Eric Lock
STOR 890: Environmental Statistics
04/28/09

Computational Techniques for Spatial Logistic Regression

Certain health outcomes depend on environmental factors that vary geographically (e.g. pollutants), and Epidemiological studies accounting for spatial variability are becoming increasingly popular. Often, the outcome measured is binary (e.g. whether or not a certain disease is diagnosed). Logistic regression with a built-in spatial function is useful for assessing risk at an individual level. However, large sample sizes can lead to computational difficulties. Here, we give a synopsis of various computational techniques for spatial logistic regression discussed in Paciorek, 2007[1].

The general model considered is one where the response, Y_i, is a Bernoulli variable with corresponding probability p_i. For example, p_i may represent the probability that individual *i* is diagnosed with a certain disease. It is assumed that p_i depends on a vector of covariates x_i (e.g. age, sex or other demographic variables), and a location s_i in two-dimensional space. The logit of p_i is represented as a standard linear model of the covariates x_i, plus a function *g* of the spatial location, parameterized by θ:

$$Y_i \sim \text{Ber}(p(\boldsymbol{x_i}, \boldsymbol{s_i})),$$
$$\text{logit}(p(\boldsymbol{x_i}, \boldsymbol{s_i})) = \boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta} + g(\boldsymbol{s_i}; \boldsymbol{\theta}),$$

Paciorek focuses on various forms and estimation techniques for the smooth spatial function, *g*.

1.) Generalized Linear Mixed Models (GLMM)

One approach is to treat the spatial function *g* as a random effect in a GLMM, $\boldsymbol{g_s = Zu}$. If we assume Cov(***u***) = $\sigma_u^2 \boldsymbol{I}$, then Cov(**Zu**) = $\sigma_u^2 \boldsymbol{Z}\boldsymbol{Z}^{\boldsymbol{t}}$, so that the matrix **Z** defines the spatial covariance structure. It is suggested to construct **Z** based on the Matern covariance:

$$C(\tau) = \frac{1}{\Gamma(v)2^{v-1}}\left(\frac{2\sqrt{v}\tau}{\rho}\right)^{v} \mathcal{K}_v\left(\frac{2\sqrt{v}\tau}{\rho}\right)$$

where τ is the distance between two spatial locations.

To reduce the computational burden, one could build the covariance structure based on a set of pre-specified knot locations κ_k, k=1,…,K, with K < n. **Z** is then constructed as $\boldsymbol{Z} = \boldsymbol{\psi}\boldsymbol{\Omega}^{-1/2}$, where

$$\boldsymbol{\psi} = (C(||\boldsymbol{s_i} - \boldsymbol{k_k}||)_{\boldsymbol{i}=\mathbf{1},\ldots,\boldsymbol{n};\boldsymbol{k}=\mathbf{1},\ldots,\boldsymbol{K}}, \text{ and}$$

$$\boldsymbol{\Omega} = (C(||\boldsymbol{k_i} - \boldsymbol{k_j}||)_{\boldsymbol{i}=\mathbf{1},\ldots,\boldsymbol{k};\boldsymbol{j}=\mathbf{1},\ldots,\boldsymbol{K}} \ .$$

If one fixes ρ and ν in advance, estimation of σ_u^2 is computationally manageable and several IWLS procedures exist for fitting such a GLMM. Kamman and Wand[2] suggest taking ν = 1.5 and $\rho = \max_{i,j=1,\ldots,n} \|s_i - s_j\|$. This simplifies model estimation, but as $\boldsymbol{g_s}$ is only parameterized by σ_u^2, fixing ρ and ν sacrifices model flexibility.

2.) Bayesian GLMM

Another approach, similar to (1), also takes a random effect model $\boldsymbol{g_s = Zu}$ but gives a prior distribution $\boldsymbol{u} \sim \boldsymbol{N}(\mathbf{0}, \sigma_u^2 \boldsymbol{I})$ for the basis coefficients. The model can then be fit via MCMC. In practice, on could include ρ and ν of the Matern covariance in the MCMC. However, in this case recalculating the covariance matrix at each iteration can be computationally difficult.

3.) Bayesian Spectral Basis (SB)

Another approach uses basis functions and the spectral representation of a Gaussian Process (GP). Here, the function *g* is approximates on a grid, $\mathbf{s}^{\#}$, of size K=k_1 X k_2. Again we represent *g* as $\boldsymbol{g}_{s^{\#}} = \boldsymbol{Zu}$, but here we let **Z** be a matrix of orthogonal SB functions, while ***u*** is a vector of complex-valued basis coefficients, $u_m = a_m + b_m$. The Fourier basis is suggested for the matrix **Z**.

Let the basis coefficients **a** and **b** have the prior distribution

$$\begin{pmatrix} a \\ b \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\theta)$$

Then, the diagonal covariance matrix of the coefficients Σ_θ can be constructed based on the spectral density of a GP covariance function. For the Matern covariance, the spectral density, evaluated at frequency **ω**, is

$$f_{(\rho,\nu)}(\omega) = \frac{\Gamma(\nu + d/2)(4\nu)^\nu}{\pi^{d/2}\Gamma(\nu)(\pi\rho)^{2\nu}} \cdot \left(\frac{4\nu}{(\pi\rho)^2} + \omega^{\mathrm{T}}\omega\right)^{-(\nu+d/2)}$$

For m=(m_1, m_2), let $V(a_{m1,m2}) = V(b_{m1,m2}) = \frac{1}{2} f(\omega_{m1}^1, \omega_{m2}^2)$.

Each observation location is then mapped to its nearest grid location through a transformation **P,** and then multiplied by σ, the standard deviation of the process: $\boldsymbol{g}_s = \sigma \boldsymbol{P} \boldsymbol{g}_{\boldsymbol{s}^{\#}}$. Model parameters σ, ρ and ν are given prior distributions and estimated via MCMC. The matrix **Z** need not be formed at each iteration, and the operation **Zu** can be done quickly using Fast Fourier Transforms, which helps computational efficiency.

Application

The methods discussed are applied to individual-level leukemia and brain cancer data in Kaohsiung, Taiwan. The metropolitan area contains four petroleum/petrochemical complexes, raising concern over possible adverse health effects. Individuals are geocoded to the location where they have resided for the longest time. The leukemia analysis contains 787 individuals, of whom 206 have been diagnosed with leukemia. The brain cancer analysis contains 576 individuals, of which 165 have been diagnosed with brain cancer.

A map of geocoded addresses is shown in Figure 1. Open circles correspond to cases, while closed circles correspond to controls. The boundaries of the four petrochemical complexes are given by a thick black line.

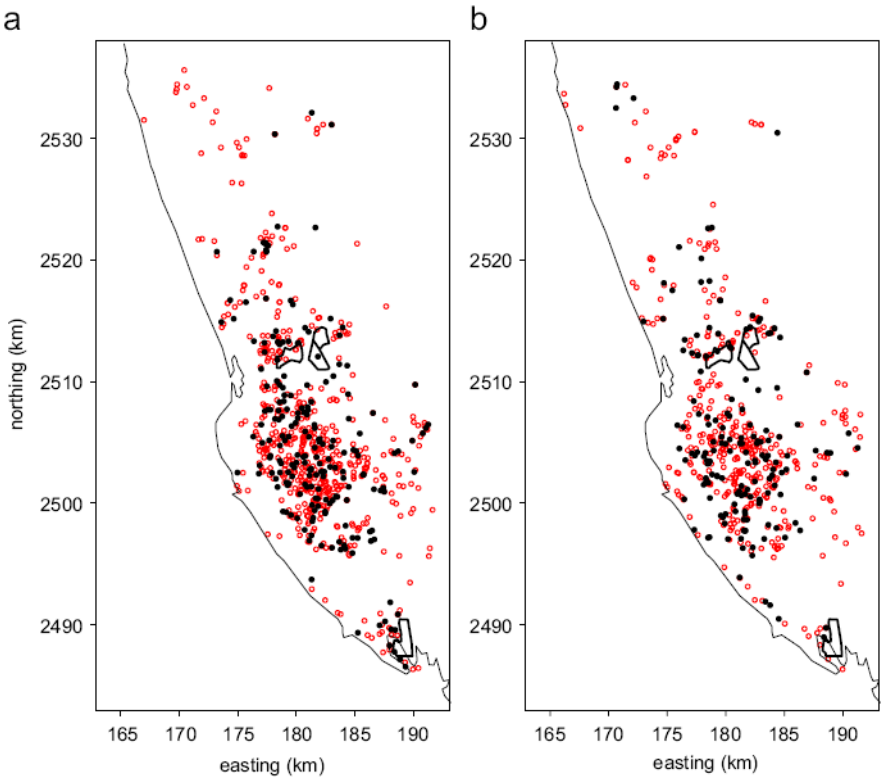

Figure 1: Location of cases and controls for (a) leukemia and (b) brain cancer

Figure 2 gives the posterior mean probability of being a case for each grid cell using the spectral basis method (3). Estimates for both are generally lower in the Northern region, further from the petrochemical complexes, and higher in the South.

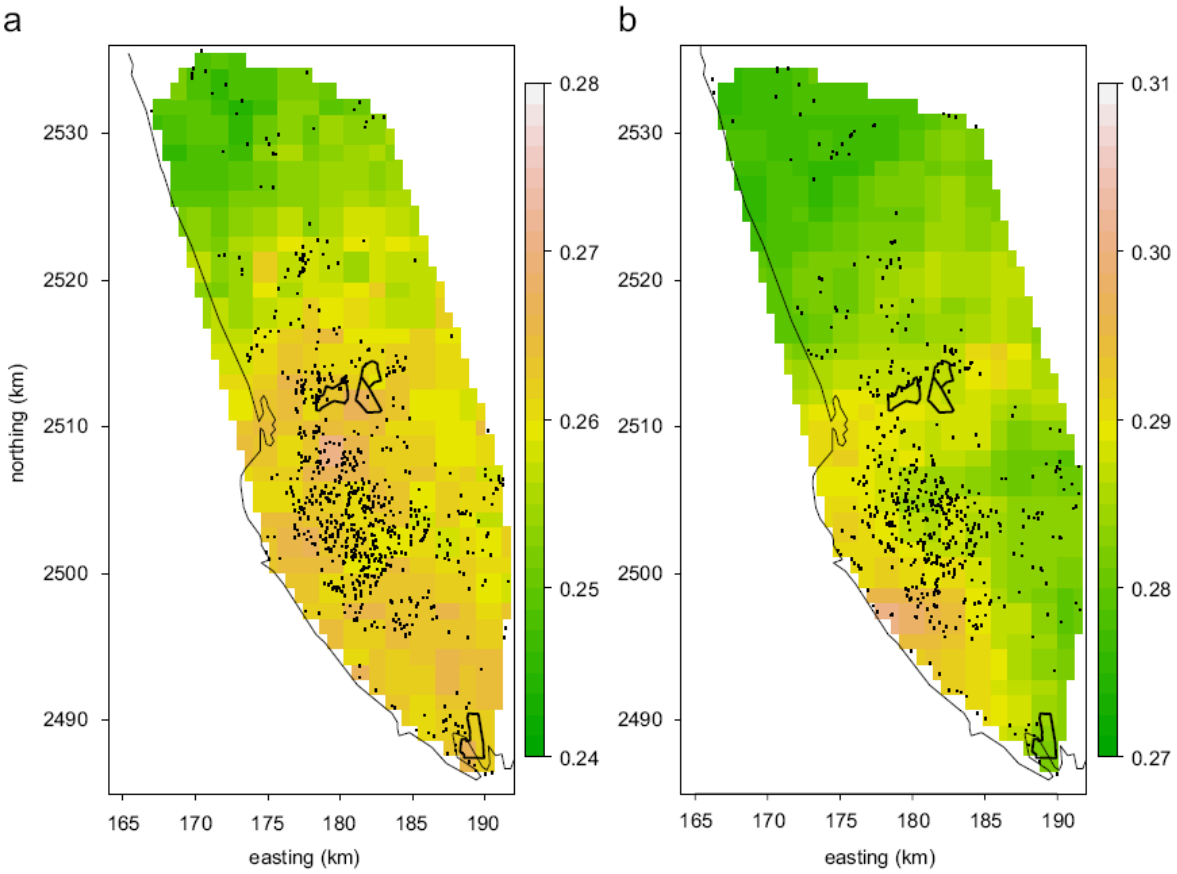

Figure 2: Posterior probability of a case for (a) leukemia and (b) brain cancer

A measure of deviance is used to evaluate various spatial models, including those described in (1) (PL-PQL) and (3) (SB). The model 'Null' corresponds to a non-spatial model. Results are given in Table 1

Table 1: Deviance for various spatial models, and a nonspatial model ('null')

	Leukemia	Brain cancer
PL–GCV	894.5	680.4
PL–PQL	891.7	677.2
SB	891.9	677.9
MRF	892.5	677.8
Null	891.5	678.0

Note that the Null model beats all other models for the leukemia data, and model differences are negligible among the brain cancer data. In fact, the null hypothesis that risk is constant spatially cannot be rejected in either case. This may reflect the absence of a true spatial effect, or an insufficient sample size.

References

[1]Paciorek (2007), *Computational Statistics and Data Analysis,* 51, 3631-3653

[2] Kammann, E.,Wand, M. (2003) *Applied Statistics,* 52, 1–18.