

Estimating auto-models with missing data

Marco Bee, Giuseppe Espa

Department of Economics, University of Trento, e-mail: marco.bee@unitn.it,
giuseppe.espa@unitn.it

Abstract: We consider the problem of estimating the parameters of Auto-models when some values of the target variable are missing at random but the auxiliary information is known for the same areas. In a complete-data setup, the parameters are estimated using the pseudo-maximum likelihood method. With missing data, the problem can be tackled by means of the EM algorithm. In general, however, the conditional expectation of the complete pseudo-likelihood function is analytically intractable. Thus, we implement the E-step by means of Monte Carlo simulation. We give the details for the Logistic Auto-logistic and the Binomial Auto-binomial models.

Keywords: Spatial Missing Data, Monte Carlo EM Algorithm, Binomial Auto-binomial Model, Pseudo-Likelihood.

1. Introduction

In spatial statistics the classic techniques for the treatment of missing-data problems have to be adapted to spatially dependent data. We focus on the missing-data problem in the Logistic Auto-logistic and Binomial Auto-binomial models. It is well-known that for these models the likelihood function cannot be written in closed form. Hence we will use the EM algorithm in the setup of a Maximum Pseudo-Likelihood approach.

2. Methodology

A spatial random process $\tilde{Y}$ on a $(k \times k)$ grid $\mathcal{D} \subset \mathbb{Z}^2$ can be described by a matrix of random processes. Each element of $\tilde{Y}$ is characterized by two indexes: $\tilde{Y}_{ij}$ ($i, j = 1, \dots, k$). However, it is common to stack the columns of the random field $\tilde{Y}$ on top of each other; in this manner, we obtain a random vector $\mathbf{Y} = \text{vec}(\tilde{Y})$ and a single index is sufficient to identify each dependent variable Y_i ($i = 1, \dots, k^2$). We can now define the Binomial Auto-binomial Model (BAM).

Definition 1 Let $Y_i = \sum_{s=1}^K W_{is}$, where W_{is} are independent Bernoulli (π_i) rvs for any $s \in \{1, \dots, B\}$, with $B \in \mathbb{N}$. A random field $\mathbf{Y}(\mathbf{s})$, $\mathbf{s} \in \mathcal{D}$, $\mathcal{D} \subset \mathbb{Z}^2$ is called a Binomial Auto-binomial Model (BAM) if the conditional distribution of Y_i given Y_j , $j \in C(i)$, is given by

$$P(Y_i = y_i | Y_j = y_j, j \neq i, j \in C\{i\}) = \binom{B}{y_i} \pi_i^{y_i} (1 - \pi_i)^{B-y_i},$$

where

$$\pi_i = P(W_{is} = 1 | W_{js} = w_{js}, j \neq i, j \in C\{i\}) = \frac{\exp\{w_{is}(\alpha_i + \boldsymbol{\gamma}'\mathbf{x}_i + \beta \sum_{j \in C\{i\}} w_{js})\}}{1 + \exp\{\alpha_i + \boldsymbol{\gamma}'\mathbf{x}_i + \beta \sum_{j \in C\{i\}} w_{js}\}}, \quad (1)$$

with W_{is} conditionally independent of $\{W_{js}, j \notin C(i)\}$, $w_{is} \in \{0, 1\}$, $\mathbf{x}_i = (x_1, \dots, x_m)'$ the vector of the m auxiliary variables for area i and $C(i)$ the neighborhood set of cell i .

Applying the ML method to the BAM is usually impossible, because the Y_i s are dependent and their joint distribution is not computable. We can, however, treat the observations as if they were independent. This approach allows to obtain a “likelihood-type” function as the product of the conditional densities $PL = \prod P(y_i | \text{all other } y_j\text{'s})$. PL, considered as a function of the parameters, is called *Pseudo-Likelihood* (PL), and is maximized with respect to the parameters. In this setup PL is the likelihood function of a binomial model, so that maximization can be performed with standard techniques. This methodology, known as *Maximum Pseudo-Likelihood* (MPL), was first introduced by Besag (1975, 1977).

If the data consist of a single realization of the spatial process, and the dimension k tends to infinity, consistency of the MPLEs of the parameters of a Markov random field on $\mathbb{Z}^d$ ($d \geq 2$) was first proved by Geman and Graffigne (1987) under the assumption, clearly satisfied here, that the range of values taken by the process is finite.

When the MLE method can be applied, a common way of tackling a missing-data problem is based on the EM algorithm (Dempster *et al.*, 1977). The algorithm iterates until convergence two steps, called E (Expectation) and M (Maximization). The first one is given by the conditional expectation of the so called complete likelihood function, i.e. the hypothetical likelihood function that would be available if the missing data were observed; the M-step consists in maximizing this conditional expectation. Assume now that, for some cells, Y_i is missing at random, whereas the auxiliary information is available. We would need to write down the complete and observed (incomplete) likelihood functions but, as mentioned above, the observations are not independent, so that even when no observations are missing, we cannot write the likelihood function. We therefore resort to the MPL estimation methodology and work with the complete and incomplete pseudo-likelihood functions:

$$PL_c = \prod_{i=1}^{k^2} \pi_i^{y_i} (1 - \pi_i)^{B-y_i}, \quad PL_{obs} = \prod_{i=1}^{k_{obs}} \pi_i^{y_i} (1 - \pi_i)^{B-y_i}, \quad (2)$$

where π_i is given by (1) and k_{obs} is the number of observed data. As can be seen, in the complete-data case estimation would be based on standard MPL methods, which implies that we know how to perform the M-step: we just have to estimate a logistic model as if the observations were independent.

Denoting with $\mathbf{Y}_{mis}$ the (k_{mis}) -vector of the missing data ($k_{mis} = k^2 - k_{obs}$), with $\mathbf{Y}_{obs}$ ($k_{obs} \times 1$) the observed data and with $\mathbf{Y} = (\mathbf{Y}'_{mis}, \mathbf{Y}'_{obs})'$ ($k^2 \times 1$) the complete data, the E-step requires to compute the conditional expectation of the complete pseudo log-likelihood function $pl_c = \log(PL_c)$, given the current values of the parameters and the observed data: $E_{\boldsymbol{\pi}^{(t)}}\{pl_c(\mathbf{Y} | \mathbf{y}_{obs}, \boldsymbol{\pi}^{(t)})\}$.

As pl_c is not linear in the missing data, its conditional expectation cannot be obtained by computing the conditional expectation of the missing data $E_{\boldsymbol{\pi}^{(t)}}\{\mathbf{Y}_{mis} | \mathbf{y}_{obs}, \boldsymbol{\pi}^{(t)}\}$ and plugging it into the complete pseudo log-likelihood function. In order to compute it, we

start by writing down explicitly the conditional expectation:

$$\begin{aligned}
& E_{\boldsymbol{\pi}^{(t)}} \{pl_c(\mathbf{Y}|\mathbf{y}_{obs}, \boldsymbol{\pi}^{(t)})\} = \\
& = E_{\boldsymbol{\pi}^{(t)}} \left\{ \sum_{i=1}^{k^2} [Y_i(\alpha_i + \boldsymbol{\gamma}'\mathbf{x} + \beta \sum_{j \in C(i)} Y_j - \log(1 + e^{\alpha_i + \boldsymbol{\gamma}'\mathbf{x} + \beta \sum_{j \in C(i)} Y_j})) - \right. \\
& \quad \left. - (1 - Y_i) \log(1 + e^{\alpha_i + \boldsymbol{\gamma}'\mathbf{x} + \beta \sum_{j \in C(i)} Y_j})] \right\}. \tag{3}
\end{aligned}$$

It is clear that (3) is analytically intractable. In such cases, a solution consists in computing it by Monte Carlo simulation (Monte Carlo EM - MCEM: Wei and Tanner 1990).

As for the MC E-step in (3), it is enough to compute the expectation of each term. The expectation of the i -th term of the sum is the univariate conditional expectation of Y_i given the values of Y in the neighboring regions and the current value of the parameters. An MC estimate is given by the average of a large number (say M) of observations simulated from the univariate conditional distribution of Y_i . Now we only have to sum up these estimates for $i = 1, \dots, k^2$ to conclude the E step. The M-step consists in maximizing the “approximate complete-data pseudo log-likelihood” pl_c .

Algorithm 1 *In the BAM setup of Definition 1, assume that k_{mis} observations of the response variable Y are missing. The t -th iteration of the EM algorithm for the estimation of the parameters of (1) is given by the following two steps:*

- *E-step: for each $i \in \{1, \dots, k_{mis}\}$, simulate $y_{i1}, \dots, y_{iB} \sim \text{Bin}(B; \pi_i^{(t)})$;*
- *M-step: perform MPL estimation using the observed response variables y_i ($i = 1, \dots, k_{obs}$), the simulated response variables y_{ij} ($i = 1, \dots, k_{mis}$, $j = 1, \dots, M$), and the auxiliary variables $\mathbf{x}$; the last auxiliary variable, given by $\sum_{l \in C(i)} y_l$, is formed by summing the observed (when available) or the simulated values.*

The observed pseudo-likelihood is the likelihood of a logistic model, so that the first and second derivatives of its logarithm can be found in closed form. Thus, the information matrix corresponding to the observed pseudo-likelihood is given by $\mathbf{I}(\boldsymbol{\theta}) = \mathbf{X}'\mathbf{W}\mathbf{X}$, where $\mathbf{W} = \text{diag}(N_1\pi_1(1 - \pi_1), \dots, N_k\pi_{k_{obs}}(1 - \pi_{k_{obs}}))$, $\mathbf{X}$ is the $(k_{obs} \times m)$ matrix of the explanatory variables and $N_1 = \dots = N_{k_{obs}} = 1$. The observed information matrix $\mathbf{I}(\hat{\boldsymbol{\theta}})$ is obtained by substituting $\boldsymbol{\theta}$ with the MPLE $\hat{\boldsymbol{\theta}}$ in $\mathbf{I}(\boldsymbol{\theta})$, and asymptotic standard errors are computed as $\text{se}(\hat{\boldsymbol{\theta}}) = \sqrt{\text{diag}(\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1})}$. However, the limiting distribution of any test statistic based on these standard errors is unknown; a possible solution to this difficulty consists in using Monte Carlo simulation.

References

- Bee, M. and Espa, G. (2008) A Monte Carlo EM algorithm for the estimation of a logistic auto-logistic model with missing data, *Letters in Spatial and Resource Sciences*, 1, 45-54.
- Besag, J. (1975) Statistical analysis of non-lattice data, *The Statistician*, 24, 179-195.
- Besag, J. (1977) Efficiency of pseudo-likelihood estimators for simple Gaussian fields, *Biometrika*, 64, 616-618.

- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
- Geman, S. and Graffigne, C. (1987) Markov random field image models and their applications to computer vision. In Gleason, A.M. (ed.), *Proceedings of the International Congress of Mathematicians*, vol. II, pp. 1496-1517. American Mathematical Society.
- Haining, T. (2003) *Spatial Data Analysis: Theory and Practice*, Cambridge University Press, Cambridge.
- Wei, G.C.G. and Tanner, M.A. (1990) A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithm, *Journal of the American Statistical Association*, 85, 699-704.