

The EM algorithm for air quality spatio-temporal models ¹

L'algoritmo EM per modelli spazio-temporali di qualità dell'aria

Alessandro Fassò²

Dipartimento di Ingegneria dell'Informazione e Metodi Matematici

Università di Bergamo

e-mail: alessandro.fasso@unibg.it

Michela Cameletti

Dipartimento di Ingegneria dell'Informazione e Metodi Matematici

Università di Bergamo

e-mail: michela.cameletti@unibg.it

Riassunto: In questo lavoro presentiamo l'utilizzo dell'algoritmo EM in relazione alla stima dei parametri incogniti di un modello gerarchico per dati spazio-temporali di qualità dell'aria.

Keywords: Hierarchical spatio-temporal model, EM algorithm, air quality data.

1. Introduction

In Fassò and Nicolis (2006), Fassò *et al.* (2005) and Fassò *et al.* (2007) it has been proposed a hierarchical model for air quality data collected by the heterogeneous network of the entire Po valley. Generalizing this approach, in Cameletti (2007) a one-dimensional hidden state hierarchical model has been proposed for modelling the relation between observed PM10 concentrations and emission inventory data; in particular, it considers a spatio-temporal structure and an improved parameter estimation based on the EM algorithm (McLachlan and Krishnan (1997), Xu and Wikle (2007)). In Fassò and Cameletti (2007) the previous model has been extended considering a p -dimensional latent variable. This paper, which refers to Fassò and Cameletti (2007) for the details and the notation, focuses on the estimation procedure using the EM algorithm.

2. The model and the implementation of the algorithm

Let $Z_t = \{Z(s_1, t), \dots, Z(s_n, t)\}'$ be the observed spatio-temporal process at time $t = 1, \dots, T$ and at n spatial locations; for example, Z_t can be the vector of daily PM10 concentrations measured in n monitoring stations. The final model is:

$$\begin{aligned} Z_t &= X_t\beta + K_tY_t + e_t & e_t &\sim Niid[0, \Sigma_e] \\ Y_t &= GY_{t-1} + \eta_t & \eta_t &\sim Niid[0, \Sigma_\eta] \end{aligned} \quad (1)$$

¹Work partially supported by PRIN project n.2006131039 and Regione Piemonte project CIPE 2004.

²Address of correspondence: Via Marconi, 5 - 24044 Dalmine (BG)

where X_t is a $(n \times d)$ matrix of covariates, K_t is a $(n \times p)$ matrix of ones and Y_t is a p -dimensional unobserved temporal process. The covariance matrix Σ_e is defined by an isotropic, weakly stationary spatial covariance function $C(h)$ which depends on a set θ of parameters and on the euclidean distance h . Moreover we have that $C(0) = \sigma_\omega^2(1 + b)$. The second equation of system (1) defines the temporal dynamic of the latent variable where $G = \text{diag}(g_1, \dots, g_p)$ is the transition matrix and η_t is a p -dimensional gaussian white noise process with covariance matrix $\Sigma_\eta = \text{diag}(\sigma_{\eta_1}^2, \dots, \sigma_{\eta_p}^2)$. Thus the vector of the unknown parameters is defined by $\Psi = \{\beta, b, \sigma_\omega^2, \theta, G, \Sigma_\eta, \mu_0, \Sigma_0\}$, where μ_0 and Σ_0 are the parameters of the initial gaussian state Y_0 .

At the generic iteration $k = 0, 1, 2, \dots$ the EM algorithm performs the following steps:

1. Given $\Psi^{(k)}$ as starting values, we calculate $y_t^T, y_{t-1}^T, P_t^T, P_{t-1}^T$ and $P_{t,t-1}^T$ for $t = 0, \dots, T$ using the Kalman smoother formulas; then these quantities are rewritten in a compact form using the notation S_{11}, S_{10} and S_{00} .
2. Using the so-called closed forms we update a subset of parameters: $\mu_0^{(k+1)} = y_0^T$,

$$g_i^{(k+1)} = \frac{S_{10;ii}}{S_{00;ii}}, \quad \sigma_{\eta_i}^{2(k+1)} = \frac{S_{11;ii} - 2g_i^{(k+1)}S_{10;ii} + g_i^{2(k+1)}S_{00;ii}}{T}, \quad \sigma_\omega^{2(k+1)} = \frac{\text{tr}(\Sigma^{*-1}B)}{Tn} \text{ and}$$

$$\beta^{(k+1)} = \left[\sum_{t=1}^T \left(X_t' \Sigma_e^{(k)-1} X_t \right) \right]^{-1} \left\{ \sum_{t=1}^T \left[X_t' \Sigma_e^{(k)-1} (z_t - K y_t^T) \right] \right\}, \text{ where}$$

$$B = \sum_{t=1}^T \left[K P_t^T K' + (z_t - X_t \beta^{(k+1)} - K y_t^T) (z_t - X_t \beta^{(k+1)} - K y_t^T)' \right].$$
3. Using the semi-closed forms based on the Newton-Raphson algorithm and given explicitly in Fassò and Cameletti (2007) we calculate $\theta^{(k+1)}$ and $b^{(k+1)}$.

Note that the covariance matrix Σ_0 is considered constant. The steps 1-3 are iteratively repeated until some convergence criteria are jointly met. It is well known that the algorithm performance is sensitive to the initial values $\Psi^{(0)}$. For this reason, we propose the use of a preliminary analysis based on a regression model and a combined temporal and geostatistical analysis of its residuals. Moreover note that the EM algorithm is better than other numerical optimization methods for its implementation simplicity and for the properties of the obtained estimates.

References

- Cameletti M. (2007) *Modelli spazio-temporali per dati ambientali*, Ph.D. thesis, University of Milano Bicocca.
- Fassò A. and Cameletti M. (2007) A general spatio-temporal model for environmental data, *Working Paper GRASPA* n. 27, www.graspa.org.
- Fassò A., Cameletti M. and Nicolis O. (2007) Air quality monitoring using heterogeneous networks, *Environmetrics*, 18, 3.
- Fassò A. and Nicolis O. (2006) Modelling dynamics and uncertainty in assessment of quality standards for fine particulate matters, *Statistica e Applicazioni*, 4.
- Fassò A., Nicolis O. and Cameletti M. (2005) A statistical approach to heterogeneous monitoring networks, in: *Proceedings of Ecomondo Conference, Rimini 26-29 October 2005*, Maggioli Editore, volume 2, 629–636.
- McLachlan G.J. and Krishnan T. (1997) *The EM Algorithm and Extensions*, Wiley.
- Xu K. and Wikle C.K. (2007) Estimation of parameterized spatio-temporal dynamic models, *Journal of Statistical Inference and Planning*, 137, 567–588.