Two-step centered spatio-temporal auto-logistic regression model

Anne Gégout-Petit, Shuxian Li

Two-step centered spatio-temporal auto-logistic regression model

Anne Gégout-Petit[1] & Shuxian Li[2]

[1] *Université de Lorraine, Institut Elie Cartan, Faculté des Sciences et Technologies, B.P. 70239 54506 VANDOEUVRE LES NANCY CEDEX*
[2] *UMR SAVE INRA, Bordeaux Sciences Agro*

Abstract. In our study, we focus on spatio-temporal causal auto-logistic model and proposed a two-step-centered parametrization version of it. We study the existence of the joint law according to the conditional marginals. The simulation study show that the one-step model can not reflect the temporal data structure when both spatial and temporal dependance are strong, while for the two-step model, there is an adequate agreement between the data structure and the temporal large-scale structure. The results of estimation for simulated lattices over years were performed by expectation-maximization (EM) pseudo-likelihood in two stages. They show that under the two-step centered parametrization, the inference for parameters of both temporal and spatial regressions are accurate, while under one-step centered parametrization their inference are always conflicting.

Keywords: Spatio-temporal modelling; autologistic model; Markov random field; large-scale, small-scale model structure; binary response.

1 Introduction

Since spatial and spatio-temporal binary data are commonly existing in nature, the models on such data had drawn large interests of scientists from various fields such as ecology, epidemiology and image analysis, during past years. 40 years ago, Besag [2] firstly proposed an auto-logistic model for spatial binary data, assuming a simple dependence on surrounding neighbors. This model was proved to be very useful and then was extended in order to integrate the regression on the covariates [6,7] and temporal dependance as in [9].

However, the non-centered parametrization of the auto-logistic regression models present parameter interpretation difficulties across varying levels of statistical dependence. This problem has been first pointed out by Caragea [3] who proposed a centered parametrization for spatial auto-logistic regression model to overcome this difficulty. Wang [8] developed a centered spatio-temporal auto-logistic regression model but (probably due to constrain about existence of the joint law) this model depends both on the past and future and is not well adapted to real world problems modelling.

In this paper, we propose a two-step centered auto-logistic causal model and will show its advantage over the existing centered parametrization for the auto-logistic models only depending on the past. We tackle the existence of the joint distribution but present the expectation-maximization pseudo-likelihood estimation in two steps. The paper is organized as follows: first of all, after a brief introduction about auto-logistic models and the interest to center them in some cases, we will present the two-step centered auto-logistic model. Then we show several

simulations to compare the spatio-temporal auto-logistic model depending on past under one-step and two-step centered parametrization. Afterward we present the results of estimation via maximization of the pseudo-likelihood.

2 Centered autologistic models

We are interested by a set of variables $\mathbf{Z} = \{Z_i : i = 1, ..., n\}$ indexed by a lattice with n points denoted by $S = \{1, \ldots, n\}$. A spatial auto-model for $\mathbf{Z}$ is given by the n marginal conditional laws of the Z_i's given the other variables $\{Z_j, j \neq i]\}$. If Z_i is binary, we speak about auto-logistic model and we have to model the $p_i = \mathbb{P}(Z_i = 1|Z_j, j \neq i)$ for $1 \leq i \leq n$. Hammesley-Clifford theorem (see for instance [2] or [4]) makes the link between auto-model and Markov Random Field and gives some conditions for the existence of the joint law. If probability p_i does not depend on the whole set of variables $\{Z_j, j \neq i]\}$ but only on the neighborhood of i denoted by N_i, we have $p_i = \mathbb{P}(Z_i = 1|Z_j, j \in N_i)$ and the N_i's are link to the structure of cliques of the Markov Random Field.

In 1997, Gumpertz followed by many other papers, introduce covariates in the model leading to

$$p_i = \frac{exp(\mathbf{X}_i^T\boldsymbol{\beta} + \sum_{j\in N_i}\beta_{ij}Z_j)}{1 + exp(\mathbf{X}_i^T\boldsymbol{\beta} + \sum_{j\in N_i}\beta_{ij}Z_j)} \qquad \Leftrightarrow \qquad logit(p_i) = \underbrace{\mathbf{X}_i^T\boldsymbol{\beta}}_{largescale} + \underbrace{\sum_{j\in N_i}\beta_{ij}Z_j}_{smallscale}$$

where large-scale component is the overall level of a process given also by covariates and the small-scale component take into account high-order parts of the data structure like variances or covariances. Owever, Caragea [3] made the remark that if we put

$$c_i = \frac{exp(\mathbf{X}_i^T\boldsymbol{\beta})}{1 + exp(\mathbf{X}_i^T\boldsymbol{\beta})} \qquad \text{(independent case)},$$

then $log[\frac{p_i/(1-p_i)}{c_i/(1-c_i)}] = \sum_{j\in N_i}\beta_{ij}Z_j$, and odds of $Z_i = 1$ relative to the independence model increases for any nonzero neighbors, and can never decrease. It is not reasonable if most of neighbors are zeros and could bias the realizations towards 1. For this reason, they proposed the centered model (1) in 2009 that allows the p_i's to increase or decrease around the mean behavior (large scale) of the model.

$$logit(p_i) = \mathbf{X}_i^T\boldsymbol{\beta} + \rho\sum_{j\in N_i}(Z_j - \frac{exp(\mathbf{X}_j^T\boldsymbol{\beta})}{1 + exp(\mathbf{X}_j^T\boldsymbol{\beta})}) \tag{1}$$

We study now the opportunity to propose an autologistic spatio-temporal model that is centered in two steps, the first one according to the current covariates and the second one according to the regression on the past. It is given by the marginal conditional laws $[Z_{i,t}|Z_{j,t}; j \neq i, \mathbf{Z}_{t-1}, \mathbf{X}_t] = [Z_{i,t}|Z_{j,t} : j \in N_i, Z_{i,t-1}, \mathbf{X}_t]$ and where the status at point i is influenced by the past at point i and the the current neighborhood status of N_i. More precisely:

$$logit(p_{i,t}) = \underbrace{\mathbf{X}_{i,t}^T\boldsymbol{\beta}}_{\text{reg covariates}} + \underbrace{\rho_2 Z_{i,t-1}}_{\text{reg past}} + \underbrace{\rho_1 \sum_{j\in N_i}\left(Z_{j,t} - \frac{exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta} + \rho_2 Z_{ij,t-1})}{1+exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_2 Z_{j,t-1})}\right)}_{\text{auto-reg centered}} \tag{2}$$

To prove the existence of the joint law of the $Z_{i,t}$, we have to consider the process as a Markov chain of Markov field. Framework of Guyon-Hardouin [5] is useful for this and allows us to express the transition probabilities from value $\mathbf{y}$ to value $\mathbf{z}$ according to the covariates at time t. They are given by

$$\mathbb{P}(\mathbf{y},\mathbf{z} \mid \mathcal{F}_t^X) = C(y,\mathcal{F}_t^X)\exp(\underbrace{\sum_{i\in S} z_i \mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_1 \mathbb{1}_{j\in N_i} z_i z_j}_{\text{instantaneous}} + \underbrace{\sum_{i\in S} z_i(\rho_2 y_i - \rho_1 \sum_{j\in N_i}\frac{exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta} + \rho_2 y_j)}{1+exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta} + \rho_2 y_j)})}_{\text{past}})$$

3 Simulation and inference

We will use Perfect sampling preceded by a Gibbs sampling step for the burn-in to simulate data and compare the large scale structure measured by the mean overall the points of the lattice in model (2) and the following one-step model

$$logit(p_{i,t}) = \mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_2 Z_{i,t-1} + \rho_1 \sum_{j\in N_i}\left(Z_{j,t} - \frac{exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta})}{1+exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta})}\right) \tag{3}$$

Simulation were performed on a 20×20 lattice and we find that this centering is fairly accurate to capture the small-scale effect.

For inference, because of the intractable constant in joint distribution, we deal with the pseudo-likelihood and a Expectation Maximisation algorithm to infer the parameters. We give in Tables 1 and 2 the result of inference and see that the two-step centered model is more suitable for estimation.

	β_1	β_2	ρ_1	ρ_2
mean	-1.7917(-1.5)	0.2322(0.5)	0.1941(0.8)	0.8400(0.8)
variance	0.0221	0.0013	0.0421	0.0159

Table 1: One-step centered model simulation estimated by PL, (true values in the brackets).

	β_1	β_2	ρ_1	ρ_2
mean	-1.4835(-1.5)	0.5167(0.5)	0.7858(0.8)	0.7994(0.8)
variance	0.0117	0.0057	0.0061	0.0084

Table 2: two-step centered model simulation estimated by PL, (true values in the brackets).

4 Discussion

In this chapter, we developed a two-step centered spatio-temporal auto-logistic regression model to fit the binary spatio-temporal data over a lattice with exogenous covariates. The marginal data structure of the model under such centered parametrization can better reflect the large-scale structure, and in this way, the correct interpretation of the parameters is ensured. The objective of the work is to use it for the analysis of epidemiological data about esca disease for vines. But lattice in this context are very large (around 2000 vines over 12 years).We have to develop more adapted methods of inference. Here we benefited the easy calculation of pseudo-likelihood function to obtain a numerical estimation of the model which is statistically imperfect. In perspective, several methods based on approximation likelihood could be tried to estimate our models. Recently, Bee [1] proposed a AMLE (Approximate Maximize Likelihood Estimation) which is very interesting for our case. The advantage of AMLE method is since the only requirement of approximating maximum likelihood estimates is the ability to simulate the model to be estimated. It does not need to evaluate the likelihood function, only sufficient statistics but no joint distribution need to be clarified. When the inference of this model will be improved and adapted to large data set, we will apply the model to esca data in order to test the dependence on neighbors and the effects of environmental covariates.

Bibliographie

[1] Bee, M., Espa, G., and Giuliani, D. (2015). *Approximate maximum likelihood estimation of the autologistic model.* Computational Statistics & Data Analysis, 84:14-26.
[2] Besag, J. (1974). *Spatial interaction and the statistical analysis of lattice systems.* Journal of the Royal Statistical Society. Series B (Methodological), 192-236
[3] Caragea, P. C. and Kaiser, M. S. (2009). *Autologistic models with interpretable parameters.* Journal of agricultural, biological, and environmental statistics, 14(3):281-300
[4] Gaetan, C. and Guyon, X. (2008). Modélisation et statistique spatiales, volume 63. Springer.
[5] Guyon, X. and Hardouin, C. (2002). Markov chain markov field dynamics: models and statistics. Statistics: A Journal of Theoretical and Applied Statistics, 36(4):339-363.
[6] Gumpertz, M. L., Graham, J. M., and Ristaino, J. B. (1997). *Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variables on disease presence.* Journal of Agricultural, Biological, and Environmental Statistics, 131-156
[7] Huffer, F. W. and Wu, H. (1998). *Markov chain monte carlo for autologistic regression models with application to the distribution of plant species.* Biometrics, 509-524
[8] Wang, Z. and Zheng, Y. (2013). Analysis of binary data via a centered spatial-temporal autologistic regression model. *Environmental and Ecological Statistics*, 20(1): 37-57.
[9] Zhu, J., Huang, H.-C., andWu, J. (2005). *Modeling spatial-temporal binary data using markov random fields.* Journal of Agricultural, Biological, and Environmental Statistics, 10(2): 212-225.