Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 3631 – 3653
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Computational techniques for spatial logistic regression with large data sets

Christopher J. Paciorek*

Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA

Received 6 October 2005; received in revised form 29 August 2006; accepted 7 November 2006
Available online 1 December 2006

Abstract

In epidemiological research, outcomes are frequently non-normal, sample sizes may be large, and effect sizes are often small. To relate health outcomes to geographic risk factors, fast and powerful methods for fitting spatial models, particularly for non-normal data, are required. I focus on binary outcomes, with the risk surface a smooth function of space, but the development herein is relevant for non-normal data in general. I compare penalized likelihood (PL) models, including the penalized quasi-likelihood (PQL) approach, and Bayesian models based on fit, speed, and ease of implementation.

A Bayesian model using a spectral basis (SB) representation of the spatial surface via the Fourier basis provides the best tradeoff of sensitivity and specificity in simulations, detecting real spatial features while limiting overfitting and being reasonably computationally efficient. One of the contributions of this work is further development of this underused representation. The SB model outperforms the PL methods, which are prone to overfitting, but is slower to fit and not as easily implemented. A Bayesian Markov random field model performs less well statistically than the SB model, but is very computationally efficient. We illustrate the methods on a real data set of cancer cases in Taiwan.

The success of the SB with binary data and similar results with count data suggest that it may be generally useful in spatial models and more complicated hierarchical models.

Keywords: Bayesian statistics; Disease mapping; Fourier basis; Generalized linear mixed model; Geostatistics; Risk surface; Spatial statistics; Spectral basis

1. Introduction

Epidemiological investigations that assess how health outcomes are related to risk factors that vary geographically are becoming increasingly popular. This paper is motivated by an ongoing epidemiological study in Kaohsiung, Taiwan, a center of petrochemical production, for which administrative data suggest excess cancer deaths amongst residents less than 20 years old living within 3 km of a plant (Pan et al., 1994). A full case–control study is in progress to investigate this suspected link between plant emissions and cancer, with the residences of individuals being geocoded to allow for spatial modelling. Such individual-level or point-referenced data are becoming increasingly common in epidemiology with the use of geographic information systems (GIS) and geocoding of addresses. In contrast to health

* Tel.: +1 617 4324912; fax: +1 617 4325619.
E-mail addresses: paciorek@alumni.cmu.edu, paciorek@hsph.harvard.edu
URL: http://www.biostat.harvard.edu/~paciorek.

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.11.008

data aggregated into regions, often fit using Markov random field (MRF) models (Waller et al., 1997; Best et al., 1999; Banerjee et al., 2004), analysis of individual-level data (often called geostatistics) avoids both ecological bias (Greenland, 1992; Richardson, 1992; Best et al., 2000) and reliance on arbitrary regional boundaries, but introduces computational difficulties.

In this paper, I focus on methods for fitting models for Bernoulli response data with the outcome a binary variable indicating disease or health status, but my development is relevant for non-normal data in general. The specific model I investigate is a logistic regression,

$$\begin{aligned} &Y_i \sim \mathrm{Ber}(p(\boldsymbol{x_i}, \boldsymbol{s_i})), \\ &\mathrm{logit}(p(\boldsymbol{x_i}, \boldsymbol{s_i})) = \boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta} + g(\boldsymbol{s_i}; \boldsymbol{\theta}), \end{aligned} \tag{1}$$

where $Y_i,\ i = 1, \ldots, n$, is the binary status of the ith subject; $g(\cdot; \boldsymbol{\theta})$ is a smooth function, parameterized by $\boldsymbol{\theta}$, of the spatial location of subject i, $\boldsymbol{s_i} \in \Re^2$; and $\boldsymbol{x_i}$ is a vector of additional individual-level covariates of interest. One application of this model is to the analysis of case–control data. For binary outcomes with the logit link, the use of a retrospective case–control design in place of random sampling from the population increases the power for assessing relative risk based on the covariates, including any spatial effect, but prevents one from estimating the absolute risk of being a case (Prentice and Pyke, 1979; Elliott et al., 2000; Diggle, 2003, pp. 133–134).

There have been several approaches to modelling the smooth function, $g(\cdot; \boldsymbol{\theta})$, each with a variety of parameterizations. One basic distinction is between deterministic and stochastic representations. In the former, (1) is considered a generalized additive model (GAM) (Hastie and Tibshirani, 1990; Wood, 2006), e.g., using a thin plate spline or radial basis function representation for $g(\cdot; \boldsymbol{\theta})$, with the function estimated via a penalized approach. The stochastic representation considers the smooth function to be random, either as a collection of correlated random effects (Ruppert et al., 2003), i.e., a generalized linear mixed model (GLMM), or as an equivalent stochastic process, such as in kriging (Cressie, 1993), which takes $g(\cdot; \boldsymbol{\theta})$ to be a Gaussian process (GP). Of course in the Bayesian paradigm, the unknown function is always treated as random with a (perhaps implicit) prior distribution over functions. Note that from this perspective, the GAM can be expressed in an equivalent Bayesian representation, and there are connections between the thin plate spline and stochastic process approaches (Cressie, 1993; Nychka, 2000) and also between thin plate splines and mixed model representations (Ruppert et al., 2003).

While models of form (1) have a simple structure, fitting them can be difficult for non-Gaussian responses. If the response were Gaussian, there are many methods, both classical and Bayesian, for estimating $\boldsymbol{\beta}$, $g(\cdot; \boldsymbol{\theta})$, and $\boldsymbol{\theta}$. Most methods rely on integrating $g(\cdot; \boldsymbol{\theta})$ out of the model to produce a marginal likelihood or posterior. In the non-Gaussian case, this integration cannot be done analytically, which leads to substantial difficulty in fitting the model because of the high-dimensional quantities that need to be estimated. Initial efforts to fit similar models have focused on approximating the integral in the GLMM framework or fitting the model in a Bayesian fashion. In the case of binary data, the approximations used in the GLMM framework may be poor (Breslow, 2003) and standard Markov chain Monte Carlo (MCMC) techniques for the Bayesian model exhibit slow mixing (Christensen et al., 2006). More recent efforts have attempted to overcome these difficulties; this paper focuses on comparing promising methods and making recommendations relevant to practitioners.

My goal in this paper is to investigate methods for fitting models of form (1). I describe a set of methods (Section 2) that hold promise for good performance with Bernoulli and other non-Gaussian data and large samples (hundreds to thousands of individuals) and detail their implementation in Section 3. One of these methods, the spectral basis (SB) approach of Wikle (2002), has seen limited use; I develop the approach in this context by simplifying the model structure and devising an effective MCMC sampling scheme. I compare the performance of the methods on simulated epidemiological data (Section 4) as well as preliminary data from the Taiwan case–control study (Section 5). In evaluating the methods, my primary criteria are the accuracy of the fit, speed of the fitting method, and ease of implementation, including the availability of software. I close by discussing extensions of the models considered here and considering the relative merits and shared limitations of the methods.

2. Overview of methods

The methods to be discussed fall into two categories: penalized likelihood (PL) models fit via iteratively weighted least squares (IWLS) and Bayesian models fit via MCMC, but there are connections between all of them. First I describe