Statistica Sinica **21** (2011), 1785-1806
doi:http://dx.doi.org/10.5705/ss.2009.292

QUASI-DEVIANCE FUNCTIONS FOR SPATIALLY CORRELATED DATA

Pei-Sheng Lin

National Health Research Institutes and National Chung Cheng University

Abstract: This paper introduces a quasi-deviance function for model selection of given spatial data. The proposed deviance function involves only the mean and covariance of responses, and therefore avoids the difficulty of specifying a full-likelihood function. We show that, under certain regularity conditions, the deviance function with quasi-likelihood estimating equation has a limiting chi-squared distribution. The asymptotic quadratic form of the deviance function provides a consistent method for selecting the true model. We also conduct simulations to evaluate the performance of the proposed method, and use the East Lansing Woods data to illustrate the application.

Key words and phrases: Deviance function, model selection, quasi-likelihood estimate, spatial data.

1. Introduction

In practice, model selection means selecting parameters in an attempt to create a model of optimal complexity for the given data. For generalized linear models with spatial observations, the task of model selection may not be easy because the presence of correlation can render traditional methods inappropriate. In this paper, we propose a deviance function for generalized linear models in multi-dimensional space with a conjunction to quasi-likelihood (QL) functions.

Current model selection methods are mostly designed for independent data or based on an assumption of full likelihoods (e.g., Shao (1993); Qian, Gabor, and Gupta (1996); Shao (1996); Zhang and Huang (2008); Wasserman and Roeder (2009)). However, for correlated data in a multi-dimensional space, the full likelihood may be difficult to specify because associations among observations can be intractable (McCullagh (1991)). Even with a correctly specified likelihood function, computational intensity caused by complicated interactions can make estimation procedures infeasible. A deviance function involving only the first two moments of responses may therefore be appealing for model selection of given correlated data.

When data are repeatedly measured, Pan (2001) may have been the first to discuss model selection. Pan combined generalized estimating equations (GEEs) and independent quasi-likelihood to propose an Akaike information type of criterion. Wang and Qu (2009) later developed a Bayesian information type of criterion based on a quadratic inference function (Qu, Lindsay, and Li (2000)) in the GEE setting. In the quadratic inference function, the inverse of the working correlation matrix is decomposed into a linear combination of basis matrices. However, for spatially correlated data, such decomposition may not be easy since the distance function between observation sites could be any L_p norm function rather than the L_1 norm considered in longitudinal data.

To balance computational intensity and the accommodation of covariance structure, we use an artificial log-likelihood to develop a selection criterion for spatial models. Instead of considering an artificial 'independent' likelihood through integration (e.g., Qian, Gabor, and Gupta (1996); Pan (2001)), we use a projected likelihood ratio to develop an inference function. Let $\boldsymbol{Y} = \{Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)\}^T$ be observations drawn from a random field, where $\boldsymbol{s}_i \in R^d,\ d \geq 2$, denotes the location of the ith observation. Assume that $L(\boldsymbol{\xi})$ and $L(\boldsymbol{\beta})$ are log-likelihood functions associated with parameter sets $\boldsymbol{\xi}$ and $\boldsymbol{\beta}$, respectively. For correlated data, Hanfelt and Liang (1995) proposed an approximate likelihood ratio

$$D(\boldsymbol{\xi}, \boldsymbol{\beta}) = 0.5E\{\boldsymbol{Q}(\boldsymbol{\xi}, \boldsymbol{Y})\}^T \boldsymbol{V}^{-1}(\boldsymbol{\xi})\boldsymbol{Q}(\boldsymbol{\xi}, \boldsymbol{Y}) - 0.5E\{\boldsymbol{Q}(\boldsymbol{\beta}, \boldsymbol{Y})\}^T \boldsymbol{V}^{-1}(\boldsymbol{\beta})\boldsymbol{Q}(\boldsymbol{\beta}, \boldsymbol{Y}),$$

where $\boldsymbol{V} = \mathrm{cov}(\boldsymbol{Y})$ and $\boldsymbol{Q}(\cdot, \boldsymbol{Y})$ denotes an estimating equation. The approximate likelihood $D(\boldsymbol{\xi}, \boldsymbol{\beta})$ is regarded as a projection of the log-likelihood ratio into a vector space spanned by a linear combination of estimating equations. Specifically, if we focus on an optimal estimating equation, referred to as a QL estimating equation as discussed in Section 2.1, this projection function becomes

$$D(\boldsymbol{\xi}, \boldsymbol{\beta}) = 0.5\{\boldsymbol{\theta}(\boldsymbol{\xi}) - \boldsymbol{\theta}(\boldsymbol{\beta})\}^T \left[\boldsymbol{V}^{-1}(\boldsymbol{\xi})\{\boldsymbol{Y} - \boldsymbol{\theta}(\boldsymbol{\xi})\} + \boldsymbol{V}^{-1}(\boldsymbol{\beta})\{\boldsymbol{Y} - \boldsymbol{\theta}(\boldsymbol{\beta})\}\right], \quad (1.1)$$

where $\boldsymbol{\theta}(\cdot)$ represents the mean function of $\boldsymbol{Y}$. Equation (1.1) was also proposed by Li (1993) as a projected likelihood ratio on a linear function of observations.

The concept of the projected likelihood (1.1) can be traced back to the work of McLeish and Small (1992). Li (1993) later showed that the inference function can be regarded as a projected likelihood ratio which has first-order equivalence to a QL function. Hanfelt and Liang (1995) then extended these works to propose two approximate likelihood ratios as alternatives to the Wald test. There are several properties, provided by Hanfelt and Liang, that to make the projected likelihood attractive. For example, the function is anti-symmetric, as is the log-likelihood. When the quasi-score function has multiple roots, the deviance function can distinguish the correct solution from incorrect ones. Nevertheless, the projected likelihood proposed by Hanfelt and Liang was based on a generalized estimating equation for longitudinal data. For a generalized linear model with covariates expressed by spatial coordinates, the corresponding covariance structure is usually more complicated and increases the difficulty in doing statistical inference.

In this paper, we combine the projected likelihood ratio function $D(\boldsymbol{\xi}, \boldsymbol{\beta})$ in (1.1) with the QL estimate as a deviance function for spatial data. For convenience, we refer to the proposed deviance as quasi-deviance (QDEV), a deviance function with QL estimates. We note that the deviance (1.1) is linear with regard to a given covariance structure. One benefit of using linear deviance is that it allows us to study the behavior of $D(\boldsymbol{\xi}, \boldsymbol{\beta})$ locally at an alternative parameter through the non-central moments. In the next section, we provide a central limit theorem for spatial data with a continuous location index. We show that, under some regularity conditions, the QDEV function in a multi-dimensional space has a quadratic expression asymptotically similar to a maximum likelihood ratio test. The similarity between the QDEV statistic and the maximum likelihood ratio test is used to develop a model selection procedure. We show the proposed criterion is consistent for model selection in Section 3, and conduct some simulation studies in Section 4. The data analysis to illustrate the proposed method is given in Sections 5. Discussion about future development is in Section 6.

2. Quasi-deviance Function

2.1. Quasi-likelihood estimates for spatial data

Assume that the marginal model connects the mean response $\theta_i = E\{Y(\boldsymbol{s}_i)\}$ and an explanatory vector $\boldsymbol{x}_i$ through a generalized linear model $g(\theta_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T \in R^q$ is a vector of parameters of interest. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ be the design matrix and $g(\boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{\beta}$.

Assumption 1. (a) The first two orders of derivatives of $g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ are continuous in the parameter space. In addition, some constants c_0 and c_1 exist such that $|g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})| \leq c_0$ and $|\partial g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})/\partial\beta_j| \leq c_1$, for $i = 1, \ldots, n$ and $j = 1, \ldots, q$. (b) There exists a smooth variance function $V(\cdot)$ such that $\mathrm{var}\{Y(\boldsymbol{s}_i)\} = V\{g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})\}$.

We develop a central limit theorem for a random field with a continuous location index. Take a "standardized" distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ to be

$$r(\boldsymbol{s}_i, \boldsymbol{s}_j) = \frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_p}{\min\limits_{i,j} \|\boldsymbol{s}_i - \boldsymbol{s}_j\|_p},$$

where $\|\cdot\|_p$ is an L_p norm function. Note that the use of standardized distance here is only for convenience of theoretical derivation. Let $B_{\boldsymbol{s},r}$ denote the number of observations in a d-dimensional ball $B(\boldsymbol{s}, r)$ centered at $\boldsymbol{s}$ with radius r. Let Ω_n be a strictly increasing sequence of subsets in R^d, and let $|\Omega_n|$ denote the cardinality of Ω_n. Assume that $|\Omega_n| = O(n^d)$.

Assumption 2. There exists a constant $c_0 > 0$ such that $B_{\boldsymbol{s},r} - B_{\boldsymbol{s},r-1} \le c_0 r^{d-1}$ for any $\boldsymbol{s}$ in the study region, $r > 1$.

To explain why Assumption 2 is reasonable, we note that the smallest distance between observations is standardized to be 1. So, with the number of observations in $B(\boldsymbol{s}, r)$ proportional to its volume $c_r r^d$, where c_r is a constant, the number of observations in $B(\boldsymbol{s}, r)$ but not in $B(\boldsymbol{s}, r-1)$ is about $r^d - (r-1)^d \doteq r^{d-1}$.

Assumption 3. We have max $E\{Y^2(\boldsymbol{s})\} < \infty$. The correlation model satisfies $\alpha(r) = \mathrm{corr}\{Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)\} = o(r^{-d})$, where r is the standardized distance between sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. Further, $\mathrm{corr}\{Y(\boldsymbol{s}_{i_1})Y(\boldsymbol{s}_{i_2}), Y(\boldsymbol{s}_{j_1})Y(\boldsymbol{s}_{j_2})\} = o\{(r^*)^{-d}\}$, where $r^* = \inf\{r(\boldsymbol{s}_i^*, \boldsymbol{s}_j^*) : \boldsymbol{s}_i^* = \boldsymbol{s}_{i_1}$ or $\boldsymbol{s}_{i_2}$, $\boldsymbol{s}_j{*} = \boldsymbol{s}_{j_1}$ or $\boldsymbol{s}_{j_2}\}$, so the correlation between $Y(\boldsymbol{s}_{i_1})Y(\boldsymbol{s}_{i_2})$ and $Y(\boldsymbol{s}_{j_1})Y(\boldsymbol{s}_{j_2})$ depends on the shortest distance between the location sets $\{\boldsymbol{s}_{i_1}, \boldsymbol{s}_{i_2}\}$ and $\{\boldsymbol{s}_{j_1}, \boldsymbol{s}_{j_2}\}$.

Lemma 1. *If the responses $Y(\boldsymbol{s})$ satisfy Assumptions* 2 *and* 3*, then Z_n/ν_n converges in distribution to $N(0,1)$, where $Z_n = \sum_{\boldsymbol{s}\in\Omega_n}\{Y(\boldsymbol{s}) - \theta(\boldsymbol{s})\}$ and $\nu_n^2 = \mathrm{var}(Z_n)$.*

The proof of Lemma 1 is given in the Appendix. Note that in Lemma 1, observations can be in any location of R^d. Compared with other central limit theorems built for gridded data (e.g., Guyon (1995); Lin (2008)), Lemma 1 is more flexible for statistical inference. Let $\dot{\boldsymbol{\theta}}(\boldsymbol{\beta})$ denote the $n \times q$ derivative matrix of $\boldsymbol{\theta}$ with respect to $\boldsymbol{\beta}$, and $\dot{\boldsymbol{\theta}}^T(\boldsymbol{\beta})$ be the transpose matrix of $\dot{\boldsymbol{\theta}}(\boldsymbol{\beta})$. Assumption 1 ensures that $\boldsymbol{\theta}(\boldsymbol{\beta})$ and $\dot{\boldsymbol{\theta}}(\boldsymbol{\beta})$ are bounded. We take $\boldsymbol{U}(\boldsymbol{\beta}) = \dot{\boldsymbol{\theta}}^T(\boldsymbol{\beta})\boldsymbol{V}^{-1}(\boldsymbol{\beta})\{\boldsymbol{Y} - \boldsymbol{\theta}(\boldsymbol{\beta})\}$ as a quasi-score function. For correlated data in multi-dimensional space, the QL estimating equation

$$\boldsymbol{U}(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \dot{\boldsymbol{\theta}}^T(\hat{\boldsymbol{\beta}})\boldsymbol{V}^{-1}(\hat{\boldsymbol{\beta}})\{\boldsymbol{Y} - \boldsymbol{\theta}(\hat{\boldsymbol{\beta}})\} = \boldsymbol{0} \tag{2.1}$$

is useful for estimating unknown parameters; here $\hat{\boldsymbol{\beta}}$ is referred as the QL estimate. Let $\boldsymbol{\beta}_0$ denote the true parameter. Using an argument similar to Lin (2008), we have the following result.

Lemma 2. *Under Assumptions* 1−3 *and Lemma* 1*, the QL estimate is consistent and*

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \;\to\; N_q\{\boldsymbol{0}, \boldsymbol{I}^{-1}(\boldsymbol{\beta}_0)\}, \tag{2.2}$$

in distribution, where $\boldsymbol{I}(\boldsymbol{\beta}_0)$ is the limit of $n^{-1}\dot{\boldsymbol{\theta}}^T(\boldsymbol{\beta}_0)\boldsymbol{V}^{-1}(\boldsymbol{\beta}_0)\dot{\boldsymbol{\theta}}(\boldsymbol{\beta}_0)$ as $n \to \infty$.

Since the QL estimate has the asymptotical covariance matrix $\boldsymbol{I}^{-1}(\boldsymbol{\beta}_0)$, we refer to the QL estimating equation as an optimal estimating equation, after Godambe (1991). We note that in (2.1), only the mean and covariances are specified. The QL estimating equation thus provides a flexible tool for analyzing multi-dimensional data.

2.2. Limit distribution of the quasi-deviance function

We study asymptotic properties of the QDEV function. Let $\nabla_{\boldsymbol{\xi}} D(\boldsymbol{\xi}, \boldsymbol{\beta}_0)$ and $\nabla^2_{\boldsymbol{\xi}} D(\boldsymbol{\xi}, \boldsymbol{\beta}_0)$ be the gradient vector and second-order derivative matrix of $D(\boldsymbol{\xi}, \boldsymbol{\beta}_0)$ with respect to $\boldsymbol{\xi}$, respectively. We can show that

$$E_{\boldsymbol{\beta}_0}\{\nabla_{\boldsymbol{\xi}} D(\boldsymbol{\xi}, \boldsymbol{\beta}_0)\} = \mathbf{0}$$

and

$$E_{\boldsymbol{\beta}_0}[\{\nabla_{\boldsymbol{\xi}} D(\boldsymbol{\xi}, \boldsymbol{\beta}_0)\}\{\nabla_{\boldsymbol{\xi}} D(\boldsymbol{\xi}, \boldsymbol{\beta}_0)\}^T] = -E_{\boldsymbol{\beta}_0} \nabla^2_{\boldsymbol{\xi}} D(\boldsymbol{\xi}, \boldsymbol{\beta}_0).$$

Therefore, the linear deviance behaves like a log-likelihood function locally at the true parameter value $\boldsymbol{\beta}_0$. The following theorems establish a relationship between the QDEV function and a likelihood ratio test.

Theorem 1. *If $\hat{\boldsymbol{\beta}}$ is the QL estimate, under some regularity conditions, $2D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$ converges in distribution to χ^2_q.*

More generally, take $\boldsymbol{\xi} = (\boldsymbol{\psi}^T, \boldsymbol{\beta}^T)^T$ and are interested in testing $H : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ versus $K : \boldsymbol{\psi} \neq \boldsymbol{\psi}_0$. Let $\hat{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{\beta}}$ be the corresponding QL estimates for $\boldsymbol{\xi}$ (under the hypothesis K) and $\boldsymbol{\beta}$ (under the hypotheses H), respectively.

Theorem 2. *Under the null hypothesis $H : \boldsymbol{\psi} = \boldsymbol{\psi}_0$, we have*

$$2D(\hat{\boldsymbol{\xi}}, \tilde{\boldsymbol{\beta}}) = n(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T \boldsymbol{I}^{-1}_{\psi_0\psi_0}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) + o_p(1), \tag{2.3}$$

where $\hat{\boldsymbol{\psi}}$ is the QL estimate of $\boldsymbol{\psi}$, and $\boldsymbol{I}^{-1}_{\psi_0\psi_0}$ is the inverse of the covariance matrix of $\hat{\boldsymbol{\psi}}$; $2D(\hat{\boldsymbol{\xi}}, \tilde{\boldsymbol{\beta}})$ converges in distribution to a χ^2_r distribution, where $r = dim(\boldsymbol{\psi})$ denotes the difference of dimensions between H and K.

Equation (2.3) implies that, under H, $2D(\hat{\boldsymbol{\xi}}, \tilde{\boldsymbol{\beta}})$ is asymptotically the same as a Wald test statistic (Cox and Hinkley (1974, p.314)). Let $\boldsymbol{\xi}_k = (\boldsymbol{\psi}_k^T, \boldsymbol{\beta}^T)^T$ be the true parameter under the alternative hypothesis K.

Theorem 3. *Under the alternative hypothesis $K : \boldsymbol{\psi}_k = \boldsymbol{\psi}_0 + n^{-1/2}\boldsymbol{\delta}$, we have*

$$2D(\hat{\boldsymbol{\xi}}, \tilde{\boldsymbol{\beta}}) = n(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_k)^T \boldsymbol{I}^{-1}_{\psi_k\psi_k}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_k) + \boldsymbol{\delta}^T \boldsymbol{I}^{-1}_{\psi_k\psi_k} \boldsymbol{\delta} + o_p(1); \tag{2.4}$$

$2D(\hat{\boldsymbol{\xi}}, \tilde{\boldsymbol{\beta}})$ converges in distribution to a noncentral χ^2_r distribution with noncentrality parameter $\boldsymbol{\delta}^T \boldsymbol{I}^{-1}_{\psi_k\psi_k} \boldsymbol{\delta}$, where $\boldsymbol{\psi}_k$ is the true value of $\boldsymbol{\psi}$ and $r = dim(\boldsymbol{\psi})$.

It follows from (2.4) that the power of the QDEV statistic increases with the sample size n and the absolute value of $\boldsymbol{\psi}_k - \boldsymbol{\psi}_0$. Furthermore, if we set $\boldsymbol{\psi}_0 = \mathbf{0}$, then the asymptotic quadratic form of (2.4) implies that using the QDEV function for model selection would choose correct models with probability approaching to 1. We discuss this issue in the next section.

3. Model Selection by Quasi-Deviance Functions

3.1. Projected likelihood ratio tests for variable selection

We describe how to use the QDEV criterion to select a subset of covariates consisting of all non-zero regression coefficients. Let

$$\mathcal{M} = \{\mathcal{M}_\tau : g(\boldsymbol{\theta}_\tau) = \boldsymbol{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}_\tau\}$$

be a collection of candidate models, where τ is a subset of $\{1, \ldots, q\}$ consisting of the indices of the covariates that are included in the candidate model $\mathcal{M}_\tau$. Also, $\boldsymbol{\beta}_\tau$ is a vector containing the components of $\boldsymbol{\beta}$ indexed by the integers in τ, $\boldsymbol{X}_\tau = (\boldsymbol{x}_{1,\tau}, \ldots, \boldsymbol{x}_{n,\tau})^T$ contains the columns of $\boldsymbol{X}$ indexed by the integers in τ, and $\boldsymbol{\theta}_\tau = (\theta_{\tau,1}, \ldots, \theta_{\tau,n})^T$ is the assumed mean of $\boldsymbol{Y}$ under the model $\mathcal{M}_\tau$.

Let $\mathcal{M}_{\tau_1}$ and $\mathcal{M}_{\tau_2}$ be two models with the corresponding parameter sets $\boldsymbol{\beta}_{\tau_2}$ and $\boldsymbol{\beta}_{\tau_1}$. Assume that the cardinalities of τ_1 and τ_2 are $q_1 \le q$ and $q_2 \le q$, respectively. For each model $\mathcal{M}_{\tau_j}$, we first obtain the QL estimates $\hat{\boldsymbol{\beta}}_{\tau_j}$ from

$$\boldsymbol{U}(\hat{\boldsymbol{\beta}}_{\tau_j}) = \dot{\boldsymbol{\theta}}_{\tau_j}^T(\hat{\boldsymbol{\beta}}_{\tau_j})\boldsymbol{V}_{\tau_j}^{-1}(\hat{\boldsymbol{\beta}}_{\tau_j})\{\boldsymbol{Y} - \boldsymbol{\theta}_{\tau_j}(\hat{\boldsymbol{\beta}}_{\tau_j})\} = \mathbf{0}, \quad j = 1, 2.$$

Since the quasi-score function $\boldsymbol{U}(\hat{\boldsymbol{\beta}}_{\tau_j})$ is nonlinear, we use the Newton-Raphson iteration,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\tau_j}^{(k+1)} = \hat{\boldsymbol{\beta}}_{\tau_j}^{(k)} &+ \left[\dot{\boldsymbol{\theta}}_{\tau_j}^T\{\hat{\boldsymbol{\beta}}_{\tau_j}^{(k)}\}\boldsymbol{V}_{\tau_j}^{-1}\{\hat{\boldsymbol{\beta}}_{\tau_j}^{(k)}\}\dot{\boldsymbol{\theta}}_{\tau_j}\{\hat{\boldsymbol{\beta}}_{\tau_j}^{(k)}\}\right]^{-1} \\ &\times \dot{\boldsymbol{\theta}}_{\tau_j}^T\{\hat{\boldsymbol{\beta}}_{\tau_j}^{(k)}\}\boldsymbol{V}_{\tau_j}^{-1}\{\hat{\boldsymbol{\beta}}_{\tau_j}^{(k)}\}\left[\boldsymbol{Y} - \boldsymbol{\theta}_{\tau_j}\{\hat{\boldsymbol{\beta}}_{\tau_j}^{(k)}\}\right],\end{aligned}$$

to derive a numerical solution, where the index (k) denotes the kth iteration.

We now wish to test

$$H_s: \ \mathcal{M}_{\tau_1} \text{ and } \mathcal{M}_{\tau_2} \text{ are not discriminated.}$$

A comparison between the competing models are classified into the following cases.

Case 1. Strictly non-nested models

We say two models M_{τ_1} and M_{τ_2} are strictly non-nested models if $\tau_1 \cap \tau_2 = \emptyset$. To test the alternative hypotheses that $K_s : \mathcal{M}_{\tau_2}$ is correct or $K_s' : \mathcal{M}_{\tau_1}$ is correct, we use the following theorem.

Theorem 4. *Assume that responses satisfy the assumptions of Lemma* 2. *If*

$$\liminf_{n\to\infty} n^{-1}\boldsymbol{\theta}^T(\boldsymbol{\beta}_{\tau_1})\boldsymbol{V}^{-1}(\boldsymbol{\beta}_{\tau_1})\boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_1}) > \mathbf{0}$$

and

$$\liminf_{n\to\infty} n^{-1}\boldsymbol{\theta}^T(\boldsymbol{\beta}_{\tau_2})\boldsymbol{V}^{-1}(\boldsymbol{\beta}_{\tau_2})\boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_2}) > \mathbf{0},$$

then

(i) *under* K_s, $D(\hat{\boldsymbol{\beta}}_{\tau_2}, \hat{\boldsymbol{\beta}}_{\tau_1}) \to \infty$ *almost surely,*

(ii) *under* K_s', $D(\hat{\boldsymbol{\beta}}_{\tau_2}, \hat{\boldsymbol{\beta}}_{\tau_1}) \to -\infty$ *almost surely.*

Proof. We first assume that M_{τ_2} is correct. Since M_{τ_1} and M_{τ_2} are disjoint, it is obvious that $\boldsymbol{\beta}_{\tau_2} \neq \mathbf{0}$ and $\boldsymbol{\beta}_{\tau_1} = \mathbf{0}$. Since the QL estimates $\hat{\boldsymbol{\beta}}_{\tau_1}$ and $\hat{\boldsymbol{\beta}}_{\tau_2}$ are consistent and the deviance function is continuous, it follows from Slutsky's Theorem that $D(\hat{\boldsymbol{\beta}}_{\tau_2}, \hat{\boldsymbol{\beta}}_{\tau_1}) \to D(\boldsymbol{\beta}_{\tau_2}, \boldsymbol{\beta}_{\tau_1})$ in probability. We decompose $D(\boldsymbol{\beta}_{\tau_2}, \boldsymbol{\beta}_{\tau_1})$ as $D_1 + D_2$, where $D_1 = \{\boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_2}) - \boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_1})\}^T\{\boldsymbol{V}^{-1}(\boldsymbol{\beta}_{\tau_2}) + \boldsymbol{V}^{-1}(\boldsymbol{\beta}_{\tau_1})\}\{\boldsymbol{Y} - \boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_2})\}$ and $D_2 = \{\boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_2}) - \boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_1})\}^T\boldsymbol{V}^{-1}(\boldsymbol{\beta}_{\tau_1})\{\boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_2}) - \boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_1})\}$. Note that D_1 is a linear combination of responses and, applying Lemma 1, that $n^{-1/2}D_1$ converges to a normal distribution with mean zero. So, $n^{-1}D(\boldsymbol{\beta}_{\tau_2}, \boldsymbol{\beta}_{\tau_1}) = n^{-1}D_1 + n^{-1}D_2$ converges in probability to a positive value by the assumption of Theorem 4, which implies result (i). Result (ii) can be shown in similar fashion.

Testing H_s is equivalent to testing whether the expected value of the log-likelihood ratio is zero (Vuong (1989)). In this case, we can use an argument similar to that of the proof of Theorem 4 and the work by Hanfelt and Liang to show that $D(\hat{\boldsymbol{\beta}}_{\tau_2}, \hat{\boldsymbol{\beta}}_{\tau_1}) \to 0$ almost surely. For strictly non-nested models, we would therefore accept H_s if the absolute value of $2D(\hat{\boldsymbol{\beta}}_{\tau_2}, \hat{\boldsymbol{\beta}}_{\tau_1})$ is less than $\chi^2_{q_0}(1-\alpha/2)$, where q_0 is the absolute value of $q_2 - q_1$ and $\chi^2_{q_0}(1-\alpha/2)$ denotes the $(1-\alpha/2)$th quantile of a chi-squared distribution with degrees of freedom q_0.

Case 2. Nested models

Two models are called nested models if $\tau_1 \subset \tau_2$. In this case, we consider the alternative hypothesis to be K_N : $\mathcal{M}_{\tau_2}$ is more significant than $\mathcal{M}_{\tau_1}$.

Theorem 5. *Assume that the conditions of Theorem* 4 *are satisfied.*

(i) *Under* H_s, $2D(\hat{\boldsymbol{\beta}}_{\tau_2}, \hat{\boldsymbol{\beta}}_{\tau_1}) \to \chi^2_{q_0}$ *in distribution, where* $q_0 = q_2 - q_1$.

(ii) *Under* K_N, $2D(\hat{\boldsymbol{\beta}}_{\tau_2}, \hat{\boldsymbol{\beta}}_{\tau_1}) \to \infty$ *almost surely.*

Theorem 5 is an immediate consequence of Theorems 2 and 3. By Theorem 5, we choose $\mathcal{M}_{\tau_2}$ if

$$2D(\hat{\boldsymbol{\beta}}_{\tau_2}, \hat{\boldsymbol{\beta}}_{\tau_1}) > \chi^2_{q_0}(1-\alpha). \tag{3.1}$$

3.2. Asymptotic properties of the selection procedure

Let $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,q})^T$ be the true parameters, and $\tau^* = \{j : \beta_{0,j} \neq 0\}$. We refer to $\mathcal{M}_{\tau^*}$ as the true model. According to the values of $\boldsymbol{\beta}_0$, each candidate model $\mathcal{M}_\tau$ can be assigned to one of the following categories (Shao (1993)):

C1. At least one nonzero component $\beta_{0,j}$ is not in $\boldsymbol{\beta}_\tau$ or, equivalently, $\tau^* \cap \tau^c \neq \emptyset$, where τ^c represents the complement set of τ.

C2. $\boldsymbol{\beta}_\tau$ contains all nonzero components of $\boldsymbol{\beta}_0$ or, equivalently, $\tau^* \subset \tau$.

A model in C1 is clearly an incorrect model. When $\mathcal{M}_\tau$ is in C2, Shao (1996) called $\mathcal{M}_\tau$ a correct model since $\boldsymbol{x}_{i,\tau*}^T \boldsymbol{\beta}_{\tau*} = \boldsymbol{x}_{i,\tau}^T \boldsymbol{\beta}_\tau$.

Theorem 6. *Assume that the set of candidate models has an intersection with* C2. *A model selection procedure based on a stepwise procedure assures that, with probability tending to* 1*, the selected model is in* C2.

Proof. To see why Theorem 6 stands, let $\mathcal{M}_{\tau_1}$ be a candidate model in C1. For the given τ_1, we define $\omega = \tau_1^c \cap \tau^*$ and $\tau_2 = \xi \cup \tau_1$. We note that $\tau^* \subset \tau_2$, and hence $\mathcal{M}_{\tau_2}$ belongs to C2. Also, by the definition of C1, we know that $\omega \neq \emptyset$, and therefore $\tau_1 \subset \tau_2$. Note that the true values of $\boldsymbol{\beta}_\omega$ are non-zeros since $\omega \subset \tau^*$. It follows from the above discussion that $2D(\hat{\boldsymbol{\beta}}_{\tau_2}, \hat{\boldsymbol{\beta}}_{\tau_1}) \to \infty$ as n tends to be infinity. So, for any model $\mathcal{M}_{\tau_1} \in$ C1, another candidate model $\mathcal{M}_{\tau_2} \in$ C2 exists such that

$$P\{2D(\hat{\boldsymbol{\beta}}_{\tau_2}, \hat{\boldsymbol{\beta}}_{\tau_1}) > \chi^2_{q_0}(1-\alpha)\} \ \to\ 1 \text{ as } n \to \infty,$$

no matter what value of α is pre-specified. The probability that a selected model is not a correct model therefore approaches zero as $n \to \infty$.

There may be, as indicated by Shao (1993), more than one correct model. To derive consistency of the proposed selection method, we further assume that the candidate models consist of a sequence of nested models that includes the true model.

Theorem 7. *Let $\mathcal{M}_{\tau_1} \subset \cdots \subset \mathcal{M}_{\tau_m}$ be a sequence of models that contains the true model $\mathcal{M}_{\tau^*}$. If we let $\alpha \to 0$ in criterion* (3.1)*, then the probability of selecting the true model tends to 1 as $n \to \infty$. That is,*

$$P\{\mathcal{M}_{\tau^*} \equiv \mathcal{M}_{\hat{\tau}}\} \ \to\ 1 \textit{ as } n \to \infty, \tag{3.2}$$

where $\hat{\tau}$ is the index set for the parameters of the selected model.

Proof. To show (3.2) is equivalent to showing that $P\{\tau^* = \hat{\tau}\}$ approaches 1 as $n \to \infty$. From Theorem 6 and the definition of C2, we know that $P\{\tau^* \subset \hat{\tau}\} \to$

1 as $n \to \infty$. It thus remains to show that $P\{\hat{\tau} \subset \tau^*\} \to 1$ as $n \to \infty$. Note that the event $\{\hat{\tau} \cap (\tau^*)^c \neq \emptyset\}$ means that we reject the hypothesis $H : \beta_{0,k} = 0$ for some $k \in \{1, \ldots, q\}$ when, in fact, $\beta_{0,k} = 0$. Therefore, $P\{\hat{\tau} \cap (\tau^*)^c \neq \emptyset\}$ can be regarded as a Type I error rate, which is controlled within α by criterion (3.1). That is, $P\{\hat{\tau} \cap (\tau^*)^c \neq \emptyset\} \leq \alpha$. If we let $\alpha \to 0$, then $P\{\hat{\tau} \cap (\tau^*)^c \neq \emptyset\} \to 0$, and therefore $P\{\hat{\tau} \subset \tau^*\} \to 1$ as $n \to \infty$. We thus have the consistency property forthe selection procedure.

4. Simulations

We conducted simulations to evaluate performance of the proposed deviance function. In the simulations, spatial errors were generated from a multivariate normal distribution indexed by an $m \times m$ lattice. Specifically, we took $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, where $n = m^2$, from a stationary Gaussian process with mean 0, variance 1, and correlation $\rho_{i,j} = \rho^{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2}$. We set $\rho = 0.3$, 0.5, or 0.7. We also simulated covariates $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, x_{i,3})^T$, $i = 1, \ldots, n$, independently from a standard normal distribution. For given $\boldsymbol{\epsilon}$ and $\boldsymbol{x}_i$, responses were simulated as follows.

We first generated count responses based on a hierarchical generalized linear model. Given ϵ_i, the response Y_i was independently Poisson with mean $E\{Y_i; \epsilon_i\} = \exp(\beta_0^* + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i)$. The model then had (Heagerty and Lumley (2000))

$$\theta_i = E(Y_i) = \exp\{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}\}, \tag{4.1}$$

$$\mathrm{var}(Y_i) = \theta_i + \theta_i^2\{\exp(1) - 1\} \quad \text{and} \quad \mathrm{cov}(Y_i, Y_j) = \theta_i \theta_j\{\exp(\rho_{ij}) - 1\}, \tag{4.2}$$

where $\beta_0 = 0.5 + \beta_0^*$ and $\rho_{ij} = \mathrm{corr}(\epsilon_i, \epsilon_j)$. We fixed β_0 at 0.5, and set $(\beta_1, \beta_2, \beta_3) = (1, 1, 0)$ to evaluate the size of the proposed method. We also set $(\beta_1, \beta_2, \beta_3)$ to be (1,0.5,0.1) and (0.2,0.2,0.2) to evaluate the power. For each simulation setting of the Poisson model, we generated 10×10 and 15×15 lattices with 500 replicates.

Before employing the QDEV function, one first needs to estimate the correlation function $\rho_{i,j}$. In spatial statistics, the empirical variogram

$$\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(s_i, s_j) \in N(h)} \{Y(s_i) - Y(s_j)\}^2,$$

where $N(h)$ denotes all the pairs (s_i, s_j) with distance h, is commonly used to measure spatial correlation. For a given model, we thus used an iterative procedure proposed by Lin (2008) to obtain a correlation estimate $\hat{\rho}_{i,j}$ from the variogram and QL estimates $\hat{\boldsymbol{\beta}}$ for the parameters. These estimates $\hat{\rho}_{i,j}$ and $\hat{\boldsymbol{\beta}}$ were then plugged into (4.1) and (4.2) to get a 'working' mean function and 'working' covariance matrix. We call the QDEV function from the working mean and covariance matrix a working QDEV function. The QDEV function under the exact covariance matrix was also computed for a comparison. We found that, in simulations, the values of working and exact QDEV functions were very close. Therefore, we only present the simulation results from the working QDEV function in the following.

To use the working QDEV function for model selection, we take a forward selection procedure. Specifically, we first estimate all the parameters by a QL estimating equation under a full model. By (2.2), we can obtain an asymptotic z-value for each estimated parameter. Then, starting from a model with only the intercept being included, we sequentially add variables to the model, one at a time, from the one with the largest significant z-value to the one with the least significant z-value, based on (3.1). We refer to this method as method QDEV. Then, to study the sensitivity of α to the consistency of method QDEV, we consider $\alpha = 0.01$ and $\alpha = 0.05$ in (3.1). In the following tables, method QDEV_1 and method QDEV_2 are used to denote method QDEV with $\alpha = 0.01$ and $\alpha = 0.05$, respectively.

For the simulation study of model (4.1), we also considered other information criteria. Since a maximum likelihood estimate for the generalized linear mixed model (4.1) is usually infeasible (Breslow and Clayton (1993)), we used a penalized quasi-likelihood (PQL)

$$\sum_{i=1}^{N} \exp\{y_i \log(\theta_i) - \theta_i\} - 0.5\boldsymbol{\epsilon}'\boldsymbol{V}^{-1}\boldsymbol{\epsilon}$$

as a likelihood function to construct the Akaike information criterion (AIC) and Bayesian information criterion (BIC). We refer to these two methods as method PQL_A and method PQL_B, respectively. We also computed AIC and BIC based on a generalized linear model under the independence assumption, termed method AIC_i and method BIC_i, respectively.

Tables 1−3 show the simulation results for model (4.1) under mild, moderate, and strong correlation, respectively. In the tables, we use symbol x_1 to denote the selected model consisting of only one covariate. Overall, the performance of method QDEV was very good, with most estimated probabilities of selecting the true models above 0.90. The only exception occurred on 10 × 10 lattices with $(\beta_0, \beta_1, \beta_2, \beta_3) = (0.5, 1, 1, 0)$. In this case, the estimated probability of selecting a parsimonious correct model by method QDEV was around 0.75 to 0.8 (but the estimated probabilities of selecting correct models were above 0.95). Nevertheless, the estimated probability of selecting the true model at $(\beta_0, \beta_1, \beta_2, \beta_3) = (0.5, 1, 1, 0)$ increased from 0.75 on a 10 × 10 lattices to 0.90 on

Table 1. The estimated probabilities of selection methods for Poisson models under $\rho = 0.3$ with 500 replicates. Methods QDEV_1 and QDEV_2 are based on the working QDEV function with $\alpha = 0.01$ and 0.05, respectively. Darkened values denote the estimated probabilities of selecting the true models.

		Models							
		10 × 10 lattice				15 × 15 lattice			
$(\beta_0, \beta_1, \beta_2, \beta_3)$	Methods	x_1	x_1, x_2	x_1, x_3	x_1, x_2, x_3	x_1	x_1, x_2	x_1, x_3	x_1, x_2, x_3
(0.5,1,1,0)	QDEV_1	0.000	**0.806**	0.024	0.170	0.000	**0.922**	0.014	0.064
	QDEV_2	0.002	**0.752**	0.032	0.214	0.004	**0.906**	0.016	0.074
	AIC_i	0.000	**0.232**	0.018	0.750	0.000	**0.224**	0.012	0.764
	BIC_i	0.000	**0.402**	0.014	0.584	0.000	**0.334**	0.008	0.658
	PQL_A	0.194	**0.482**	0.172	0.152	0.024	**0.614**	0.014	0.348
	PQL_B	0.048	**0.622**	0.034	0.296	0.034	**0.732**	0.016	0.218
(0.5,1,0.5,0.1)	QDEV_1	0.018	0.058	0.010	**0.914**	0.010	0.050	0.016	**0.924**
	QDEV_2	0.014	0.052	0.008	**0.926**	0.006	0.038	0.010	**0.946**
	AIC_i	0.000	0.262	0.004	**0.734**	0.002	0.174	0.004	**0.820**
	BIC_i	0.006	0.392	0.000	**0.602**	0.000	0.334	0.004	**0.662**
	PQL_A	0.154	0.398	0.118	**0.330**	0.132	0.460	0.084	**0.324**
	PQL_B	0.182	0.478	0.092	**0.248**	0.164	0.510	0.076	**0.250**
(0.5,0.2,0.2,0.2)	QDEV_1	0.032	0.040	0.020	**0.908**	0.008	0.024	0.026	**0.942**
	QDEV_2	0.026	0.034	0.018	**0.922**	0.004	0.022	0.016	**0.958**
	AIC_i	0.042	0.166	0.148	**0.644**	0.006	0.064	0.076	**0.854**
	BIC_i	0.100	0.198	0.228	**0.474**	0.018	0.124	0.106	**0.752**
	PQL_A	0.308	0.228	0.200	**0.264**	0.138	0.258	0.246	**0.358**
	PQL_B	0.480	0.196	0.172	**0.152**	0.216	0.244	0.242	**0.298**

15 × 15 lattices, no matter by method QDEV_1 ($\alpha = 0.01$) or method QDEV_2 ($\alpha = 0.05$). This simulation result may provide evidence for 'consistency' of method QDEV in finite samples.

The performance of the other methods, on the other hand, was relatively poor. By using an approximate Z-test to compare the proportions of selecting true models, we found that method QDEV was significantly better than the other methods in the simulation. Another interesting point to observe is that the power of method AIC_i was quite good in a comparison with methods BIC_i, PQL_A, and PQL_B. Although the estimated size of method AIC_i was not good, this method still had high probability of selecting correct models. We also found that the power of methods PQL_A and PQL_B was quite bad. Since the PQL method is designed for longitudinal data, mis-specification of covariance structures in methods PQL_A and PQL_B may cause estimation bias for spatial data.

We also simulated continuous responses on 15 × 15 lattices from the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i. \tag{4.3}$$

Table 2. The estimated probabilities of selection methods for Poisson models under $\rho = 0.5$ with 500 replicates. Methods QDEV_1 and QDEV_2 are based on the working QDEV function with $\alpha = 0.01$ and 0.05, respectively. Darkened values denote the estimated probabilities of selecting the true models.

		Models							
		10 × 10 lattice				15 × 15 lattice			
$(\beta_0, \beta_1, \beta_2, \beta_3)$	Methods	x_1	x_1, x_2	x_1, x_3	x_1, x_2, x_3	x_1	x_1, x_2	x_1, x_3	x_1, x_2, x_3
(0.5,1,1,0)	QDEV_1	0.000	**0.822**	0.026	0.152	0.000	**0.928**	0.014	0.058
	QDEV_2	0.002	**0.776**	0.034	0.188	0.004	**0.910**	0.012	0.074
	AIC_i	0.002	**0.294**	0.014	0.690	0.000	**0.272**	0.008	0.720
	BIC_i	0.002	**0.394**	0.010	0.594	0.002	**0.342**	0.004	0.652
	PQL_A	0.198	**0.456**	0.216	0.130	0.016	**0.614**	0.014	0.356
	PQL_B	0.042	**0.684**	0.032	0.242	0.032	**0.708**	0.024	0.236
(0.5,1,0.5,0.1)	QDEV_1	0.012	0.062	0.016	**0.910**	0.012	0.042	0.014	**0.932**
	QDEV_2	0.010	0.052	0.012	**0.926**	0.004	0.034	0.006	**0.956**
	AIC_i	0.008	0.270	0.006	**0.716**	0.000	0.224	0.000	**0.776**
	BIC_i	0.004	0.396	0.004	**0.596**	0.000	0.328	0.002	**0.670**
	PQL_A	0.166	0.424	0.082	**0.328**	0.100	0.460	0.080	**0.360**
	PQL_B	0.198	0.494	0.082	**0.226**	0.122	0.544	0.092	**0.242**
(0.5,0.2,0.2,0.2)	QDEV_1	0.038	0.046	0.022	**0.894**	0.032	0.024	0.026	**0.918**
	QDEV_2	0.036	0.040	0.022	**0.902**	0.018	0.022	0.024	**0.936**
	AIC_i	0.048	0.156	0.144	**0.652**	0.008	0.058	0.062	**0.872**
	BIC_i	0.086	0.234	0.220	**0.460**	0.018	0.128	0.108	**0.746**
	PQL_A	0.304	0.212	0.242	**0.242**	0.116	0.240	0.240	**0.404**
	PQL_B	0.444	0.220	0.204	**0.132**	0.188	0.262	0.244	**0.306**

For (4.3), the AIC was computed based on a full-likelihood function. We used method QDEV_2 ($\alpha = 0.05$) with a forward selection procedure similar to that in the simulation study of (4.1). Table 4 shows the simulation result. In Table 4, we find that method QDEV is slightly better than AIC. Particularly, the performance of method QDEV seems to improve as the correlation increases. This suggests that the proposed method works well for both discrete and continuous data.

5. Data Analysis

The proposed method is applied to the Lansing Wood data (Diggle (1983)) to examine whether the hickory has mutualism or repulsion with some of the other tree species. The original data set consists of 2,143 trees, among which are 702 hickories, 513 maples, 448 white oaks, 345 red oaks, and 135 black oaks, with a marked location for each tree in a 19.6 acre square plot. We follow Fingleton (1986) to divide the study region into 24 × 24 quadrats. For quadrat i, $i = 1, \ldots, 576$, we label the location by $\boldsymbol{s}_i = (r_i, c_i)$, where r_i and c_i denote the corresponding row and column numbers of quadrat i, respectively. The image

Table 3. The estimated probabilities of selection methods for Poisson models under $\rho = 0.7$ with 500 replicates. Methods QDEV_1 and QDEV_2 are based on the working QDEV function with $\alpha = 0.01$ and 0.05, respectively. Darkened values denote the estimated probabilities of selecting the true models.

		Models							
		10 × 10 lattice				15 × 15 lattice			
$(\beta_0, \beta_1, \beta_2, \beta_3)$	Methods	x_1	x_1, x_2	x_1, x_3	x_1, x_2, x_3	x_1	x_1, x_2	x_1, x_3	x_1, x_2, x_3
(0.5,1,1,0)	QDEV_1	0.000	**0.858**	0.004	0.138	0.000	**0.926**	0.022	0.052
	QDEV_2	0.004	**0.804**	0.008	0.184	0.002	**0.912**	0.018	0.068
	$\mathrm{AIC_i}$	0.004	**0.300**	0.010	0.686	0.000	**0.214**	0.008	0.778
	$\mathrm{BIC_i}$	0.000	**0.384**	0.012	0.604	0.000	**0.386**	0.012	0.602
	$\mathrm{PQL_A}$	0.168	**0.522**	0.180	0.130	0.022	**0.632**	0.026	0.320
	$\mathrm{PQL_B}$	0.046	**0.700**	0.030	0.224	0.028	**0.726**	0.012	0.234
(0.5,1,0.5,0.1)	QDEV_1	0.032	0.030	0.010	**0.914**	0.014	0.028	0.016	**0.942**
	QDEV_2	0.020	0.018	0.010	**0.942**	0.004	0.022	0.006	**0.968**
	$\mathrm{AIC_i}$	0.002	0.354	0.006	**0.638**	0.000	0.234	0.000	**0.766**
	$\mathrm{BIC_i}$	0.004	0.448	0.006	**0.542**	0.000	0.362	0.002	**0.638**
	$\mathrm{PQL_A}$	0.120	0.440	0.092	**0.348**	0.122	0.476	0.050	**0.352**
	$\mathrm{PQL_B}$	0.198	0.522	0.064	**0.216**	0.140	0.550	0.068	**0.242**
(0.5,0.2,0.2,0.2)	QDEV_1	0.036	0.034	0.026	**0.904**	0.016	0.048	0.024	**0.912**
	QDEV_2	0.028	0.032	0.022	**0.928**	0.012	0.032	0.010	**0.946**
	$\mathrm{AIC_i}$	0.060	0.148	0.140	**0.652**	0.010	0.064	0.070	**0.856**
	$\mathrm{BIC_i}$	0.142	0.200	0.200	**0.458**	0.022	0.100	0.074	**0.804**
	$\mathrm{PQL_A}$	0.374	0.208	0.204	**0.214**	0.146	0.250	0.244	**0.360**
	$\mathrm{PQL_B}$	0.522	0.168	0.180	**0.130**	0.182	0.252	0.254	**0.312**

Table 4. The estimated probabilities of selection methods in a linear model. Method QDEV_2 is computed by a working QDEV with $\alpha = 0.05$, and method AIC is computed under a maximum likelihood estimate. The true model is marked by $*$ with simulation results based on 500 replicates on a 15 × 15 lattice.

		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
$(\beta_0, \beta_1, \beta_2, \beta_3)$	Models	QDEV_2	AIC	QDEV_2	AIC	QDEV_2	AIC
(0.5,1,1,0)	x_1	0.004	0.018	0.008	0.012	0.000	0.008
	$^*x_1, x_2$	0.834	0.670	0.850	0.724	0.904	0.768
	x_1, x_3	0.004	0.026	0.018	0.014	0.008	0.014
	x_1, x_2, x_3	0.158	0.286	0.124	0.250	0.088	0.210
(0.5,1,0.5,0.1)	x_1	0.000	0.008	0.000	0.000	0.000	0.000
	x_1, x_2	0.084	0.098	0.036	0.080	0.004	0.064
	x_1, x_3	0.010	0.022	0.008	0.012	0.000	0.006
	$^*x_1, x_2, x_3$	0.906	0.872	0.956	0.908	0.996	0.930
(0.5,0.2,0.2,0.2)	x_1	0.004	0.004	0.000	0.004	0.000	0.008
	x_1, x_2	0.018	0.018	0.014	0.026	0.002	0.030
	x_1, x_3	0.018	0.020	0.012	0.024	0.006	0.022
	$^*x_1, x_2, x_3$	0.960	0.958	0.974	0.946	0.992	0.940

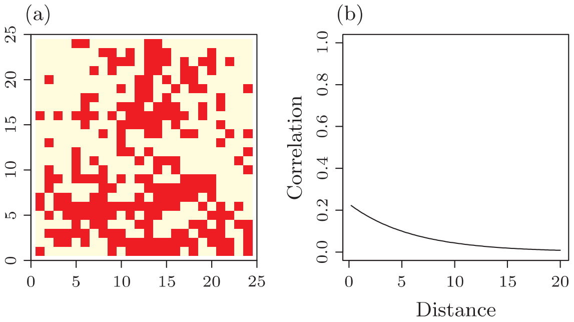

Figure 1. (a) Image plot for the hickories. (b) The fitted exponential correlation model $0.23\exp(-\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|_2/5.6)$ for the hickories. The degree of darkness represents the number of trees, with scale from 0 (white) to 8 (black).

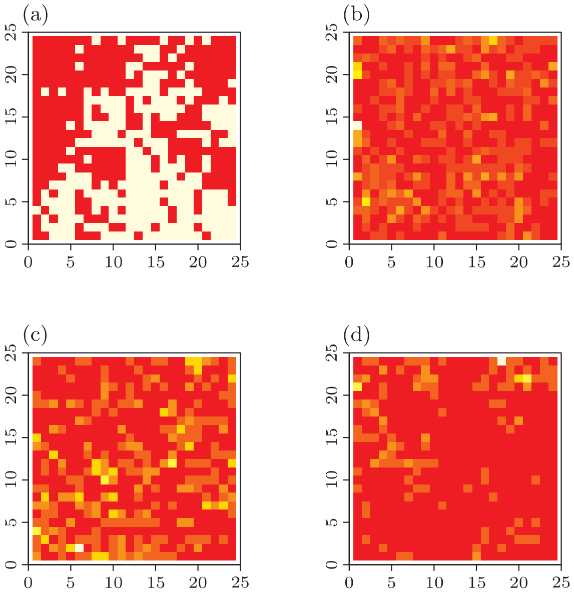

Figure 2. Image plots for (a) the maples, (b) the white oaks, (c) the red oaks, and (d) the black oaks in the Lansing Woods data. The degree of darkness represents the number of trees, with scale from 0 (white) to 9 (black).

plots for the numbers of trees from a re-scaled data set in Software R are shown in Figure 1(a) and Figure 2.

To model the observations, let Y_i denote the number of hickories at site $\boldsymbol{s}_i$. We assume that Y_i follows a Poisson generalized linear mixed model with the mean function

$$\theta_i = E\{Y_j : x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}\} = \exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4}) \quad (5.1)$$

Table 5. Analysis of deviance (Dev) for the Lansing Wood data. The estimates with standard deviations in parentheses listed in the table are 10 times the true values.

Models	Estimate					Reduction	Dev (df)	p-value
	β_0	β_1	β_2	β_3	β_4			
$\mathcal{M}_0$	3.63	0	0	0	0	$\mathcal{M}_1 \to \mathcal{M}_0$	37.4 (2)	0.00
	(1.60)	–	–	–	–			
$\mathcal{M}_1$	5.60	-1.60	-1.69	0	0	$\mathcal{M}_2 \to \mathcal{M}_1$	0.87 (1)	0.35
	(1.44)	(0.36)	(0.39)	–	–	$\mathcal{M}_3 \to \mathcal{M}_1$	3.25 (1)	0.07
$\mathcal{M}_2$	5.90	-1.65	-1.73	-0.38	0	$\mathcal{M}_4 \to \mathcal{M}_2$	3.27 (1)	0.07
	(1.48)	(0.37)	(0.40)	(0.38)	–			
$\mathcal{M}_3$	5.90	-1.60	-1.72	0	-1.01	$\mathcal{M}_4 \to \mathcal{M}_3$	0.87 (1)	0.35
	(1.45)	(0.37)	(0.40)	–	(0.66)			
$\mathcal{M}_4$	6.44	-1.67	-1.73	-0.38	-1.01	$\mathcal{M}_4 \to \mathcal{M}_1$	4.20 (2)	0.12
	(1.45)	(0.37)	(0.40)	(0.38)	(0.65)			

and the covariance structure given by (4.2), where $x_{i,1}$, $x_{i,2}$, $x_{i,3}$, and $x_{i,4}$ denote the numbers of maples, white oaks, red oaks and black oaks at site $\boldsymbol{s}_i$, respectively. In the following analysis, the intercept is always in the model.

To find a parsimonious correct model for the numbers of hickories, we use a selection procedure based on the working QDEV function with criterion (3.1). (Details about the working QDEV function can be found in Section 4.) The selection procedure is stepwise with $\alpha = 0.05$ in each add-in or drop-out step. Under the full model (5.1), an iterative procedure gives an estimate for the correlation by $\text{corr}(Y_i, Y_j) = 0.23 \exp\{-\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2/5.6\}$. The estimated correlation function is shown in Figure 1(b). Note that the estimated correlation model satisfies the assumptions of asymptotic normality shown in Section 2.

We show some selection results in Table 5. There, $\mathcal{M}_0$ denotes the model consisting of only the intercept, and $\boldsymbol{x}_j = (x_{1,j}, \ldots, x_{576,j})^T$, $j = 1, \ldots, 4$. We use $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$ to denote the models consisting of $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_4\}$ and $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4\}$, respectively. Table 5 shows the QL estimates with the corresponding standard deviations for model $\mathcal{M}_j$, $j = 0, \ldots, 4$. We observe that, no matter which model is specified, both $\hat{\beta}_1$ and $\hat{\beta}_2$ have significant χ^2-values around 7.11 and 7.33, respectively, compared to the χ^2_1 table.

We also computed the deviance reduction for some pairs of nested models listed in Table 5. A comparison of deviances between models $\mathcal{M}_3$ and $\mathcal{M}_1$ suggests that adding covariate $\boldsymbol{x}_3$ to $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ could slightly increase the significance of the model. Since the estimates of β_1, β_2, and β_4 in model $\mathcal{M}_3$ are negative, we conclude that, based on the results shown in Table 5, the numbers of maples and white oaks significantly defer the growth of hickories, and that the black oak also causes some slight repulsion to the hickories.

Finally, it is also interesting to note that

$$\begin{aligned} &\text{Dev}(\mathcal{M}_4,\mathcal{M}_1) \doteq \text{Dev}(\mathcal{M}_4,\mathcal{M}_3)+\text{Dev}(\mathcal{M}_3,\mathcal{M}_1),\\ \text{and}\quad &\text{Dev}(\mathcal{M}_4,\mathcal{M}_1) \doteq \text{Dev}(\mathcal{M}_4,\ \mathcal{M}_2)+\text{Dev}(\mathcal{M}_2,\mathcal{M}_1). \end{aligned}$$

These results suggest that the additivity of the QDEV functions could hold when the parameter values are small, see the discussion.

6. Discussion

We have proposed a generalized deviance function with a combination of the QL estimating equation as model selection criterion for spatially correlated data. Under certain mixing conditions, the proposed QDEV function has an asymptotic chi-squared distribution and the related model selection could be consistent. We also conducted simulations to compare the proposed method with other criteria. Simulations showed that, in most cases, the proposed method had higher probabilities of selecting the true models than did the other criteria.

In the GEE setting, many people use AIC to compare different correlation structures (Pan and Connett (2002); Cui and Qian (2007)). We also use the working QDEV function with correlation estimated from the variogram as the selection criterion. Nevertheless, when responses satisfy the assumptions of Lemma 1, in a small simulation study, we found that using exponential or spherical correlation models in the working QDEV function would not make a significant difference in the selection result. So, the proposed method may be unsuitable for choosing a covariance structure.

To study additivity of the deviance function, Li (1993) claimed that, when the models are not very different, this property may approximately hold. We have seen in our data analysis that the asymptotic additivity may be valid in some situations. To discuss the additivity property in detail, we assume that $\tau_1 \subset \tau_2 \subset \tau$ are some index subsets of the parameters. Then, equation (2.4) implies that $2D(\hat{\boldsymbol{\beta}}_\tau,\hat{\boldsymbol{\beta}}_{\tau_1}) \doteq 2D(\hat{\boldsymbol{\beta}}_\tau,\hat{\boldsymbol{\beta}}_{\tau_2}) + 2D(\hat{\boldsymbol{\beta}}_{\tau_2},\hat{\boldsymbol{\beta}}_{\tau_1})$, if $\boldsymbol{I}^{-1}_{\boldsymbol{\beta}_{\tau_1}\boldsymbol{\beta}_{\tau_1}} \doteq \boldsymbol{I}^{-1}_{\boldsymbol{\beta}_{\tau_2}\boldsymbol{\beta}_{\tau_2}}$. However, theoretical details for an exact condition of asymptotic additivity may be complicated since the error rate may need to be computed.

In Section 3, we presented the asymptotic behavior of the QDEV function under strictly non-nested and nested models. However, for overlapping models, the proposed method may not be easy to apply. One possible approach to generalizing the proposed method to overlapping models would come from the variance test in Vuong (1989). Besides, our model selection in this paper is mainly built on a stepwise procedure, but when the number of covariates is large, this approach may cause computational burden. A more efficient method would be to develop an AIC-type method for spatial data, as Pan (2001) had done for longitudinal data. This needs lots of theoretical work.

Acknowledgement

The author would like to thank editor, an associate editor and two anonymous referees for their constructive comments toward improving the paper.

Appendix

Proof of Lemma 1. Since $\alpha(r) = o(r^{-d})$, we can find some $\epsilon > 0$ such that $\alpha(r) = O(r^{-d-\epsilon})$. We then note that

$$\sum_{\boldsymbol{s}\in\Omega_n} \text{cov}\{Y_0, Y(\boldsymbol{s})\} \leq c_1 \sum_{r=1}^{\infty} r^d r^{-1-d-\epsilon} < \infty$$

for some $c_1 > 0$. Therefore, $\text{var}(S_n) = O(|\Omega_n|)$. For convenience, we assume that $O(|\Omega_n|) = O(n^d)$. Let k be a number satisfying $0.5 > k > d/2(d+\epsilon)$, and $m_n = O(n^k)$. Then, for $n \to \infty$,

$$\alpha(m_n)|\Omega_n|^{1/2} \to 0, \quad \text{and} \quad m_n^{-d}|\Omega_n|^{1/2} \to \infty. \tag{A.1}$$

For a given $\boldsymbol{s} \in \Omega_n$, we partition Z_n as $Z_{\boldsymbol{s},n} + Z^*_{\boldsymbol{s},n}$ by m_n, where $Z_{\boldsymbol{s},n} = \sum_{\|\boldsymbol{s}-\boldsymbol{s}'\|\leq m_n} Y(\boldsymbol{s}')$. Without loss of generality, we assume that $E\{Y(\boldsymbol{s})\} = 0$. Then,

$$\text{var}(Z_n) = E\left\{\sum_{\boldsymbol{s}\in\Omega_n} Y(\boldsymbol{s})Z_{\boldsymbol{s},n}\right\} + E\left\{\sum_{\boldsymbol{s}\in\Omega_n} Y(\boldsymbol{s})Z^*_{\boldsymbol{s},n}\right\}.$$

Let $\phi_n^2 = \text{var}(\sum_{\boldsymbol{s}\in\Omega_n} Z_{\boldsymbol{s},n})$, $\bar{Z}_n = Z_n/\phi_n$ and $\bar{Z}_{\boldsymbol{s},n} = Z_{\boldsymbol{s},n}/\phi_n$. Since

$$E\left\{\sum_{\boldsymbol{s}\in\Omega_n} Y(\boldsymbol{s})Z^*_{\boldsymbol{s},n}\right\} = \sum_{\boldsymbol{s}\in\Omega_n} \sum_{\|\boldsymbol{s}-\boldsymbol{s}'\|>m_n} \text{cov}\{Y(\boldsymbol{s}), Y(\boldsymbol{s}')\} \leq c_2 n^d \sum_{r=m_m}^{\infty} r^d r^{-d-\epsilon}, \tag{A.2}$$

which is $o(n^d)$ for some $c_2 > 0$, we have $\phi_n^2 = \text{var}(Z_n)\{1+o(1)\} = |\Omega_n|\{1+o(1)\}$. So, to show the asymptotic normality of Z_n/ν_n is equivalent to showing that of $\bar{Z}_n$.

The central limit theorem follows from proving Stein's condition

$$\lim_{n\to\infty} E\left\{(i\lambda - \bar{Z}_n)e^{i\lambda\bar{Z}_n}\right\} = 0. \tag{A.3}$$

Since $E\{Y^2(\boldsymbol{s})\} < \infty$ and $\text{cov}(Y_0, Z_n) < \infty$, it suffices to show (A.3) for bounded random variables (Guyon (1995)). We follow the procedure of Bolthausen (1982) to decompose $(i\lambda - \bar{Z}_n)e^{i\lambda\bar{Z}_n}$ to

$$i\lambda e^{i\lambda\bar{Z}_n}\left\{1 - \phi_n^{-2}\sum_{\boldsymbol{s}\in\Omega_n} Y(\boldsymbol{s})Z_{\boldsymbol{s},n}\right\} - \phi_n^{-1}e^{i\lambda\bar{Z}_n}\sum_{\boldsymbol{s}\in\Omega_n} Y(\boldsymbol{s})(1 - i\lambda\bar{Z}_{\boldsymbol{s},n} - e^{-i\lambda\bar{Z}_{\boldsymbol{s},n}})$$
$$-\phi_n^{-1}\sum_{\boldsymbol{s}\in\Omega_n} Y(\boldsymbol{s})e^{i\lambda(\bar{Z}_n - \bar{Z}_{\boldsymbol{s},n})}. \tag{A.4}$$

Let D_1, D_2, and D_3 be the first, second, and third items of (A.4). We note that

$$E(D_1^2) = \lambda^2 \phi_n^{-4} \sum_{\{\boldsymbol{s}\in\Omega_n, \|\boldsymbol{s}-\boldsymbol{s}'\|\le m_n\}}\sum \sum_{\{\boldsymbol{l}\in\Omega_n, \|\boldsymbol{l}-\boldsymbol{l}'\|\le m_n\}}\sum \mathrm{cov}\{Y(\boldsymbol{s})Y(\boldsymbol{s}'), Y(\boldsymbol{l})Y(\boldsymbol{l}')\}.$$

It then follows from Assumptions 2 and 3 that some $c_3 > 0$ exists such that

$$E(D_1^2) \le c_3\lambda^2\, \phi_n^{-4}\, |\Omega_n|\, m_n^{2d} \sum_{r^*=1}^{\infty} (r^*)^d \alpha(r^*) = \lambda^2 O(|\Omega_n|^{-1} m_n^{2d}),$$

which is $o(1)$ by (A.1). For the asymptotic behavior of D_2, recall that $Y(\boldsymbol{s})$ is assumed to be a bounded random variable. We can thus find some constants c_4 and c_5 such that

$$\bar{Z}_{\boldsymbol{s},n} \le c_4\,(1+\cdots+m_n^{d-1})|\Omega_n|^{-1/2} \le c_5|\Omega_n|^{-1/2}m_n^d \to 0, \qquad \text{(A.5)}$$

as $n \to \infty$. So, by Taylor's expansion, $|1 - i\lambda\bar{Z}_{\boldsymbol{s},n} - e^{-i\lambda\bar{Z}_{\boldsymbol{s},n}}| \le c_6\lambda^2\bar{Z}^2_{\boldsymbol{s},n}$ for some $c_6 > 0$. Then

$$E|D_2| \le c_6\lambda^2|\Omega_n|^{1/2} \max_{\boldsymbol{s}\in\Omega_n} E(\bar{Z}_{\boldsymbol{s},n})^2.$$

Using an argument similar to (A.2) and (A.5) with the above inequality gives $E|D_2| \le c_7|\Omega_n|^{1/2}m_n^{-d-1}$ for some $c_7 > 0$. It thus follows from (A.1) that $E|D_2| \to 0$ as $n \to \infty$. Finally, we can use an argument, similar to the proof of Theorem 1 of Lin (2008), to show that $E(D_3) \to 0$ as $n \to \infty$. Equation (A.3) then follows from the above results.

Proof of Theorem 1. Since $\nabla_{\boldsymbol{\psi}} D(\boldsymbol{\psi}, \boldsymbol{\beta}_0)|_{\boldsymbol{\psi}=\boldsymbol{\beta}_0} = \dot{\boldsymbol{\theta}}^T(\boldsymbol{\beta}_0)\boldsymbol{V}^{-1}(\boldsymbol{\beta}_0)\{Y - \boldsymbol{\theta}(\boldsymbol{\beta}_0)\}$ and $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \boldsymbol{O}_p(n^{-1/2})$, we have

$$\begin{aligned}\nabla_{\boldsymbol{\psi}} D(\boldsymbol{\psi}, \boldsymbol{\beta}_0)|_{\boldsymbol{\psi}=\boldsymbol{\beta}_0} = {} & \dot{\boldsymbol{\theta}}^T(\hat{\boldsymbol{\beta}})\boldsymbol{V}^{-1}(\hat{\boldsymbol{\beta}})\{\boldsymbol{Y} - \boldsymbol{\theta}(\hat{\boldsymbol{\beta}})\} + \dot{\boldsymbol{\theta}}^T(\hat{\boldsymbol{\beta}})\boldsymbol{V}^{-1}(\hat{\boldsymbol{\beta}})\{\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\theta}(\boldsymbol{\beta}_0)\} \\ & + \boldsymbol{O}_p(1), \qquad \text{(A.6)}\end{aligned}$$

which is almost surely equal to $\dot{\boldsymbol{\theta}}^T(\hat{\boldsymbol{\beta}})\boldsymbol{V}^{-1}(\hat{\boldsymbol{\beta}})\{\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\theta}(\boldsymbol{\beta}_0)\} + \boldsymbol{O}_p(1)$ by (2.1). A first-order Taylor expansion on $\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\theta}(\boldsymbol{\beta}_0)$ gives

$$\nabla_{\boldsymbol{\psi}} D(\boldsymbol{\psi}, \boldsymbol{\beta}_0)|_{\boldsymbol{\psi}=\boldsymbol{\beta}_0} = \dot{\boldsymbol{\theta}}^T(\hat{\boldsymbol{\beta}})\boldsymbol{V}^{-1}(\hat{\boldsymbol{\beta}})\dot{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \boldsymbol{O}_p(1). \qquad \text{(A.7)}$$

Also, $\nabla^2_{\boldsymbol{\psi}} D(\boldsymbol{\psi}, \boldsymbol{\beta}_0)|_{\boldsymbol{\psi}=\boldsymbol{\beta}_0} = -\dot{\boldsymbol{\theta}}^T(\boldsymbol{\beta}_0)\boldsymbol{V}^{-1}(\boldsymbol{\beta}_0)\dot{\boldsymbol{\theta}}(\boldsymbol{\beta}_0) + \boldsymbol{O}_p(n^{1/2})$. Since

$$\begin{aligned}D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = {} & D(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0) + \left\{\nabla_{\boldsymbol{\psi}} D(\boldsymbol{\psi}, \boldsymbol{\beta}_0)|_{\boldsymbol{\psi}=\boldsymbol{\beta}_0}\right\}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ & + 0.5(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\left\{\nabla^2_{\boldsymbol{\psi}} D(\boldsymbol{\psi}, \boldsymbol{\beta}_0)|_{\boldsymbol{\psi}=\boldsymbol{\beta}_0}\right\}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2), \text{(A.8)}\end{aligned}$$

combining (A.6)−(A.8) gives

$$\begin{aligned}D(\hat{\boldsymbol\beta},\boldsymbol\beta_0) &= (\hat{\boldsymbol\beta}-\boldsymbol\beta_0)^T\{\dot{\boldsymbol\theta}^T(\hat{\boldsymbol\beta})\boldsymbol V^{-1}(\hat{\boldsymbol\beta})\dot{\boldsymbol\theta}(\hat{\boldsymbol\beta}) - 0.5\,\dot{\boldsymbol\theta}^T(\boldsymbol\beta_0)\boldsymbol V^{-1}(\boldsymbol\beta_0)\dot{\boldsymbol\theta}(\boldsymbol\beta_0)\}(\hat{\boldsymbol\beta}-\boldsymbol\beta_0)\\ &\quad +O_p(n^{-1/2}).\end{aligned}$$

Since $\dot{\boldsymbol\theta}^T(\hat{\boldsymbol\beta})\boldsymbol V^{-1}(\hat{\boldsymbol\beta})\dot{\boldsymbol\theta}(\hat{\boldsymbol\beta}) = \dot{\boldsymbol\theta}^T(\boldsymbol\beta_0)\boldsymbol V^{-1}(\boldsymbol\beta_0)\dot{\boldsymbol\theta}(\boldsymbol\beta_0)+\boldsymbol O_p(n^{1/2})$, the above equation is

$$D(\hat{\boldsymbol\beta},\boldsymbol\beta_0) = 0.5\,(\hat{\boldsymbol\beta}-\boldsymbol\beta_0)^T\{\dot{\boldsymbol\theta}^T(\boldsymbol\beta_0)\boldsymbol V^{-1}(\boldsymbol\beta_0)\dot{\boldsymbol\theta}(\boldsymbol\beta_0)\}(\hat{\boldsymbol\beta}-\boldsymbol\beta_0) + O_p(n^{-1/2}),$$

which gives the desired result.

Proof of Theorem 2. Assume that $\hat{\boldsymbol\xi} = (\hat{\boldsymbol\psi}^T,\hat{\boldsymbol\beta}^T)^T$ and $\boldsymbol\xi_H = (\boldsymbol\psi_0^T,\boldsymbol\beta^T)^T$. Without loss of generality, take $\boldsymbol\psi_0 = \boldsymbol 0$. We extend $\tilde{\boldsymbol\beta}$ to be $(\boldsymbol\psi_0^T,\tilde{\boldsymbol\beta}^T)^T$. The deviance $2D(\hat{\boldsymbol\xi},\tilde{\boldsymbol\beta})$ can be then decomposed as

$$\{\boldsymbol\theta(\hat{\boldsymbol\xi})-\boldsymbol\theta(\boldsymbol\xi_H)+\boldsymbol\theta(\boldsymbol\xi_H)-\boldsymbol\theta(\tilde{\boldsymbol\beta})\}^T\left[\boldsymbol V^{-1}(\hat{\boldsymbol\xi})\{\boldsymbol Y-\boldsymbol\theta(\hat{\boldsymbol\xi})\}+\boldsymbol V^{-1}(\tilde{\boldsymbol\beta})\{\boldsymbol Y-\boldsymbol\theta(\tilde{\boldsymbol\beta})\}\right]. \tag{A.9}$$

Under the null hypothesis H, using techniques similar to the proofs of Theorems 1−3 of Lin (2008) for $\boldsymbol U(\hat{\boldsymbol\xi})$ and $\boldsymbol U(\tilde{\boldsymbol\beta})$ gives consistency of $\hat{\boldsymbol\xi}$ and $\tilde{\boldsymbol\beta}$. Also, $n^{1/2}(\hat{\boldsymbol\xi}-\boldsymbol\xi_H)$ and $n^{1/2}(\tilde{\boldsymbol\beta}-\boldsymbol\xi_H)$ converge in distribution to some multivariate normal distributions. Since $\dot{\boldsymbol\theta}(\tilde{\boldsymbol\beta})\boldsymbol V^{-1}(\tilde{\boldsymbol\beta})\{\boldsymbol Y-\boldsymbol\theta(\tilde{\boldsymbol\beta})\} = \boldsymbol 0$, we can show that $\{\boldsymbol\theta(\boldsymbol\xi_H)-\boldsymbol\theta(\tilde{\boldsymbol\beta})\}^T\boldsymbol V^{-1}(\tilde{\boldsymbol\beta})\{\boldsymbol Y-\boldsymbol\theta(\tilde{\boldsymbol\beta})\} = O_p(n^{-1/2})$. Similarly, $\{\boldsymbol\theta(\boldsymbol\xi_H)-\boldsymbol\theta(\hat{\boldsymbol\xi})\}^T\boldsymbol V^{-1}(\hat{\boldsymbol\xi})\{\boldsymbol Y-\boldsymbol\theta(\hat{\boldsymbol\xi})\} = O_p(n^{-1/2})$. Expression (A.9) can be simplified to

$$\begin{aligned}&\{\boldsymbol\theta(\hat{\boldsymbol\xi})-\boldsymbol\theta(\boldsymbol\xi_H)\}^T\boldsymbol V^{-1}(\tilde{\boldsymbol\beta})\{\boldsymbol Y-\boldsymbol\theta(\tilde{\boldsymbol\beta})\}+\{\boldsymbol\theta(\boldsymbol\xi_H)-\boldsymbol\theta(\tilde{\boldsymbol\beta})\}^T\boldsymbol V^{-1}(\hat{\boldsymbol\xi})\{\boldsymbol Y-\boldsymbol\theta(\hat{\boldsymbol\xi})\}\\ &\quad +O_p(n^{-1/2}).\end{aligned}$$

Let

$$\boldsymbol G\begin{pmatrix}\boldsymbol\psi\\ \boldsymbol\beta\end{pmatrix} = \boldsymbol V^{-1}\begin{pmatrix}\boldsymbol\psi\\ \boldsymbol\beta\end{pmatrix}\left\{\boldsymbol Y-\boldsymbol\theta\begin{pmatrix}\boldsymbol\psi\\ \boldsymbol\beta\end{pmatrix}\right\}$$

and $\Delta_{\boldsymbol\beta}\boldsymbol G\begin{pmatrix}\boldsymbol\psi_0\\ \boldsymbol\beta\end{pmatrix}$ be the derivative array of $\boldsymbol G\begin{pmatrix}\boldsymbol\psi_0\\ \boldsymbol\beta\end{pmatrix}$ with respect to $\boldsymbol\beta$. An asymptotic expression for $2D(\hat{\boldsymbol{\xi}},\tilde{\boldsymbol{\beta}})$ with order $O_p(n^{-1/2})$ is

$$\begin{aligned}&\left\{\begin{pmatrix}\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0\\ \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\end{pmatrix}^T\dot{\boldsymbol{\theta}}^T\begin{pmatrix}\boldsymbol{\psi}_0\\ \boldsymbol{\beta}\end{pmatrix}+\boldsymbol{o}_p(n^{-1/2})\right\}\\ &\quad\cdot\left[\boldsymbol{G}\begin{pmatrix}\boldsymbol{\psi}_0\\ \boldsymbol{\beta}\end{pmatrix}+\left\{\Delta_{\boldsymbol{\beta}}\boldsymbol{G}\begin{pmatrix}\boldsymbol{\psi}_0\\ \boldsymbol{\beta}\end{pmatrix}\right\}^T\cdot\begin{pmatrix}\boldsymbol{0}\\ \tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\end{pmatrix}+\boldsymbol{o}_p(n^{-1/2})\right]\\ &-\left\{\begin{pmatrix}\boldsymbol{0}\\ \tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\end{pmatrix}^T\dot{\boldsymbol{\theta}}^T\begin{pmatrix}\boldsymbol{\psi}_0\\ \boldsymbol{\beta}\end{pmatrix}+\boldsymbol{o}_p(n^{-1/2})\right\}\\ &\quad\cdot\left[\boldsymbol{G}\begin{pmatrix}\boldsymbol{\psi}_0\\ \boldsymbol{\beta}\end{pmatrix}+\left\{\Delta_{\boldsymbol{\beta}}\boldsymbol{G}\begin{pmatrix}\boldsymbol{\psi}_0\\ \boldsymbol{\beta}\end{pmatrix}\right\}^T\cdot\begin{pmatrix}\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0\\ \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\end{pmatrix}+\boldsymbol{o}_p(n^{-1/2})\right].\end{aligned}\tag{A.10}$$

After tedious algebra on (A.10), we obtain

$$2D(\hat{\boldsymbol{\xi}},\tilde{\boldsymbol{\beta}})=\begin{pmatrix}\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0\\ \hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\end{pmatrix}^T\dot{\boldsymbol{\theta}}^T\begin{pmatrix}\boldsymbol{\psi}_0\\ \boldsymbol{\beta}\end{pmatrix}\boldsymbol{V}^{-1}\begin{pmatrix}\boldsymbol{\psi}_0\\ \boldsymbol{\beta}\end{pmatrix}\dot{\boldsymbol{\theta}}\begin{pmatrix}\boldsymbol{\psi}_0\\ \boldsymbol{\beta}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0\\ \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\end{pmatrix}+o_p(1).\tag{A.11}$$

Let

$$\boldsymbol{I}(\boldsymbol{\psi}_0,\boldsymbol{\beta})=\lim_{n\to\infty}\frac{1}{n}\dot{\boldsymbol{\theta}}^T\begin{pmatrix}\boldsymbol{\psi}_0\\ \boldsymbol{\beta}\end{pmatrix}\boldsymbol{V}^{-1}\begin{pmatrix}\boldsymbol{\psi}_0\\ \boldsymbol{\beta}\end{pmatrix}\dot{\boldsymbol{\theta}}\begin{pmatrix}\boldsymbol{\psi}_0\\ \boldsymbol{\beta}\end{pmatrix}.$$

We note that, under H, $\boldsymbol{I}(\boldsymbol{\psi}_0,\boldsymbol{\beta})$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\xi}}$ from (2.2). Decompose the information matrix $\boldsymbol{I}(\boldsymbol{\psi}_0,\boldsymbol{\beta})$ as

$$\boldsymbol{I}(\boldsymbol{\psi}_0,\boldsymbol{\beta})=\begin{pmatrix}\boldsymbol{I}_{\psi_0\psi_0} & \boldsymbol{I}_{\psi_0\beta}\\ \boldsymbol{I}_{\beta\psi_0} & \boldsymbol{I}_{\beta\beta}\end{pmatrix},$$

according to the partition of $(\boldsymbol{\psi}_0,\boldsymbol{\beta})$. Since

$$n^{1/2}\begin{pmatrix}\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0\\ \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\end{pmatrix}\ \to\ N\left\{\boldsymbol{0},\begin{pmatrix}\boldsymbol{I}_{\psi_0\psi_0} & \boldsymbol{I}_{\psi_0\beta}\\ \boldsymbol{I}_{\beta\psi_0} & \boldsymbol{I}_{\beta\beta}\end{pmatrix}^{-1}\right\}\quad\text{in distribution,}$$

we can obtain an asymptotic relationship

$$\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}+\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{I}_{\beta\psi_0}(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0),\tag{A.12}$$

with order $\boldsymbol{O}_p(n^{-1/2})$, for the restricted QL estimate $\tilde{\boldsymbol{\beta}}$ from the property of multivariate normal densities (Cox and Hinkley (1974, p.308)). Plugging (A.12) into (A.11) gives

$$2D(\hat{\boldsymbol{\xi}},\tilde{\boldsymbol{\beta}})=n(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0)^T(\boldsymbol{I}_{\psi_0\psi_0}-\boldsymbol{I}_{\psi_0\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{I}_{\beta\psi_0})(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0)+o_p(1).$$

The desired result then follows from the equation $\boldsymbol{I}_{\psi_0\psi_0}^{-1} = \boldsymbol{I}_{\psi_0\psi_0} - \boldsymbol{I}_{\psi_0\beta}\boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{I}_{\beta\psi_0}$.

Proof of Theorem 3. To show Theorem 3, we first note that $n^{1/2}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}_H)$ converges to a non-central normal distribution. With a modification of (A.10), we obtain a formula similar to (A.11) under the alternative hypothesis K:

$$2D(\hat{\boldsymbol{\xi}}, \tilde{\boldsymbol{\beta}}) = \begin{pmatrix} \hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_k \\ \hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} \end{pmatrix}^T \dot{\boldsymbol{\theta}}^T\begin{pmatrix} \boldsymbol{\psi}_k \\ \boldsymbol{\beta} \end{pmatrix}\boldsymbol{V}^{-1}\begin{pmatrix} \boldsymbol{\psi}_k \\ \boldsymbol{\beta} \end{pmatrix}\dot{\boldsymbol{\theta}}\begin{pmatrix} \boldsymbol{\psi}_k \\ \boldsymbol{\beta} \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_k \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \end{pmatrix}$$
$$+ \begin{pmatrix} \boldsymbol{\psi}_k - \boldsymbol{\psi}_0 \\ \hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} \end{pmatrix}^T \dot{\boldsymbol{\theta}}^T\begin{pmatrix} \boldsymbol{\psi}_k \\ \boldsymbol{\beta} \end{pmatrix}\boldsymbol{V}^{-1}\begin{pmatrix} \boldsymbol{\psi}_k \\ \boldsymbol{\beta} \end{pmatrix}\dot{\boldsymbol{\theta}}\begin{pmatrix} \boldsymbol{\psi}_k \\ \boldsymbol{\beta} \end{pmatrix}\begin{pmatrix} \boldsymbol{\psi}_k - \boldsymbol{\psi}_0 \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \end{pmatrix} + o_p(1). \quad \text{(A.13)}$$

Since

$$n^{1/2}\begin{pmatrix} \hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_k \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \end{pmatrix} \rightarrow N\left\{\boldsymbol{0}, \begin{pmatrix} \boldsymbol{I}_{\psi_k\psi_k} & \boldsymbol{I}_{\psi_k\beta} \\ \boldsymbol{I}_{\beta\psi_k} & \boldsymbol{I}_{\beta\beta} \end{pmatrix}^{-1}\right\} \quad \text{in distribution,}$$

we obtain an asymptotic relationship similar to (A.12)

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \boldsymbol{I}_{\beta\beta}^{-1}\boldsymbol{I}_{\beta\psi_k}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_k) + \boldsymbol{O}_p(n^{-1/2}). \quad \text{(A.14)}$$

Plugging (A.14) into (A.13), with simplification, then leads to the desired result.

References

Bolthausen, E. (1982). On the central limit theorem for stationary mixing random fields. *Ann. Probab.* **10**, 1047-1050.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **89**, 9-24.

Cox, D. R. and Hinkley (1974). *Theoretical Statistics.* Chapman and Hall, New York.

Cui, J. and Qian, G. (2007). Selection of working correlation structure and best model in GEE analysis of longitudinal data. *Comm. Statist. Simulation Comput.* **36**, 987-996.

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns.* Academic Press, London.

Fingleton, B. (1986). Analyzing cross-classified data with inherent spatial dependence. *Geog. Anal.* **18**, 48-61.

Godambe, V. P. (1991). *Estimating Functions.* Oxford University Press, Oxford.

Guyon, X. (1995). *Random Fields on a Network.* Springer, New York.

Heagerty, P. J. and Lumley, T. (2000). Window subsampling of estimating functions with application to regression models. *J. Amer. Statist. Assoc.* **95**, 197-211.

Hanfelt, J. J. and Liang, K.-Y. (1995). Approximate likelihood ratios for generalized estimating functions. *Biometrika* **82**, 461-477.

Li, B. (1993). A deviance function for the quasi-likelihood method. *Biometrika* **80**, 741-753.

Lin, P.-S. (2008). Estimating equation for spatially correlated data in multi-dimensional space. *Biometrika* **95**, 847-858.

McLeish, D. L. and Small, C. G. (1992). A projected likelihood function for semiparametric models. *Biometrika* **79**, 93-102.

McCullagh, P. (1991). Quasi-likelihood and estimating functions. In *Statistical Theory and Modelling* (Edited by D. V. Hinkley, N. Reid and E. J. Snell), 265-286. Chapman and Hall, New York.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120-125.

Pan, W. and Connett, J. E. (2002). Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statist. Sinica*, **12**, 475-490.

Qian, Q., Gabor, G. and Gupta, R. P. (1996). Generalized linear model selection by the predictive least quasi-deviance criterion. *Biometrika* **83**, 41-54.

Qu, A., Lindsay, B. G. and Li, B. (2000). Improving generalized estimating equations using quadratic inference functions. *Biometrika*, **87**, 823-836.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.

Shao, J. (1996). An asymptotic theory for linear model selection. *J. Amer. Statist. Assoc.* **91**, 655-665.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypothesis. *Econometrica* **57**, 307-333.

Wang, L. and Qu, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *J. Roy. Statist. Soc. Ser. B* **71**, 177-190.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Ann. Statist.* **37**, 2178-2201.

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

Division of Biostatistics and Bioinformatics, National Health Research Institutes, Zhunan, Miaoli 350, Taiwan.

Department of Mathematics, National Chung Cheng University, Taiwan.

E-mail: pslin@nhri.org.tw; pslin@math.ccu.edu.tw

(Received November 2009; accepted July 2010)