Model Selection for Semiparametric Bayesian Models with Application to Overdispersion

Jinfang Wang and Yiping Tang
Department of Mathematics and Informatics, Chiba University
1-33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan
E-mail: wang@math.s.chiba-u.ac.jp

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 Bayesian 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: Bayes linear methods, generalized linear models, Kullback-Leibler information, overdispersion, quasi-likelihood

1 Introduction

Let $y_i\,(i = 1, \ldots, n)$ be one-dimensional independent response variables having mean μ_i and variance $\phi\, v(\mu_i)$, where $v(\cdot)$ is a known positive function of the mean μ_i and ϕ is an unknown constant dispersion parameter. Although the y_i's may be either discrete or continuous, in this paper we shall mainly be interested in responses which are either proportions or counts. As in the usual generalized linear models (GLM; McCullagh and Nelder, 1989), we assume in addition that $g(\mu_i) = \eta_i$, where $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ is a linear predictor, with $\boldsymbol{x}_i = (x_1, \ldots, x_p)'$ being a p-vector of covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ a p-vector of unknown repression parameters, and $g(\cdot)$ a known (monotonic) link function. 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 $\widehat{\boldsymbol{\beta}}$ 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 ideal to the semiparametric setting with the full distributional assumption replaced by the assumptions on the mean and variance of the response variables.

Let $H(y)$ denote the true but unknown distribution function with density function $h(y)$. Let $F(y)$ be any distribution function having mean μ and variance $\phi\, v(\mu)$, the corresponding density (probability) function being denoted by $f(\cdot)$. Let $g(\mu) = \boldsymbol{x}'\boldsymbol{\beta}$, where $\boldsymbol{x}$ is a p-vector of covariates and $\boldsymbol{\beta}$ a p-vector of regression parameters. Let $\mathcal{F}$ denote the space of all density functions satisfying these two assumptions on the mean and variance. We assume that there exists a unique density function f^* which has minimum Kullback-Leibler information with the true density function h compared with any density function $f(\cdot)$ belonging to $\mathcal{F}$. Given the covariate vector $\boldsymbol{x}$, the Kullback-Leibler information between h and f^* depends only on $\boldsymbol{\beta}$ and ϕ. Let

$$w(\boldsymbol{\beta},\, \phi) = \int \log\left(\frac{h(y)}{f^*(y)}\right) h(y)\, dy = \inf_{f\in\mathcal{F}} \int \log\left(\frac{h(y)}{f(y)}\right) h(y)\, dy \tag{2.1}$$

We look for the value $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*,\, \phi^*)$, which minimizes $w(\boldsymbol{\beta},\, \phi)$ of (2.1). Let $\boldsymbol{t} = (t_1, t_2)'$, and define

$$\ell(y, \boldsymbol{\theta}, \boldsymbol{t}) = \log(1 + \boldsymbol{t}'\boldsymbol{m}(y, \boldsymbol{\theta}))$$

where $\boldsymbol{m}(y, \boldsymbol{\theta})$ is a two-dimensional vector defined by

$$\boldsymbol{m}(y, \boldsymbol{\theta}) = (\mu(\boldsymbol{\beta}),\, (y - \mu(\boldsymbol{\beta}))^2 - \phi\, v(\mu(\boldsymbol{\beta})))'\,.$$

Using the results in convex analysis (e.g., Kitamura, 2006), it can be shown that the minimizer $\boldsymbol{\theta}^*$ of (2.1) is given by

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \left\{ \sup_{\boldsymbol{t}\in\boldsymbol{T}(\boldsymbol{\theta})} E_H\left[\ell(Y,\boldsymbol{\theta},\boldsymbol{t})\right] \right\} \tag{2.2}$$

where Θ is the parameter space for $\boldsymbol{\theta}$, and $\boldsymbol{T}(\boldsymbol{\theta})$ a two-dimensional region defined by

$$\boldsymbol{T}(\boldsymbol{\theta}) = \left\{ \boldsymbol{t} : \boldsymbol{t}'\boldsymbol{m}(y,\boldsymbol{\theta}) > -1 \right\}.$$

The value $\boldsymbol{\theta}^*$ is not computable since it involves the expectation with respect to the true unknown density function $h(\cdot)$. One natural way to approximate $\boldsymbol{\theta}^*$ is by using the value $\widehat{\boldsymbol{\theta}}^*$ which minimizes the empirical version of the expected semiparametric Kullback-Leibler information:

$$\widehat{\boldsymbol{\theta}}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i,\boldsymbol{\theta},\boldsymbol{t}\left(\boldsymbol{\theta}\right)\right) \right\}, \tag{2.3}$$

where $\boldsymbol{t}(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$ and is defined by

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

Since the right-hand side of (2.4) again depends on the unknown true density function $h(\cdot)$, we shall approximate the value of $\boldsymbol{t}(\boldsymbol{\theta})$, given $\boldsymbol{\theta}$, using the plug-in principle:

$$\widehat{\boldsymbol{t}}(\boldsymbol{\theta}) = \arg\sup_{\boldsymbol{t}\in\mathbb{R}^2} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell(y_i,\boldsymbol{\theta},\boldsymbol{t}) \right\}. \tag{2.5}$$

Now we state the main results of the paper.

PROPOSITION 2.1. *Let $k = \dim(\boldsymbol{\theta})$ be the dimension of $\boldsymbol{\theta}$. Define*

$$\mathrm{SIC} = \sum_{i=1}^{n} \ell\left(y_i,\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{t}}(\widehat{\boldsymbol{\theta}})\right) + k. \tag{2.6}$$

where $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}},\hat{\phi})'$, with $\widehat{\boldsymbol{\beta}}$ being the quasi-likelihood estimator and $\hat{\phi}$ the estimator from the usual Pearson residual. Under suitable regularity conditions, SIC *is an asymptotically unbiased estimator of*

$$E_H\left[\mathrm{SKL}(\widehat{\boldsymbol{\theta}})\right] = E_H\left[\int \ell\left(y,\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{t}}(\widehat{\boldsymbol{\theta}})\right) h(y)\, dy\right] \tag{2.7}$$

We shall call SIC *of (2.6) the semiparametric information criterion in the rest of the paper.*

Sometimes we may also have information available for the parameter $\boldsymbol{\theta}$ a priori before observing the data. By imposing the restrictions on the first two moments on $\boldsymbol{\theta}$, Godambe (1999) advocated the use of optimum Bayesian estimating functions. More detailed accounts on optimum Bayesian estimating functions can be found in Small and Wang (2003). If we let $\boldsymbol{m}(y,\boldsymbol{\theta})$ to denote the optimum Bayesian estimating functions, we can then extend the theories obtained so far to construct a semiparametric Bayesian information criterion. We will report these results on the conference.

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 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] Small, C. G. and Wang, J. (2003). *Numerical Methods for Nonlinear Estimating Equations*. Clarendon Press: Oxford.

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