数理解析研究所講究録
第1910巻 2014年 20-28

Model Selection Based on Quasi-likelihood with Application to Overdispersed Data

Yiping Tang and Jinfang Wang
Graduate School of Science, Chiba University
1-33 Yayoi-cho, Inage-ku, Chiba 263-8522 Japan
E-mail: tou.kihei@gmail.com

Abstract

In analyzing complicated data, we are often unwilling or not confident to impose a parametric model for the data-generating structure. One important example is data analysis for proportional or count data with overdispersion. The obvious advantage of assuming full parametric models is that one can resort to likelihood analyses, for instance, to use AIC or BIC to choose the most appropriate regression models. For overdispersed proportional data, possible parametric models include the Beta-binomial models, the double exponential models, etc. In this paper, we extend the generalized linear models by replacing the full parametric models with a finite number of moment restrictions on both the data and the structural parameters. For such semiparametric statistical models, we propose a method for selecting the best possible regression model in the semiparametric model class. We will apply the proposed model selection technique to overdispersed data. We will demonstrate the use of the proposed semiparametric information criterion using the well-known data on germination of *Orobanche*.

Key Words: Akaike information criterion, Generalized linear models, Kullback-Leibler information, Overdispersion, Quasi-likelihood

1 Introduction

Let $y_i\,(i=1,\ldots,n)$ be independent response variables which may be proportions or counts with mean μ_i and variance $\phi V(\mu_i)$. As in the usual generalized linear models (GLM; McCullagh and Nelder, 1989), we assume in addition that $E[y_i]=\mu_i=t^{-1}(\eta_i)$, where $\eta_i=\boldsymbol{x}_i'\boldsymbol{\beta}$, $t^{-1}(\cdot)$ is the inverse link function, $\boldsymbol{x}_i$ is a q vector of covariates, $\boldsymbol{\beta}$ is a q vector of the unknown parameters $\boldsymbol{\beta}'=(\beta_1,\ldots,\beta_q)$ and ϕ does not depend on $\boldsymbol{\beta}$. Let $\boldsymbol{X}$ be the $n\times q$ matrix of covariates with the ith row being $\boldsymbol{X}_{(i)}$, $\boldsymbol{\theta}=(\boldsymbol{\beta}',\phi)'\in\Theta\subset\mathbb{R}^{q+1}$ and $q<n$.In many situations, it is difficult to impose a specific distributional assumption on y_i as in the classical GLM. In such cases we may use the quasi-likelihood method (Wedderburn, 1974) to estimate the regression parameters $\boldsymbol{\beta}$. Under weak regularity conditions, it can be shown that the quasi-likelihood estimator $\widehat{\boldsymbol{\theta}}$ is consistent and asymptotically normally distributed (Moore, 1986).

However, works on model selection based on quasi-likelihoods are rather limited. There are mainly two approaches for existing works in this area. The first approach is to restrict the choice of the variance function $V(\cdot)$ so that one can integrate the quasi-score function to obtain the scalar quasi-likelihood function and formally use AIC with the likelihood function being replaced by the quasi-likelihood function (e.g., Lebreton, *et al.* (1992) , Anderson, Burnham and White (1994) , Qian, Gabor and Gupta (1996), Pan (2001)). The second approach makes no restriction on the choice of the variance functions, instead the quasi-score vector is projected on a subspace of estimating functions so that the projection becomes integrable (e.g., McLeish and Small (1992) , Li (1993) , Hanfelt and Liang (1995)), consequently one may use the first approach to formally construct the information criterion (e.g., Lin (2011)). In this paper we shall propose a different approach by projecting the true distribution function onto the subspace of probability distribution functions satisfying the first and the second moment assumptions. This projection will be used as the semiparametric predictive distribution for the underlying statistical model. Using the obtained semiparametric predictive distribution, we then formally extend the idea underlying the derivation of the information criterion AIC to construct a new semiparametric information criterion SIC. We shall demonstrate the use of the proposed semiparametric information criterion SIC using the well-known data on germination of *Orobanche*.

2 Semiparametric Model Selection Criteria

Akaike (1973) proposed a model selection criterion called AIC which has been widely used in many areas of applications. Derivation of AIC is based upon minimizing the Kullback-Leibler information between the predictive distribution and the true distribution generating the data. In this section we will extend this idea to the semiparametric setting with the full distributional assumption replaced by the assumptions on the mean and variance of the response variables.

We propose semiparametric model selection criteria SIC and SIC_T. They apply generally to data no matter whether they are discrete or continuous. For convenience, we suppose that the data in the subsequent paragraphs are continuous. If they are discrete then integration is replaced by summation.

2.1 Semiparametric models

We assume that the true probability density function of $\boldsymbol{Y}$ is $h(y)$ and the true distribution function is $H(y)$. We will relax the assumption of a complete specification of the probability model, only using some general form of the variance function about the mean as used in GLM. Now suppose that the statistician makes the assumption that the random variables $y_i\,(i = 1,\ldots,n)$ have distribution $G(\cdot)$, density function $g(\cdot)$, mean $\mu_i(\boldsymbol{\beta})$ and variance $\phi V(\mu_i(\boldsymbol{\beta}))$. The first and the second moments can be combined in the following way:

$$m(y_i,\boldsymbol{\theta}) = m(y_i,\boldsymbol{\beta},\phi) = \begin{pmatrix} y_i - \mu_i(\boldsymbol{\beta}) \\ (y_i - \mu_i(\boldsymbol{\beta}))^2 - \phi V(\mu_i(\boldsymbol{\beta})) \end{pmatrix} = \begin{pmatrix} m_{1i} \\ m_{2i} \end{pmatrix}$$

By assumption, we have that

$$E_G\left[m\left(y_i,\boldsymbol{\theta}\right)\right]=\int m(y_i,\boldsymbol{\theta})\ dG(y)=\mathbf{0}\ .$$

Let $\boldsymbol{M}$ denote the set of all possible distribution functions, let

$$\mathcal{G}(\boldsymbol{\theta})=\left\{G\in\boldsymbol{M}:\int m(y,\boldsymbol{\theta})\ dG(y)=\mathbf{0}\right\}.$$

Define $\mathcal{G}=\bigcup_{(\boldsymbol{\theta}\in\Theta)}\mathcal{G}(\boldsymbol{\theta})$ as the set of all possible distribution functions that are restricted by the moments conditions. The class $\mathcal{G}$ is the statistical model we shall consider in this paper.

2.2 Kullback-Leibler information

Extending the idea of AIC, we shall generalize the data-based model selection approach based also on the Kullback-Leibler information. We use the Kullback-Leibler information to measure the closeness between $H(y)$ and the assumed semiparametric model $G(\cdot)$. Let g vary in the statistical model class $\mathcal{G}$. We consider the constrained minimization problem

$$\begin{aligned} v(\boldsymbol{\beta},\phi)= \quad & \inf_{g\in\mathcal{G}}\int\log\left(\frac{h(y)}{g(y)}\right)h(y)\ dy \\ & \text{subject to}\ \int m(y,\boldsymbol{\beta},\phi)g(y)\ dy=0\ \text{ and }\ \int g(y)\ dy=1. \end{aligned} \tag{2.1}$$

Using the results in convex analysis (e.g., Kitamura, 2006), the solution to infinite dimensional optimization problem $v(\boldsymbol{\beta},\phi)$ can be equivalently written as the solution to the finite dimensional optimization problem $v^*(\boldsymbol{\beta},\phi)$ as follows:

$$v^*(\boldsymbol{\beta},\phi)=\sup_{\lambda\in\mathbb{R},\boldsymbol{r}\in\mathbb{R}^2}\left[1+\lambda+\int\log\left(-\lambda-\boldsymbol{r}'m\left(y,\boldsymbol{\beta},\phi\right)\right)h(y)\ dy\right]. \tag{2.2}$$

Since we are only interested in the maximization of $v^*(\boldsymbol{\beta},\phi)$ with respect to $\boldsymbol{\beta}$ and ϕ, we can ignore other terms and proceed to define the semiparametric information of the statistical model $\mathcal{G}$ as follows:

DEFINITION 2.1 *The semiparametric Kullback-Leibler information for quasi-likelihood function is defined as follows:*

$$\mathrm{SKL}(\boldsymbol{\theta})=\sup_{\boldsymbol{r}\in\mathbb{R}^2}E_H\left[\rho(y,\boldsymbol{\theta},\boldsymbol{r})\right]=\sup_{\boldsymbol{r}\in\mathbb{R}^2}\int\rho(y,\boldsymbol{\theta},\boldsymbol{r})h(y)\ dy, \tag{2.3}$$

where $\rho(y,\boldsymbol{\theta},\boldsymbol{r})=\log\left(1+\boldsymbol{r}'m(y,\boldsymbol{\beta},\phi)\right)$.

Since the true distribution function $H(\cdot)$ is unknown, we use the empirical distribution of the data to estimate the true distribution function $H(\cdot)$.

2.3 Regularity conditions

In this paper, we shall use SKL($\boldsymbol{\theta}$) to select an appropriate semiparametric model. We will consider both the correctly specified case and the misspecified case. A semiparametric model is called correctly specified if the moments assumed evaluated at a particular value of the parameter are equal to the moments of the underlying true distribution. First, we state the regularity conditions which will be used in the sequel to prove the asymptotic unbiasedness of the proposed information criteria.

C1 Define $\boldsymbol{\theta}_* = (\boldsymbol{\beta}_*', \phi_*)'$ as the pseudo-true value, such that the semiparametric Kullback-Leibler information SKL($\boldsymbol{\theta}$) is minimized at $\boldsymbol{\theta}_*$.

C2 $\rho(y, \boldsymbol{\theta}, \boldsymbol{r})$ is twice continuously differentiable in $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{r} \in \mathbb{R}^2$.

Note that **C2** implies that

$$\tilde{\boldsymbol{r}}(\boldsymbol{\theta}) = \arg\sup_{\boldsymbol{r}\in\mathbb{R}^2} \int \rho(y, \boldsymbol{\theta}, \boldsymbol{r}) h(y)\, dy \tag{2.4}$$

is continuous with respect to $\boldsymbol{\theta}$. Furthermore, from **C1** we know that there is a unique $\tilde{\boldsymbol{r}}_*$ corresponding to $\boldsymbol{\theta}_*$:

$$\tilde{\boldsymbol{r}}_* = \arg\sup_{\boldsymbol{r}\in\mathbb{R}^2} \int \rho(y, \boldsymbol{\theta}_*, \boldsymbol{r}) h(y)\, dy. \tag{2.5}$$

C3 The problem of $\inf_{\boldsymbol{\theta}\in\Theta}$ SKL($\boldsymbol{\theta}$) has a unique saddle point solution $(\boldsymbol{\theta}_*', \tilde{\boldsymbol{r}}_*')$, where $\tilde{\boldsymbol{r}}_*$ is defined in (2.5). When the model $\mathcal{G}$ is correctly specified, there exists a unique $(\boldsymbol{\theta}_*', \tilde{\boldsymbol{r}}_*')$ satisfying SKL$(\boldsymbol{\theta}_*) = \mathbf{0}$.

C4 The following law of large numbers holds:

$$\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{r}\in\mathbb{R}^2} \left| \frac{1}{n}\sum_{i=1}^{n} \rho(y_i, \boldsymbol{\theta}, \boldsymbol{r}) - E_H\left[\rho(y, \boldsymbol{\theta}, \boldsymbol{r})\right] \right| = o_p(1).$$

2.4 The expected information

When we consider SKL($\boldsymbol{\theta}$) as defined in (2.3), the minimizer of SKL($\boldsymbol{\theta}$) can be viewed as the solution to the following saddle point problem:

$$\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta}\in\Theta} \sup_{\boldsymbol{r}\in\mathbf{L}(\boldsymbol{\theta})} E_H\left[\rho(y, \boldsymbol{\theta}, \boldsymbol{r})\right], \tag{2.6}$$

where Θ denotes the parameter space for $\boldsymbol{\theta}$, $\mathbf{L}(\boldsymbol{\theta}) = \{\boldsymbol{r} : \boldsymbol{r}' m(y, \boldsymbol{\theta}) \in \mathcal{I}\}$, and $\mathcal{I}$ is an open interval containing zero.

Next we describe the process of solving the above saddle point problem. Firstly, the expectation of $\rho(y, \boldsymbol{\theta}, \boldsymbol{r})$ is maximized for given $\boldsymbol{\theta}$, so that the $\boldsymbol{\theta}_*$ satisfies the following condition:

$$E_H\left[\rho_1(y, \boldsymbol{\theta}, \boldsymbol{r})\, m(y, \boldsymbol{\theta})\right] = \mathbf{0},$$

where $\rho_1(y, \boldsymbol{\theta}, \boldsymbol{r}) = 1/(1 + \boldsymbol{r}' m(y, \boldsymbol{\theta}))$ is the first derivative of ρ with respect to $\boldsymbol{r}' m(y, \boldsymbol{\theta})$.

C5 Assume that $\int \rho(y, \boldsymbol{\theta}, \boldsymbol{r}) h(y)\, dy$ and $\int \rho_1(y, \boldsymbol{\theta}, \boldsymbol{r}) h(y)\, dy$ are differentiable under the integral sign.

Secondly, Since $\boldsymbol{\theta}_*$ is the minimizer of the quantity $E_H\left[\rho(y, \boldsymbol{\theta}, \boldsymbol{r})\right]$, for a proper $\boldsymbol{r}$, implying that

$$E_H\left[\rho_1\left(y, \boldsymbol{\theta}, \boldsymbol{r}\right) M(y, \boldsymbol{\theta})' \boldsymbol{r}(\boldsymbol{\theta})\right] = \mathbf{0},$$

where $M(y, \boldsymbol{\theta}) = \partial m(y, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$.

Now denote the matrix of the first derivative of $\rho(y, \boldsymbol{\theta}, \boldsymbol{r})$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{r}$ by

$$\boldsymbol{\Upsilon}(\boldsymbol{\theta}, \boldsymbol{r}) = \begin{pmatrix} \rho_1\left(y, \boldsymbol{\theta}, \boldsymbol{r}\right) M(y, \boldsymbol{\theta})' \boldsymbol{r} \\ \rho_1(y, \boldsymbol{\theta}, \boldsymbol{r}) m(y, \boldsymbol{\theta}) \end{pmatrix},$$

The values $(\boldsymbol{\theta}_*, \tilde{\boldsymbol{r}}(\boldsymbol{\theta}_*))$ must satisfy the following first-order condition:

$$E_H\left[\boldsymbol{\Upsilon}\left(\boldsymbol{\theta}_*, \tilde{\boldsymbol{r}}\left(\boldsymbol{\theta}_*\right)\right)\right] = \mathbf{0}. \tag{2.7}$$

C6 Define

$$\mathrm{Q} = E_H\left[\frac{\partial \boldsymbol{\Upsilon}(\boldsymbol{\theta}_*, \tilde{\boldsymbol{r}}_*)}{\partial(\boldsymbol{\theta}', \boldsymbol{r}')'}\right], \quad \mathrm{S} = E_H\left[\boldsymbol{\Upsilon}(\boldsymbol{\theta}, \tilde{\boldsymbol{r}}_*) \boldsymbol{\Upsilon}(\boldsymbol{\theta}, \tilde{\boldsymbol{r}}_*)'\right].$$

Assume that Q and S are finite and S is invertible.

However, $\boldsymbol{\theta}_*$ is not computable because it involves the expectation with respect to the true density function $h(y)$. One way is to use the value to minimize the empirical version of the expected semiparametric Kullback-Leibler information:

$$\widehat{\boldsymbol{\theta}}_* = \arg\min_{\boldsymbol{\theta} \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho\left(y_i, \boldsymbol{\theta}, \tilde{\boldsymbol{r}}\left(\boldsymbol{\theta}\right)\right) \right\}, \tag{2.8}$$

where $\tilde{\boldsymbol{r}}(\boldsymbol{\theta})$ is defined in (2.4). To approximate $\tilde{\boldsymbol{r}}(\boldsymbol{\theta})$, we use the empirical distribution function in place of the true distribution function $H(\cdot)$ to obtain

$$\widehat{\boldsymbol{r}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{r} \in \mathbb{R}^2} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho(y_i, \boldsymbol{\theta}, \boldsymbol{r}) \right\}. \tag{2.9}$$

We can show that $(\widehat{\boldsymbol{\theta}}_*, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}}_*))$ are consistent and asymptotically normally distributed along similar lines of Chen et al. (2007), we omit the proof here. The calculation of $(\widehat{\boldsymbol{\theta}}_*, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}}_*))$ is instable and usually difficult to compute. We assume that $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \widehat{\phi})'$ with $\widehat{\boldsymbol{\beta}}$ being the quasi-likelihood estimator and $\widehat{\phi}$ the estimator from the usual Pearson residual. Since the $\widehat{\boldsymbol{\theta}}$ is also a consistent estimator, it is reasonable to use $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}}))$ instead. It is easy to show that $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}}))$ holds the properties of consistency and asymptotic normality too.

DEFINITION 2.2 *In the statistical model $\mathcal{G}$, the expected semiparametric information is defined as follows:*

$$E_H[\mathrm{SKL}(\widehat{\boldsymbol{\theta}})] = E_H\left[\int \rho(y, \widehat{\boldsymbol{\theta}}, \tilde{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}})) h(y)\, dy\right]. \tag{2.10}$$

For later use, we write explicitly the components of $\boldsymbol{\Upsilon}(\boldsymbol{\theta}, \boldsymbol{r})$ as follows:

$$M(y, \boldsymbol{\theta}) = \frac{\partial m(y, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \begin{pmatrix} M_1 & M_3 \\ M_2 & M_4 \end{pmatrix},$$

where

$$M_1 = \frac{\partial m_1}{\partial \boldsymbol{\beta}'} = -\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}, \qquad M_2 = \frac{\partial m_2}{\partial \boldsymbol{\beta}'} = 2(\mu_i(\boldsymbol{\beta}) - y)\frac{\partial \mu_i}{\partial \boldsymbol{\beta}'} - \phi\frac{\partial V(\mu_i(\boldsymbol{\beta}))}{\partial \mu_i}\frac{\partial \mu_i}{\partial \boldsymbol{\beta}'},$$
$$M_3 = \frac{\partial m_1}{\partial \phi} = 0, \qquad M_4 = \frac{\partial m_2}{\partial \phi} = -V(\mu_i(\boldsymbol{\beta})).$$

2.5 The proposed information criteria

DEFINITION 2.3 *Let k be the dimension of $\boldsymbol{\theta}$, then we define the following information criterion:*

$$\text{SIC} = \sum_{i=1}^{n} \rho\left(y_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}})\right) + k. \tag{2.11}$$

We shall call this information criterion the semiparametric information criterion, or SIC *for short.*

The proposed criterion SIC has the following properties.

PROPOSITION 2.1 *Suppose that the model is correctly specified. Further suppose that* **(C1)**–**(C6)** *hold. Then* SIC *is an asymptotically unbiased estimator of the expected information.*

For possibly misspecified model, we propose the following criterion for model selection.

DEFINITION 2.4 *Consider a possibly misspecified semiparametric model $\mathcal{G}$ as defined in Section 2.1. Let*

$$\widehat{\mathrm{S}} = \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{\Upsilon}_i\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}})\right)\boldsymbol{\Upsilon}_i\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}})\right)'\right], \quad \widehat{\mathrm{Q}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial \boldsymbol{\Upsilon}_i\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}})\right)}{\partial(\boldsymbol{\theta}', \boldsymbol{r}')'}\right].$$

Then we defined the following information criterion

$$\text{SIC}_T = \sum_{i=1}^{n} \rho\left(y_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}})\right) - \text{trace}(\widehat{\mathrm{S}}\widehat{\mathrm{Q}}^{-1}). \tag{2.12}$$

The proposed criterion SIC_T has the following properties.

PROPOSITION 2.2 *Under the regularity conditions* **(C1)** – **(C6)** *listed in Section 2.3 and 2.4,* SIC_T *is an asymptotically unbiased estimator of the expected information.*

SIC_T can be understood as a semiparametric extension of the Takeuchi information criterion for parametric likelihood analysis (Takeuchi, 1976).

3 Data Analysis

Now we illustrate the use of SIC using a well-known data set *Orobanche* on germination of the seed variates, which was considered in Crowder (1978, Table 3). In this experiment, two types of seeds from Orobanche (O.aegyptiaca75 and O.aegyptiaca73) were tested for germination when they were given with two different root extracts (Bean and cucumber). The goal of the experiment was to determine the effect of the root extract on inhibiting the growth of the parasitic plant. There is suspicion of overdispersion for this data set and a number of authors have used this data set to study methodologies extending the classical GLM to take into consideration the problem of overdispersion. For instance, this data was analyzed by Breslow and Clayton (1993) to illustrate the use of the generalized linear mixed models for overdispersion. In these studies, however, the authors have not considered the aspects of model selection. Now we shall reanalyze it by applying the model selection technique proposed in the previous section based on the quasi-likelihood.

Now let y_i be the proportions of germinated seeds, and n_i the seeds in the ith data set $(i = 1, \ldots, 21)$. Suppose that $E[y_i] = \pi_i$ and $var[y_i] = \phi v(\pi_i) = \phi\pi_i(1-\pi_i)/n_i$, where ϕ is the overdispersion parameter. Further, we shall suppose that $\text{logit}(\pi_i) = \log(\pi_i/(1-\pi_i)) = \boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{x}_i = (1, x_{1i}, x_{2i}, x_{1i}x_{2i})'$, where $x_{1i} = 1$ if the root extract is cucumber, otherwise 0 if the root extract is Bean; and $x_{2i} = 1$ if the type of seeds is O.aegyptiaca75, otherwise 0 if the type of seeds is O.aegyptiaca73. We shall compare the folowing five candidate models:

(1) $\text{logit}(\pi_i) = \beta_0$

(2) $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{1i}$

(3) $\text{logit}(\pi_i) = \beta_0 + \beta_2 x_{2i}$

(4) $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$

(5) $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$

Table 1: Model selection criteria for the *Orobanche* data

Models	(1)	(2)	(3)	(4)	(5)
SIC	3.821	2.732	5.971	3.881	5.521
SIC_T	0.551	1.260	0.305	1.860	3.380
AIC_{BB}	133.0	120.4	133.1	119.7	117.5

We computed the values of the various information criteria and the results are summarized in Table 1. The proposed SIC has the smallest value for model (2) among all candidate

models. SIC ranks the models in the following way: (2) > (1) > (4) > (5) > (3). This result is consistent with the results obtained in Crawley (2005, p.255-260). Crawley (2005) used quasi-likelihood method to analyze the *Orobanche* data. He applied the F-test to compare the full model (5) with other simpler models. He found that there is no compelling evidence that the types of seeds and the interaction should be kept in the model, and the minimal adequate model is model (2). In Table 1, we have also shown the values of AIC_{BB}, the values of AIC based on the beta-binomial model. AIC_{BB} has minimum value for model (5), the most complicated model considered here. In Table 1 the values SIC_T are also shown; SIC_T is a modified version of SIC which is not discussed here. SIC_T is a corresponding version of the Takeuchi type information criteria in full likelihood analysis.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, B. N., Petrov and F., Csaki (eds), 267–281, Budapest: Akademiai Kiado.

[2] Anderson, D. R., Burnham, K. P. and White, G. C. (1994). AIC model selection in overdispersed capture-recapture data. *Ecology*, **75**, 1780–1793.

[3] Breslow, N E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J Am. Statist. Ass.*, **88**, 9–25.

[4] Crawley, M. J. (2005). *Statistics: an introduction using R.* John Wiley & Sons, Ltd.

[5] Crowder, M. J. (1978). Beta-binomial anova for proportions. *Appl. Statist.*, **27**, 34–37.

[6] Godambe, V. P. (1999). Linear Bayes and optimal estimation. *Annals of the Institute of Statistical Mathematics*, **51**, 201–215.

[7] Hanfelt, J. J. and Liang, K.-Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika*, **82**, 461– 477.

[8] Lebreton, J. D., K. P. Burnham, J. Clobert., and Anderson, D. R. (1992). Modeling survival and testing biological hypothesis using marked animals: a unified approach with case studies. *Ecological Monographs*, **62**, 67–118.

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

[10] Lin, P.-S. (2011). Quasi-deviance functions for spatially correlated data. *Statistica Sinica*, **21**, 1785–1806.

[11] McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models.* Chapman and Hall: London.

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

[13] Moore, D. F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika*, **73**, 283–288.

[14] Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120–5.

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

[16] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.