

Estimation by stochastic EM in a class of spatial factor models

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

Marco Minozzo¹

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 coregionalization (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 $n_i(\mathbf{x}_k)$, $i = 1, \dots, m$, $k = 1, \dots, K$, be a set of geo-referenced frequency counts relative to m regionalized variables, gathered at K

¹Indirizzo per corrispondenza: Dipartimento di Scienze Statistiche, Università degli Studi di Perugia, Via A. Pascoli, 06100 Perugia, Italy

spatial locations $\mathbf{x}_k$. These m regionalized counts are seen as one partial realization of a set of m random functions $N_i(\mathbf{x})$, $i = 1, \dots, m$, $\mathbf{x} \in D \subseteq \mathbb{R}^2$. For these functions we assume, for any $\mathbf{x}$, and for $i \neq j$, $N_i(\mathbf{x}) \perp\!\!\!\perp N_j(\mathbf{x}) \mid Z_i(\mathbf{x})$ and $N_i(\mathbf{x}) \perp\!\!\!\perp Z_j(\mathbf{x}) \mid Z_i(\mathbf{x})$, and, for $\mathbf{x}' \neq \mathbf{x}''$, and $j = 1, \dots, m$, $N_i(\mathbf{x}') \perp\!\!\!\perp N_j(\mathbf{x}'') \mid Z_i(\mathbf{x}')$ and $N_i(\mathbf{x}') \perp\!\!\!\perp Z_j(\mathbf{x}'') \mid Z_i(\mathbf{x}')$, where $Z_i(\mathbf{x})$, $i = 1, \dots, m$, are mean zero joint stationary Gaussian processes. We also assume that, for any given i and $\mathbf{x}$, conditionally on $Z_i(\mathbf{x})$, the random variables $N_i(\mathbf{x})$ are Poisson distributed with means given by the conditional expectations $M_i(\mathbf{x}) = E[N_i(\mathbf{x}) \mid Z_i(\mathbf{x})]$, and that $\ln(M_i(\mathbf{x})) = \beta_i + Z_i(\mathbf{x})$, for some parameters β_i .

For processes $Z_i(\mathbf{x})$, let us assume the linear factor model $Z_i(\mathbf{x}) = \sum_{p=1}^P a_{ip} F_p(\mathbf{x}) + \xi_i(\mathbf{x})$, where a_{ip} are coefficients, $F_p(\mathbf{x})$ are $P \leq m$ non-observable spatial components (common factors) responsible for the cross correlation between the variables $Z_i(\mathbf{x})$, and $\xi_i(\mathbf{x})$ are non-observable spatial components (unique factors) responsible for the residual autocorrelation in the $Z_i(\mathbf{x})$ unexplained by the common factors. We assume that $F_p(\mathbf{x})$ and $\xi_i(\mathbf{x})$ are stationary Gaussian processes with $E[F_p(\mathbf{x})] = 0$ and $E[\xi_i(\mathbf{x})] = 0$, and with covariance functions $\text{Cov}[F_p(\mathbf{x}), F_p(\mathbf{x} + \mathbf{h})] = \rho(\mathbf{h})$ and $\text{Cov}[\xi_i(\mathbf{x}), \xi_i(\mathbf{x} + \mathbf{h})] = \psi_i \rho(\mathbf{h})$, where $\rho(\mathbf{h})$ is a real spatial correlation function common to all factors such that $\rho(0) = 1$ and $\rho(\mathbf{h}) \rightarrow 0$, as $|\mathbf{h}| \rightarrow \infty$, and ψ_i are non-negative real parameters. We also assume that the processes $F_p(\mathbf{x})$ and $\xi_i(\mathbf{x})$ 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 $f(\mathbf{n}_1, \dots, \mathbf{n}_m)$. 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 $\mathbf{n}_i = (n_{i1}, \dots, n_{iK})^T$, $i = 1, \dots, m$, where $n_{ik} \doteq n_i(\mathbf{x}_k)$, let us define the vectors of random variables $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{iK})^T$, $i = 1, \dots, m$, and $\mathbf{F}_p = (F_{p1}, \dots, F_{pK})^T$, $p = 1, \dots, P$, where $\xi_{ik} \doteq \xi_i(\mathbf{x}_k)$ and $F_{pk} \doteq F_p(\mathbf{x}_k)$, for $i = 1, \dots, m$, $p = 1, \dots, P$, and $k = 1, \dots, K$. Then, by defining $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\psi}, \boldsymbol{\varrho})$, where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)^T$, $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_m)^T$, with $\mathbf{a}_i = (a_{i1}, \dots, a_{iP})$, $i = 1, \dots, m$, and where $\boldsymbol{\psi} = (\psi_1, \dots, \psi_m)^T$ and $\boldsymbol{\varrho}$ is the vector of parameters of $\rho(\mathbf{h})$, the full density function of all observed and unobserved random variables is given by

$$\begin{aligned} f(\mathbf{n}_1, \dots, \mathbf{n}_m, \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_m, \mathbf{F}_1, \dots, \mathbf{F}_P) \\ &= f(\mathbf{n}_1, \dots, \mathbf{n}_m \mid \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_m, \mathbf{F}_1, \dots, \mathbf{F}_P) \cdot f(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_m, \mathbf{F}_1, \dots, \mathbf{F}_P) \\ &= \left[\prod_{i=1}^m \prod_{k=1}^K \frac{\exp(-M_{ik}) M_{ik}^{n_{ik}}}{n_{ik}!} \right] \cdot f(\boldsymbol{\xi}_1) \cdots f(\boldsymbol{\xi}_m) \cdot f(\mathbf{F}_1) \cdots f(\mathbf{F}_P), \end{aligned}$$

where $M_{ik} \doteq M_i(\mathbf{x}_k)$ depend on parameters β and $\mathbf{A}$, and $f(\xi_i)$, $i = 1, \dots, m$, and $f(\mathbf{F}_p)$, $p = 1, \dots, P$, are k -multivariate normals depending on parameters $\boldsymbol{\varrho}$ and $\boldsymbol{\psi}$.

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

- S step* – draw R_s samples $(\xi_1, \dots, \xi_m, \mathbf{F}_1, \dots, \mathbf{F}_P)^{(r)}$, $r = 1, \dots, R_s$, from the (filtered) conditional distribution $p(\xi_1, \dots, \xi_m, \mathbf{F}_1, \dots, \mathbf{F}_P | \mathbf{n}_1, \dots, \mathbf{n}_m; \boldsymbol{\theta}_s)$;
- E step* – compute

$$Q_{s+1}(\boldsymbol{\theta}, \boldsymbol{\theta}_s) = \frac{1}{R_s} \sum_{r=1}^{R_s} \ln f(\mathbf{n}_1, \dots, \mathbf{n}_m, \xi_1^{(r)}, \dots, \xi_m^{(r)}, \mathbf{F}_1^{(r)}, \dots, \mathbf{F}_P^{(r)}; \boldsymbol{\theta});$$

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

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 $p(\xi_1, \dots, \xi_m, \mathbf{F}_1, \dots, \mathbf{F}_P | \mathbf{n}_1, \dots, \mathbf{n}_m; \boldsymbol{\theta}_s)$.

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 $l(\boldsymbol{\theta}) = \ln f(\mathbf{n}_1, \dots, \mathbf{n}_m, \xi_1, \dots, \xi_m, \mathbf{F}_1, \dots, \mathbf{F}_P; \boldsymbol{\theta})$ which can be written as

$$l(\boldsymbol{\theta}) = \sum_{i=1}^m l_i(\beta_i, \mathbf{a}_i) + \sum_{i=1}^m \ln f(\xi_i; \psi_i, \boldsymbol{\varrho}) + \sum_{p=1}^P \ln f(\mathbf{F}_p; \boldsymbol{\varrho}),$$

where

$$\begin{aligned} l_i(\beta_i, \mathbf{a}_i) &= \sum_{k=1}^K \left[-M_{ik} + n_{ik} \ln(M_{ik}) - \ln(n_{ik}!) \right], \\ \ln f(\xi_i; \psi_i, \boldsymbol{\varrho}) &= -\frac{K}{2} \ln(2\pi\psi_i) - \frac{1}{2} \ln |\mathbf{P}_\rho| - \frac{1}{2\psi_i} \xi_i^T \mathbf{P}_\rho^{-1} \xi_i, \\ \ln f(\mathbf{F}_p; \boldsymbol{\varrho}) &= -\frac{K}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{P}_\rho| - \frac{1}{2} \mathbf{F}_p^T \mathbf{P}_\rho^{-1} \mathbf{F}_p, \end{aligned}$$

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

$$l_\rho(\boldsymbol{\varrho}) \propto -\frac{K}{2} \sum_{i=1}^m \ln \left(\frac{1}{K} \xi_i^T \mathbf{P}_\rho^{-1} \xi_i \right) - \frac{1}{2} (m + P) \ln |\mathbf{P}_\rho| - \frac{1}{2} \sum_{p=1}^P \mathbf{F}_p^T \mathbf{P}_\rho^{-1} \mathbf{F}_p,$$

which can be maximized numerically with respect to $\boldsymbol{\varrho}$.

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

$$\sum_{r=1}^{R_s} l^{(r)}(\boldsymbol{\theta}) = \sum_{i=1}^m \sum_{r=1}^{R_s} l_i^{(r)}(\beta_i, \mathbf{a}_i) + \sum_{i=1}^m \sum_{r=1}^{R_s} \ln f(\boldsymbol{\xi}_i^{(r)}; \psi_i, \boldsymbol{\varrho}) + \sum_{p=1}^P \sum_{r=1}^{R_s} \ln f(\mathbf{F}_p^{(r)}; \boldsymbol{\varrho}),$$

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 $\boldsymbol{\theta}_s$, 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 $\hat{\mathbf{A}}$ may be considered to allow for diverse interpretations of the underlying factors.

References

- Celeux G., Chauveau D. and Diebolt J. (1995) On stochastic versions of the EM algorithm, Technical report, N. 2514, Institut National de Recherche en Informatique et en Automatique, Rhône-Alpes, Faculté de Médecine de Grenoble, Dept. de Statistique, 38700 Latronche, France.
- Celeux G. and Diebolt J. (1985) The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, *Computational Statistics Quarterly*, 2, 73–82.
- Chilès J.P. and Delfiner P. (1999) *Geostatistics: Modeling Spatial Uncertainty*, Wiley, New York.
- Cressie N. (2000) Geostatistical methods for mapping environmental exposures, in: *Disease and Exposure Mapping*, Elliot, P. et al. (Eds.), Oxford University Press, Oxford.
- Dempster A.P., Laird N.M. and Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Diggle P.J., Moyeed R.A. and Tawn J.A. (1998) Model-based geostatistics (with Discussion), *Applied Statistics*, 47, 299–350.
- Goulard M. and Voltz M. (1992) Linear coregionalization model: tools for estimation and choice of cross-variogram matrix, *Mathematical Geology*, 24, 269–286.
- Matheron G. (1982) Pour une analyse krigéante des données régionalisées, Technical report, Publication N-732, Centre de Géostatistique, Fontainebleau, France, 22 pp.
- Minozzo M. (2002) Hierarchical spatial factor models for Poisson count data, *Proceedings of the XLI Conference of the Italian Statistical Society*, 103–106.
- Minozzo M. and Fruttini D. (2003) Loglinear spatial factor analysis: an application to diabetes mellitus complications, *Environmetrics*, to be published.
- Nielsen S.F. (2000) The stochastic EM algorithm: estimation and asymptotic results, *Bernoulli*, 6, 457–489.
- Wackernagel H. (1995) *Multivariate Geostatistics*, Springer, Berlin.
- Wei G.C.G. and Tanner M.A. (1990) A Monte Carlo implementation of the EM algorithm and poor man’s data augmentation algorithm, *Journal of the American Statistical Association*, 85, 699–704.
- Zhang H. (2002) On estimation and prediction for spatial generalized linear mixed models, *Biometrics*, 58, 129–136.