EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS
Vol. 3, No. 3, 2010, 417-434
ISSN 1307-5543 – www.ejpam.com

SPECIAL ISSUE ON
GRANGER ECONOMETRICS AND STATISTICAL MODELING
DEDICATED TO THE MEMORY OF PROF. SIR CLIVE W.J. GRANGER

Law of Iterated Logarithm and Strong Consistency in Poisson Regression Model Selection

Guogi Qian

Department of Mathematics and Statistics, University of Melbourne, VIC 3010, Australia

Abstract. In this paper we first derive a law of iterated logarithm for the maximum likelihood estimator of the parameters in a Poisson regression model. We then use this result to establish the strong consistency of a class of model selection criteria in Poisson regression model selection. We show that under some general conditions, a model selection criterion, which consists of a minus maximum log-likelihood and a penalty term, will select the simplest correct model almost surely if the penalty term increases with model dimension and has an order in between $O(\log\log n)$ and $O(n)$.

2000 Mathematics Subject Classifications: 62F12, 62J12, 60F15

Key Words and Phrases: Law of iterated logarithm; Poisson regression; Maximum likelihood estimator; Model selection; Strong consistency

1. Introduction

Poisson regression model is a widely used econometric and statistical tool for studying the relationship between a Poisson-type response variable and a set of explanatory variables. A familiar example is the analysis of contingency tables of categorical data. In addition to parameter estimation, another important inference task in Poisson regression is searching for a subset of available explanatory variables that can best explain or predict the response. This amounts to the Poisson regression model or variable selection. Many papers can be found in recent literature in the area of model selection, which deal with different models in different ways. We refer to [4] and [14] and references therein for the detailed survey.

It appears that, while it might be implied according to some general model selection principle, the Poisson regression model selection method by itself has hardly been investigated in

Email address: g.qian@ms.unimelb.edu.au

a formal and rigorous way. The lack of a formal theory for Poisson regression model selection creates uncertainty and inconvenience for people applying the method for practice. This motivates the writing of this paper which focuses on the asymptotic performance of a class of model selection criteria including AIC, BIC, Mallows C_p and stochastic complexity or minimum description length for Poisson regression models. A byproduct of this asymptotic study is the establishment of the law of iterated logarithm for the maximum likelihood estimators (MLE) in the Poisson regression models. The convergence rate of the MLE provided by the law of iterated logarithm is very useful in deriving precise approximations for likelihood function.

In the paper we first set up a model selection framework for Poisson regression models and review several general model selection criteria such as AIC [1], BIC [18] and stochastic complexity or minimum description length [16] in section 2. In section 3 we present the main results and the conditions for ensuring these results. We have shown that when the employed model is a correct one, the MLE $\hat{\beta}$ converges almost surely to the true parameter value β_0 with a rate not slower than $O(\sqrt{n^{-1}\log\log n})$. We have also shown that, for a model selection criterion consisting of the minus log-likelihood and a penalty term, it will select the simplest correct model almost surely if the penalty term is an increasing function of the model dimension and is of an order higher than $O(\log\log n)$ and lower than $O(n)$. The detailed proof of these results are given in section 4 and the appendix. The paper is concluded with a discussion given in section 5.

2. Model Selection in Poisson Regression Model

The problem to our interest is whether any component of a given explanatory vector $\mathbf{x} = (x_1, \cdots, x_p)^t$ has any effect on a response variable Y. When Y is a count variable, it is often sensible to assume a Poisson distribution for Y which has a probability function $P(Y = y) = (y!)^{-1}\mu^y e^{-\mu}$ $(y = 0, 1, 2, \cdots)$. Then the problem can be studied in the framework of a log-linear regression model which assumes a linear predictor $\eta = \mathbf{x}^t\beta$ for logarithm of the mean of Y, i.e., $\log\mu = \eta = \mathbf{x}^t\beta$, where $\beta = (\beta_1, \cdots, \beta_p)^t$ is the unknown parameter vector of interest.

Now let $Y_n = (y_1, \cdots, y_n)^t$ be the n independent observations from Y, with the corresponding explanatory vectors being $\mathbf{x}_1, \cdots, \mathbf{x}_n$. Denote $X_n = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^t$ as the design matrix. Then under the log-linear regression model considered, the distribution for y_i is Poisson(μ_i) with $\mu_i = e^{\eta_i} = e^{\mathbf{x}_i^t\beta}$; and the log-likelihood function for the parameter β is

$$\ell(\beta|Y_n, X_n) = \sum_{i=1}^{n}\{-\log y_i! + y_i\log\mu_i - \mu_i\} = -\sum_{i=1}^{n}\log y_i! + \sum_{i=1}^{n}\{y_i\mathbf{x}_i^t\beta - e^{\mathbf{x}_i^t\beta}\}. \qquad (1)$$

The maximum likelihood estimator(MLE) $\hat{\beta}$ is defined to be

$$\hat{\beta} = \arg\max_{\beta}\ell(\beta|Y_n, X_n) = \arg\min_{\beta}\sum_{i=1}^{n}\{e^{\mathbf{x}_i^t\beta} - y_i\mathbf{x}_i^t\beta\}$$

which can be solved from

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} (y_i - \mu_i)\mathbf{x}_i = 0.$$

In this paper we will show that the estimation error $||\hat{\beta} - \beta_0|| = O(\sqrt{n^{-1}\log\log n})$ almost surely under some general conditions. Here β_0 is the true value of β and $||\cdot||$ is the Euclidean norm. The Fisher information for the parameter β can be found to be

$$I_n(\beta) = -E\frac{\partial^2 \ell}{\partial \beta \partial \beta^t} = -\frac{\partial^2 \ell}{\partial \beta \partial \beta^t} = X_n^t \mathscr{U}_n X_n, \tag{2}$$

where $\mathscr{U}_n = diag\{\mu_1, \cdots, \mu_n\}$.

With the Poisson regression model $\log\mu = \mathbf{x}^t\beta$, the effect of each x variable on Y can be measured by the value of the corresponding β component. Thus if any of the β components equals 0 or is close to 0, there would be no necessity to include the corresponding x components into the model. Since the values of β can only be estimated, one needs a statistical model selection criterion to determine which x components have significant effects on Y thus should be included in the model. The maximum likelihood principle cannot serve as a model selection criterion because the maximum likelihood for the full model including all the available x components is always greater than the maximum likelihood for a sub-model using a subset of the x components. But a model selection criterion can be based on a penalised log-likelihood. Let α be a p_α-component sub-vector of $(1, 2, \cdots, p)$. Let $\mathbf{x}_\alpha$ and $\beta(\alpha)$ be the sub-vectors of $\mathbf{x}$ and β indexed by α respectively. Further let $\log\mu_\alpha = \eta_\alpha = \mathbf{x}_\alpha^t\beta_\alpha$ be a Poisson regression model containing a subset of explanatory variables given by $\mathbf{x}_\alpha$. The penalised log-likelihood based model selection criterion can be expressed as

$$S(\eta_\alpha) = -\ell(\hat{\beta}(\alpha)|Y_n, X_{n\alpha}) + C(n, \hat{\beta}(\alpha)), \tag{3}$$

where the first term is the minus maximum log-likelihood measuring the goodness of fit of model η_α, and the second term measures the complexity of the model. The matrix $X_{n\alpha}$ comprises those columns of X_n indexed by α; and $\hat{\beta}(\alpha)$ is the MLE of $\beta(\alpha)$. Under the criterion (3), those sub-models having both better goodness of fit and smaller complexity will be preferred than the others; and the best model will be the one achieving the smallest $S(\eta_\alpha)$ value. Many commonly used model selection criteria, such as AIC [1], BIC [18], C_p [7] and stochastic complexity criterion (SCC)[16, 17, 11], are of the form given by (3). For example, for AIC and C_p $C(n, \hat{\beta}(\alpha)) = p_\alpha$; for BIC $C(n, \hat{\beta}(\alpha)) = \frac{1}{2}p_\alpha \log n$; and for SCC $C(n, \hat{\beta}(\alpha)) = \frac{1}{2}\log|I_n(\hat{\beta}(\alpha))| + \sum_{i=2}^{p_\alpha}\log(|\hat{\beta}(\alpha)_i| + \varepsilon n^{-1/4})$ where $\hat{\beta}(\alpha)_i$ is the i-th component of $\hat{\beta}(\alpha)$, and ε is a specified quantity to ensure the invariance of the SCC [see 11, for details].

Assuming that the model $\log\mu = \mathbf{x}^t\beta$ is the full model which includes all the explanatory variables available and the first component of $\mathbf{x}$ is an intercept term, there will be in total $2^p - 1$ sub-models of the form $\log\mu_\alpha = \mathbf{x}_\alpha^t\beta(\alpha)$ for selection, provided that only those models having an intercept term are considered. In this paper we assume that some components of $\beta_0 = (\beta_{01}, \cdots, \beta_{0p})^t$, the true value of β, are equal to 0. We also use α to represent a Poisson regression sub-model $\log\mu_\alpha = \mathbf{x}_\alpha^t\beta(\alpha)$, which is actually a one-to-one representation. Then all the $2^p - 1$ sub-models can be classified into the following two groups:

1. $\mathscr{A}_c = \{\alpha : \beta_{0i} = 0 \quad \text{for any} \quad i \notin \alpha\}$;
2. $\mathscr{A}_w = \{\alpha : \beta_{0i} \neq 0 \quad \text{for some} \quad i \notin \alpha\}$.

Apparently, every sub-model in $\mathscr{A}_w$ is a wrong model which misses at least one x variable having non-zero effect on Y, while every sub-model in $\mathscr{A}_c$ is a correct model which includes all x variables having non-zero effects on Y. But the models in $\mathscr{A}_c$ may contain some redundant x variables having no effects on Y. An ideal model selection criterion should render the selection of the simplest correct model in $\mathscr{A}_c$ containing no redundant explanatory variables. In this paper we will show that the penalised log-likelihood based model selection criterion, under some general conditions, selects the simplest correct model with probability 1 as the sample size n goes to infinity. For simplicity of the presentation, we assume the simplest correct model in $\mathscr{A}_c$ to be unique, which is the case if all components of $\mathbf{x}$ are linearly independent of each other.

3. Conditions and Main Results

Let $\lambda_1\{S\} \leq \cdots \leq \lambda_p\{S\}$ be the p eigenvalues of a $p \times p$ symmetric matrix S. Also let $b = \frac{1}{2}\min_{1\leq i\leq p_{\alpha_0}} |\beta_0(\alpha_0)_i|$, where α_0 is the correct model in $\mathscr{A}_c$ with the smallest dimension, and $\beta_0(\alpha_0)_i$ is the i-th component of $\beta_0(\alpha_0)$. We assume $b > 0$ in this paper. Note that b is only used in the proof of Theorem 3 in this paper; and $b = 0$ represents the case where $\mathscr{A}_w$ is an empty set thereby Theorem 3 is not applicable. Further define

$$\delta_n = \sqrt{\max_{1\leq i\leq n} \mu_{0i}\mathbf{x}_i^t I_n(\beta_0)^{-1}\mathbf{x}_i} \quad \text{and} \quad \xi_n = \sqrt{\max_{1\leq i\leq n} \mathbf{x}_i^t I_n(\beta_0)^{-1}\mathbf{x}_i}$$

where $\mu_{0i} = e^{\mathbf{x}_i^t\beta_0}$ is the true value of μ_i.

The following conditions will be required in various places in proving our main results:

(C.1). $\lim_{n\to\infty} \lambda_j\{I_n(\beta_0)\} = \infty, \quad j = 1, \cdots, p$. Also there exists a constant b_0 such that $0 < \lambda_p\{I_n(\beta_0)\} \leq b_0\lambda_1\{I_n(\beta_0)\}$.

(C.2). $b_1 n \leq \lambda_p\{I_n(\beta_0)\} \leq b_2 n$ for some positive constants b_1 and b_2.

(C.3). $\xi_n\sqrt{\log\log\lambda_p\{I_n(\beta_0)\}} = o(1)$.

(C.4). $\delta_n\sqrt{\log\lambda_p\{I_n(\beta_0)\}}\sqrt{\log\log\lambda_p\{I_n(\beta_0)\}} = o(1)$.

(C.5). $\delta_n\log\lambda_p\{I_n(\beta_0)\}\sqrt{\log\log\lambda_p\{I_n(\beta_0)\}} = o(1)$.

(C.6). $\xi_n\log\lambda_p\{I_n(\beta_0)\}\sqrt{\log\log\lambda_p\{I_n(\beta_0)\}} = o(1)$.

(C.7). $\lambda_1\{X_n^t\mathscr{M}_n X_n\} \geq b_3 n$ for some constant $b_3 > 0$, where $\mathscr{M}_n = diag\{\mu_{01}e^{-2b\|\mathbf{x}_1\|}, \cdots, \mu_{0n}e^{-2b\|\mathbf{x}_n\|}\}$.

Note that these conditions are not completely independent of each other. Firstly, one can see that both (C.1) and (C.2) are implied from $b_1 n \le \lambda_1\{I_n(\beta_0)\} \le \lambda_p\{I_n(\beta_0)\} \le b_2 n$. Secondly, conditions (C.1), (C.2) and (C.7) together suggest that all the eigenvalues of $I_n(\beta_0)$ and $X_n \mathscr{M}_n X_n$ are of order $O(n)$. Thirdly, conditions (C.3) to (C.6) together with (C.2) indicate $\xi_n\sqrt{\log\log n} \to 0$, $\delta_n\sqrt{\log n \log\log n} \to 0$, $\delta_n \log n\sqrt{\log\log n} \to 0$ and $\xi_n \log n\sqrt{\log\log n} \to 0$ respectively. Finally, under condition (C.1), (C.3) is implied by (C.6); and (C.4) is implied by (C.5). Although conditions (C.1) to (C.7) may be simplified according to the preceding discussion, we prefer not to do so in order to clarify that to what extent each condition is required in the proof.

The conditions (C.1) to (C.7) are essentially about the behavior of the explanatory variables $\mathbf{x}$. Roughly speaking, they mean most of the observations $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ should be finite and stay away from 0; and if a subsequence of $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ diverges to infinity, it should do so with an appropriate rate. In fact, if we assume $\mathbf{x}$ is a random vector and $\mathbf{x}_1, \cdots, \mathbf{x}_n$ as i.i.d. observations from $\mathbf{x}$, then the following are sufficient for (C.1) to (C.7) to hold:

(S.1). The moment generating function $Ee^{\mathbf{x}^t s}$ exists for $||s|| \le ||\beta_0|| + s_0$ for some constant $s_0 > 0$. This implies that all of $E(e^{\mathbf{x}^t\beta_0}\mathbf{x}^t\mathbf{x})^\kappa$, $E(e^{\mathbf{x}^t\beta_0 - 2b\|\mathbf{x}\|}\mathbf{x}^t\mathbf{x})^\kappa$ and $E(\mathbf{x}^t\mathbf{x})^\kappa$ are finite for some $\kappa > 1$.

(S.2). $P(\mathbf{x}^t\mathbf{v} \ne 0) > 0$ for all $\mathbf{v} \ne 0$ in $\mathscr{R}^p$, which implies $Ee^{\mathbf{x}^t\beta_0}\mathbf{x}\mathbf{x}^t$, $Ee^{\mathbf{x}^t\beta_0 - 2b\|\mathbf{x}\|}\mathbf{x}\mathbf{x}^t$ and $E\mathbf{x}\mathbf{x}^t$ are all positive definite.

To see the sufficiency of (S.1) and (S.2), one can apply the strong law of large numbers for the i.i.d. random variables $\mathbf{x}_1, \cdots, \mathbf{x}_n, \cdots$ under condition (S.1), which gives the following results:

$$\frac{1}{n}X_n^t \mathscr{U}_n X_n - Ee^{\mathbf{x}^t\beta_0}\mathbf{x}\mathbf{x}^t \xrightarrow{\text{a.s.}} 0,$$
$$\frac{1}{n}X_n^t \mathscr{M}_n X_n - Ee^{\mathbf{x}^t\beta_0 - 2b\|\mathbf{x}\|}\mathbf{x}\mathbf{x}^t \xrightarrow{\text{a.s.}} 0.$$

These results together with (S.2) imply (C.1),(C.2) and (C.7). The conditions (C.3) to (C.6) are implied from (C.1), (C.2) and the fact that, under (S.1)

$$\delta_n^{2(1+\kappa')} \le \lambda_1\{I_n(\beta_0)\}^{-1-\kappa'} \sum_{j=1}^n e^{\mathbf{x}_j^t\beta_0(1+\kappa')}(\mathbf{x}_j^t\mathbf{x}_j)^{1+\kappa'} = O(n^{-\kappa'}) \quad \text{a.s. and}$$
$$\xi_n^{2(1+\kappa')} \le \lambda_1\{I_n(\beta_0)\}^{-1-\kappa'} \sum_{j=1}^n (\mathbf{x}_j^t\mathbf{x}_j)^{1+\kappa'} = O(n^{-\kappa'}) \quad \text{a.s.}$$

for some $\kappa' > 0$. In this paper we will regard the observations $\mathbf{x}_1, \cdots, \mathbf{x}_n$ as deterministic for simplicity of the presentation. There is no essential complication with random $\mathbf{x}_i$'s.

In this paper we have obtained the following results.

Theorem 1. *Suppose conditions* $(C.1)$ *to* $(C.6)$ *are satisfied. Then for any correct model* $\alpha \in \mathscr{A}_c$,

$$||\hat{\beta}(\alpha) - \beta_0(\alpha)|| = O(\sqrt{n^{-1}\log\log n}) \quad a.s.. \tag{4}$$

Further, there exists a constant $d > 0$ such that for any $\alpha \in \mathscr{A}_c$

$$\limsup_{n\to\infty} \frac{||\hat{\beta}(\alpha) - \beta_0(\alpha)||}{\sqrt{n^{-1}\log\log n}} = d \quad a.s.. \tag{5}$$

Hence the MLE $\hat{\beta}(\alpha)$ follows the law of iterated logarithm.

Theorem 2. *Under conditions $(C.1)$ to $(C.6)$, for any correct model $\alpha \in \mathscr{A}_c$,*

$$0 \leq \ell(\hat{\beta}(\alpha)|Y_n, X_{n\alpha}) - \ell(\beta_0|Y_n, X_n) = O(\log\log n) \quad a.s.. \tag{6}$$

Theorem 3. *Under conditions $(C.1)$ to $(C.7)$, for any incorrect model $\alpha \in \mathscr{A}_w$, we have*

$$\limsup_{n\to\infty} \frac{1}{n}\{\ell(\hat{\beta}(\alpha)|Y_n, X_{n\alpha}) - \ell(\beta_0|Y_n, X_n)\} < 0 \quad a.s.. \tag{7}$$

From Theorems 2 and 3 we know that the maximum log-likelihood of any correct model is almost surely greater than the unknown true log-likelihood of the full model by an amount bounded by $|O(\log\log n)|$. On the other hand, the maximum log-likelihood of any incorrect model in $\mathscr{A}_w$ is almost surely smaller than the true log-likelihood of the full model by an amount greater than τn with $\tau > 0$ when n is sufficiently large. Therefore, if we use a penalised log-likelihood based criterion of form (3) for model selection, we will almost surely select the simplest correct model in $\mathscr{A}_c$ if the penalty term $C(n, \hat{\beta}(\alpha))$ is an increasing function of the model dimension p_α and is of an order in between $O(\log\log n)$ and $O(n)$. We call a model selection criterion *strongly consistent* if it selects the simplest correct model almost surely; and *consistent* if almost surely it only selects one of the correct models. From this discussion we have the following:

Theorem 4. *For a Poisson regression model satisfying conditions $(C.1)$ to $(C.7)$, both model selection criteria BIC and SCC are strongly consistent, while AIC is consistent but not necessarily strongly consistent.*

Proof. As the Fisher information's determinant $|I(\beta(\alpha))|$ is typically of order $O(n^{p_\alpha})$, both the penalty terms of SCC and BIC are increasing functions of the model dimension p_α and are of order $O(\log n)$, it follows from Theorem 2 and Theorem 3 that both SCC and BIC are strongly consistent. AIC is not necessarily strongly consistent because its penalty term is of order $O(1)$. But AIC is clearly consistent because its criterion value for a correct model is almost surely smaller than that for any incorrect model by an amount greater than τn when n is sufficiently large.

The proof of Theorems 1 to 3 will be the focus of the next section.

4. Proof of the Results

The key to proving our main results lies on the convexity and quadratic approximation of the negative log-likelihood function, the Normal and Gamma approximations of the Poisson

probabilities, and the law of iterated logarithm for independent random variables. The idea of using the convexity property is broadly seen in establishing asymptotic representations of the M-estimators in linear models, see e.g. [15, 20, 12] among the others.

By the definition of ξ_n and conditions (C.2) and (C.3) it is easy to see that there exists a sequence of positive numbers $\{\tau_n\}$ satisfying:

$$\tau_n \uparrow \infty, \quad \tau_n \xi_n \sqrt{\log\log n} \to 0 \quad \text{and} \quad \tau_n \sqrt{n^{-1}\log\log n} \downarrow 0.$$

Using τ_n we introduce two sequences of subsets:

$$\begin{aligned} A_n &= \{\beta : ||\beta - \beta_0|| \le \tau_n \sqrt{n^{-1}\log\log n}\} \\ \partial A_n &= \{\beta : ||\beta - \beta_0|| = \tau_n \sqrt{n^{-1}\log\log n}\}. \end{aligned}$$

It is clear that $A_1 \supset A_2 \supset A_3 \supset \cdots \supset A_n$. Further we define

$$H(\beta, n) = \ell(\beta_0|Y_n, X_n) - \ell(\beta|Y_n, X_n) = \sum_{k=1}^{n} \{e^{\mathbf{x}_k^t \beta} - e^{\mathbf{x}_k^t \beta_0} - y_k \mathbf{x}_k^t (\beta - \beta_0)\},$$

and $K(t,s) = e^t - e^s - e^s(t-s)$. By these definitions it follows that

$$H(\beta, n) = \sum_{k=1}^{n} K(\mathbf{x}_k^t \beta, \mathbf{x}_k^t \beta_0) - \sum_{k=1}^{n} (y_k - \mu_{0k}) \mathbf{x}_k^t (\beta - \beta_0) \stackrel{\text{def}}{=} R_1(\beta, n) + R_2(\beta, n). \tag{8}$$

Before proving the main results we need to establish some preliminary results.

Lemma 1. *The function $K(t,s)$ defined has the following properties:*

(i). $K(t,s) \ge 0$ for any real numbers t and s.

(ii). $K(t,s)$ is strictly convex with respect to t.

(iii). For any $\Delta > 0$,

$$\frac{1}{2} e^{s-2\Delta}(t-s)^2 \le K(t,s) \le \frac{1}{2} e^{s+2\Delta}(t-s)^2 \quad \text{if } |t-s| \le \Delta.$$

The proof of Lemma 1 will be give in Appendix.

Lemma 2. *Let W be a Poisson(θ) random variable. Then for any $w \ge 0$ the following inequalities hold:*

$$P(W \le w) \le (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{(w+1-\theta)/\sqrt{\theta}} e^{-\frac{1}{2}t^2} dt, \tag{9}$$

$$P(W \le w) \ge [\Gamma(\theta+1)]^{-1} \int_0^w t^\theta e^{-t} dt. \tag{10}$$

The results of Lemma 2 were obtained by [2] which can also be found in [6, p. 102].

Lemma 3 (Law of the Iterated Logarithm)**.** *Let $\{Z_n, n \geq 1\}$ be independent random variables with $EZ_n = 0$, $EZ_n^2 = \sigma_n^2$ and $s_n^2 = \sum_{k=1}^n \sigma_k^2 \to \infty$. If $|Z_n| \leq d_n$ a.s., where $d_n = o((s_n^2/\log\log s_n^2)^{1/2})$, then*

$$\limsup_{n\to\infty} \frac{\pm\sum_{k=1}^n Z_k}{\sqrt{2s_n^2 \log\log s_n^2}} = 1 \quad a.s..$$

This lemma and its proof can be found in e.g. [3, pp. 373-374] and [9, pp. 239-246].

Lemma 4. *Under conditions (C.1), (C.2) and (C.4) to (C.6), we have*

$$\limsup_{n\to\infty} \frac{\pm\sum_{k=1}^n (y_k - \mu_{0k})x_{kj}}{\sqrt{2I_n(\beta_0)(j,j)\log\log I_n(\beta_0)(j,j)}} = 1 \quad a.s. \text{ for } j = 1, \cdots, p. \tag{11}$$

Here x_{kj} is the j-th element of $\mathbf{x}_k$ and $I_n(\beta_0)(j,j)$ is the (j,j)-th element of $I_n(\beta_0)$. Equation (11) suggests that $\{(y_k - \mu_{0k})x_{kj}, k = 1, 2, \cdots\}$ obeys the law of iterated logarithm. Accordingly, we have

$$\frac{\partial \ell}{\partial \beta}|_{\beta=\beta_0} = \sum_{k=1}^n (y_k - \mu_{0k})\mathbf{x}_k = X_n^t(Y_n - \boldsymbol{\mu}_0) = O(\sqrt{n\log\log n}) \quad a.s. \tag{12}$$

where $\boldsymbol{\mu}_0 = (\mu_{01}, \cdots, \mu_{0n})^t$ is the true mean vector.

Proof. The result (12) is obvious from (11) and condition (C.2). Hence we only need to prove (11). Without losing generality we assume all $x_{kj} > 0$.

Using the information that $y_k \sim$ Poisson(μ_{0k}) and the definition of $I_n(\beta_0)$ it is easy to verify that for $j = 1, \cdots, p$

$$E(y_k - \mu_{0k})x_{kj} = 0, \tag{13}$$

$$\sum_{k=1}^n E((y_k - \mu_{0k})x_{kj})^2 = \sum_{k=1}^n \mu_{0k}x_{kj}^2 = I_n(\beta_0)(j,j) \to \infty \tag{14}$$

as $n \to \infty$ by condition (C.1).

From now on we proceed to show that for $j = 1, \cdots, p$

$$|(y_n - \mu_{0n})x_{nj}| \leq o(d_{nj}) \quad \text{a.s.} \tag{15}$$

where $d_{nj} = \sqrt{I_n(\beta_0)(j,j)/\log\log I_n(\beta_0)(j,j)} \to \infty$ by (14). For any $\varepsilon > 0$, it is easy to see that

$$P\{|(y_n - \mu_{0n})x_{nj}| > \varepsilon d_{nj}\} \leq P\{y_n > \mu_{0n} + \varepsilon d_{nj}x_{nj}^{-1}\} + P\{y_n < \mu_{0n} - \varepsilon d_{nj}x_{nj}^{-1}\}. \tag{16}$$

Applying (9) of Lemma 2 we have

$$P\{y_n < \mu_{0n} - \varepsilon d_{nj}x_{nj}^{-1}\} \leq (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{(1-\varepsilon d_{nj}x_{nj}^{-1})/\sqrt{\mu_{0n}}} e^{-\frac{1}{2}t^2} dt. \tag{17}$$

Note that from condition (C.1) and the inequality $\lambda_1\{I_n(\beta_0)\} \le I_n(\beta_0)(j,j) \le \lambda_p\{I_n(\beta_0)\}$ we have

$$\begin{aligned}\mu_{0n}x_{nj}^2 &\le \mu_{0n}\mathbf{x}_n^t\mathbf{x}_n \le \lambda_p\{I_n(\beta_0)\}\mu_{0n}\mathbf{x}_n^t I_n(\beta_0)^{-1}\mathbf{x}_n \\ &\le \lambda_p\{I_n(\beta_0)\}\delta_n^2 \le \frac{\lambda_p\{I_n(\beta_0)\}}{\lambda_1\{I_n(\beta_0)\}}\frac{I_n(\beta_0)(j,j)}{\log\log I_n(\beta_0)(j,j)}\delta_n^2\log\log\lambda_p\{I_n(\beta_0)\} \\ &\le b_0 d_{nj}^2\delta_n^2\log\log\lambda_p\{I_n(\beta_0)\}. \end{aligned} \tag{18}$$

Thus when $\mu_{0n} \ge 1$,

$$\frac{1-\varepsilon d_{nj}x_{nj}^{-1}}{\sqrt{\mu_{0n}}} = \frac{1}{\sqrt{\mu_{0n}}} - \varepsilon\sqrt{\frac{d_{nj}^2}{\mu_{0n}x_{nj}^2}} \le 1 - \frac{\varepsilon}{\sqrt{b_0\delta_n^2\log\log\lambda_p\{I_n(\beta_0)\}}} \to -\infty$$

by condition (C.4). This implies that

$$\frac{1-\varepsilon d_{nj}x_{nj}^{-1}}{\sqrt{\mu_{0n}}} \le -\frac{1}{2}\varepsilon(b_0\delta_n^2\log\log\lambda_p\{I_n(\beta_0)\})^{-\frac{1}{2}} \quad \text{when } n \text{ is sufficiently large.} \tag{19}$$

From (17), (19) and a well-known inequality

$$\int_a^\infty e^{-\frac{1}{2}t^2}dt < \frac{1}{a}e^{-\frac{1}{2}a^2} \quad \text{for all } a>0$$

[see 3, p.49], it follows that when $\mu_{0n} \ge 1$ and n is sufficiently large,

$$\begin{aligned}P\{y_n < \mu_{0n} - \varepsilon d_{nj}x_{nj}^{-1}\} &\le (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{-\frac{1}{2}\varepsilon(b_0\delta_n^2\log\log\lambda_p\{I_n(\beta_0)\})^{-\frac{1}{2}}} e^{-\frac{1}{2}t^2}dt \\ &< (2\pi)^{-1}2\varepsilon^{-1}(b_0\delta_n^2\log\log\lambda_p\{I_n(\beta_0)\})^{\frac{1}{2}}e^{-\frac{1}{8}\varepsilon^2(b_0\delta_n^2\log\log\lambda_p\{I_n(\beta_0)\})^{-1}} \\ &\le (\log\lambda_p\{I_n(\beta_0)\})^{-\frac{1}{2}}(\lambda_p\{I_n(\beta_0)\})^{-2} \end{aligned} \tag{20}$$

where the last inequality follows from condition (C.4).

When $\mu_{0n} < 1$, it follows from (18) and conditions (C.1) and (C.4) that

$$\mu_{0n} - \varepsilon d_{nj}x_{nj}^{-1} = \sqrt{\mu_{0n}}(\sqrt{\mu_{0n}} - \varepsilon\sqrt{\frac{d_{nj}^2}{\mu_{0n}x_{nj}^2}}) \le \sqrt{\mu_{0n}}(1-\varepsilon(b_0\delta_n^2\log\log\lambda_p\{I_n(\beta_0)\})^{-\frac{1}{2}}) < 0$$

if n is sufficiently large, which suggests $P\{y_n < \mu_{0n} - \varepsilon d_{nj}x_{nj}^{-1}\} = 0$ if n is sufficiently large. Therefore the result (20) is true for any $\mu_{0n} > 0$, which implies that

$$\sum_{n=1}^\infty P\{y_n < \mu_{0n} - \varepsilon d_{nj}x_{nj}^{-1}\} < \infty \tag{21}$$

if further condition (C.2) holds.

Now let $\{c_n : 0 < c_n \leq \frac{1}{2}, n = 1, 2, \cdots\}$ be a sequence of numbers to be determined. Applying the result (10) of Lemma 2 and the property for the Gamma function $[\Gamma(\theta + 1)]^{-1} \int_0^\infty t^\theta e^{-t} dt = 1$, we have

$$\begin{aligned} P\{y_n > \mu_{0n} + \varepsilon d_{nj} x_{nj}^{-1}\} &\leq [\Gamma(\mu_{0n}+1)]^{-1} \int_{\mu_{0n}+\varepsilon d_{nj} x_{nj}^{-1}}^{\infty} t^{\mu_{0n}} e^{-t} dt \\ &\leq e^{-c_n(\mu_{0n}+\varepsilon d_{nj} x_{nj}^{-1})} [\Gamma(\mu_{0n}+1)]^{-1} \int_{\mu_{0n}+\varepsilon d_{nj} x_{nj}^{-1}}^{\infty} t^{\mu_{0n}} e^{-(1-c_n)t} dt \\ \leq e^{-c_n(\mu_{0n}+\varepsilon d_{nj} x_{nj}^{-1})} (1-c_n)^{-(\mu_{0n}+1)} &= \frac{1}{1-c_n} \left(\frac{e^{-c_n}}{1-c_n} \right)^{\mu_{0n}} e^{-\varepsilon c_n d_{nj} x_{nj}^{-1}} \\ \leq 2 \left(\frac{1-c_n+\frac{1}{2}c_n^2}{1-c_n} \right)^{\mu_{0n}} e^{-\varepsilon c_n d_{nj} x_{nj}^{-1}} &= 2 \left(1 + \frac{c_n^2}{2(1-c_n)} \right)^{\mu_{0n}} e^{-\varepsilon c_n d_{nj} x_{nj}^{-1}}. \end{aligned} \tag{22}$$

We take $c_n = \min\{\frac{1}{2}, \mu_{0n}^{-\frac{1}{2}}\}$ so $\frac{c_n^2}{2(1-c_n)} \leq \min\{\frac{1}{4}, \mu_{0n}^{-1}\}$. By considering the two cases $\mu_{0n} \leq 4$ and $\mu_{0n} > 4$ separately and using the property that $(1 + \frac{1}{a})^a \uparrow e$ as $a \uparrow \infty$, it is easy to see that

$$\left(1 + \frac{c_n^2}{2(1-c_n)} \right)^{\mu_{0n}} \leq \max\{ \left(\frac{5}{4} \right)^4, e\} = e. \tag{23}$$

Applying (18) one can show that when n is sufficiently large

$$\varepsilon d_{nj} \mu_{0n}^{-\frac{1}{2}} x_{nj}^{-1} \geq \varepsilon b_0^{-\frac{1}{2}} (\delta_n \sqrt{\log\log \lambda_p\{I_n(\beta_0)\}})^{-1} \geq 2 \log \lambda_p\{I_n(\beta_0)\} \tag{24}$$

under conditions (C.1) and (C.5). In the same way as proving (18) one can show that under condition (C.1)

$$x_{nj}^2 \leq b_0 d_{nj}^2 \xi_n^2 \log\log \lambda_p\{I_n(\beta_0)\}. \tag{25}$$

By (25) and condition (C.6) it follows that

$$\frac{1}{2} \varepsilon d_{nj} x_{nj}^{-1} \geq \frac{1}{2} \varepsilon b_0^{-\frac{1}{2}} (\xi_n \sqrt{\log\log \lambda_p\{I_n(\beta_0)\}})^{-1} \geq 2 \log \lambda_p\{I_n(\beta_0)\} \tag{26}$$

when n is sufficiently large. By (24), (26) and the fact that $c_n = \min\{\frac{1}{2}, \mu_{0n}^{-\frac{1}{2}}\}$ it follows that when n is sufficiently large,

$$\varepsilon c_n d_{nj} x_{nj}^{-1} \geq 2 \log \lambda_p\{I_n(\beta_0)\}. \tag{27}$$

Now from (22), (23) and (27) we have

$$P\{y_n > \mu_{0n} + \varepsilon d_{nj} x_{nj}^{-1}\} \leq 2e \lambda_p\{I_n(\beta_0)\}^{-2} \tag{28}$$

when n is sufficiently large. From (28) and condition (C.2) it follows that

$$\sum_{n=1}^{\infty} P\{y_n > \mu_{0n} + \varepsilon d_{nj} x_{nj}^{-1}\} < \infty. \tag{29}$$

Following the results (16), (21) and (29) we have

$$\sum_{n=1}^{\infty} P\{|(y_n - \mu_{0n})x_{nj}| > \varepsilon d_{nj}\} < \infty.$$

Hence by the Borel-Cantelli lemma,

$$P\{|(y_n - \mu_{0n})x_{nj}| > \varepsilon d_{nj} \quad \text{occurs infinitely often}\} = 0 \quad \text{for any } \varepsilon > 0,$$

which implies that (15) is true. Since (13) to (15) are true, the result (11) is followed by applying Lemma 3 for the independent random variables $\{(y_k - \mu_{0k})x_{kj}, k = 1, 2, \cdots\}$.

Proof. (Theorem 1) Clearly it is sufficient to prove (4) only for the full model:

$$||\hat{\beta} - \beta_0|| = O(\sqrt{n^{-1}\log\log n}) \quad \text{a.s.} \tag{30}$$

Applying result (iii) of Lemma 1 with $t = \mathbf{x}_k^t\beta$, $s = \mathbf{x}_k^t\beta_0$ and $\Delta = |\mathbf{x}_k^t\beta - \mathbf{x}_k^t\beta_0|$, it follows that

$$K(\mathbf{x}_k^t\beta, \mathbf{x}_k^t\beta_0) \geq \frac{1}{2} e^{-2|\mathbf{x}_k^t(\beta-\beta_0)|}\mu_{0k}[\mathbf{x}_k^t(\beta - \beta_0)]^2. \tag{31}$$

Following the definition of ξ_n and condition (C.2) one can find that

$$\max_{1\leq k\leq n} ||\mathbf{x}_k||^2 \leq \lambda_p\{I_n(\beta_0)\} \max_{1\leq k\leq n} \mathbf{x}_k^t I_n(\beta_0)^{-1}\mathbf{x}_k = \lambda_p\{I_n(\beta_0)\}\xi_n^2 \leq b_2 n\xi_n^2. \tag{32}$$

Thus by (32) and Cauchy-Schwarz inequality,

$$\max_{1\leq k\leq n} |\mathbf{x}_k^t(\beta-\beta_0)|I(\beta\in\partial A_n) \leq \max_{1\leq k\leq n} ||\mathbf{x}_k|| \cdot ||\beta-\beta_0||I(\beta\in\partial A_n) \leq \sqrt{b_2}\xi_n\tau_n\sqrt{\log\log n} \tag{33}$$

where $I(\beta \in \partial A_n)$ is an indicator function indicating that only those β in ∂A_n will be under consideration. (This type of definition for the indicator function will be used in the rest of the paper.) It follows from (8) and (31) to (33) that

$$\begin{aligned} R_1(\beta, n)I(\beta\in\partial A_n) &\geq \frac{1}{2} e^{-2\max_{1\leq k\leq n}|\mathbf{x}_k^t(\beta-\beta_0)|} \sum_{k=1}^{n} \mu_{0k}[\mathbf{x}_k^t(\beta-\beta_0)]^2 I(\beta\in\partial A_n) \\ &\geq \frac{1}{2} e^{-2\sqrt{b_2}\xi_n\tau_n\sqrt{\log\log n}}(\beta-\beta_0)^t I_n(\beta_0)(\beta-\beta_0)I(\beta\in\partial A_n) \\ &\geq \frac{1}{2} e^{-2\sqrt{b_2}\xi_n\tau_n\sqrt{\log\log n}}\lambda_1\{I_n(\beta_0)\}||\beta-\beta_0||^2 I(\beta\in\partial A_n). \end{aligned} \tag{34}$$

By conditions (C.1) and (C.2) and the fact that $\tau_n \xi_n \sqrt{\log\log n} \to 0$, it comes after (34) that there exists a constant $b_4 > 0$ such that

$$R_1(\beta, n) I(\beta \in \partial A_n) \geq b_4 \tau_n^2 \log\log n. \tag{35}$$

On the other hand, by result (12) of Lemma 4 and (8),

$$\begin{aligned} |R_2(\beta, n)| I(\beta \in \partial A_n) &\leq ||\sum_{k=1}^{n} (y_k - \mu_{0k}) \mathbf{x}_k|| \cdot ||\beta - \beta_0|| I(\beta \in \partial A_n) \\ &= O(\sqrt{n \log\log n}) \tau_n \sqrt{n^{-1} \log\log n} = O(1) \tau_n \log\log n \quad \text{a.s..} \end{aligned} \tag{36}$$

Knowing (35) and (36) we can find a constant $b_5 > 0$ so that

$$H(\beta, n) I(\beta \in \partial A_n) = \{R_1(\beta, n) + R_2(\beta, n)\} I(\beta \in \partial A_n) \geq b_5 \tau_n^2 \log\log n \quad \text{a.s..} \tag{37}$$

It is easy to see that $H(\beta, n)$ is convex by Lemma 1 and that $H(\beta_0, n) = 0$. This and (37) suggests that the MLE $\hat{\beta}$ which also minimizes $H(\beta, n)$ must be inside the subset A_n almost surely; namely

$$||\hat{\beta} - \beta_0|| \leq \tau_n \sqrt{n^{-1} \log\log n} \quad \text{a.s..} \tag{38}$$

Equation (38) implies (30) because the sequence $\{\tau_n\}$ can be chosen to diverge as slowly as possible.

We now proceed to prove (5) for the full model. Suppose (5) does not hold for the full model. This implies that

$$||\hat{\beta} - \beta_0|| = o(\sqrt{n^{-1} \log\log n}) \quad \text{a.s.,} \tag{39}$$

knowing that (30) is true. By applying result (iii) of Lemma 1, (8) and (32) it follows that

$$\begin{aligned} R_1(\hat{\beta}, n) &\leq \frac{1}{2} e^{2 \max_{1 \leq k \leq n} |\mathbf{x}_k^t (\hat{\beta} - \beta_0)|} \sum_{k=1}^{n} \mu_{0k} [\mathbf{x}_k^t (\hat{\beta} - \beta_0)]^2 \\ \leq \frac{1}{2} e^{2\sqrt{b_2 n} \xi_n ||\hat{\beta} - \beta_0||} (\hat{\beta} - \beta_0)^t I_n(\beta_0) (\hat{\beta} - \beta_0) &\leq \frac{1}{2} e^{2\sqrt{b_2 n} \xi_n ||\hat{\beta} - \beta_0||} \lambda_p \{I_n(\beta_0)\} ||\hat{\beta} - \beta_0||^2. \end{aligned} \tag{40}$$

Thus by (39) and conditions (C.2) and (C.3) we have $R_1(\hat{\beta}, n) = o(1) \log\log n$ a.s.. Corresponding to (36) it can be seen that $R_2(\hat{\beta}, n) = o(1) \log\log n$ a.s. under the assumption of (39). Hence we have that under assumption (39)

$$H(\hat{\beta}, n) = R_1(\hat{\beta}, n) + R_2(\hat{\beta}, n) = o(1) \log\log n \quad \text{a.s..} \tag{41}$$

On the other hand, from (11) of Lemma 4 we know there exists a sequence of positive integers $\{n_i \uparrow \infty\}$ such that

$$\lim_{i \to \infty} \frac{\sum_{k=1}^{n_i} (y_k - \mu_{0k}) x_{k1}}{\sqrt{2 I_{n_i}(\beta_0)(1,1) \log\log I_{n_i}(\beta_0)(1,1)}} = 1 \quad \text{a.s..}$$

Thus when n_i is sufficiently large,

$$\frac{\sum_{k=1}^{n_i}(y_k-\mu_{0k})x_{k1}}{\sqrt{2I_{n_i}(\beta_0)(1,1)\log\log I_{n_i}(\beta_0)(1,1)}} \geq \frac{1}{2} \quad \text{a.s..} \tag{42}$$

Now define a $p\times 1$ vector $\tilde{\beta}_{n_i}$ with $\tilde{\beta}_{n_i}(j)=\beta_{0j}$ for $j=2,\cdots,p$ and

$$\tilde{\beta}_{n_i}(1)=\frac{b_1}{4b_0b_2}\sqrt{\frac{2\log\log I_{n_i}(\beta_0)(1,1)}{I_{n_i}(\beta_0)(1,1)}}+\beta_{01}.$$

Then from (42) it follows that, when n_i is sufficiently large

$$\begin{aligned} R_2(\tilde{\beta}_{n_i},n_i) &= \sum_{k=1}^{n_i}(\mu_{0k}-y_k)\mathbf{x}_k^t(\tilde{\beta}_{n_i}-\beta_0)=\sum_{k=1}^{n_i}(\mu_{0k}-y_k)x_{k1}(\tilde{\beta}_{n_i}(1)-\beta_{01}) \\ &\leq -\frac{1}{2}\sqrt{2I_{n_i}(\beta_0)(1,1)\log\log I_{n_i}(\beta_0)(1,1)}\cdot\frac{b_1}{4b_0b_2}\sqrt{\frac{2\log\log I_{n_i}(\beta_0)(1,1)}{I_{n_i}(\beta_0)(1,1)}} \\ &= -\frac{b_1}{4b_0b_2}\log\log I_{n_i}(\beta_0)(1,1) \quad \text{a.s..} \end{aligned} \tag{43}$$

Note that by conditions (C.1) and (C.2)

$$b_2 n \geq \lambda_p\{I_n(\beta_0)\} \geq I_n(\beta_0)(1,1) \geq \lambda_1\{I_n(\beta_0)\} \geq \frac{b_1}{b_0}n \tag{44}$$

and accordingly

$$2\log\log n \geq \log\log I_n(\beta_0)(1,1) \geq \frac{1}{2}\log\log n \quad \text{when } n \text{ is sufficiently large.} \tag{45}$$

It follows that

$$\frac{\sqrt{b_1}}{2\sqrt{b_0}b_2}\sqrt{n_i^{-1}\log\log n_i} \geq \tilde{\beta}_{n_i}(1)-\beta_{01} \geq \frac{b_1}{4b_0b_2\sqrt{b_2}}\sqrt{n_i^{-1}\log\log n_i} \tag{46}$$

when n_i is sufficiently large. Using (46), (32) and the fact that $\xi_n\sqrt{\log\log n}\to 0$ one can show that

$$\max_{1\leq k\leq n_i}|\mathbf{x}_k^t(\tilde{\beta}_{ni}-\beta_0)| \leq \max_{1\leq k\leq n_i}||\mathbf{x}_k||(\tilde{\beta}_{n_i}(1)-\beta_{01}) \leq \frac{1}{2}\sqrt{\frac{b_1}{b_0b_2}}\xi_{n_i}\sqrt{\log\log n_i} \leq \frac{1}{2}\log 2 \tag{47}$$

when n_i is sufficiently large. Similar to proving (40), by (44) and (47) one can see that

$$R_1(\tilde{\beta}_{n_i},n_i) \leq \lambda_p\{I_{n_i}(\beta_0)\}||\tilde{\beta}_{n_i}-\beta_0||^2 \leq b_2 n_i\frac{b_1^2}{16b_0^2b_2^2}\frac{2\log\log I_{n_i}(\beta_0)(1,1)}{I_{n_i}(\beta_0)(1,1)}$$

$$\leq \frac{b_1}{8b_0b_2}\log\log I_{n_i}(\beta_0)(1,1) \quad \text{when } n_i \text{ is sufficiently large.} \tag{48}$$

Thus, by (43), (48) and (45) it follows that when n_i is sufficiently large,

$$\begin{aligned} H(\tilde{\beta}_{n_i}, n_i) &= R_1(\tilde{\beta}_{n_i}, n_i) + R_2(\tilde{\beta}_{n_i}, n_i) \\ \leq -\frac{b_1}{8b_0b_2}\log\log I_{n_i}(\beta_0)(1,1) &\leq -\frac{b_1}{16b_0b_2}\log\log n_i \quad \text{a.s..} \end{aligned} \tag{49}$$

Since $\hat{\beta} \equiv \hat{\beta}_n$ is the MLE that minimizes $H(\beta, n)$, inferring from (49) we have

$$H(\hat{\beta}_{n_i}, n_i) \leq H(\tilde{\beta}_{n_i}, n_i) \leq -\frac{b_1}{16b_0b_2}\log\log n_i \quad \text{a.s.} \tag{50}$$

when n_i is sufficiently large. Clearly, (50) is contradictory to (41) which suggests that (39) is wrong. Therefore (5) is true for the full model and consequently so for the other correct models in $\mathscr{A}_c$.

Proof. (Theorem 2) As in proving Theorem 1, we only need to prove (6) for the full model which is equivalent to

$$0 \geq H(\hat{\beta}, n) = R_1(\hat{\beta}, n) + R_2(\hat{\beta}, n) = O(\log\log n) \quad \text{a.s.} \tag{51}$$

by the definition of $H(\beta, n)$. The inequality part of (51) is obvious because $\hat{\beta}$ is the MLE of β. Note that result (40) is also valid here, so by Theorem 1 and conditions (C.2) and (C.3)

$$0 \leq R_1(\hat{\beta}, n) \leq \frac{1}{2}e^{2\sqrt{b_2 n}\xi_n||\hat{\beta}-\beta_0||}\lambda_p\{I_n(\beta_0)\}||\hat{\beta}-\beta_0||^2 = O(\log\log n) \quad \text{a.s..} \tag{52}$$

On the other hand, by result (12) of Lemma 4 and (4) of Theorem 1 we have

$$|R_2(\hat{\beta}, n)| \leq ||\sum_{k=1}^{n}(y_k - \mu_{0k})\mathbf{x}_k|| \cdot ||\hat{\beta} - \beta_0|| = O(\log\log n) \quad \text{a.s..} \tag{53}$$

By (52) and (53) it follows that $|H(\hat{\beta}, n)| = O(\log\log n)$ a.s. which suffices the proof of the theorem.

Proof. (Theorem 3) First we extend the $p_\alpha \times 1$ vector $\hat{\beta}(\alpha)$, the MLE of β_α, to a $p \times 1$ vector $\hat{\beta}^*(\alpha)$ by inserting $p - p_\alpha$ 0's into $\hat{\beta}(\alpha)$ in such a way that the sub-vector of $\hat{\beta}^*(\alpha)$ indexed by α is equal to $\hat{\beta}(\alpha)$. Then it is easy to see that proving (7) is equivalent to proving

$$\liminf_{n\to\infty} n^{-1}H(\hat{\beta}^*(\alpha), n) > 0 \quad \text{a.s. for any incorrect model } \alpha \in \mathscr{A}_w. \tag{54}$$

Define

$$A_0 = \{\beta : ||\beta - \beta_0|| \leq \frac{1}{2}\min_{1\leq i\leq p_{\alpha_0}} |\beta_0(\alpha_0)_i| = b\}.$$

Clearly A_0 is a compact set; and for any incorrect model $\alpha \in \mathscr{A}_w$ we have $\hat{\beta}^*(\alpha) \notin A_0$ because $||\hat{\beta}^*(\alpha) - \beta_0|| \geq 2b$. Moreover, by Theorem 1, the MLE $\hat{\beta}$ for the full model is an interior point of A_0 almost surely when n is sufficiently large. Since $H(\beta, n)$ is convex with respect to β, it follows that

$$H(\hat{\beta}^*(\alpha), n) \geq \inf_{\beta \in \partial A_0} H(\beta, n)$$

where ∂A_0 is the boundary of A_0. Now it is sufficient to prove

$$\liminf_{n\to\infty} \inf_{\beta \in \partial A_0} n^{-1} H(\beta, n) > 0 \quad \text{a.s.} \tag{55}$$

in order to prove (54). Using result (iii) of Lemma 1, (8), condition (C.7) and Cauchy-Schwarz inequality one can show that

$$\begin{aligned} R_1(\beta, n) I(\beta \in \partial A_0) &\geq \frac{1}{2} \sum_{k=1}^{n} e^{-2||\mathbf{x}_k|| \cdot ||\beta - \beta_0||} \mu_{0k} [\mathbf{x}_k^t (\beta - \beta_0)]^2 I(\beta \in \partial A_0) \\ &= \frac{1}{2} (\beta - \beta_0)^t X_n^t \mathscr{M}_n X_n (\beta - \beta_0) I(\beta \in \partial A_0) \\ &\geq \frac{1}{2} \lambda_1 \{X_n^t \mathscr{M}_n X_n\} ||\beta - \beta_0||^2 I(\beta \in \partial A_0) \geq \frac{1}{2} b_3 b^2 n, \end{aligned}$$

which suggests

$$\inf_{\beta \in \partial A_0} R_1(\beta, n) \geq \frac{1}{2} b_3 b^2 n. \tag{56}$$

Following the same line as proving (36) one can show that

$$\sup_{\beta \in \partial A_0} |R_2(\beta, n)| = O(\sqrt{n \log\log n}) \quad \text{a.s.}. \tag{57}$$

Since $\inf_{\beta \in \partial A_0} H(\beta, n) \geq \inf_{\beta \in \partial A_0} R_1(\beta, n) - \sup_{\beta \in \partial A_0} |R_2(\beta, n)|$ by (8), it follows from (56) and (57) that (55) is true and consequently (54) is true.

5. Discussion

The asymptotic results obtained in this paper are based on the assumption that the response variable follows a Poisson distribution and the candidate models under consideration for selection are based on those available explanatory variables. In practice, there may exist some lurking variables which also affect the response variable. In this situation, a mixture Poisson distribution may be introduced for modeling the response variable, in which an over-dispersion parameter is used to account for the effects of the lurking variables. We refer to [8, section 6.2.3] for details of the over-dispersion log-linear models. Then the model selection procedure can still focus on those available explanatory variables. Provided that a result similar to that in Lemma 2 can be obtained for the mixture Poisson probability (which may be shown by imitating the proof in [2]), it seems the same results as of Theorems 1 to 4 can also

be established for the over-dispersion log-linear models using the same methods employed in this paper.

It is also possible to extend our asymptotic results to the Poisson regression models with the link functions other than the log-link, which is still under our investigation but shall be presented somewhere else.

In addition to the model selection criteria studied here, there are many other approaches for model selection, e.g. the hierarchical Bayesian approach [5] and LASSO [19] etc., for which our results may not be applicable.

Finally, other than determining a model selection criterion and assessing its asymptotic performance, there is a computational issue on how to execute a model selection procedure from many possible candidate models. This becomes especially important when the number of candidate models, often of the magnitude 2^p, is enormous. However, a thorough investigation of this issue is beyond the scope of this paper. We refer to [10] and [13] for some results in this area.

References

[1] H Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csáki, editors, *Proceedings of the Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akadémia Kiadó.

[2] H Bohman. Two inequalities for poisson distributions. *Skandinavisk Aktuarietidskrift*, 46:47–52, 1963.

[3] Y Chow and H Teicher. *Probability Theory: independence, interchangeability, martingales*. Springer, New York, 3 edition, 1997.

[4] E George. *Statistics in the 21st Century*, chapter The variable selection problem, pages 350–358. Chaptman & Hall/CRC, 2002.

[5] E George and R McCullock. Approaches for bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

[6] N Johnson and S Kotz. *Discrete Distributions*. Houghton Mifflin Company, Boston, Massachusetts, 1969.

[7] C Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

[8] P McCullagh and J Nelder. *Generalized Linear Models*. Chapman & Hall, London, 2 edition, 1989.

[9] V Petrov. *Limit Theorems of Probability Theory: sequences of independent random variables*. Oxford University Press, 1995.

[10] G Qian. Computations and analysis in robust regression model selection using stochastic complexity. *Computational Statistics*, 14:293–314, 1999.

[11] G Qian and H Künsch. Some notes on rissanen's stochastic complexity. *IEEE Transactions on Information Theory*, 44:782–786, 1998.

[12] G Qian and Y Wu. Strong limit theorems on model selection in generalized linear regression with binomial responses. *Statisticsa Sinica*, 16:1335–1365, 2006.

[13] G Qian and X Zhao. On time series model selection involving many candidate arma models. *Computational Statistics & Data Analysis*, 51:6180–6196, 2007.

[14] C Rao and Y Wo. *Model Selection*, volume 38 of *IMS Lecture Notes - Monograph Series*, chapter On model selection (with discussion), pages 1–64. Institute of Mathematical Statistics, Beachwood, Ohio, 2001.

[15] C Rao and L Zhao. Linear representation of m-estimates in linear models. *The Canadian Journal of Statistics*, 20:359–368, 1992.

[16] J Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Co. Pte. Ltd., Singapore, 1989.

[17] J Rissanen. Fisher information and stochastic complexity. *IEEE Transactions Information Theory*, 42:40–47, 1996.

[18] G Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[19] R Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*.

[20] Y Wu and M Zen. A strongly consistent linear model selection procedure based on m-estimation. *Probability Theory and Related Fields*, 113:599–625, 1999.

Appendix

Proof. (Lemma 1) Since

$$K(t,s) = e^t - e^s - e^s(t-s),$$

it is easy to see that

$$K(s,s) = 0, \quad K'_t(t,s) = e^t - e^s, \quad K'_t(s,s) = 0 \quad \text{and} \quad K''_t(t,s) = e^t > 0.$$

Hence, $K(t,s)$ is strictly convex with respect to t, and $K(t,s) \geq 0$ with $K(t,s) = 0$ only if $t = s$. Now let

$$F(t,s) = K(t,s) - \frac{1}{2}e^{s-2\Delta}(t-s)^2$$
$$G(t,s) = K(t,s) - \frac{1}{2}e^{s+2\Delta}(t-s)^2$$

For any real numbers s, t and $\Delta > 0$, suppose $|t-s| \leq \Delta$. Then we have

$$s - 2\Delta < s - \Delta \leq t \leq s + \Delta < s + 2\Delta.$$

It is clear that

$$\begin{gathered} F(s,s) = 0, \quad F_t'(t,s) = e^t - e^s - e^{s-2\Delta}(t-s), \\ F_t'(s,s) = 0 \quad \text{and} \quad F_t''(t,s) = e^t - e^{s-2\Delta} > 0 \end{gathered}$$

Therefore, $F(t,s) \geq 0$ with $F(t,s) = 0$ only if $t = s$, namely

$$K(t,s) \geq \frac{1}{2} e^{s-2\Delta}(t-s)^2.$$

Similarly, we can show that $G(t,s) \leq 0$ with $G(t,s) = 0$ only if $t = s$, namely

$$K(t,s) \leq \frac{1}{2} e^{s+2\Delta}(t-s)^2.$$

This concludes the proof.