A spatial model for rare binary events

September 17, 2015

1 Introduction

The goal in binary regression is to relate a set of covariates with the response using a link function. Two very commonly used types of binary regression include logistic regression and probit regression. One limitation to these link functions is that they assume the data are symmetric which may not always be the case. The complementary log-log (cloglog) link function is an asymmetric alternative to logistic and probit regression. More recently, Wang and Dey (2010) introduced the generalized extreme value (GEV) link function for rare binary data. The GEV link function introduces a new shape parameter to the link function that controls the degree of asymmetry. The cloglog link is a special case of the GEV link function when the shape parameter is 0.

Want to make the case in this paragraph that spatial logistic and probit models are not appropriate because asymptotic dependence is 0. Spatial logistic and probit models are commonly presented using a hierarchical model citation. In the hierarchical framework, spatial dependence is accounted for by assuming that there exists an underlying latent Gaussian process, and conditioned on this process, observations are independent. However, for extremal distributions, Gaussian processes do not appropriately capture spatial dependence because there is no asymptotic dependence regardless of the strength of the correlation in the bulk of the data. Instead, when modeling spatial extremes, a max-stable process is preferred because it allows for asymptotic dependence citation. Therefore, we choose to incorporate a latent max-stable process to model the spatial dependence.

Paragraph outlining the structure of the paper

2 Spatial dependence for binary regression

2.1 Non-rare data

In this section we present the spatial logistic and probit models.

In the case that n is large, low-rank predictive process models can be used to ease the computation.

3 Rare binary data

In this section we extend the GEV link function to allow for spatial dependence. Let $Y_i \in \{0, 1\}$ be the binary response at spatial location $\mathbf{s}_i \in \mathcal{D}$, We assume that $Y_i = I(Z_i > 0)$ where $I(\cdot)$ is an indicator function, and the marginal distribution of Z_i is

$$Z_i \sim \text{GEV}(\mathbf{X}_i\boldsymbol{\beta}, 1, \xi) \tag{1}$$

where $\mathbf{X}_i$ be the associated p-vector of covariates with first element equal to one for the intercept. We fix the shape to be 1 for identifiability. Therefore, the marginal probability of an event is

$$\pi_i = 1 - \exp\left[-\left(1 - \xi\mathbf{X}_i\boldsymbol{\beta}\right)^{-1/\xi}\right] \tag{2}$$

as in Wang and Dey (2010).

To incorporate spatial dependence into the model, we consider the hierarchical max-stable process of Reich and Shaby (2012). The spatial dependence is determined by the joint distribution of $\mathbf{Z} = (Z_1, \ldots, Z_n)$,

$$G(\mathbf{z}) = \text{P}[Z_1 < z_1, \ldots, Z_n < z_n] = \exp\left\{-\sum_{l=1}^{L}\left[\sum_{i=1}^{n}\left(\frac{w_l(\mathbf{s}_i)}{z_i}\right)^{1/\alpha}\right]^{\alpha}\right\}, \tag{3}$$

where $\mathbf{z} = (z_1, \ldots, z_n)$, $w_l(\mathbf{s}_i)$ are a set of weights that determine the spatial dependence structure and are discussed further in Section 3.1, and $\alpha \in (0, 1)$ determines the strength of dependence, with α near zero giving strong dependence and $\alpha = 1$ giving joint independence. This is a special case of the multivariate GEV distribution with asymmetric Laplace dependence function (Tawn, 1990). One nice feature to this hierarchical model is that the lower-dimensional marginal distributions also follow a multivariate extreme value distribution. More importantly, at a single site i, the marginal distribution gives $P(Y_i = 1) = 1 - \exp\left\{-\frac{1}{z_i}\right\}$ which is the same as the marginal distributions given by Wang and Dey (2010).

3.1 Weight functions

Many weight functions are possible, but the weights must be constrained so that $\sum_{l=1}^{L} w_l(\mathbf{s}_i) = 1$ for all $i = 1, \ldots, n$ to preserve the marginal GEV distribution. The weights $w_l(\mathbf{s}_i)$ in (3) should vary smoothly across space to induce spatial dependence. For example, Reich and Shaby (2012) take the weights to be scaled Gaussian kernels with knots $\mathbf{v}_l$, that is

$$w_l(\mathbf{s}_i) = \frac{\exp\left[-0.5\left(||\mathbf{s}_i - \mathbf{v}_l||/\rho\right)^2\right]}{\sum_{j=1}^{L} \exp\left[-0.5\left(||\mathbf{s}_i - \mathbf{v}_j||/\rho\right)^2\right]}. \tag{4}$$

The kernel bandwidth $\rho > 0$ determines the spatial range of the dependence, with large ρ giving long-range dependence and vice versa.

4 Joint distribution

The joint likelihood of Y is generally computationally challenging to compute. In this paper, we combine the random effect representation with MCMC; however, when exact expressions of the likelihood are available, better methods could be used. In section 4.1, we give an exact expression in the case where there are only

two spatial locations which is useful for constructing a pairwise composite likelihood and studying spatial dependence. For more than two locations, we are also able to compute the exact likelihood when the number of locations is large but the number of events is small, as might be expected for very rare events.

4.1 Bivariate distribution

Then in a bivariate setting, the probability of observing a joint exceedances as a function of α is

$$\mathrm{P}(Y_i = 1, Y_j = 1) = 1 - \exp\left\{-\frac{1}{z_i}\right\} - \exp\left\{-\frac{1}{z_j}\right\} + \exp\left\{-\sum_{l=1}^{L}\left[\left(\frac{w_l(\mathbf{s}_i)}{z_i}\right)^{1/\alpha} + \left(\frac{w_l(\mathbf{s}_i)}{z_j}\right)^{1/\alpha}\right]^{\alpha}\right\} \tag{5}$$

5 Quantifying spatial dependence

I still need to incorporate Brian's suggestions here In the literature on extremes, one common metric to describe the bivariate dependence is the χ statistic of Coles et al. (1999). The χ statistic between two observations z_1 and z_2 is given by

$$\chi(\mathbf{s}_1, \mathbf{s}_2) = \lim_{c \to \infty} = P(Z_1 > c | Z_2 > c). \tag{6}$$

However, in this latent variable approach, $\lim_{c\to\infty}$ may not be the most reasonable metric because the observed data are a series of zeros and ones. Therefore, we chose the κ statistic of Cohen (1960) defined by

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{7}$$

where $P(A)$ is the joint probability of agreement and $P(E)$ is the joint probability of agreement under an assumption of independence. We believe this measure of dependence to be reasonable because,

$$\lim_{\beta_0 \to \infty} \kappa(h) = \chi(h) = 2 - \vartheta(\mathbf{s}_i, \mathbf{s}_j) \tag{8}$$

where β_0 is the intercept from $\mathbf{X}^T\boldsymbol{\beta}$ and $\vartheta(\mathbf{s}_i, \mathbf{s}_j) = \sum_{l=1}^{L} \left[w_l(\mathbf{s}_i)^{1/\alpha} + w_l(\mathbf{s}_j)^{1/\alpha} \right]^{\alpha}$ is the pairwise extremal coefficient given by Reich and Shaby (2012) (see Appendix A.2). In the case of complete dependence, $\kappa = 1$, and in the case of complete independence, $\kappa = 0$.

6 Computation

For small K we can evaluate the likelihood directly. In the random effects model, the expression for the joint density conditional on θ is

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} \left[\exp\left\{ \sum_{l=1}^{L} A_l \left(\frac{w_l(\mathbf{s}_i)}{z_i} \right)^{1/\alpha} \right\} \right]^{1-Y_i} \left[1 - \exp\left\{ \sum_{l=1}^{L} A_l \left(\frac{w_l(\mathbf{s}_i)}{z_i} \right)^{1/\alpha} \right\} \right]^{Y_i}. \tag{9}$$

7 Simulation study

For our simulation study, we generate $n_m = 100$ datasets under 4 different settings to explore the impact of rareness of observations, sample size, and knot spacing. We consider two degrees of rareness $\pi = 0.01, 0.05$ and two sample sizes $n_s = 1000, 2000$. For the different knot spacings, we use knots in $[0, 1] \times [0, 1]$ on a 21×21 grid and 31×31 grid. For each dataset, we fit the model using three different methods, spatial logistic regression, spatial probit regression, and the proposed spatial GEV method. In each case, we fit the model using Bayesian methods with proper, but fairly uninformative priors. For each method, we fit

the model using 75% of the observations as a training set, and the remaining observations are used as a validation set to assess the model's predictive power.

8 Data analysis

For the data analysis, we consider data from the eBirds dataset, a citizen-based observation network of bird sitings in the United States (Sullivan et al., 2009). The data are publicly available from `http://ebird.org`. We use data from 2002, and focus specifically on cattle egrets and vesper sparrows.

9 Conclusions

Acknowledgments

A Appendices

A.1 Derivation of the likelihood

We use the hierarchical max-stable spatial model given by Reich and Shaby (2012). If at each margin, $Z_i \sim \text{GEV}(1,1,1)$, then $Z_i|\theta_i \overset{indep}{\sim} \text{GEV}(\theta, \alpha\theta, \alpha)$. As defined in section 6, we reorder the data such that $Y_1 = \ldots = Y_K = 1$, and $Y_{K+1} = \ldots = Y_n = 0$. Then the joint likelihood conditional on the random effect θ is

$$
\begin{aligned}
P(Y_1 = y_1, \ldots, Y_n = y_n) &= \prod_{i \le K} \left\{ 1 - \exp\left[- \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \right\} \prod_{i > K} \exp\left[- \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \\
&= \exp\left[- \sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] - \exp\left[- \sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \sum_{i=1}^{K} \exp\left[- \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \\
&\quad + \exp\left[- \sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \sum_{1 < i < j \le K} \left\{ \exp\left[- \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} - \left(\frac{\theta_j}{z_j} \right)^{1/\alpha} \right] \right\} \\
&\quad + \cdots + (-1)^K \exp\left[- \sum_{i=1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \qquad (10)
\end{aligned}
$$

93 Finally marginalizing over the random effect, we obtain

$$
\begin{aligned}
P(Y_1 = y_1, \ldots, Y_n = y_n) &= \int G(\mathbf{z}|\mathbf{A}) p(\mathbf{A}|\alpha) d\mathbf{A}. \\
&= \int \exp\left[- \sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] - \exp\left[- \sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \sum_{i=1}^{K} \exp\left[- \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \\
&\quad + \exp\left[- \sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \sum_{1 < i < j \le K} \left\{ \exp\left[- \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} - \left(\frac{\theta_j}{z_j} \right)^{1/\alpha} \right] \right\} \\
&\quad + \cdots + (-1)^K \exp\left[- \sum_{i=1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] p(\mathbf{A}|\alpha) d\mathbf{A}. \qquad (11)
\end{aligned}
$$

94 Consider the first term in the summation,

$$\int \exp\left\{-\sum_{i=K+1}^{n}\left(\frac{\theta_i}{z_i}\right)^{1/\alpha}\right\} p(\mathbf{A}|\alpha)d\mathbf{A} = \int \exp\left\{-\sum_{i=K+1}^{n}\left(\frac{\left[\sum_{l=1}^{L} A_l w_l(\mathbf{s}_i)^{1/\alpha}\right)^{\alpha}\right]^{1/\alpha}}{z_i}\right\} p(\mathbf{A}|\alpha)d\mathbf{A}$$

$$= \int \exp\left\{-\sum_{i=K+1}^{n}\sum_{l=1}^{L} A_l\left(\frac{w_l(\mathbf{s}_i)}{z_i}\right)^{1/\alpha}\right\} p(\mathbf{A}|\alpha)d\mathbf{A}$$

$$= \exp\left\{-\sum_{l=1}^{L}\left[\sum_{i=K+1}^{n}\left(\frac{w_l(\mathbf{s}_i)}{z_i}\right)^{1/\alpha}\right]^{\alpha}\right\}. \tag{12}$$

The remaining terms in equation (11) are straightforward to obtain, and after integrating out the random effect, the joint density is the density given in (**??**).

A.2 Derivation of the χ statistic

$$\begin{aligned}
\chi &= \lim_{p\to 0} \mathrm{P}(Y_i = 1|Y_j = 1) \\
&= \lim_{p\to\infty} \frac{p + p - \left(1 - \exp\left\{-\sum_{l=1}^{L}\left[(-\log(1-p)w_l(\mathbf{s}_i))^{1/\alpha} + (-\log(1-p)w_l(\mathbf{s}_j))^{1/\alpha}\right]^{\alpha}\right\}\right)}{p} \\
&= \lim_{p\to 0} \frac{2p - \left(1 - \exp\left\{\log(1-p)\sum_{l=1}^{L}\left[w_l(\mathbf{s}_i)^{1/\alpha} + w_l(\mathbf{s}_j)^{1/\alpha}\right]^{\alpha}\right\}\right)}{p} \\
&= \lim_{p\to 0} \frac{2p - \left(1 - (1-p)^{\sum_{l=1}^{L}\left[(w_l(\mathbf{s}_i))^{1/\alpha} + (w_l(\mathbf{s}_j))^{1/\alpha}\right]^{\alpha}}\right)}{p} \\
&= \lim_{p\to 0} 2 - \sum_{l=1}^{L}\left[w_l(\mathbf{s}_i)^{1/\alpha} + w_l(\mathbf{s}_j)^{1/\alpha}\right]^{\alpha}(1-p)^{-1+\sum_{l=1}^{L}\left[w_l(\mathbf{s}_i)^{1/\alpha} + w_l(\mathbf{s}_j)^{1/\alpha}\right]^{\alpha}} \\
&= 2 - \sum_{l=1}^{L}\left[w_l(\mathbf{s}_i)^{1/\alpha} + w_l(\mathbf{s}_j)^{1/\alpha}\right]^{\alpha}.
\end{aligned} \tag{13}$$

References

Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**, 37–46.

Coles, S., Heffernan, J. and Tawn, J. (1999) Dependence Measures for Extreme Value Analyses. *Extremes*, **2**, 339–365.

Reich, B. J. and Shaby, B. A. (2012) A hierarchical max-stable spatial model for extreme precipitation. *The Annals of Applied Statistics*, **6**, 1430–1451.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D. and Kelling, S. (2009) eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, **142**, 2282–2292.

Tawn, J. A. (1990) Modelling multivariate extreme value distributions. *Biometrika*, **77**, 245–253.

Wang, X. and Dey, D. K. (2010) Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *The Annals of Applied Statistics*, **4**, 2000–2023.