Fitting spatial hurdle models using a hierarchical likelihood method

Md. Moudud Alam

Lecturer in Statistics
Dalarna University.
e-mail: maa@du.se

Aug. 25, 2014.

This is a joint work with Shengqiang Guo. Thanks to Anna Skarin, SLU Uppsala, for kindly allowing us an access to the reindeer data.

Outline of the presentation

1. Main message

2. A motivating real data problem

3. Spatial hurdle mixed model and it's estimation

4. Simulation study

5. Preliminary results from reindeer pellet group data

Main Message

Spatial hurdle models with Gaussian CAR and SAR random effects can be fitted by using model fitting algorithm for ordinary generalized linear mixed models (GLMM), after some minor modifications.

A motivating real data problem

- In order to assess the effect of windmills on reindeer habitat preference a survey on reindeer pellets was conducted in Storliden (northern mountain) area in Sweden during 2009–2012.
- In the transact survey, no pellet was observed in about 85% of the sampling plots.
- In analysing this such data a natural choice would be to fit some kind of zero inflated model that also takes spatial correlation into account.

Spatial hurdle mixed model

A spatial hurdle mixed model can be presented as follows:

1. Given $\mathbf{u}_0 = \{u_{0,i}\}_{i=1}^{n}$ and $\mathbf{u}_1$ the response follows

   $$Pr\left(y_i = y | u_{0,i}, u_{1,i}, X_i\right) = \begin{cases} \mu_{0,i} & if\;\; y = 0 \\ \left(1 - \mu_{0,i}\right) TP\left(y; \mu_{1,i}\right) & if\;\; y = 1, \ldots \end{cases}$$

   where TP is a 0 truncated Poisson pmf.

2. $g_j\left(\mu_{j,i}\right) = \mathbf{X_i}\beta_j + Z_i\mathbf{u}_j;\;\; j = 0, 1$
3. $\mathbf{u}_j \sim N_n\left(\mathbf{0}, \boldsymbol{\Sigma}_j\right)$; $\mathbf{u}_0 \perp \mathbf{u}_1$; $\boldsymbol{\Sigma}_j = \tau_j\left(\mathbf{I} - \rho_j\mathbf{D}\right)^{-1}$ with known $\mathbf{D}$ gives CAR while $\boldsymbol{\Sigma}_j = \tau_j\left(\mathbf{I} - \rho_j\mathbf{D}\right)^{-1}\left(\mathbf{I} - \rho_j\mathbf{D}\right)^{-1}$ gives SAR structure for random effects.
4. A popular choice for $\mathbf{D}$ is the so called neighbourhood matrix.

Estimation of model parameters

1. It can be shown that hurdle log-likelihood factors into two parts

$$l = l_0 + l_1$$

   where l_0 is a log-likelihood of spatial binary mixed models containing parameter in μ^0 and $\mathbf{\Sigma}_0$ and l_1 is a log-likelihood of a spatial TP mixed model containing parameter in μ^1 and $\mathbf{\Sigma}_1$.
2. Already existing software packages, e.g. `hglm` package in `R`, can be used to obtain an estimate of the parameters in l_0.
3. In this talk, we will focus only on the other part, l_1.
4. Before that, let us look into a trick to reformulate a generalized linear mixed model (GLMM) with CAR/SAR random effects into a GLMM with independent but heteroskedastic random effects.

Reformulating spatial mixed model as independent random effects

1. Let $\mathbf{V}$ be the matrix of eigenvectors and $\mathbf{\Lambda} = diag\{\lambda_i\}$ be a diagonal matrix with corresponding eigenvalues of $\mathbf{D}$ then
    1. $\tau\left(\mathbf{I}-\rho\mathbf{D}\right)^{-1} = Vdiag\{\frac{\tau}{1-\rho\lambda_i}\}V^T$
    2. $\tau\left(\mathbf{I}-\rho\mathbf{D}\right)^{-1}\left(\mathbf{I}-\rho\mathbf{D}\right)^{-1} = Vdiag\{\frac{\tau}{(1-\rho\lambda_i)^2}\}V^T$
2. By reformulating the linear predictors of the above model as $\eta_{j,i} = X_i\beta + Z_i^*\mathbf{u}_j^*$ where $Z_i^* = Z_i\mathbf{V}$ we have $\mathbf{u}_j^* \sim N_n\left(\mathbf{0}, diag\{\phi_{j,i}\}\right)$ where $\nu\left(\phi_{j,i}\right) = \theta_{0,j} + \theta_{1,j}\lambda_i$, $\nu\left(\right)$ is an inverse function for CAR (and inverse-square-root for SAR), $\theta_{0,j} = \frac{1}{\tau_j}$ (and $= \frac{1}{\sqrt{\tau_j}}$ for SAR) and $\theta_{1,j} = \frac{-\rho}{\tau_j}$ (and $= \frac{-\rho}{\sqrt{\tau_j}}$ for SAR).
3. Therefore, any computational method that can fit fixed effects in the dispersion parameters of a GLMM can be used to fit the spatial GLMM.
4. Next we see the workings of EQL algorithm fitting this model.

Step 1: H-likelihood estimate of β_1 and $\mathbf{u}_1$

For fixed $\theta_{1,0}$, $\theta_{1,1}$ and an initial value for $\beta_1^{(0)}$ and $\mathbf{u}_1^{*(0)}$ the algorithm is as follows:

1. Calculate a weight matrix $\mathbf{W} = diag\{\mathbf{W_1}, \mathbf{W_2}\}$ where $$\mathbf{W}_1 = diag\{\mu_1^{(0)} + \frac{\mu_1^{(0)}\left(\exp\left[\mu_1^{(0)}\right] - \mu_1^{(0)}\left(\exp\left[\mu_1^{(0)}\right] - 1\right)\right)}{\left(\exp\left[\mu_1^{(0)}\right] - 1\right)^2}\}$$ and $$\mathbf{W}_1 = diag\{\frac{1}{\phi_{1,i}}\}$$
2. Calculate working response, $S = (\mathbf{s}, \mathbf{0})^T$ where $\mathbf{0}$ is an n-vector of 0's and $$\mathbf{s} = \eta_1^{(0)} + \mathbf{W}_1^{-1}\left(y - \mu_1^{(0)} - \frac{\mu_1^{(0)}}{\left(\exp\left[\mu_1^{(0)}\right] - 1\right)^2}\right)$$
3. Update $\beta_1^{(0)}$ and $\mathbf{u}_1^{*(0)}$ by fitting a linear model with S as the response, $\mathbf{X}^* = \begin{pmatrix} \mathbf{X} & \mathbf{Z}^* \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$ as the design matrix and $\mathbf{W}$ as the weight matrix.

Step 2: EQL estimate of $\theta_{1,0}$ and $\theta_{1,1}$

Given an estimate of $\mathbf{u}_1^*$ and corresponding hat values $h_{1,i}$ from the linear model in Step 1 the EQL estimating algorithm for $\theta_{1,0}$ and $\theta_{1,1}$ is as follows.

1. Calculate $\xi_i = \frac{u_{1,i}^{*2}}{1-h_{1,i}}$.
2. Fit a Gamma model with ξ as the response, eigenvalues of **D** as the covariate, $\nu()$ as the link function and $\frac{1-h_{1,i}}{2}$ as the weights.
3. Then iterate between Step 1 and Step 2 until convergence.
4. As an alternative to EQL one can also maximize the profile likelihood directly to obtain an estimate of $\theta_{1,0}$ and $\theta_{1,1}$.

Simulation setup

In order to study the finite sample properties of the EQL estimator for spatial TP we set up a series of simulation studies, as follows

- We chose two **D** matrices, small and large. Small one is constructed after the districts of Scotland (n=56) and the large one after the survey location in Storliden (n=357).
- We simulate 2 covariates independently from $N(0, 0.8)$ and $N(0, 0.5)$.
- Fixed effects parameter are chosen as $\beta_1 = (1.2, 0.3, -0.5)$.
- Spatial dispersion and dependence parameters are chosen as $(\tau_1, \rho_1) = (1.5, -0.2)$ and $(0.4, -0.2)$.
- For each location,$i = 1, 2, \ldots, n$, we simulate $k = 2, 5, 10, 30$ and 100 observations.
- We use 500 Monte Carlo iterations (non-convergence is compensated by additional simulation).
- We focus only on CAR random effects.

Simulation Results 1: Mean and MSE of EQL estimates with small **D**

K	$\beta_{1,0} = .12$	$\beta_{1,1} = 0.3$	$\beta_{1,2} = -0.5$	$\theta_{1,0} = 0.67$	$\theta_{1,1} = 0.13$
2	1.30*	0.30	-0.50	0.51*	0.11*
	(0.152)	(0.004)	(0.013)	(0.066)	(0.006)
5	1.28*	0.30	-0.50	0.72*	0.16*
	(0.183)	(0.001)	(0.004)	(0.047)	(0.007)
10	1.24*	0.30	-0.50	0.75*	0.16*
	(0.046)	($<$0.001)	(0.001)	(0.040)	(0.007)
30	1.22*	0.30	-0.50	0.74	0.15*
	(0.013)	($<$0.001)	($<$0.001)	(0.028)	(0.005)
100	1.22	0.30	-0.50	0.75*	0.15*
	(0.069)	($<$0.001)	($<$0.001)	(0.036)	(0.006)

Values within parenteses show MSE; $*$:significant at 5% level.

Simulation Results 2: Mean and MSE of Profile likelihod estimates with small **D**

K	$\beta_{1,0} = .12$	$\beta_{1,1} = 0.3$	$\beta_{1,2} = -0.5$	$\theta_{1,0} = 0.67$	$\theta_{1,1} = 0.13$
2	1.35*	0.30	-0.50	0.93*	0.19*
	(0.064)	(0.004)	(0.013)	(0.138)	(0.015)
5	1.29*	0.30	-0.50	0.83*	0.16*
	(0.056)	($<$0.001)	(0.003)	(0.066)	(0.008)
10	1.24*	0.30	-0.50	0.79*	0.16*
	(0.016)	($<$0.001)	(0.001)	(0.046)	(0.007)
30	1.22*	0.30	-0.50	0.75*	0.15*
	(0.013)	($<$0.001)	($<$0.001)	(0.029)	(0.005)

Simulation Results 3: Performance of EQL with small **D** but moderate $\theta_{1,j}$'s

K	$\beta_{1,0} = .12$	$\beta_{1,1} = 0.3$	$\beta_{1,2} = -0.5$	$\theta_{1,0} = 2.5$	$\theta_{1,1} = 0.5$
5	1.23*	0.30	-0.50	3.20*	0.73*
	(0.006)	(0.002)	(0.004)	(2.12)	(0.284)
30	1.21*	0.30	-0.50	2.82*	0.58*
	(0.004)	($<$0.001)	(0.001)	(0.547)	(0.088)
100	1.20	0.30	-0.50	2.76*	0.55*
	(0.004)	($<$0.001)	($<$0.001)	(0.435)	(0.069)

Performance of EQL with large **D**

K	$\beta_{1,0} = .12$	$\beta_{1,1} = 0.3$	$\beta_{1,2} = -0.5$	$\theta_{1,0} = 0.67$	$\theta_{1,1} = 0.13$
5	1.27*	0.30	-0.50	0.66	0.14*
	(0.017)	($<$0.001)	($<$0.001)	(0.010)	(0.001)
100	1.21	0.30	-0.50	0.68*	0.14*
	(0.008)	($<$0.001)	($<$0.001)	(0.004)	($<$0.001)

A brief description of reindeer pellete count data

- Study location: Storliden mountain, north of Sweden. Study area: 25 km^2; 8 windmills were built in the study area in 2011.
- Study area was marked with some transacts. Distance between 2 neighbouring transacts is 300 m.
- Plots of 2.18 m radius were created on transacts at 100 m apart.
- Reindeer pellets observed within each plot were recorded as 1 pellet group = 20 pellets.
- Survey was conducted between 2009 and 2012 during May-June each year.
- Pellet counts data were supplemented with forest age structure and other geographic data from external sources.

Results from data analysis

Coef.	Binary model($Pr(y=0)$)		Trunc. Poisson	
	Est.	Std. Err.	Est.	Std. Err.
Intercept	1.13*	0.38	-0.17	0.30
Dist. Windmills (100m)	0.01	0.03	0.02	0.03
Year (2009)	–	–	–	–
– (2010)	1.44*	0.32	-1.01	0.58
– (2011)	1.56*	0.35	-1.44	0.68
– (2012)	3.80*	0.66	0.66	1.17
Forage Type (Excellent)	–	–	–	–
– (Very good)	-0.40	0.38	-0.42	0.50
– (Not so good)	-0.04	0.23	-0.21	0.24
Dist. Wml. & Year (Intr.)				
– (2010)	-0.10*	0.03	0.06	0.05
– (2011)	-0.06	0.04	0.07	0.05
– (2012)	-0.12*	0.06	-0.06	0.11
θ_0	1.24	0.17	3.45	0.63
θ_1	-0.19	0.03	-0.65	0.04

Possible future work

1. Higher order correction might improve the properties of the HL estimators.
2. Additional overdispersion via a Gamma random effect might be fitted.
3. More covariates might be considered in the model for reindeer data.
4. Different specifications of $\mathbf{D}$ can be tried.

Thank You