G.E. Kelly

A model for spatial binary data

G. E. Kelly[1], J. Pan[2]

[1]School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland

[2]School of Mathematics, University of Manchester, Oxford Road, Manchester M13 9PL, U.K.

Abstract

This is a description of modelling spatial binary data that uses existing methodology in a unique way. Firstly, a model for a realization of a binary random field is considered where the correlations satisfy the Fréchet–Hoeffding bounds. The binary variables are related to latent variables that have a Matérn spatial correlation via a multivariate probit model. Thus, both the marginal means of the binary variables and their spatial distances contribute to their correlations. Secondly, a profile likelihood approach to maximizing the likelihood is carried out as a full likelihood approach is computationally not feasible. The model is fitted to TB infection data in cattle herds. Minimum and maximum values of the binary correlations are estimated and associated Euclidean distances recorded.

Keywords: Binary variables; Bovine TB; Latent variables; Profile likelihood; Spatial correlation

1 Introduction

Consider a multivariate binary response vector $Y = (Y_1, \ldots, Y_n)$ with specified marginal means $\mu_i = E(Y_i), (i = 1, \ldots, n)$. For binary data where observations are spatially correlated, a simple parametric form for the correlation structure, such as is used for Gaussian random fields, is not readily

available. Limits to Pearson correlations between Bernoulli random variables are well known. Valid bivariate correlations r_{ij} satisfy the well known Fréchet–Hoeffding bound (McDonald, 1993)

$$r_{ij} \leq \min\{\psi_i/\psi_j, \psi_j/\psi_i\} = \bar{r}_{ij},$$

where $\psi_i = (\mu_i/(1-\mu_i))^{1/2}$.
Several authors have addressed the problem of modelling correlated binary variables but did not take the above constraint into account. The two main approaches used are conditional models and generalized estimating equations (GEE). A review may be found in De Oliveira (2020). Here we consider a likelihood approach via multivariate probit models. A multivariate probit model was proposed by Chaganty and Joe (2004) for samples of longitudinal binary data but they have not been used in the context of spatial binary data where the observed response is a single n-dimensional vector.

2 A spatial model

Let $\{Y(s) : s \epsilon D\}$ be a binary random field where for any $s \epsilon D$, $Y(s)$ takes two values, coded as 0 and 1. Let $\{Z(s) : s \epsilon D\}$ be an unobserved Gaussian random field with mean function $\nu(s)$ and covariance function $C(s,u) = \sigma^2\rho(s,u)$ where $\rho(s,u)$ is a correlation function. We let $Y(s) = I(Z(s) > 0)$ and without loss of generality set $\sigma^2 = 1$ (De Oliveira, 2020). Let $\nu(s) = \Phi^{-1}(\mu(s))$, where $\mu(s) = E(Y(s))$. We assume the mean response $\mu(s)$ is associated with the measurements of explanatory variables X through a link function $\Phi^{-1}(\mu) = X\beta$.
Consider a realisation of this model, $Y = (Y_1, \ldots, Y_n)$ where $Y_i = I(Z_i - v_i \geq -v_i = -\beta^T x_i), i = 1, \ldots, n$, D $\subset$ R^2 and denote the Euclidean distance between observations Y_i and Y_j by $d(i,j)$. There is a 1:1 correspondence between the correlations of the Z's and the correlations of the Y's.

$$\text{corr}(Y_i, Y_j) = r_{ij} = \frac{\Phi_2(v_i, v_j, \rho_{ij}) - \Phi(v_i)\Phi(v_j)}{[\Phi(v_i)(1-\Phi(v_i))\Phi(v_j)(1-\Phi(v_j))]^{1/2}} \qquad (1)$$

where $\Phi_2(\omega_1, \omega_2; \rho_{ij})$ and $\Phi(\omega)$ denote the standardized bivariate normal with correlation $\Sigma = (\rho_{ij})_{n\times n}$ and the univariate standard normal distribution functions respectively. The latent correlation is assumed to have a form such as $\rho(s,u) = \exp(-d/\gamma)$ where $d = \|s-u\|$, then $\rho_{ij} = \exp(-d(i,j)/\gamma)$ or

more generally we can consider a Matérn latent correlation with scale ϕ and smoothness κ parameters. Thus, a valid spatial correlation function for the binary variables is established. We can then compute the log likelihood of the data $Pr(Y = y; X, \beta, \Sigma)$, $y\epsilon\{0,1\}^n$, based on the multivariate probit model, since $Y_i = I(Z_i - v_i \leq v_i = \beta' x_i), i = 1, \ldots, n$, as in Chaganty and Joe (2004). The likelihood is given by

$$\int_{A(y_1)} \ldots \int_{A(y_n)} \left(\frac{1}{2\pi}\right)^{n/2} |\Sigma|^{-1/2} \exp\{(z - X\beta)^T \Sigma^{-1} (z - X\beta)\} dz$$

$$\begin{array}{rcll} A(y_i) & = & (0, \infty], & \text{if } y_i = 1 \\ . & = & (-\infty, 0), & \text{if } y_i = 0 \\ & & i = 1, \ldots, n. & \end{array}$$

The integrand has a standard multivariate normal distribution with correlation matrix Σ.

The negative log-likelihood can be minimized for the unknown parameters, ϕ, κ and β. The Hessian can also be obtained and inverted to get standard errors of the parameter estimates. As the full likelihood may not be very smooth in the parameters, the likelihood is maximized in two stages. Let θ_1 denote the spatial parameter κ and let θ_2 denote ϕ and the β parameters and let $\theta = (\theta_1, \theta_2)$. Given $L(\theta_1, \theta_2)$ we compute the profile likelihood of θ_1 as

$$L(\theta_1) = \max_{\theta_2} L(\theta_1, \theta_2).$$

The maximum with respect to θ_1, $\hat{\theta_1}$ can then be found. We then have predicted values of our binary variables Y_i. Equation (1) can then be used to estimate the $\text{corr}(Y_i, Y_j)$.

θ_1 is related to the correlation of the latent variables but not to their mean $\nu(s)$. Therefore θ_1 and β are orthogonal and their maximum likelihood estimates and associated standard errors are asymptotically independent (Cox and Reid, 1987). Therefore, θ_1 was set equal to $\hat{\theta_1}$, and an estimated likelihood of θ_2 is $L_e(\theta_2) = L(\hat{\theta_1}, \theta_2)$ and asymptotically it can be treated as a standard likelihood to obtain estimates of the standard errors of the ϕ, β parameters (Pawitan, 2001, Section 10.7).

The Genz-Bretz randomized quasi-Monte-Carlo procedure Genz and Bretz (2009) was used to calculate multivariate normal probabilities using the `mvtnorm` library in `R`.

Genz uses a sequence of three transformations to transform the original integral into an integral over a unit hypercube, the rectangle probability can

then be successfully evaluated via importance sampling based on a (d-1) - variate standard uniform random sample. Genz later improved on this with the use of a randomized quasi Monte Carlo method with the use of antithetic variates.
Asymptotic and small-sample efficiency calculations by Genz show that this method is nearly as efficient as maximum likelihood for fully specified multivariate normal copula-based models. In addition, Nikoloulopoulos (2013) showed this method is highly efficient for a high dimensional discrete response with aggregated data up to dimension d=225.

3 Example: Cattle data

In Ireland and the UK, bovine bTB infects cattle and wildlife badgers (*Meles meles linnaeus*) and badgers contribute to the spread of the disease in cattle and perhaps vice versa. Here data are drawn from the Four Area Project (FAP), a formal badger removal project undertaken in four counties in Ireland from September 1997 to August 2002 (Griffin et al., 2005).

- In the FAP, badgers were pro-actively removed from the removal areas and in the matching reference areas culling was minimal.
- Badger information is not available for reference areas.
- For illustrative purposes we consider one area only, the removal area of Cork, an area of approximately 400 km ^{2}.
- The GIS coordinates for cattle herds was taken as the centroid of the main parcel of land where the herd was located.
- A herd is recorded as TB positive if any cattle is tested positive and similarly a sett is infected if any badger in it tests positive.
- For these data we wish to establish if bTB incidence in cattle is spatially correlated.

Data for the combined first two culling years are considered as the majority of badgers were captured in this period. There were 417 cattle herds with 12.2% TB positive. A spatial model with Matérn correlation function as described above was then fitted, with covariates previous history of infection in

the herd (0/1) and log(herd size) and log(herd size)× easting. The maximum likelihood estimator had values $\phi = 0.1412(0.2242)$ and $\kappa = 3.0$. The β estimates are given in Table 1 together with the approximate standard errors. The standard errors in Table 1 do not account for κ being estimated. They

Table 1: Predictors of TB infection in cattle herds using a spatial model

	Estimate	S.E.[a]
Intercept	-1.5199	0.1025
ph[b]	0.5308	0.2413
loghsc[c]	7.1324	0.0162
loghsc × x[d]	-5.1677	0.0620

[a]standard error; [b]previous history of TB infection in the herd;
[c]log of herd size; [d]x, easting GIS coordinate of the herd location.

are also considerably smaller than those in an independence model perhaps because considerable variation has been accounted for by spatial variation. The correlation for the binary variables had a max = 0.5491 (with a distance of 0.18km) and a min of almost 0 (achieved by several pairs of herds). The values are consistent with the fact the correlation between the binary variables is lower than the corresponding correlation between the latent variables, as they should be (Chaganty and Joe, 2004). The practical range for the binary data was estimated to be 0.91km. Note that correlations in the binary data may not strictly decrease with distance while the correlations between the latent variables do.
Note also that when sett size is accounted for, no spatial correlation or any correlation between badger setts was found. The estimate of ϕ was zero indicating an independence model.

4 Discussion

As noted by both De Oliveira (2020) and Diggle and Ribeiro (2007) the inference about the binary realization depends heavily on the correlation structure of the underlying Gaussian random field. Although a grid search was used here to obtain optimal values of the Matérn latent correlation, it is not feasible to carry out a grid search for the entire parameter space or to estimate parameters of the spatial field simultaneously with the fixed effects.

Our results are consistent with Moustakas and Evans (2017), that found badgers mainly spread the disease locally while cattle infected both locally and across longer distances. They report the mean distance that an infected badger individual spreads TB is 0.92 km year^{-1}, SD = 0.62 and the mean distance that an infected cattle individual contributes into the spread of the disease is 2.34 km year^{-1}, SD = 0.98.
Spatial association of infection persisted during the proactive badger culling period in cattle herds in Cork. Possible explanations, include residual (persistent but undetected) infection in cattle, and ongoing herd-to-herd transmission. Griffin et al. (2005) found proactive culling of badgers decreases TB incidence in cattle herds. However, the scale and direction of culling remains an important issue and establishing the magnitude and range of spatial correlation may help to inform this issue.

References

Chaganty, N. R. and Joe, H. (2004). Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society* B **66**, 851–60.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditionalinference. *Journal of the Royal Statistical Society* B **49**, 1–39.

De Oliveira, V. (2020). Models for Geostatistical Binary Data: Properties and Connections. *American Statistician* **74**, 72–79.

Diggle, P. and Ribeiro, P. (2007). *Model-based geostatistics*. New York:Springer.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Heidelberg:Springer-Verlag, Lecture Notes in Statistics, Vol. 195.

Griffin, J., Williams, D. H., Kelly, G. E.,Clegg, T.,O Boyle, I.,Collins, J. D. and More, S. J. (2005). The impact of badger removal on the control of tuberculosis in cattle herds in Ireland. *Preventive Veterinary Medicine* **67**, 236–66.

McDonald, B.W. (1993). Estimating logistic regression parameters for bivariate binary data. *Journal of the Royal Statistical Society* B **55**, 391–97.

Nikoloulopoulos, A. K. (2013). On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood, *Journal of Statistical Planning and Inference* **143**,1923–37.

Moustakas, A. and Evans, M. R. (2004). A big-data spatial, temporal and network analysis of bovine tuberculosis between wildlife (badgers) and cattle, *Stochastic and Environmental Research and Risk Assessment* **31**, 315—28

Pawitan, Y. (2001). *In all likelihood*. Oxford: Oxford University Press.