optimality properties and more efficient procedures, such as likelihood based procedures, should be considered. Assuming that the number of factors $P$ and the spatial autocorrelation function $\rho(h)$ have already been chosen, likelihood inference would require the maximization of the marginal density function of the observed random variables $f(n_1, \ldots, n_m)$. However, since this marginal density is not available, and since the integration required in the E step of the EM algorithm (Dempster et al., 1977) would not be easy, we have to resort to some stochastic versions of the EM algorithm, like the stochastic EM (StEM) algorithm or the Monte Carlo EM (MCEM) algorithm (see, for instance, Celeux and Diebolt (1985), Wei and Tanner (1990), Celeux et al. (1995), Nielsen (2000) and Zhang (2002)).

For a given set of data $n_i = (n_{i1}, \ldots, n_{iK})^T$, $i = 1, \ldots, m$, where $n_{ik} = n_i(x_k)$, let us define the vectors of random variables $\xi_i = (\xi_{i1}, \ldots, \xi_{iK})^T$, $i = 1, \ldots, m$, and $F_p = (F_{p1}, \ldots, F_{pK})^T$, $p = 1, \ldots, P$, where $\xi_{ik} = \xi_i(x_k)$ and $F_{pk} = F_p(x_k)$, for $i = 1, \ldots, m$, $p = 1, \ldots, P$, and $k = 1, \ldots, K$. Then, by defining $\theta = (\beta, A, \psi, \varphi)$, where $\beta = (\beta_1, \ldots, \beta_m)^T$, $A = (a_1, \ldots, a_m)^T$, with $a_i = (a_{i1}, \ldots, a_{ip})$, $i = 1, \ldots, m$, and where $\psi = (\psi_1, \ldots, \psi_m)^T$ and $\varphi$ is the vector of parameters of $\rho(h)$, the full density function of all observed and unobserved random variables is given by

\[
f(n_1, \ldots, n_m, \xi_1, \ldots, \xi_m, F_1, \ldots, F_p) = f(n_1, \ldots, n_m | \xi_1, \ldots, \xi_m, F_1, \ldots, F_p) \cdot f(\xi_1, \ldots, \xi_m, F_1, \ldots, F_p)
\]

\[
= \prod_{i=1}^{m} \prod_{k=1}^{K} \exp \left(-\frac{M_{ih} M_{ik}}{\sum_{i=1}^{m} M_{ih}}\right) \cdot f(\xi_i) \cdot f(F_1) \cdot \cdots \cdot f(F_p),
\]
where $M_{ik} \equiv M_i(x_k)$ depend on parameters $\beta$ and $A$, and $f(\xi_i)$, $i = 1, \ldots, m$, and $f(F_p)$, $p = 1, \ldots, P$, are $k$-multivariate normals depending on parameters $\varrho$ and $\psi$.

Thus, assuming that the current guess for the parameters at the $s$th step is given by $\theta_s$, and that $R_s$ is a fixed positive integer, the MCEM algorithm can be implemented as follows:

**S step** – draw $R_s$ samples $(\xi_1, \ldots, \xi_m, F_1, \ldots, F_P)^{(r)}$, $r = 1, \ldots, R_s$, from the (filtered) conditional distribution $p(\xi_1, \ldots, \xi_m, F_1, \ldots, F_P|n_1, \ldots, n_m; \theta_s)$;

**E step** – compute

$$Q_{s+1}(\theta, \theta_s) = \frac{1}{R_s} \sum_{r=1}^{R_s} \ln f(n_1, \ldots, n_m; \xi_1^{(r)}, \ldots, \xi_m^{(r)}, F_1^{(r)}, \ldots, F_P^{(r)}; \theta);$$

**M step** – take as the new guess $\theta_{s+1}$ the value of $\theta$ which maximizes $Q_{s+1}(\theta, \theta_s)$.

Choosing $R_s = 1$ this procedure reduces to the StEM algorithm, whereas for $R_s$ very large it approximates the deterministic EM algorithm. Moreover, a simulated annealing version could be obtained by choosing a slowly increasing sequence $R_s \rightarrow \infty$, as $s \rightarrow \infty$.

Although computationally intensive, the sampling involved in the S step of the algorithm can be accomplished by Markov chain Monte Carlo (MCMC) simulation taking as target distribution the conditional distribution $p(\xi_1, \ldots, \xi_m, F_1, \ldots, F_P|n_1, \ldots, n_m; \theta_s)$.

On the other hand, the maximization in the M step can be carried out numerically. In the case of the StEM algorithm, that is, assuming $R_s = 1$, we have to maximize the full loglikelihood function $l(\theta) = \ln f(n_1, \ldots, n_m; \xi_1, \ldots, \xi_m, F_1, \ldots, F_P; \theta)$ which can be written as

$$l(\theta) = \sum_{i=1}^{m} l_i(\beta_i, a_i) + \sum_{i=1}^{m} \ln f(\xi_i; \psi_i, \varrho) + \sum_{p=1}^{P} \ln f(F_p; \varrho),$$

where

$$l_i(\beta_i, a_i) = \sum_{k=1}^{K} \left[ -M_{ik} + n_{ik} \ln(M_{ik}) - \ln(n_{ik}) \right],$$

$$\ln f(\xi_i; \psi_i, \varrho) = -\frac{K}{2} \ln(2\pi \psi_i) - \frac{1}{2} \ln |P_\rho| - \frac{1}{2\psi_i} \xi_i^T P_\rho^{-1} \xi_i,$$

$$\ln f(F_p; \varrho) = -\frac{K}{2} \ln(2\pi) - \frac{1}{2} \ln |P_\rho| - \frac{1}{2} F_p^T P_\rho^{-1} F_p, \quad P_\rho = |\rho(x_i - x_k)|_{(k,i)}.$$

Proceeding in steps, the values of the parameters $\beta_i$ and $a_i$ maximizing $l(\theta)$, for every $i = 1, \ldots, m$, can be obtained by considering the terms $l_i(\beta_i, a_i)$. Analytic maximization leads to a system of $P + 1$ nonlinear equations which can be solved iteratively. Then, maximization of $l(\theta)$ with respect to the parameters $\psi_i$ can be accomplished by considering the terms $\ln f(\xi_i; \psi_i, \varrho)$, separately for every $i = 1, \ldots, m$. Equating to zero the first derivatives of $\ln f(\xi_i; \psi_i, \varrho)$ with respect to $\hat{\psi}_i$ leads to $\hat{\psi}_i = (1/K)\xi_i^T P_\rho^{-1} \xi_i$, which depend on $\varrho$. Substituting the $\hat{\psi}_i$ just obtained into $l(\theta)$ leads to the profile likelihood

$$l_\rho(\varrho) = -\frac{K}{2} \sum_{i=1}^{m} \ln \left( \frac{1}{K} \xi_i^T P_\rho^{-1} \xi_i \right) - \frac{1}{2} (m + P) \ln |P_\rho| - \frac{1}{2} \sum_{p=1}^{P} F_p^T P_\rho^{-1} F_p,$$
which can be maximized numerically with respect to $\theta$.

In the case of the MCEM, that is, with $R_s > 1$, the M step of the algorithm requires the maximization of the ‘average’ likelihood

$$\sum_{r=1}^{R_s} L^{(r)}(\theta) = \sum_{i=1}^{m} \sum_{r=1}^{R_s} \ell^{(r)}(\beta_i, a_i) + \sum_{i=1}^{m} \sum_{r=1}^{R_s} \ln f(\xi_i^{(r)}; \psi_i, \theta) + \sum_{p=1}^{P} \sum_{r=1}^{R_s} \ln f(F_p^{(r)}; \theta),$$

which can be accomplished by proceeding similarly to the case of the StEM algorithm ($R_s = 1$). Whereas for the case in which $R_s$ is small the algorithm does not converge pointwise and estimates of the parameters have to be obtained by considering some summary statistic of the values $\theta_s$, for $R_s$ large (or $R_s$ increasing) the algorithm is guaranteed to converge to a local maximum of the likelihood. Once estimates of the parameters have been obtained, being the model considered non-identifiable, different rotations of the matrix $\hat{A}$ may be considered to allow for diverse interpretations of the underlying factors.

**References**


