MODEL COMPARISON FOR DEPENDENT GENERALIZED LINEAR MODEL

SHOICHI EGUCHI

ABSTRACT. The classical Bayesian information criterion (BIC) is derived through the stochastic expansion of marginal likelihood function under suitable regularity condition when models are correctly specified. However, despite of its popularity, mathematical validity of BIC for possibly misspecified models with complicated dependence structure is often ignored. Thus it is important to extend the reach of the classical BIC with rigorous theoretical foundation with allowing model misspecification and asymptotic mixed normality of estimator. In this paper, we will prove the stochastic expansion of marginal quasi-likelihood function associated with a class of possibly misspecified generalized linear models for dependent data.

1. Introduction

Generalized linear model (GLM, McCullagh and Nelder [22]) is an extension of a linear regression model. This model depends on canonical parameter and dispersion parameter, where the former is represented by the link function determined by the conditional distribution of the response variable given the explanatory variable. Moreover, GLM has, among others, the following applications and extensions:

- Actuarial Science (Antonio and Beirlant [3], Haberman and Renshaw [14]):
  - Insurance pricing, loss reserving, estimating claim settlement values, territorial rating, modeling accident frequencies.
- GLMixedM in risk management (McNeil and Wendin [23]):
  - Credit Risk.
- Generalized Additive Models (Berg [4], Hastie and Tibshirani [15]):
  - Predictive modeling, real estate appraisal.

We consider data $(y_j, x_j)_{j=1}^n = (y_j, x_{j,1}, \dots, x_{j,p})_{j=1}^n$, where y_j's and x_j's are realizations of the response variables $\mathbf{Y}_n = (Y_1, \dots, Y_n)'$ and the explanatory variables $\mathbf{X}_n = (X_1, \dots, X_n)'$, respectively, where the notation $'$ means the transpose. Furthermore, we will assume that the conditional distribution of $\mathbf{Y}_n$ given $\mathbf{X}_n$ is given by a GLM. Then the conditional distribution is assumed to belong to an exponential family, for example normal, binomial, Poisson and so on. In this paper, we will a result about the stochastic expansion the stochastic expansion of marginal quasi-likelihood function associated with a class of possibly misspecified GLMs for dependent data. Based on the expansion, we propose the quasi-Bayesian information criterion, which is the extension of the generalized BIC given by Luv and Liu [20].

Suppose that we are given M Bayesian candidate models $\mathfrak{M}_1, \dots, \mathfrak{M}_M$. Each $\mathfrak{M}_m$ is described by $\left\{\left(\mathfrak{p}_m, \pi_m(\theta), \mathbb{H}_{m,n}(\theta)\right) \middle| \theta \in \Theta_m\right\}$, where $\mathfrak{p}_m$ is the non-zero prior relative occurrence probability of mth-model among the M Bayesian models, π_m is the prior-probability density on Θ_m and $\mathbb{H}_{m,n}$ is the logarithmic quasi-likelihood function. The conventional Bayesian principle of model selection for $\mathfrak{M}_1, \dots, \mathfrak{M}_M$ is to choose the model that is most likely in terms of the posterior probability, i.e. to choose model $\mathfrak{M}_{m_0}$ such that $m_0 = \mathrm{argmax}_{m \in \{1,\dots,M\}} P(\mathfrak{M}_m | \mathbf{y}_n)$, where

$$P(\mathfrak{M}_m | \mathbf{y}_n) = \frac{\left(\int_{\Theta_m} \exp\{\mathbb{H}_{m,n}(\theta)\}\pi_m(\theta)d\theta\right)\mathfrak{p}_m}{\sum_{i=1}^M \left(\int_{\Theta_i} \exp\{\mathbb{H}_{i,n}(\theta)\}\pi_i(\theta)d\theta\right)\mathfrak{p}_i},$$

where $\int_{\Theta_m} \exp\{\mathbb{H}_{m,n}(\theta)\}\pi_m(\theta)d\theta$ is called the marginal quasi-likelihood function. When the prior plausibilities on the M competing models would be equal, we select the model that maximizes the marginal

Date: January 7, 2016.
Key words and phrases. Asymptotic Bayesian model comparison, quasi-likelihood, dependent data, model misspecification, generalized linear model.

quasi-likelihood function; even if the prior probabilities are not equal, we can trivially correct the selection manner by the factors $\mathfrak{p}_m$. Hence we focus on the logarithm of the marginal quasi-likelihood function

$$\log\left(\int_\Theta \exp\{\mathbb{H}_{m,n}(\theta)\}\pi_m(\theta)d\theta\right)$$

as the principle of model selection.

As was explained in [20], another interpretation of model selection is possible through the Kullback-Leibler divergence (KL divergence). The KL divergence between the true conditionnal model g_n and the marginal quasi-likelihood function $\int_\Theta \exp\{\mathbb{H}_{m,n}(\theta)\}\pi_m(\theta)d\theta$ is given by

$$I\left(g_n; \int_\Theta \exp\{\mathbb{H}_{m,n}(\theta)\}\pi_m(\theta)d\theta\right) = E[\log g_n(\mathbf{Y}_n|\mathbf{X}_n)] + E\left[-\int_\Theta \exp\{\mathbb{H}_{m,n}(\theta)\}\pi_m(\theta)d\theta\right], \quad (1)$$

where the expectation is taken with respect to the true distribution G_n. Because of (1), we see that $-\int_\Theta \exp\{\mathbb{H}_{m,n}(\theta)\}\pi_m(\theta)d\theta$ is an unbiased estimator of $I\big(g_n; \int_\Theta \exp\{\mathbb{H}_n(\theta;\cdot)\}\pi(\theta)d\theta\big)$ except for a constant term free of θ. Note that (1) holds true regardless of whether or not the true model is in the set of candidate models, implying that Bayesian principle of model selection can be restated as choosing the model that minimizes the KL divergence of the marginal quasi-likelihood function from the true distribution.

In particular, assume that $\mathbf{X}_n$ is absent and that $\mathbb{H}_{m,n}(\theta) = \sum_{j=1}^n \log f_{m,n}(y_j;\theta)$ for the case of independent observations with correctly specified regular models, then Schwarz [24] showed that the marginal quasi-likelihood, $\log(\int_\Theta \exp\{\mathbb{H}_{m,n}(\theta)\}\pi_m(\theta)d\theta)$ admits the stochastic expansion

$$\log\left(\int_\Theta \exp\{\mathbb{H}_{m,n}(\theta)\}\pi_m(\theta)d\theta\right) = \sum_{j=1}^n \log f_{m,n}(y_j;\hat{\theta}_{m,n}^{\mathrm{MLE}}) - \frac{p}{2}\log n + O_p(1), \quad (2)$$

with $\hat{\theta}_{m,n}^{\mathrm{MLE}}$ denoting the maximum likelihood estimator of θ, under some regularity conditions. Due to (2), we obtain the classical Bayesian information criterion for model selection:

$$\mathrm{BIC} = -2\sum_{j=1}^n \log f_{m,n}(y_j;\hat{\theta}_{m,n}^{\mathrm{MLE}}) + p\log n.$$

In the past, many authors have investigated the information criteria for model selection in various settings; see, for example, Burnham and Anderson [6] for an account of these developments. Bozdogan [5] showed that Akaike information criterion (AIC, Akaike [1], [2]) has a positive probability of overestimating the true dimension. Casella *et al.* [7] and Fasen and Kimmig [13] as well as the references therein studied the model selection consistency of BIC. Moreover, various extensions of AIC and BIC have been proved; for example, the extended BIC for large model spaces (Chen and Chen [9]), the generalized information criteria (Konishi and Kitagawa [18]), the generalized BIC in misspecified GLMs for independent data (Lv and Liu [20]) and the information criteria in the case of dependent data (e.g. Sei and Komaki [25] and Uchida [26]).

The rest of the paper is organized as follows. In Section 2, we describe our working model, notations and assumptions. We also discuss the asymptotic properties of the quasi-maximum likelihood estimator. Section 3 presents the stochastic expansion of the logarithmic marginal quasi-likelihood in possibly misspecified GLM for dependent data and the consistency of the model selection with respect to the optimal model. In Section 4, we illustrate the performance of model selection criterion in both correctly specified and misspecified models. Section 5 presents the proofs of our results.

2. Quasi-maximum likelihood estimation of dependent GLM

Let $\mathbf{Y}_n = (Y_1,\dots,Y_n)'$ be the n-dimensional random vector and $\mathbf{X}_n = (X_1,\dots,X_n)'$ be the $n\times p$ random time series. We write $X_j = (X_{j,1},\dots,X_{j,p})'$ for any j. We assume that the unknown true distribution of $(\mathbf{X}_n,\mathbf{Y}_n)$ has the density g_n with respect to some dominating σ-finite measure:

$$g_n(\mathbf{x}_n,\mathbf{y}_n) = g_n(\mathbf{x}_n)g_n(\mathbf{y}_n|\mathbf{x}_n),$$

where $\mathbf{x}_n = (x_1,\dots,x_n)'$, $x_j = (x_{j,1},\dots,x_{j,p})'$ and $\mathbf{y}_n = (y_1,\dots,y_n)'$.

2.1. **Model setup.** We assume possibly misspecified M candidate models to estimate the true model G_n. Each candidate model has density function

$$f_{m,n}(\mathbf{x}_n,\mathbf{y}_n;\theta)=f_n(\mathbf{x}_n)f_{m,n}(\mathbf{y}_n|\mathbf{x}_n;\theta)=f_n(\mathbf{x}_n)\prod_{j=1}^{n}f_{m,n,j}(y_j|x_j;\theta)$$

with $\theta=(\theta_1,\dots,\theta_{p_m})\in\Theta_m$, where the mth parameter space $\Theta_m\subset\mathbb{R}^{p_m}$ is a bounded convex domain and $p_m\le p$. This model assumes that $Y_1,\dots,Y_n$ are $(X_1,\dots,X_n)$-conditionally independent and that each $(X_1,\dots,X_n)$-conditional distribution of Y_j depends on only X_j. We also assume that the true unknown distribution of $\mathbf{X}_n$ does not depend on the parameter and the candidate model. Therefore, we consider only the true conditional distribution of $\mathbf{Y}_n$ given $\mathbf{X}_n$ and use GLM $\mathfrak{M}_m$ as our working model, with respect to some dominating measure:

$$f_{m,n}(\mathbf{y}_n|\mathbf{x}_n;\theta)=\prod_{j=1}^{n}f_{m,n,j}(y_j|x_j;\theta)=\prod_{j=1}^{n}\exp\big(y_jx_j'\theta-b_m(x_j'\theta)+c_m(y_j)\big),\tag{3}$$

where, for brevity, we write $x_j'\theta=\sum_{i=1}^{p_m}x_{j,d_i(m)}\theta_i$ with $\{d_1(m),\dots,d_{p_m}(m)\}\subset\{1,\dots,p\}$ for any m, $b_m(\cdot)$ and $c_m(\cdot)$ are determined by each assumed conditional distribution of $\mathbf{Y}_n$ given $\mathbf{X}_n$ and $b_m(\cdot)$ is a sufficiently smooth convex function defined on $\mathbb{R}$; for example, $b_m(\theta)=\theta^2/2$ (Gaussian regression) and $b_m(\theta)=\log(1+e^\theta)$ (Logistic regression). Moreover, we assume that $b_1(\theta)=\cdots=b_M(\theta)$ and $c_1(y)=\cdots=c_M(y)$. For any n-dimensional random vector $\mathbf{Z}_n$ whose conditional distribution given X_n is (3), the characteristic function is given by

$$\begin{aligned}&\int e^{i\mathbf{t}'\mathbf{z}_n}\prod_{j=1}^{n}\exp\big(z_jx_j'\theta-b_m(z_j'\theta)+c_m(z_j)\big)d\mathbf{z}_n\\&=\int\exp\big((i\mathbf{t}+\mathbf{x}_n\theta)'\mathbf{z}_n-b_m(\mathbf{x}_n\theta)+c_m(\mathbf{z}_n)\big)d\mathbf{z}_n\\&=\exp\big(b_m(i\mathbf{t}+\mathbf{x}_n\theta)-b_m(\mathbf{x}_n\theta)\big)\int\exp\big((i\mathbf{t}+\mathbf{x}_n\theta)'\mathbf{z}_n-b_m(i\mathbf{t}+\mathbf{x}_n\theta)+c_m(\mathbf{z}_n)\big)d\mathbf{z}_n\\&=\exp\big(b_m(i\mathbf{t}+\mathbf{x}_n\theta)-b_m(\mathbf{x}_n\theta)\big),\end{aligned}$$

where $\mathbf{t}$ is an n-dimensional vector. Because of this characteristic function, we have in the correctly specified case

$$E_\theta[Z_j|X_j]=\partial b_m(X_j'\theta),$$
$$V_\theta[Z_j|X_j]=\partial^2 b_m(X_j'\theta),$$

where $\partial b_m(x)=\frac{\partial}{\partial\theta}b_m(\theta)\big|_{\theta=x}$.

Since any candidate model $\mathfrak{M}_m$ is possibly misspecified and $c_m(\cdot)$ of (3) is independent of θ, we may and do define the logarithmic marginal quasi-likelihood function $\mathbb{H}_{m,n}$ by

$$\mathbb{H}_{m,n}(\theta)=\sum_{j=1}^{n}\big(Y_jX_j'\theta-b_m(X_j'\theta)\big).\tag{4}$$

Any random mapping $\hat{\theta}_{m,n}$ such that

$$\hat{\theta}_{m,n}\in\underset{\theta\in\Theta}{\operatorname{argmax}}\,\mathbb{H}_{m,n}(\theta)$$

is called the quasi-maximum likelihood estimator (QMLE) associated with $\mathbb{H}_{m,n}$. Clearly, when b is differentiable, $\hat{\theta}_{m,n}$ is the solution to the quasi-score function

$$\partial_\theta\mathbb{H}_{m,n}(\theta)=\sum_{j=1}^{n}\big(Y_j-\partial b_m(X_j'\theta)\big)X_j=0,$$

where $\partial_\theta=\partial/\partial\theta$.

For notational brevity, from now on we will omit the model index "m" from the notation.

2.2. **Asymptotic behavior of the QMLE.** Fahrmeir and Kaufmann [12] studied the consistency and asymptotic normality of the maximum likelihood estimator in correctly specified GLMs. In this section, we will show that the asymptotic properties of the QMLE in misspecified GLMs with the dependent observations.

Let $\mathcal{F}_j = \sigma(Y_i, X_i; i \le j)$ denote the σ-field representing the data information at stage j. If a and b satisfy $a \le Cb$ for some constant $C > 0$, we write $a \lesssim b$. We introduce the following conditions:

Assumption 2.1. *For some constant $C \ge 0$ and $C' \ge 0$, (i)* $\max_{i\in\{1,2,3\}} |\partial^i b(x)| \lesssim 1 + |x|^C$,

(ii) $E\big[|Y_j|^3 | \mathcal{F}_{j-1} \vee \sigma(X_j)\big] \lesssim 1 + |X_j|^{C'}$ *a.s. for any* $j \in \mathbb{N}$,

(iii) $\sup_{j\in\mathbb{N}} E\big[|X_j|^{3C+C'+3}\big] < \infty$.

Assumption 2.2. *There exists a measurable function $F : \mathbb{R}^p \to \mathbb{R}$ such that $E[Y_j|\mathcal{F}_{j-1} \vee \sigma(X_j)] = F(X_j)$ for every $j \in \mathbb{N}$.*

Assumption 2.3. *Denote $\zeta_j = (X_j, Y_j)$ for any j. For some $c > 0$,*

$$\alpha(k) \le c^{-1} e^{-ck}$$

for all $k \in \mathbb{N}$, where

$$\alpha(k) := \sup_{j\in\mathbb{N}} \sup_{\substack{A\in\sigma(\zeta_i; i\le j)\\ B\in\sigma(\zeta_i; i\ge j+k)}} |P[A \cap B] - P[A]P[B]|.$$

When Assumption 2.3 holds, $\{\zeta_j; j = 1, 2, \dots\}$ is called exponential α-mixing. In particular, Assumption 2.3 implies that $\psi_j := \big(Y_j - F(X_j)\big)X_j$, $j \in \mathbb{N}$, is exponential α-mixing.

Assumption 2.4. *There exists a non-degenerate probability measure ν such that the following holds.*

(i) $\displaystyle\sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{j=1}^n \big(F(X_j)X_j'\theta - b(X_j'\theta)\big) - \int \big(F(x)x'\theta - b(x'\theta)\big)\nu(dx) \right| \xrightarrow{P} 0.$

(ii) $\displaystyle\sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{j=1}^n \partial^2 b(X_j'\theta)X_jX_j' - \int \partial^2 b(x'\theta)xx'\nu(dx) \right| \xrightarrow{P} 0.$

Remark 2.5. *The β-mixing coefficients of $\{\zeta_j\}$ are defined by*

$$\beta(k) := \sup_{j\in\mathbb{N}} E\left[\sup_{B\in\sigma(\zeta_i; i\ge j+k)} \big|P\big(B|\sigma(\zeta_i; i \le j)\big) - P(B)\big| \right].$$

If $\beta(k) = O(e^{-ak})$ for some $a > 0$ and for all $k \in \mathbb{N}$, then $\{\zeta_j\}$ is called exponential β-mixing (e.g. Davydov [10] and Liebscher [19]). The exponential β-mixing property implies the exponential α-mixing property. When we replace Assumption 2.3 by the condition that $\{\zeta_j\}$ is exponential β-mixing under some appropriate moment condition, the following conditions follow on applying an obvious discrete-time counterpart of Masuda [21, Lemma 4.3]*: (i) For some constant $\beta_1 > 0$ and $q_1 > 0$,*

$$\sup_{n>0} E\left[\left(n^{\beta_1} \sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{j=1}^n \big(F(X_j)X_j'\theta - b(X_j'\theta)\big) - \int \big(F(x)x'\theta - b(x'\theta)\big)\nu(dx) \right| \right)^{q_1} \right] < \infty.$$

(ii) For some constant $\beta_2 > 0$ and $q_2 > 0$,

$$\sup_{n>0} E\left[\left(n^{\beta_2} \sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{j=1}^n \partial^2 b(X_j'\theta)X_jX_j' - \int \partial^2 b(x'\theta)xx'\nu(dx) \right| \right)^{q_2} \right] < \infty.$$

Then, if q_1 and q_2 can be taken large enough, we may deduce almost surely

$$\sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{j=1}^n \big(F(X_j)X_j'\theta - b(X_j'\theta)\big) - \int \big(F(x)x'\theta - b(x'\theta)\big)\nu(dx) \right| \to 0,$$

$$\sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{j=1}^n \partial^2 b(X_j'\theta)X_jX_j' - \int \partial^2 b(x'\theta)xx'\nu(dx) \right| \to 0.$$

Assumption 2.4 (i) gives that $\frac{1}{n}\sum_{j=1}^n \big(F(X_j)X_j'\theta - b(X_j'\theta)\big) = \int \big(F(x)x'\theta - b(x'\theta)\big)\nu(dx) + o_p(1)$ for each θ. Under Assumptions 2.1-2.4, we have

$$\begin{aligned}\frac{1}{n}\mathbb{H}_n(\theta) &= \frac{1}{n}\sum_{j=1}^n \psi_j'\theta + \frac{1}{n}\sum_{j=1}^n \big\{\big(F(X_j)X_j'\theta - b(X_j'\theta)\big)\big\}\\ &= O_p\Big(\frac{1}{\sqrt{n}}\Big) + \frac{1}{n}\sum_{j=1}^n \big\{\big(F(X_j)X_j'\theta - b(X_j'\theta)\big)\big\} \xrightarrow{P} \int \big(F(x)x'\theta - b(x'\theta)\big)\nu(dx) =: \mathbb{H}_0(\theta), \quad (5)\end{aligned}$$

where the notation $\xrightarrow{P}$ means the convergence in probability. Here, the proof of the tightness of $\{\frac{1}{\sqrt{n}}\sum_{j=1}^n \psi_j\}$ is given in the proof of Lemma 5.2 (i).

Since $b(\cdot)$ is a convex function, $-\partial_\theta^2\mathbb{H}_0(\theta) = \int \partial^2 b(x'\theta)xx'\nu(dx)$ is a positive definite for any $\theta \in \Theta$. Thus, the equation

$$\partial_\theta \mathbb{H}_0(\theta) = \int \big(F(x) - \partial b(x'\theta)\big)x\nu(dx) = 0$$

admits a unique solution. Then we may define the *optimal* parameter θ_0 as the unique maximizer of $\mathbb{H}_0(\theta)$:

$$\{\theta_0\} = \operatorname*{argmax}_{\theta\in\Theta} \mathbb{H}_0(\theta)$$

Assumption 2.4 (ii) gives that $\frac{1}{n}\sum_{j=1}^n \partial^2 b(X_j'\theta)X_jX_j' = \int \partial^2 b(x'\theta)xx'\nu(dx) + o_p(1)$ for each θ. In particular, the quasi-observed information is given by $\Gamma_n := -\frac{1}{n}\partial_\theta^2\mathbb{H}_n(\theta_0) = \frac{1}{n}\sum_{j=1}^n \partial^2 b(X_j'\theta)X_jX_j'$, so that Γ_n satisfies the equation

$$\Gamma_n = \Gamma_0 + o_p(1),$$

where $\Gamma_0 := \int \partial^2 b(x'\theta_0)xx'\nu(dx)$.

Theorem 2.6. *Under Assumptions 2.1-2.4, the QMLE satisfies*

$$\hat\theta_n \xrightarrow{P} \theta_0$$

as $n \to \infty$.

Assumption 2.7. *(i) $\{X_j; j = 1, 2, \ldots\}$ is strictly stationary.*

(ii) For some $\Sigma_0 > 0$, $\displaystyle \frac{1}{n}E\Bigg[\Bigg\{\sum_{j=1}^n \big(Y_j - \partial b(X_j'\theta_0)\big)X_j\Bigg\}\Bigg\{\sum_{j=1}^n \big(Y_j - \partial b(X_j'\theta_0)\big)X_j\Bigg\}'\Bigg] \to \Sigma_0.$

Theorem 2.8. *Under Assumptions 2.1-2.4 and 2.7, the asymptotic distribution of the QMLE is normal:*

$$\sqrt{n}(\hat\theta_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Gamma_0^{-1}\Sigma_0\Gamma_0^{-1}).$$

When the candidate model is correctly specified, then $F(x) = \partial b(x'\theta_0)$, and $\Sigma_0 = \Gamma_0$, i.e. $\sqrt{n}(\hat\theta_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Gamma_0^{-1})$.

3. Quasi-Bayesian information criterion for dependent GLM

3.1. **Stochastic expansion.** We use the GLM as our working model to choose the optimal model, so we consider the stochastic expansion of the marginal quasi-likelihood in GLM.

Assumption 3.1. $\displaystyle \frac{1}{\sqrt{n}}\sum_{j=1}^n \big(F(X_j) - \partial b(X_j'\theta_0)\big)X_j = O_p(1).$

Assumption 3.2. *There exists a function $\underline{b} : \mathbb{R}^p \to (0,\infty)$, (i) for any x, $\inf_{\theta\in\Theta} \partial^2 b(x'\theta) \geq \underline{b}(x)$,*

(ii) for some constant $\lambda_0 > 0$, $\displaystyle \limsup_{n\to\infty} P\Bigg[\lambda_{\min}\Bigg(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\Bigg) < \lambda_0\Bigg] = 0$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalues of a given matrix.

The next theorem shows the asymptotic behavior of the log marginal quasi-likelihood function.

Theorem 3.3. *Assume that Assumptions 2.1-2.4, 3.1 and 3.2 hold and that the following conditions are satisfied:*

(i) $\pi(\theta_0) > 0,\ \sup_{\theta\in\Theta}\pi(\theta) < \infty$.

(ii) *For every* $M > 0$, $\sup_{|u|<M}\left|\pi\left(\theta_0 + \frac{u}{\sqrt{n}}\right) - \pi(\theta_0)\right| \to 0$ *as* $n \to \infty$.

(iii) $\log\pi(\hat{\theta}_n) - \log\pi(\theta_0) = o_p(1)$.

Then we have the expansion

$$\begin{aligned}
&\log\left(\int_\Theta \exp\{\mathbb{H}_n(\theta)\}\pi(\theta)d\theta\right)\\
&= \sum_{j=1}^n \left(Y_jX_j'\hat{\theta}_n - b(X_j'\hat{\theta}_n)\right) - \frac{p}{2}\log n + \frac{p}{2}\log 2\pi - \frac{1}{2}\log\det\left(\frac{1}{n}\sum_{j=1}^n \partial^2 b(X_j'\hat{\theta}_n)X_jX_j'\right) + \log\pi(\hat{\theta}_n) + o_p(1)\\
&= \sum_{j=1}^n \left(Y_jX_j'\hat{\theta}_n - b(X_j'\hat{\theta}_n)\right) + \frac{p}{2}\log 2\pi - \frac{1}{2}\log\det\left(\sum_{j=1}^n \partial^2 b(X_j'\hat{\theta}_n)X_jX_j'\right) + \log\pi(\hat{\theta}_n) + o_p(1).
\end{aligned}$$

Remark 3.4. *Suppose that we replace Assumptions 2.4 and 3.2 (ii) by the following conditions:*

(i) $\frac{1}{n}\sum_{j=1}^n \left(F(X_j)X_j'\theta - b(X_j'\theta)\right) \to \int \left(F(x)x'\theta - b(x'\theta)\right)\nu(dx)$ *almost surely as* $n\to\infty$, *uniformly in* $\theta\in\Theta$.

(ii) $\frac{1}{n}\sum_{j=1}^n \partial^2 b(X_j'\theta)X_jX_j' \to \int \partial^2 b(x'\theta)xx'\nu(dx)$ *almost surely as* $n\to\infty$, *uniformly in* $\theta\in\Theta$.

(iii) *For some constant* $\lambda_0 > 0$, $P\left[\limsup_{n\to\infty}\lambda_{\min}\left(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\right) < \lambda_0\right] = 0$.

Then, as in Cavanaugh and Neath [8, Section 3], we can show that the log marginal quasi-likelihood function almost surely satisfies the expansion similar to Theorem 3.3, i.e. almost surely

$$\begin{aligned}
\log\left(\int_\Theta \exp\{\mathbb{H}_n(\theta)\}\pi(\theta)d\theta\right) &= \sum_{j=1}^n \left(Y_jX_j'\hat{\theta}_n - b(X_j'\hat{\theta}_n)\right) - \frac{p}{2}\log n + \frac{p}{2}\log 2\pi\\
&\quad - \frac{1}{2}\log\det\left(\frac{1}{n}\sum_{j=1}^n \partial^2 b(X_j'\hat{\theta}_n)X_jX_j'\right) + \log\pi(\hat{\theta}_n) + o(1).
\end{aligned}$$

Due to Theorem 3.3, we define the quasi-Bayesian information criterion (QBIC) and BIC for dependent GLM by

$$\begin{aligned}
\mathrm{QBIC} &= -2\sum_{j=1}^n \left(Y_jX_j'\hat{\theta}_n - b(X_j'\hat{\theta}_n)\right) + \log\det\left(\sum_{j=1}^n \partial^2 b(X_j'\hat{\theta}_n)X_jX_j'\right),\\
\mathrm{BIC} &= -2\sum_{j=1}^n \left(Y_jX_j'\hat{\theta}_n - b(X_j'\hat{\theta}_n)\right) + p\log n.
\end{aligned}$$

Let $\mathrm{QBIC}^{(1)},\dots,\mathrm{QBIC}^{(M)}$ be the QBIC for each candidate model. We calculate $\mathrm{QBIC}^{(1)},\dots,\mathrm{QBIC}^{(M)}$ and select the best model $\mathfrak{M}_{m_0}$ having the minimum-QBIC value:

$$m_0 = \underset{m\in\{1,\dots,M\}}{\mathrm{argmin}}\ \mathrm{QBIC}^{(m)}.$$

We can also select the best model by using BIC in a similar manner. As directly seen by the definition, the QBIC have more computational load than the BIC. Since the QBIC involves the observed-information matrix quantity, which is directly computed from data, the QBIC would more effectively take data dependence into account.

3.2. **Model selection consistency.** Let $\Theta_i \subset \mathbb{R}^{p_i}$ and $\Theta_j \subset \mathbb{R}^{p_j}$ be the parameter space associated with $\mathfrak{M}_i$ and $\mathfrak{M}_j$, respectively. If $p_i < p_j$ and there exist a matrix $A \in \mathbb{R}^{p_j \times p_i}$ with $A'A = I_{p_i \times p_i}$ as well as a $c \in \mathbb{R}^{p_j}$ such that $\mathbb{H}_{i,n}(\theta) = \mathbb{H}_{j,n}(A\theta + c)$ for all $\theta \in \Theta_i$, we say that Θ_i is nested in Θ_j.

Under Assumptions 2.1-2.4, when m_0 satisfies

$$\{m_0\} = \underset{m \in \{1,\dots,M\}}{\operatorname{argmax}}\ \mathbb{H}_{m,0}(\theta_{m,0}) = \underset{m \in \{1,\dots,M\}}{\operatorname{argmax}} \int \big(F(x)x'\theta_{m,0} - b_m(x'\theta_{m,0})\big)\nu(dx),$$

we say that $\mathfrak{M}_{m_0}$ is the *optimal* model.

Theorem 3.5. *Assume that Assumptions 2.1-2.4, 3.1 and 3.2 are satisfied and that there exists a unique $m_0 \in \{1,\dots,M\}$ such that $\mathfrak{M}_{m_0}$ is the optimal model. For any fixed $m \in \{1,\dots,M\}\backslash\{m_0\}$, if Θ_{m_0} is nested in Θ_m, or $\mathbb{H}_{m,0}(\theta) \neq \mathbb{H}_{m_0,0}(\theta_{m_0,0})$ for any $\theta \in \Theta_m$, then*

$$\lim_{n\to\infty} P\big[\mathrm{QBIC}^{(m_0)} - \mathrm{QBIC}^{(m)} < 0\big] = 1.$$

This theorem implies that BIC also has the (weak) consistency for the model selection.

4. Examples and Simulation results

In this section, we conduct simulations to evaluate finite sample performance of the model selection by using QBIC, BIC and formal AIC (fAIC). Since we do not deal with the theoretical part of AIC in this paper, we use the word fAIC as AIC, i.e. fAIC of mth model is defined by

$$\mathrm{fAIC}^{(m)} = -2\mathbb{H}_{m,n}(\hat{\theta}_{m,n}) + 2p_m.$$

Let θ^* be the true value. We here set the initial value as the value generated from uniform distribution $U(\theta^* - 1,\ \theta^* + 1)$ to use `optim` at software R for numerical optimization.

4.1. **Model selection in correctly specified model.** We assume that the explanatory variables $X_{j,1}, \dots, X_{j,4}$ are given by

$$\begin{aligned}
&X_{j,1} = 1\ (j \geq 1),\\
&X_{1,2} = 1,\ X_{j,2} = 0.5X_{j-1,2} + \epsilon_{j,2},\ (j \geq 2),\\
&X_{1,3} = 0,\ X_{j,3} = -0.7X_{j-1,3} + \epsilon_{j,3},\ (j \geq 2),\\
&X_{1,4} = -1,\ X_{j,3} = 0.8X_{j-1,4} + \epsilon_{j,4},\ (j \geq 2),
\end{aligned}$$

where the error vector $(\epsilon_2, \epsilon_3, \epsilon_4) \sim N(0, \Sigma)$ with $\Sigma = (0.5^{|k-\ell|})_{k,\ell=1,2,3}$. Moreover, the response variable Y_j is obtained from the true model defined by the linear logistic regression model

$$Y_j \sim B\left(1, \frac{\exp(X_j'\theta^*)}{1 + \exp(X_j'\theta^*)}\right), \tag{6}$$

where the true value $\theta^* = (0, -3, 0, 1)$. We consider this model for the following combination of X_j:

Model 1 : $X_j = (X_{j,1}, X_{j,2}, X_{j,3}, X_{j,4})$; **Model 2** : $X_j = (X_{j,1}, X_{j,2}, X_{j,3})$;
Model 3 : $X_j = (X_{j,1}, X_{j,2}, X_{j,4})$; **Model 4** : $X_j = (X_{j,1}, X_{j,3}, X_{j,4})$; **Model 5** : $X_j = (X_{j,2}, X_{j,3}, X_{j,4})$;
Model 6 : $X_j = (X_{j,1}, X_{j,2})$; **Model 7** : $X_j = (X_{j,1}, X_{j,3})$; **Model 8** : $X_j = (X_{j,1}, X_{j,4})$;
Model 9 : $X_j = (X_{j,2}, X_{j,3})$; **Model 10** : $X_j = (X_{j,2}, X_{j,4})$; **Model 11** : $X_j = (X_{j,3}, X_{j,4})$;
Model 12 : $X_j = X_{j,1}$; **Model 13** : $X_j = X_{j,2}$; **Model 14** : $X_j = X_{j,3}$; **Model 15** : $X_j = X_{j,4}$.

Then the true model is Model 10, and Models 1, 3, 5 contain the true model.

In the present situation, the function b defined in (3) is given by $b(\theta) = \log(1 + e^{\theta})$. We simulate the number of the model selected by using QBIC, BIC and fAIC among the candidate Models 1-15 over 10000 simulations. For example, in the case of Model 1, QBIC, BIC and fAIC given by

$$\mathrm{QBIC} = -2\sum_{j=1}^{n}\left\{Y_j\sum_{i=1}^{4}X_{j,i}\hat{\theta}_i - \log\left(1 + \exp\left(\sum_{i=1}^{4}X_{j,i}\hat{\theta}_i\right)\right)\right\} + \log\det\left(\sum_{j=1}^{n}\frac{\exp\left(\sum_{i=1}^{4}X_{j,i}\hat{\theta}_i\right)X_jX_j'}{\left(1 + \exp\left(\sum_{i=1}^{4}X_{j,i}\hat{\theta}_i\right)\right)^2}\right)$$

TABLE 1. The number of models selected by QBIC, BIC and fAIC in Section 4.1 over 10000 simulations for various n (1-15 represent the models, and the true model is Model 10)

Criteria	$n=50$														
	1	2	3	4	5	6	7	8	9	10*	11	12	13	14	15
QBIC	1489	65	2084	0	1260	201	0	0	56	4666	0	0	172	0	0
BIC	99	18	531	0	562	222	0	0	86	7720	0	0	762	0	0
fAIC	479	55	1310	0	1424	192	0	0	93	6242	0	0	205	0	0
Criteria	$n=100$														
	1	2	3	4	5	6	7	8	9	10*	11	12	13	14	15
QBIC	298	2	1483	0	989	10	0	0	4	7206	0	0	8	0	0
BIC	19	1	323	0	397	15	0	0	8	9179	0	0	58	0	0
fAIC	347	1	1380	0	1367	7	0	0	2	6895	0	0	1	0	0
Criteria	$n=200$														
	1	2	3	4	5	6	7	8	9	10*	11	12	13	14	15
QBIC	86	0	910	0	616	0	0	0	0	8388	0	0	0	0	0
BIC	5	0	235	0	222	0	0	0	0	9538	0	0	0	0	0
fAIC	281	0	1314	0	1414	0	0	0	0	6991	0	0	0	0	0

$$\mathrm{BIC} = -2\sum_{j=1}^{n}\left\{Y_j\sum_{i=1}^{4}X_{j,i}\hat{\theta}_i - \log\left(1+\exp\left(\sum_{i=1}^{4}X_{j,i}\hat{\theta}_i\right)\right)\right\} + 4\log n,$$

$$\mathrm{fAIC} = -2\sum_{j=1}^{n}\left\{Y_j\sum_{i=1}^{4}X_{j,i}\hat{\theta}_i - \log\left(1+\exp\left(\sum_{i=1}^{4}X_{j,i}\hat{\theta}_i\right)\right)\right\} + 4\times 2.$$

Table 1 summarizes the comparison results of the frequency of the model selection. Model 10 is selected with high frequency for all criterions and n. Moreover, the probability that Model 10 is selected by QBIC and BIC becomes higher as n becomes larger. In Table 2, the differences between the true value and the estimators in specified models are getting small when n gets increased. From these results, we can observe the consistency of the estimators and the model selection consistency of QBIC and BIC.

4.2. **Model selection in misspecified model.** We use the same conditions as in the previous section except for the true model. In this simulation, the response variable Y_j is obtained from the true model defined by

$$Y_j \sim B\Big(1, \Phi(X_j'\theta^*)\Big),$$

where $\Phi(x) = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\exp(-\frac{t^2}{2})dt$. Then Models 1-15 are misspecified models.

From Table 3, we obtain similar results even though the candidate models do not include the true model. Table 4 summarizes the mean and the standard deviation of estimators in each model. Since the optimal parameter value is not given here, we can not see the differences between the optimal parameter value and the estimators, although the standard deviations become smaller when n become larger.

4.3. **Model selection in univariate time series model.** Let $X_j = (Z_j, Z_{j-1}, \dots, Z_{j-(p-1)})'$ be the explanatory vector for any $j \in \{1, \dots, n\}$, where for every $i \in \{2, \dots, n\}$, Z_i is given by

$$Z_{-n+2} = \cdots = Z_0 = 0,\ Z_1 = 1,\ Z_i = 0.6Z_{i-1} + \epsilon_i,$$

where $\epsilon_i \sim N(0,1)$. The response variable Y_j is obtained from the true model defined by

$$Y_j \sim B\left(1, \frac{\exp(X_j'\theta^*)}{1+\exp(X_j'\theta^*)}\right), \tag{7}$$

where the true value $\theta^* = (3, -1, 2, 1)$. For simplicity, we here focus on the hierarchical models as the candidate models:

$$\textbf{Model 1}: X_j = (Z_j);\ \textbf{Model 2}: X_j = (Z_j, Z_{j-1});\ \textbf{Model 3}: X_j = (Z_j, Z_{j-1}, Z_{j-2});$$
$$\textbf{Model 4}: X_j = (Z_j, Z_{j-1}, Z_{j-2}, Z_{j-3});\ \textbf{Model 5}: X_j = (Z_j, Z_{j-1}, Z_{j-2}, Z_{j-3}, Z_{j-4});\ \cdots.$$

Then the true mode is Model 4.

TABLE 2. The mean and the standard deviation (s.d.) of estimator $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ and $\hat{\theta}_4$ in each model for various n (1-15 represent the models, and the true parameter $(\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) = (0, -3, 0, 1)$)

		$n = 50$				$n = 100$				$n = 200$			
		$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
1	mean	-0.0793	-8.3409	-0.0219	2.7990	0.0004	-3.3727	-0.0057	1.1266	0.0023	-3.1642	0.0004	1.0542
	s.d.	8.1943	35.1378	5.9571	12.8200	0.3895	0.8918	0.2889	0.3744	0.2425	0.4946	0.1807	0.2123
2	mean	-0.0505	-2.7481	0.1061	–	-0.0425	-2.1878	0.0867	–	-0.0167	-2.0734	0.0856	–
	s.d.	2.0544	7.9339	1.3571	–	0.4653	0.5332	0.2141	–	0.3146	0.3336	0.1395	–
3	mean	0.0355	-5.7372	–	1.8913	0.0001	-3.2941	–	1.0993	0.0021	-3.1332	–	1.0441
	s.d.	4.6176	23.3197	–	8.1816	0.3763	0.8210	–	0.3508	0.2397	0.4804	–	0.2078
4	mean	-0.0999	–	-0.2581	0.3139	-0.0451	–	-0.2364	0.2852	-0.0168	–	-0.2250	0.2746
	s.d.	0.4791	–	0.2318	0.3193	0.3109	–	0.1453	0.1940	0.2100	–	0.0976	0.1264
5	mean	–	-5.6381	0.0219	1.8650	–	-3.2928	-0.0052	1.0989	–	-3.1336	0.0003	1.0442
	s.d.	–	22.5995	3.7461	7.8695	–	0.8333	0.2792	0.3465	–	0.4835	0.1787	0.2057
6	mean	-0.0635	-2.4621	–	–	-0.0429	-2.1293	–	–	-0.0169	-2.0324	–	–
	s.d.	1.8364	4.1725	–	–	0.4578	0.5074	–	–	0.3127	0.3232	–	–
7	mean	-0.1086	–	-0.1960	–	-0.0518	–	-0.1808	–	-0.0217	–	-0.1723	–
	s.d.	0.4591	–	0.2093	–	0.3188	–	0.1351	–	0.2227	–	0.0921	–
8	mean	-0.1006	–	–	0.2681	-0.0453	–	–	0.2483	-0.0170	–	–	0.2415
	s.d.	0.4660	–	–	0.3022	0.3063	–	–	0.1875	0.2081	–	–	0.1230
9	mean	–	-2.3773	0.1058	–	–	-2.1041	0.0878	–	–	-2.0342	0.0860	–
	s.d.	–	5.2112	1.0538	–	–	0.5008	0.2061	–	–	0.3263	0.1372	–
10*	mean	–	-4.2068	–	1.3952	–	-3.2211	–	1.0741	–	-3.1037	–	1.0344
	s.d.	–	13.1535	–	4.3787	–	0.7702	–	0.3259	–	0.4699	–	0.2013
11	mean	–	–	-0.2546	0.3124	–	–	-0.2350	0.2855	–	–	-0.2243	0.2747
	s.d.	–	–	0.2218	0.2851	–	–	0.1424	0.1840	–	–	0.0967	0.1230
12	mean	-0.0695	–	–	–	-0.0368	–	–	–	-0.0179	–	–	–
	s.d.	0.3186	–	–	–	0.2548	–	–	–	0.1936	–	–	–
13	mean	–	-2.6475	–	–	–	-2.5688	–	–	–	-2.5278	–	–
	s.d.	–	0.3694	–	–	–	0.3052	–	–	–	0.2789	–	–
14	mean	–	–	-0.1525	–	–	–	-0.1528	–	–	–	-0.1517	–
	s.d.	–	–	0.1694	–	–	–	0.1218	–	–	–	0.0901	–
15	mean	–	–	–	0.5703	–	–	–	0.5507	–	–	–	0.5394
	s.d.	–	–	–	0.2553	–	–	–	0.2478	–	–	–	0.2477

TABLE 3. The number of models selected by QBIC, BIC and fAIC in Section 4.2 over 10000 simulations for various n (1-15 represent the models)

Criteria	$n = 100$														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
QBIC	965	0	2025	0	1321	0	0	0	0	5689	0	0	0	0	0
BIC	41	0	435	0	398	0	0	0	0	9125	0	0	1	0	0
fAIC	443	0	1452	0	1538	0	0	0	0	6567	0	0	0	0	0

Criteria	$n = 200$														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
QBIC	223	0	1338	0	915	0	0	0	0	7524	0	0	0	0	0
BIC	9	0	278	0	274	0	0	0	0	9439	0	0	0	0	0
fAIC	349	0	1436	0	1414	0	0	0	0	6801	0	0	0	0	0

Criteria	$n = 300$														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
QBIC	108	0	1009	0	694	0	0	0	0	8189	0	0	0	0	0
BIC	5	0	190	0	216	0	0	0	0	9589	0	0	0	0	0
fAIC	295	0	1352	0	1388	0	0	0	0	6965	0	0	0	0	0

We simulate the number of the model selected by using QBIC, BIC and fAIC among the candidate models over 10000 simulations. First, we calculate $\mathrm{QBIC}^{(1)}$ and $\mathrm{QBIC}^{(2)}$. If $\mathrm{QBIC}^{(1)} < \mathrm{QBIC}^{(2)}$, Model 1 is selected as the best model. When $\mathrm{QBIC}^{(1)} \geq \mathrm{QBIC}^{(2)}$, we calculate $\mathrm{QBIC}^{(3)}$ and compare $\mathrm{QBIC}^{(2)}$ with $\mathrm{QBIC}^{(3)}$. We repeat similar procedures to stop at the best model. Furthermore, we select the best model by BIC and fAIC in a similar manner.

Table 5 summarizes the comparison results of the frequency of the model selection. The best model is searched among Models 1-11 for all cases. Model 4 is selected with the highest frequency as the best model. Moreover, the frequency that Model 4 is selected by QBIC and BIC is getting higher when n gets increased. From this result, we can observe that QBIC and BIC have the consistency for model selection.

TABLE 4. The mean and the standard deviation (s.d.) of estimator $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ and $\hat{\theta}_4$ in each model for various n (1-15 represent the models)

		$n=100$				$n=200$				$n=300$			
		$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
1	mean	0.0020	7.6472	-0.0028	-2.5525	-0.0031	5.8988	-0.0009	-1.9640	0.0038	5.6861	-0.0006	-1.8975
	s.d.	1.7319	14.4366	1.2513	5.0473	0.3440	1.1541	0.2541	0.4378	0.2599	0.8285	0.1914	0.3118
2	mean	0.0596	2.8857	-0.1171	–	0.0255	2.6805	-0.1079	–	0.0176	2.6080	-0.1066	–
	s.d.	0.5917	0.7518	0.2335	–	0.3873	0.4610	0.1501	–	0.3069	0.3574	0.1168	–
3	mean	0.0104	6.7840	–	-2.2605	-0.0030	5.7925	–	-1.9290	0.0036	5.6262	–	-1.8772
	s.d.	1.0092	9.8524	–	3.4170	0.3351	1.0856	–	0.4150	0.2561	0.8056	–	0.3038
4	mean	0.0525	–	0.2555	-0.3135	0.0218	–	0.2448	-0.2959	0.0157	–	0.2415	-0.2931
	s.d.	0.3315	–	0.1409	0.2024	0.2209	–	0.0938	0.1314	0.1781	–	0.0756	0.1048
5	mean	–	6.7809	0.0061	-2.2594	–	5.7915	-0.0008	-1.9285	–	5.6255	-0.0007	-1.8771
	s.d.	–	9.5017	0.8186	3.3869	–	1.0916	0.2479	0.4130	–	0.8008	0.1886	0.3009
6	mean	0.0594	2.7967	–	–	0.0256	2.6226	–	–	0.0177	2.5586	–	–
	s.d.	0.5801	0.7108	–	–	0.3839	0.4437	–	–	0.3053	0.3487	–	–
7	mean	0.0619	–	0.1933	–	0.0276	–	0.1866	–	0.0189	–	0.1838	–
	s.d.	0.3385	–	0.1298	–	0.2324	–	0.0877	–	0.1902	–	0.0712	–
8	mean	0.0527	–	–	-0.2728	0.0220	–	–	-0.2591	0.0158	–	–	-0.2574
	s.d.	0.3260	–	–	0.1951	0.2182	–	–	0.1276	0.1761	–	–	0.1019
9	mean	–	2.7283	-0.1163	–	–	2.6104	-0.1073	–	–	2.5642	-0.1064	–
	s.d.	–	0.6923	0.2218	–	–	0.4446	0.1466	–	–	0.3506	0.1151	–
10	mean	–	6.2763	–	-2.0919	–	5.6925	–	-1.8959	–	5.5676	–	-1.8575
	s.d.	–	6.7882	–	2.4910	–	1.0309	–	0.3934	–	0.7797	–	0.2936
11	mean	–	–	0.2535	-0.3120	–	–	0.2438	-0.2954	–	–	0.2409	-0.2929
	s.d.	–	–	0.1375	0.1922	–	–	0.0927	0.1277	–	–	0.0751	0.1028
12	mean	0.0457	–	–	–	0.0218	–	–	–	0.0165	–	–	–
	s.d.	0.2639	–	–	–	0.1971	–	–	–	0.1677	–	–	–
13	mean	–	2.7922	–	–	–	2.7190	–	–	–	2.6821	–	–
	s.d.	–	0.3799	–	–	–	0.3080	–	–	–	0.2692	–	–
14	mean	–	–	0.1636	–	–	–	0.1630	–	–	–	0.1631	–
	s.d.	–	–	0.1182	–	–	–	0.0885	–	–	–	0.0756	–
15	mean	–	–	–	-0.5569	–	–	–	-0.5398	–	–	–	-0.5376
	s.d.	–	–	–	0.2432	–	–	–	0.2439	–	–	–	0.2466

TABLE 5. The number of models selected by QBIC, BIC and fAIC in Section 4.3 over 10000 simulations for various n (1-11 represent the models, and the true model is Model 4)

Criteria	$n=100$										
	1	2	3	4*	5	6	7	8	9	10	11
QBIC	2814	0	670	4739	1240	385	113	28	8	2	1
BIC	4144	0	1732	3934	176	14	0	0	0	0	0
fAIC	0	0	594	6091	2195	764	259	69	23	4	1
Criteria	$n=200$										
	1	2	3	4*	5	6	7	8	9	10	11
QBIC	1458	0	136	7278	962	149	16	1	0	0	0
BIC	2148	0	585	7089	175	3	0	0	0	0	0
fAIC	0	0	40	6599	2344	753	210	35	16	3	0
Criteria	$n=300$										
	1	2	3	4*	5	6	7	8	9	10	11
QBIC	812	0	14	8324	787	55	7	1	0	0	0
BIC	1267	0	116	8447	168	2	0	0	0	0	0
fAIC	0	0	0	6775	2261	744	177	32	10	1	0

In Table 6, the differences between the true value and the estimators in specified models (Models 4-6) become smaller as n becomes larger, and the standard deviations have the same tendency. Hence, the consistency of the estimators can be observed.

Remark 4.1. *If $\{Z_j; j = 1, 2, \ldots\}$ is a Markov chain of finite order, the situation of this section is included in the original model setting given in Section 2.*

Remark 4.2. *If we assume the time series structure of $\{Z_j\}$, such as the autoregressive structure, we can treat the choice of the time-lag p only by observation data of $\{Z_j\}$. However, we here use the GLM as our working model and solely focus on the contribution of $\{Z_j\}$ to $\mathbf{Y}_n$ through the conditional distribution, so that the Bayesian model selection is possible even if the distribution of $\{Z_j\}$ itself is not explicitly specified.*

TABLE 6. The mean and the standard deviation (s.d.) of estimator $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$, $\hat{\theta}_4$, $\hat{\theta}_5$ and $\hat{\theta}_6$ in each model for various n (1-6 represent the models, and the true parameter $(\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) = (3, -1, 2, 1)$)

		$n = 100$					
		$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$
1	mean	2.5215	–	–	–	–	–
	s.d.	0.2898	–	–	–	–	–
2	mean	1.6753	0.2916	–	–	–	–
	s.d.	0.4177	0.2991	–	–	–	–
3	mean	2.9516	-0.9883	2.5467	–	–	–
	s.d.	0.9033	0.5533	0.8261	–	–	–
4*	mean	3.7809	-1.2687	2.5303	1.2561	–	–
	s.d.	5.8555	2.2454	4.1243	2.2728	–	–
5	mean	4.2422	-1.4158	2.8469	1.4337	-0.0701	–
	s.d.	9.0993	3.4290	6.5145	3.7175	2.0928	–
6	mean	4.5331	-1.5110	3.0487	1.5316	-0.0649	-0.0209
	s.d.	10.9366	4.2403	8.0397	4.3018	2.3890	2.2203
		$n = 200$					
		$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$
1	mean	2.5039	–	–	–	–	–
	s.d.	0.2862	–	–	–	–	–
2	mean	1.6091	0.2974	–	–	–	–
	s.d.	0.2727	0.2027	–	–	–	–
3	mean	2.7420	-0.9105	2.3747	–	–	–
	s.d.	0.4869	0.3359	0.4473	–	–	–
4*	mean	3.2226	-1.0696	2.1499	1.0757	–	–
	s.d.	0.6336	0.3882	0.5101	0.3467	–	–
5	mean	3.2675	-1.0841	2.1792	1.0929	-0.0025	–
	s.d.	0.6578	0.3992	0.5268	0.3995	0.2995	–
6	mean	3.2940	-1.0896	2.1954	1.1048	-0.0039	-0.0007
	s.d.	0.6801	0.4120	0.5439	0.4087	0.3582	0.3068
		$n = 300$					
		$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$
1	mean	2.5039	–	–	–	–	–
	s.d.	0.2857	–	–	–	–	–
2	mean	1.5956	0.2923	–	–	–	–
	s.d.	0.2189	0.1613	–	–	–	–
3	mean	2.6921	-0.8964	2.3322	–	–	–
	s.d.	0.3762	0.2630	0.3416	–	–	–
4*	mean	3.1360	-1.0462	2.0945	1.0472	–	–
	s.d.	0.4719	0.2966	0.3812	0.2703	–	–
5	mean	3.1620	-1.0548	2.1121	1.0569	-0.0019	–
	s.d.	0.4800	0.3003	0.3874	0.3093	0.2319	–
6	mean	3.1808	-1.0590	2.1227	1.0684	-0.0070	0.0019
	s.d.	0.5004	0.3111	0.4054	0.3221	0.2790	0.2390

5. Proofs

We will make use of the following lemmas for the proofs of theorems. Recall that ψ_j is given by $\psi_j = \big(Y_j - F(X_j)\big)X_j$ for all $j \in \mathbb{N}$.

Lemma 5.1. *Assume that Assumption 2.3 is satisfied and that* $\sup_{j\in\mathbb{N}} \|\psi_j\|_2 < \infty$*, then*

$$\sup_{n>0} \frac{1}{n} E\left[\sup_{1\le i\le n}\left|\sum_{j=1}^{i}\psi_j\right|^2\right] < \infty.$$

Lemma 5.1 follows from a direct application of Yoshida [28, Lemma 4]. We write $\Delta_n = \frac{1}{\sqrt{n}}\partial_\theta \mathbb{H}_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\big(Y_j - \partial b(X_j'\theta_0)\big)X_j$

Lemma 5.2. *Assume that Assumptions 2.1-2.4 and 3.1 are satisfied, then the following claims are established:*

(i) $\Delta_n = O_p(1)$.

(ii) $\displaystyle\sup_{\theta\in\Theta}\left|\frac{1}{n\sqrt{n}}\partial_\theta^3\mathbb{H}_n(\theta)\right| = o_p(1)$.

Proof. (i)

$$\begin{aligned}
\sup_{j\in\mathbb{N}} E[|\psi_j|^2] &= \sup_{j\in\mathbb{N}} E\big[\big|\big(Y_j - F(X_j)\big)X_j\big|^2\big] \\
&\le \sup_{j\in\mathbb{N}} E\big[\big|Y_j - F(X_j)\big|^2\big|X_j\big|^2\big] \\
&\lesssim \sup_{j\in\mathbb{N}} E\big[\big(|Y_j|^2 + \big|F(X_j)\big|^2\big)\big|X_j\big|^2\big] \\
&\le \sup_{j\in\mathbb{N}} E\big[\big(|Y_j|^2 + E[|Y_j|^2|\mathcal{F}_{j-1}\vee\sigma(X_j)]\big)\big|X_j\big|^2\big] \\
&\lesssim \sup_{j\in\mathbb{N}} E\big[(1+|X_j|^{C'})\big|X_j\big|^2\big] < \infty. \qquad (8)
\end{aligned}$$

Because of this inequality, we can apply Lemma 5.1 to obtain

$$\sup_{n>0} E\left[\left|\frac{1}{\sqrt{n}}\sum_{j=1}^n \psi_j\right|^2\right] \le \sup_{n>0}\frac{1}{n}E\left[\sup_{1\le i\le n}\left|\sum_{j=1}^i \psi_j\right|^2\right] < \infty.$$

Therefore, $\frac{1}{\sqrt{n}}\sum_{j=1}^n \psi_j = O_p(1)$, and Δ_n satisfies the equality

$$\Delta_n = \frac{1}{\sqrt{n}}\sum_{j=1}^n \psi_j + \frac{1}{\sqrt{n}}\sum_{j=1}^n \big(F(X_j) - \partial b(X_j'\theta_0)\big)X_j = O_p(1).$$

(ii) For some $C > 0$,

$$\begin{aligned}
\sup_{\theta\in\Theta}\left|\frac{1}{n\sqrt{n}}\partial_\theta^3\mathbb{H}_n(\theta)\right| &\le \frac{1}{n\sqrt{n}}\sup_{\theta\in\Theta}\left(\sum_{i,k,\ell=1}^p \big|\partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_\ell}\mathbb{H}_n(\theta)\big|^2\right)^{\frac{1}{2}} \\
&\lesssim \frac{1}{n\sqrt{n}}\sup_{\theta\in\Theta}\sum_{i,k,\ell=1}^p\left|\sum_{j=1}^n \partial^3 b(X_j'\theta)X_{j,i}X_{j,k}X_{j,\ell}\right| \\
&\le \frac{1}{n\sqrt{n}}\sup_{\theta\in\Theta}\sum_{i,k,\ell=1}^p\sum_{j=1}^n \big|\partial^3 b(X_j'\theta)\big|\big|X_{j,i}X_{j,k}X_{j,\ell}\big| \\
&\le \sum_{i,k,\ell=1}^p \frac{1}{n\sqrt{n}}\sum_{j=1}^n \sup_{\theta\in\Theta}\big|\partial^3 b(X_j'\theta)\big|\big|X_j\big|^3 \\
&\lesssim \sum_{i,k,\ell=1}^p \frac{1}{n\sqrt{n}}\sum_{j=1}^n (1+|X_j|^C)|X_j|^3 \\
&= \sum_{i,k,\ell=1}^p \frac{1}{n\sqrt{n}}\sum_{j=1}^n O_p(1) = o_p(1).
\end{aligned}$$

□

We write $\mathbb{U}_n(\theta_0) = \left\{u\in\mathbb{R}^p; \theta_0 + \frac{u}{\sqrt{n}} \in \Theta\right\}$ and $\mathbb{Z}_n(u) = \exp\left\{\mathbb{H}_n\left(\theta_0 + \frac{u}{\sqrt{n}}\right) - \mathbb{H}_n(\theta_0)\right\}$.

Lemma 5.3. *If Assumptions 2.1-2.4, 3.1 and 3.2 hold, then*

$$\int_{\mathbb{U}_n(\theta_0)\cap\{|u|\ge M_n\}} \mathbb{Z}_n(u)du = o_p(1)$$

for any $M_n \to \infty$.

Proof. We have that

$$\begin{aligned}
\int_{\mathbb{U}_n(\theta_0)\cap\{|u|\ge M_n\}} \mathbb{Z}_n(u)du &= \int_{\mathbb{U}_n(\theta_0)\cap\{|u|\ge M_n\}} \exp\left\{u'\Delta_n + \frac{1}{2n}u'\partial_\theta^2\mathbb{H}_n(\tilde{\theta}_n)u\right\}du \\
&= \int_{\mathbb{U}_n(\theta_0)\cap\{|u|\ge M_n\}} \exp\left\{u'\Delta_n - \frac{1}{2}u'\left(\frac{1}{n}\sum_{j=1}^n \partial^2 b(X_j'\tilde{\theta}_n)X_jX_j'\right)u\right\}du,
\end{aligned}$$

where $\tilde{\theta}_n = \theta_0 + \xi(\theta_0 + \frac{u}{\sqrt{n}} - \theta_0) = \theta_0 + \xi\frac{u}{\sqrt{n}}$ for some ξ satisfying $0 < \xi < 1$. From Assumption 3.2 (i), there exists a function $\underline{b}$ such that

$$\int_{\mathbb{U}_n(\theta_0)\cap\{|u|\geq M_n\}} \exp\left\{u'\Delta_n - \frac{1}{2}u'\left(\frac{1}{n}\sum_{j=1}^n \partial^2 b(X_j'\tilde{\theta}_n)X_jX_j'\right)u\right\}du$$
$$\leq \int_{\mathbb{U}_n(\theta_0)\cap\{|u|\geq M_n\}} \exp\left\{u'\Delta_n - \frac{1}{2}u'\left(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\right)u\right\}du$$
$$\leq \int_{|u|\geq M_n} \exp\left\{u'\Delta_n - \frac{1}{2}u'\left(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\right)u\right\}du.$$

Fix any $\epsilon > 0$. For $\lambda_0 > 0$ given in Assumption 3.2 (ii),

$$\begin{aligned}
&P\left[\int_{\mathbb{U}_n(\theta_0)\cap\{|u|\geq M_n\}} \mathbb{Z}_n(u)du > \epsilon\right] \\
&\leq P\left[\int_{|u|\geq M_n} \exp\left\{u'\Delta_n - \frac{1}{2}u'\left(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\right)u\right\}du > \epsilon\right] \\
&= P\left[\int_{|u|\geq M_n} \exp\left\{u'\Delta_n - \frac{1}{2}u'\left(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\right)u\right\}du > \epsilon; \lambda_{min}\left(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\right) < \lambda_0\right] \\
&\quad + P\left[\int_{|u|\geq M_n} \exp\left\{u'\Delta_n - \frac{1}{2}u'\left(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\right)u\right\}du > \epsilon; \lambda_{min}\left(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\right) \geq \lambda_0\right] \\
&\leq P\left[\lambda_{min}\left(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\right) < \lambda_0\right] + P\left[\int_{|u|\geq M_n} \exp\left\{u'\Delta_n - \frac{1}{2}\lambda_0 u'u\right\}du > \epsilon\right]. \qquad (9)
\end{aligned}$$

There exists a constant $K > 0$ such that

$$\begin{aligned}
&P\left[\int_{|u|\geq M_n} \exp\left\{u'\Delta_n - \frac{1}{2}\lambda_0 u'u\right\}du > \epsilon\right] \\
&= P\left[\int_{|u|\geq M_n} \exp\left\{u'\Delta_n - \frac{1}{2}\lambda_0 u'u\right\}du > \epsilon; |\Delta_n| > K\right] \\
&\quad + P\left[\int_{|u|\geq M_n} \exp\left\{u'\Delta_n - \frac{1}{2}\lambda_0 u'u\right\}du > \epsilon; |\Delta_n| \leq K\right] \\
&\leq P\Big[|\Delta_n| > K\Big] + P\left[\exp\left(\frac{\Delta_n'\Delta_n}{2\lambda_0}\right)\int_{|u|\geq M_n} \exp\left\{-\frac{\lambda_0}{2}(u - \lambda_0^{-1}\Delta_n)'(u - \lambda_0^{-1}\Delta_n)\right\}du > \epsilon; |\Delta_n| \leq K\right] \\
&= P\Big[|\Delta_n| > K\Big] + P\left[\exp\left(\frac{\Delta_n'\Delta_n}{2\lambda_0}\right)\int_{|t+\lambda_0^{-1}\Delta_n|\geq M_n} \exp\left\{-\frac{\lambda_0}{2}t't\right\}dt > \epsilon; |\Delta_n| \leq K\right] \\
&\leq P\Big[|\Delta_n| > K\Big] + P\left[\exp\left(\frac{K^2}{2\lambda_0}\right)\int_{|t|\geq M_n - \lambda_0^{-1}K} \exp\left\{-\frac{\lambda_0}{2}t't\right\}dt > \epsilon\right].
\end{aligned}$$

Because of Lemma 5.2 (i), for some N,

$$P\Big[|\Delta_n| > K\Big] + P\left[\exp\left(\frac{K^2}{2\lambda_0}\right)\int_{|t|\geq M_n - \lambda_0^{-1}K} \exp\left\{-\frac{\lambda_0}{2}t't\right\}dt > \epsilon\right] < \frac{\epsilon}{2} \qquad (10)$$

for every $n \geq N$. Due to Assumption 3.2 (ii), (9) and (10),

$$P\left[\int_{\mathbb{U}_n(\theta_0)\cap\{|u|\geq M_n\}} \mathbb{Z}_n(u)du > \epsilon\right] < \epsilon$$

for all $n \geq N$. Thus, $\int_{\mathbb{U}_n(\theta_0)\cap\{|u|\geq M_n\}} \mathbb{Z}_n(u)du$ converges to 0 in probability. □

5.1. **Proof of Theorem 2.6.** Denote $\mathbb{Y}_n(\theta) = \frac{1}{n}\big(\mathbb{H}_n(\theta) - \mathbb{H}_n(\theta_0)\big)$. Since $\hat{\theta}_n \in \mathrm{argmax}_{\theta\in\Theta}\mathbb{H}_n(\theta)$ if and only if $\hat{\theta}_n \in \mathrm{argmax}_{\theta\in\Theta}\mathbb{Y}_n(\theta)$, we consider *Argmax theorem* (van der Vaart [27, Theorem 5.56, Corollary 5.58]) for $\mathbb{Y}_n(\theta)$. Under Assumptions 2.1-2.4, it is enough to show that there exist a θ-uniform limit in probability of $\mathbb{Y}_n$ and a unique θ_0 maximizing the limit. Because of (5), we have

$$\mathbb{Y}_n(\theta) \xrightarrow{P} \int \Big(F(x)x'(\theta - \theta_0) - \big(b(x'\theta) - b(x'\theta_0)\big)\Big)\nu(dx) =: \mathbb{Y}_0(\theta).$$

From Assumptions 2.1, 2.2 and (8),

$$\begin{aligned}
\sup_{n>0} E\left[\sup_{\theta\in\Theta}|\partial_\theta \mathbb{Y}_n(\theta)|\right] &= \sup_{n>0} E\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^n \big(Y_j - \partial b(X_j'\theta)\big)X_j\right|\right] \\
&= \sup_{n>0} E\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^n \big(Y_j - F(X_j) + F(X_j) - \partial b(X_j'\theta)\big)X_j\right|\right] \\
&\le \sup_{n>0}\frac{1}{n}\sum_{j=1}^n E\big[|\psi_j|\big] + \sup_{n>0}\frac{1}{n}\sum_{j=1}^n E\left[\sup_{\theta\in\Theta}\left|\big(F(X_j) - \partial b(X_j'\theta)\big)X_j\right|\right] \\
&\le \sup_{n>0}\frac{1}{n}\sum_{j=1}^n E\big[|\psi_j|\big] + \sup_{n>0}\frac{1}{n}\sum_{j=1}^n E\left[\left(\big|F(X_j)\big| + \sup_{\theta\in\Theta}\big|\partial b(X_j'\theta)\big|\right)|X_j|\right] \\
&\lesssim \sup_{n>0}\frac{1}{n}\sum_{j=1}^n E\big[|\psi_j|\big] + \sup_{n>0}\frac{1}{n}\sum_{j=1}^n E\left[\left((1+|X_j|)^{C'} + (1+|X_j|)^{C}\right)|X_j|\right] \\
&< \infty.
\end{aligned}$$

Then for any $\epsilon > 0$ and for some $K > 0$,

$$\begin{aligned}
&P\left[\sup_{|\theta_1-\theta_2|\le\delta}\left|\big(\mathbb{Y}_n(\theta_1) - \mathbb{Y}_0(\theta_1)\big) - \big(\mathbb{Y}_n(\theta_2) - \mathbb{Y}_0(\theta_2)\big)\right| > K\right] \\
&\le P\left[\delta\sup_{\theta\in\Theta}|\partial_\theta\mathbb{Y}_n(\theta)| + \sup_{|\theta_1-\theta_2|\le\delta}|\mathbb{Y}_0(\theta_1) - \mathbb{Y}_0(\theta_2)| > K\right] \\
&\le P\left[\delta\sup_{\theta\in\Theta}|\partial_\theta\mathbb{Y}_n(\theta)| > \frac{K}{2}\right] + P\left[\sup_{|\theta_1-\theta_2|\le\delta}|\mathbb{Y}_0(\theta_1) - \mathbb{Y}_0(\theta_2)| > \frac{K}{2}\right] \\
&\lesssim \frac{\delta}{K}E\left[\sup_{\theta\in\Theta}|\partial_\theta\mathbb{Y}_n(\theta)|\right] + P\left[\sup_{|\theta_1-\theta_2|\le\delta}|\mathbb{Y}_0(\theta_1) - \mathbb{Y}_0(\theta_2)| > \frac{K}{2}\right] \\
&< \epsilon
\end{aligned}$$

as $\delta \to 0$. Hence, in view of the Arzela-Ascoli criterion we obtain

$$\sup_{\theta\in\Theta}\big|\mathbb{Y}_n(\theta) - \mathbb{Y}_0(\theta)\big| \xrightarrow{P} 0.$$

Moreover, $\partial_\theta\mathbb{Y}_0(\theta) = \partial_\theta\mathbb{H}_0(\theta)$ is satisfied for any $\theta\in\Theta$, so $\{\theta_0\} = \mathrm{argmax}_{\theta\in\Theta}\mathbb{H}_0(\theta) = \mathrm{argmax}_{\theta\in\Theta}\mathbb{Y}_0(\theta)$.

5.2. **Proof of Theorem 2.8.** It will be shown in Section 5.3 that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \Gamma_0^{-1}\Delta_n + o_p(1).$$

In view of Herrndorf [16, Theorem, Corollary 1], Δ_n converges to the normal distribution $N(0,\Sigma_0)$ in law if we show the following four conditions:

(i) $E\big[\big(Y_j - \partial b(X_j'\theta_0)\big)X_j\big] = 0$, $E\big[\big(Y_j - \partial b(X_j'\theta_0)\big)^2 X_j'X_j\big] < \infty$ for all $j\in\mathbb{N}$.
(ii) $\frac{1}{n}E\left[\left\{\sum_{j=1}^n\big(Y_j - \partial b(X_j'\theta_0)\big)X_j\right\}\left\{\sum_{j=1}^n\big(Y_j - \partial b(X_j'\theta_0)\big)X_j\right\}'\right] \to \Sigma_0$ as $n\to\infty$.
(iii) $\sum_{k\in\mathbb{N}}\alpha(k)^{\frac{1}{3}} < \infty$.
(iv) $\limsup_{n\to\infty} E\left[\left|\big(Y_n - \partial b(X_n'\theta_0)\big)X_n\right|^3\right] < \infty$.

Then, (ii) is ensured by Assumption 2.7 (ii).
(i) Because of Assumptions 2.2, 2.7 (i) and the definition of θ_0, we have

$$\begin{aligned}
E\big[\big(Y_j - \partial b(X_j'\theta_0)\big)X_j\big] &= E[\big(Y_j - F(X_j)\big)X_j + \big(F(X_j) - \partial b(X_j'\theta_0)\big)X_j] \\
&= 0 + \int\big(F(x) - \partial b(x'\theta_0)\big)x\nu(dx) \\
&= 0
\end{aligned}$$

for any $j\in\mathbb{N}$. Furthermore, from Assumption 2.1,

$$\begin{aligned}
\sup_{j\in\mathbb{N}} E\big[\big(Y_j - \partial b(X_j'\theta_0)\big)^2 X_j'X_j\big] &\lesssim \sup_{j\in\mathbb{N}} E\left[\left(|Y_j|^2 + |\partial b(X_j'\theta_0)|^2\right)|X_j|^2\right] \\
&\lesssim \sup_{j\in\mathbb{N}} E\left[\left((1+|X_j|^{C'}) + (1+|X_j|^{C})^2\right)|X_j|^2\right]
\end{aligned}$$

$$< \infty. \tag{11}$$

(iii) Assumption 2.2 gives

$$\begin{aligned}\sum_{k\in\mathbb{N}} \alpha(k)^{\frac{1}{3}} &\leq c^{-\frac{1}{3}} \sum_{k\in\mathbb{N}} e^{-\frac{1}{3}ck} \\ &= c^{-\frac{1}{3}} \frac{e^{-\frac{1}{3}c}}{1-e^{-\frac{1}{3}c}} < \infty.\end{aligned}$$

(iv) In a similar way as (11), we can show that

$$\sup_{j\in\mathbb{N}} E\Big[\big|(Y_j - \partial b(X_j'\theta_0))X_j\big|^3\Big] < \infty$$

Hence, (iv) is satisfied.

5.3. **Proof of Theorem 3.3.** In what follows, we consider the zero-extended version of $\mathbb{Z}_n$ and use the same notation:

$$\int_{\mathbb{R}^p\setminus\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)du = 0.$$

By using the change of variable $\theta = \theta_0 + \frac{u}{\sqrt{n}}$, the log marginal quasi-likelihood function becomes

$$\begin{aligned}\log\left(\int_\Theta \exp\{\mathbb{H}_n(\theta)\}\pi(\theta)d\theta\right) &= \mathbb{H}_n(\theta_0) - \frac{p}{2}\log n + \log\left\{\int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\pi\Big(\theta_0 + \frac{u}{\sqrt{n}}\Big)du\right\} \\ &= \mathbb{H}_n(\theta_0) - \frac{p}{2}\log n \\ &\quad + \log\left\{\int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\Big(\pi\Big(\theta_0 + \frac{u}{\sqrt{n}}\Big) - \pi(\theta_0)\Big)du + \pi(\theta_0)\int_{\mathbb{R}^p} \mathbb{Z}_n(u)du\right\}.\end{aligned}$$

First we consider the asymptotic behavior of $\int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\big(\pi(\theta_0 + \frac{u}{\sqrt{n}}) - \pi(\theta_0)\big)du$. Because of (ii) of Theorem 3.3, Assumption 3.2 (i) and Lemma 5.3, we can take $M > 0$ large enough so that

$$\begin{aligned}&\left|\int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\Big(\pi\Big(\theta_0 + \frac{u}{\sqrt{n}}\Big) - \pi(\theta_0)\Big)du\right| \\ &\leq \int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\left|\pi\Big(\theta_0 + \frac{u}{\sqrt{n}}\Big) - \pi(\theta_0)\right|du \\ &= \int_{\mathbb{U}_n(\theta_0)\cap\{|u|<M\}} \mathbb{Z}_n(u)\left|\pi\Big(\theta_0 + \frac{u}{\sqrt{n}}\Big) - \pi(\theta_0)\right|du + \int_{\mathbb{U}_n(\theta_0)\cap\{|u|\geq M\}} \mathbb{Z}_n(u)\left|\pi\Big(\theta_0 + \frac{u}{\sqrt{n}}\Big) - \pi(\theta_0)\right|du \\ &\leq \sup_{|u|<M}\left|\pi\Big(\theta_0 + \frac{u}{\sqrt{n}}\Big) - \pi(\theta_0)\right| \sup_{|u|<M} \mathbb{Z}_n(u) + 2\sup_{\theta\in\Theta}\pi(\theta)\int_{\mathbb{U}_n(\theta_0)\cap\{|u|\geq M\}} \mathbb{Z}_n(u)du \\ &= o_p(1)\times \sup_{|u|<M}\left\{\exp\left(u'\Delta_n - \frac{1}{2}u'\Big(\frac{1}{n}\sum_{j=1}^n \partial^2 b(X_j'\tilde{\theta}_n)X_jX_j'\Big)u\right)\right\} + O_p(1)\times o_p(1) \\ &\leq o_p(1)\times \sup_{|u|<M}\left\{\exp\left(u'\Delta_n - \frac{1}{2}u'\Big(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\Big)u\right)\right\} + o_p(1),\end{aligned}$$

where $\tilde{\theta}_n = \theta_0 + \xi\frac{u}{\sqrt{n}}$ for some ξ satisfying $0 < \xi < 1$. Since $\frac{\partial}{\partial u}\{u'\Delta_n - \frac{1}{2}u'(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j')u)\} = 0$ if and only if $u = \big(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\big)^{-1}\Delta_n$, we have

$$u'\Delta_n - \frac{1}{2}u'\Big(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\Big)u \leq \frac{1}{2}\Delta_n'\Big(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\Big)^{-1}\Delta_n.$$

From Assumption 3.2 (ii) and Lemma 5.2 (i), for any $\epsilon > 0$ and for some $L > 0$,

$$\begin{aligned}&\limsup_{n\to\infty} P\left[\sup_{|u|<M}\left\{\exp\left(u'\Delta_n - \frac{1}{2}u'\Big(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\Big)u\right)\right\} > L\right] \\ &\leq \limsup_{n\to\infty} P\left[\exp\left\{\frac{1}{2}\Delta_n'\Big(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\Big)^{-1}\Delta_n\right\} > L; \lambda_{min}\Big(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\Big) < \lambda_0\right] \\ &\quad + \limsup_{n\to\infty} P\left[\exp\left\{\frac{1}{2}\Delta_n'\Big(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\Big)^{-1}\Delta_n\right\} > L; \lambda_{min}\Big(\frac{1}{n}\sum_{j=1}^n \underline{b}(X_j)X_jX_j'\Big) \geq \lambda_0\right]\end{aligned}$$

$$\le \limsup_{n\to\infty} P\left[\lambda_{min}\left(\frac{1}{n}\sum_{j=1}^{n}\underline{b}(X_j)X_jX_j'\right) < \lambda_0\right] + \limsup_{n\to\infty} P\left[\exp\left\{\frac{1}{2\lambda_0}\Delta_n'\Delta_n\right\} > L\right]$$
$$< \epsilon, \tag{12}$$

so that $\sup_{|u|<M}\left\{\exp\left(u'\Delta_n - \frac{1}{2}u'\left(\frac{1}{n}\sum_{j=1}^{n}\underline{b}(X_j)X_jX_j'\right)u\right)\right\} = O_p(1)$. Hence, $\int_{\mathbb{U}_n(\theta_0)}\mathbb{Z}_n(u)\Big(\pi\big(\theta_0+\frac{u}{\sqrt{n}}\big) - \pi(\theta_0)\Big)$ converges to 0 in probability.

Next we will prove that $\int_{\mathbb{R}^p}\mathbb{Z}_n(u)du = \int_{\mathbb{R}^p}\exp(u'\Delta_n - \frac{1}{2}u'\Gamma_n u)du + o_p(1)$. For each $K>0$,

$$\begin{aligned}
&\left|\int_{\mathbb{R}^p}\mathbb{Z}_n(u)du - \int_{\mathbb{R}^p}\exp\left(u'\Delta_n - \frac{1}{2}u'\Gamma_n u\right)du\right| \\
&\le \int_{\mathbb{R}^p}\left|\mathbb{Z}_n(u) - \exp\left(u'\Delta_n - \frac{1}{2}u'\Gamma_n u\right)\right|du \\
&= \int_{|u|<K}\left|\mathbb{Z}_n(u) - \exp\left(u'\Delta_n - \frac{1}{2}u'\Gamma_n u\right)\right|du + \int_{|u|\ge K}\left|\mathbb{Z}_n(u) - \exp\left(u'\Delta_n - \frac{1}{2}u'\Gamma_n u\right)\right|du.
\end{aligned}$$

Due to Assumption 3.2 and Lemma 5.3, we can take K large enough so that

$$\begin{aligned}
&\int_{|u|\ge K}\left|\mathbb{Z}_n(u) - \exp\left(u'\Delta_n - \frac{1}{2}u'\Gamma_n u\right)\right|du \\
&\le \int_{|u|\ge K}\mathbb{Z}_n(u)du + \int_{|u|\ge K}\exp\left\{u'\Delta_n - \frac{1}{2}u'\left(\frac{1}{n}\sum_{j=1}^{n}\partial^2 b(X_j'\theta_0)X_jX_j'\right)u\right\}du \\
&\lesssim \int_{|u|\ge K}\exp\left\{u'\Delta_n - \frac{1}{2}u'\left(\frac{1}{n}\sum_{j=1}^{n}\underline{b}(X_j)X_jX_j'\right)u\right\}du = o_p(1).
\end{aligned}$$

In the same way as (12), for the same K,

$$\begin{aligned}
&\int_{|u|<K}\left|\mathbb{Z}_n(u) - \exp\left(u'\Delta_n - \frac{1}{2}u'\Gamma_n u\right)\right|du \\
&\le \sup_{|u|<K}\left|\exp\left(u'\Delta_n - \frac{1}{2}u'\Gamma_n u\right)\left\{\exp\left(\frac{1}{6}\sum_{i,k,\ell=1}^{p}\frac{1}{n\sqrt{n}}\partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_\ell}\mathbb{H}_n(\tilde{\theta}_n)u_iu_ku_\ell\right) - 1\right\}\right| \\
&= \sup_{|u|<K}\exp\left(u'\Delta_n - \frac{1}{2}u'\left(\frac{1}{n}\sum_{j=1}^{n}\partial^2 b(X_j'\theta_0)X_jX_j'\right)u\right)\left|\exp\left(\frac{1}{6}\sum_{i,k,\ell=1}^{p}\frac{1}{n\sqrt{n}}\partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_\ell}\mathbb{H}_n(\tilde{\theta}_n)u_iu_ku_\ell\right) - 1\right| \\
&\le \sup_{|u|<K}\exp\left(u'\Delta_n - \frac{1}{2}u'\left(\frac{1}{n}\sum_{j=1}^{n}\underline{b}(X_j)X_jX_j'\right)u\right)\left|\exp\left(\frac{1}{6}\sum_{i,k,\ell=1}^{p}\frac{1}{n\sqrt{n}}\partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_\ell}\mathbb{H}_n(\tilde{\theta}_n)u_iu_ku_\ell\right) - 1\right| \\
&\le \exp\left\{\frac{1}{2}\Delta_n'\left(\frac{1}{n}\sum_{j=1}^{n}\underline{b}(X_j)X_jX_j'\right)^{-1}\Delta_n\right\}\sup_{|u|<K}\left|\exp\left(\frac{1}{6}\sum_{i,k,\ell=1}^{p}\frac{1}{n\sqrt{n}}\partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_\ell}\mathbb{H}_n(\tilde{\theta}_n)u_iu_ku_\ell\right) - 1\right| \\
&= O_p(1)\times o_p(1) = o_p(1).
\end{aligned}$$

Therefore, we obtain that $\int_{\mathbb{R}^p}\mathbb{Z}_n(u)du = \int_{\mathbb{R}^p}\exp(u'\Delta_n - \frac{1}{2}u'\Gamma_n u)du + o_p(1)$. Moreover,

$$\begin{aligned}
\int_{\mathbb{R}^p}\exp\left(u'\Delta_n - \frac{1}{2}u'\Gamma_n u\right)du &= \exp\left(\frac{1}{2}\|\Gamma_n^{-\frac{1}{2}}\Delta_n\|^2\right)\int_{\mathbb{R}^p}\exp\Big(-\frac{1}{2}(u-\Gamma_n^{-1}\Delta_n)'\Gamma_n(u-\Gamma_n^{-1}\Delta_n)\Big)du \\
&= \exp\left(\frac{1}{2}\|\Gamma_n^{-\frac{1}{2}}\Delta_n\|^2\right)(2\pi)^{\frac{p}{2}}\det(\Gamma_n)^{-\frac{1}{2}},
\end{aligned}$$

so $\log\big(\int_\Theta \exp\{\mathbb{H}_n(\theta)\}\pi(\theta)d\theta\big)$ is given by

$$\begin{aligned}
\log\left(\int_\Theta \exp\{\mathbb{H}_n(\theta)\}\pi(\theta)d\theta\right) &= \mathbb{H}_n(\theta_0) - \frac{p}{2}\log n + \log\left\{\pi(\theta_0)\exp\left(\frac{1}{2}\|\Gamma_n^{-\frac{1}{2}}\Delta_n\|\right)(2\pi)^{\frac{p}{2}}\det(\Gamma_n)^{-\frac{1}{2}} + o_p(1)\right\} \\
&= \mathbb{H}_n(\theta_0) - \frac{p}{2}\log n + \log\pi(\theta_0) \\
&\quad + \frac{1}{2}\|\Gamma_0^{-\frac{1}{2}}\Delta_n\|^2 + \frac{p}{2}\log 2\pi + \log\det(\Gamma_n)^{-\frac{1}{2}} + o_p(1).
\end{aligned}$$

Finally we replace θ_0 by the QMLE $\hat{\theta}_n$:

$$\Delta_n = \frac{1}{\sqrt{n}}\partial_\theta\mathbb{H}_n(\theta_0)$$

$$= \big(\sqrt{n}(\hat{\theta}_n - \theta_0)\big)'\Big(- \frac{1}{n}\partial_\theta^2 \mathbb{H}_n(\check{\theta}_n)\Big),$$

where $\check{\theta}_n = \hat{\theta}_n + \eta_1(\theta_0 - \hat{\theta}_n)$ for some η_1 satisfying $0 < \eta_1 < 1$. Because of Lemma 5.2, there exists a η_2 satisfying $0 < \eta_2 < 1$ such that

$$\begin{aligned}
-\frac{1}{n}\partial_\theta^2 \mathbb{H}_n(\check{\theta}_n) &= \Gamma_n - \big(\sqrt{n}(\check{\theta}_n - \theta_0)\big)'\Big(\frac{1}{n\sqrt{n}}\partial_\theta^3 \mathbb{H}_n\big(\theta_0 + \eta_2(\check{\theta}_n - \theta_0)\big)\Big) \\
&= \Gamma_n - (1-\eta_1)\big(\sqrt{n}(\hat{\theta}_n - \theta_0)\big)'\Big(\frac{1}{n\sqrt{n}}\partial_\theta^3 \mathbb{H}_n\big(\theta_0 + \eta_2(\check{\theta}_n - \theta_0)\big)\Big) \\
&= \Gamma_0 + o_p(1).
\end{aligned}$$

Furthermore, we can show that $\frac{1}{n}\sum_{j=1}^n b(X_j'\hat{\theta}_n)X_jX_j' = \Gamma_n + o_p(1) = \Gamma_0 + o_p(1)$ in the same way, so we have

$$\begin{aligned}
\mathbb{H}_n(\theta_0) &= \mathbb{H}_n(\hat{\theta}_n) - \frac{1}{2}\big(\sqrt{n}(\hat{\theta}_n - \theta_0)\big)'\Gamma_0\big(\sqrt{n}(\hat{\theta}_n - \theta_0)\big) + o_p(1) \\
&= \mathbb{H}_n(\hat{\theta}_n) - \frac{1}{2}\big(\Gamma_0^{-1}\Delta_n\big)'\Gamma_0\big(\Gamma_0^{-1}\Delta_n\big) + o_p(1) \\
&= \mathbb{H}_n(\hat{\theta}_n) - \frac{1}{2}\|\Gamma_0^{-\frac{1}{2}}\Delta_n\|^2 + o_p(1).
\end{aligned}$$

Thus, the asymptotic behavior of the log marginal quasi-likelihood function is given by

$$\begin{aligned}
\log\left(\int_\Theta \exp\{\mathbb{H}_n(\theta)\}\pi(\theta)d\theta\right) &= \mathbb{H}_n(\hat{\theta}_n) - \frac{p}{2}\log n + \log \pi(\hat{\theta}_n) \\
&\quad + \frac{p}{2}\log 2\pi - \frac{1}{2}\log\det\left(\frac{1}{n}\sum_{j=1}^n b(X_j'\hat{\theta}_n)X_jX_j'\right) + o_p(1).
\end{aligned}$$

5.4. **Proof of Theorem 3.5.** We basically follow the scenario of Fasen and Kimmig [13].
(i) Θ_{m_0} is nested in Θ_m. Define the map $a : \Theta_{m_0} \to \Theta_m$ by $a(\theta) = A\theta + c$, where A and c satisfy that $\mathbb{H}_{m_0,n}(\theta) = \mathbb{H}_{m,n}\big(a(\theta)\big)$ for any $\theta \in \Theta_{m_0}$. Then the equation $\mathbb{H}_{m_0,0}(\theta) = \mathbb{H}_{m,0}\big(a(\theta)\big)$ is also satisfied for every $\theta \in \Theta_{m_0}$. If $a(\theta_{m_0,0}) \neq \theta_{m,0}$, $\mathbb{H}_{m_0,0}(\theta_{m_0,0}) = \mathbb{H}_{m,0}\big(a(\theta_{m_0,0})\big) < \mathbb{H}_{m,0}(\theta_{m,0})$ and assumption of the optimal model is not satisfied. Hence we have $a(\theta_{m_0,0}) = \theta_{m,0}$.

By the Taylor expansion of $\mathbb{H}_{m,n}$

$$\begin{aligned}
\mathbb{H}_{m_0,n}(\hat{\theta}_{m_0,n}) &= \mathbb{H}_{m,n}\big(a(\hat{\theta}_{m_0,n})\big) \\
&= \mathbb{H}_{m,n}(\hat{\theta}_{m,n}) - \frac{1}{2}\big\{\sqrt{n}\big(\hat{\theta}_{m,n} - a(\hat{\theta}_{m_0,n})\big)\big\}'\left(\frac{1}{n}\sum_{j=1}^n \partial^2 b_m(X_j'\tilde{\theta}_n)X_jX_j'\right)\big\{\sqrt{n}\big(\hat{\theta}_{m,n} - a(\hat{\theta}_{m_0,n})\big)\big\},
\end{aligned}$$

where $\tilde{\theta}_n = \hat{\theta}_{m,n} + \xi\big(a(\hat{\theta}_{m_0,n}) - \hat{\theta}_{m,n}\big)$ for some ξ satisfying $0 < \xi < 1$ and $\tilde{\theta}_n \xrightarrow{P} \theta_{m,0}$ as $n \to \infty$. Therefore, the difference between $\mathrm{QBIC}^{(m_0)}$ and $\mathrm{QBIC}^{(m)}$ is given by

$$\begin{aligned}
\mathrm{QBIC}^{(m_0)} - \mathrm{QBIC}^{(m)} &= \big\{\sqrt{n}\big(\hat{\theta}_{m,n} - a(\hat{\theta}_{m_0,n})\big)\big\}'\left(\frac{1}{n}\sum_{j=1}^n \partial^2 b_m(X_j'\tilde{\theta}_n)X_jX_j'\right)\big\{\sqrt{n}\big(\hat{\theta}_{m,n} - a(\hat{\theta}_{m_0,n})\big)\big\} \\
&\quad + \log\det\big(-\partial_\theta^2\mathbb{H}_{m_0,n}(\hat{\theta}_{m_0,n})\big) - \log\det\big(-\partial_\theta^2\mathbb{H}_{m,n}(\hat{\theta}_{m,n})\big).
\end{aligned}$$

We consider the behavior of the $\hat{\theta}_{m,n} - a(\hat{\theta}_{m_0,n})$. Because of the chain rule, we have

$$\begin{aligned}
\partial_\theta\mathbb{H}_{m_0,n}(\theta_{m_0,0}) &= A'\partial_\theta\mathbb{H}_{m,n}(\theta_{m,0}), \\
\partial_\theta^2\mathbb{H}_{m_0,n}(\theta) &= A'\partial_\theta^2\mathbb{H}_{m,n}\big(a(\theta)\big)A.
\end{aligned}$$

Moreover,

$$\begin{aligned}
a(\hat{\theta}_{m_0,n}) - \theta_{m,0} &= A(\hat{\theta}_{m_0,n} - \theta_{m,0}), \\
\sqrt{n}(\hat{\theta}_{m_0,n} - \theta_{m_0,0}) &= \left(-\frac{1}{n}\partial_\theta^2\mathbb{H}_{m_0,n}(\check{\theta}_n)\right)^{-1}\left(\frac{1}{\sqrt{n}}\partial_\theta\mathbb{H}_{m_0,n}(\theta_{m_0,0})\right) \\
&= \left\{A'\left(-\frac{1}{n}\partial_\theta^2\mathbb{H}_{m,n}\big(a(\check{\theta}_n)\big)\right)A\right\}^{-1}A'\left(\frac{1}{\sqrt{n}}\partial_\theta\mathbb{H}_{m,n}(\theta_{m,0})\right)
\end{aligned}$$

$$= \left\{ A' \left(\frac{1}{n} \sum_{j=1}^{n} \partial^2 b_m \big(a(\check{\theta}_n) \big) X_j X_j' \right) A \right\}^{-1} A' \left(\frac{1}{\sqrt{n}} \partial_\theta \mathbb{H}_{m,n}(\theta_{m,0}) \right)$$
$$= \Big(A'\Gamma_{m,0} A \Big)^{-1} A' \left(\frac{1}{\sqrt{n}} \partial_\theta \mathbb{H}_{m,n}(\theta_{m,0}) \right) + o_p(1),$$

where $\check{\theta}_n = \hat{\theta}_{m_0,n} + \eta(\theta_{m_0,0} - \hat{\theta}_{m_0,n})$ for some η satisfying $0 < \eta < 1$ and $a(\check{\theta}_n) \xrightarrow{P} a(\theta_{m_0,0}) = \theta_{m,0}$ as $n \to \infty$. These equalities and Theorem 2.8 give

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_{m,n} - a(\hat{\theta}_{m_0,n})) &= \sqrt{n}(\hat{\theta}_{m,n} - \theta_{m,0}) - A\sqrt{n}(\hat{\theta}_{m_0,n} - \theta_{m_0,0}) \\
&\xrightarrow{\mathcal{L}} \left\{ \Gamma_{m,0}^{-1} - A\Big(A'\Gamma_{m,0}A\Big)^{-1} A' \right\} N_{p_m}(0, \Sigma_0) \\
&= N_{p_m}\left(0, \left\{ \Gamma_{m,0}^{-1} - A\Big(A'\Gamma_{m,0}A\Big)^{-1} A' \right\} \Sigma_0 \left\{ \Gamma_{m,0}^{-1} - A\Big(A'\Gamma_{m,0}A\Big)^{-1} A' \right\} \right) \sim \mathbf{N}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&P[\mathrm{QBIC}^{(m_0)} - \mathrm{QBIC}^{(m)} < 0] \\
&= P\Bigg[\mathbf{N}' \left(\frac{1}{n} \sum_{j=1}^{n} \partial^2 b_m (X_j' \tilde{\theta}_n) X_j X_j' \right) \mathbf{N} + \log\det \left(\frac{1}{n} \sum_{j=1}^{n} \partial^2 b_{m_0} (X_j' \hat{\theta}_{m_0,n}) X_j X_j' \right) \\
&\qquad - \log\det \left(\frac{1}{n} \sum_{j=1}^{n} \partial^2 b_m (X_j' \hat{\theta}_{m,n}) X_j X_j' \right) < (p_m - p_{m_0}) \log n \Bigg] \\
&\to P\Big[\mathbf{N}'\Gamma_{m,0}\mathbf{N} + \log\det\big(\Gamma_{m_0,0}\big) - \log\det\big(\Gamma_{m,0}\big) < \infty \Big]
\end{aligned}$$

as $n \to \infty$. From Imhof [17, (1.1)], $\mathbf{N}'\Gamma_{m,0}\mathbf{N} = \sum_{j=1}^{p_m} \lambda_j \chi_j^2$ in distribution, where (χ_j^2) is a sequence of independent χ^2 random variables with one degree of freedom and λ_j are the eigenvalues of $\Gamma_{m,0}^{\frac{1}{2}} \{\Gamma_{m,0}^{-1} - A\big(A'\Gamma_{m,0}A\big)^{-1} A'\} \Sigma_0 \{\Gamma_{m,0}^{-1} - A\big(A'\Gamma_{m,0}A\big)^{-1} A'\} \Gamma_{m,0}^{\frac{1}{2}}$. Furthermore, $\log\det\big(\Gamma_{m_0,0}\big) = O(1)$ and $\log\det\big(\Gamma_{m,0}\big) = O(1)$. Hence,

$$P\big[\mathbf{N}'\Gamma_{m,0}\mathbf{N} + \log\det\big(\Gamma_{m_0,0}\big) - \log\det\big(\Gamma_{m,0}\big) < \infty\big] \geq P\left[\max_{j \in \{1,\ldots,p_m\}} \lambda_j \sum_{j=1}^{p_m} \chi_j^2 < \infty \right] = 1.$$

(ii) $\mathbb{H}_{m,0}(\theta) \neq \mathbb{H}_{m_0,0}(\theta_{m_0,0})$ for every $\theta \in \Theta_m$. Because of Lemma 5.2 (i) and the consistency of $\hat{\theta}_{m_0,n}$ and $\hat{\theta}_{m,n}$, we have

$$\begin{aligned}
\frac{1}{n}\mathbb{H}_{m_0,n}(\hat{\theta}_{m_0,n}) &= \frac{1}{n}\mathbb{H}_{m_0,n}(\theta_{m_0,0}) + o_p(1) = \mathbb{H}_{m_0,0}(\theta_{m_0,0}) + o_p(1), \\
\frac{1}{n}\mathbb{H}_{m,n}(\hat{\theta}_{m,n}) &= \frac{1}{n}\mathbb{H}_{m,n}(\theta_{m,0}) + o_p(1) = \mathbb{H}_{m,0}(\theta_{m,0}) + o_p(1).
\end{aligned}$$

Since $\mathbb{H}_{m_0,0}(\theta_{m_0,0})$ is lager than $\mathbb{H}_{m,0}(\theta_{m_0,0})$, we obtain

$$\begin{aligned}
&P[\mathrm{QBIC}^{(m_0)} - \mathrm{QBIC}^{(m)} < 0] \\
&= P\Bigg[-2\mathbb{H}_{m_0,n}(\hat{\theta}_{m_0,n}) + 2\mathbb{H}_{m,n}(\hat{\theta}_{m,n}) + \log\det \left(\frac{1}{n} \sum_{j=1}^{n} \partial^2 b_m (X_j' \hat{\theta}_{m_0,n}) X_j X_j' \right) \\
&\qquad - \log\det \left(\frac{-1}{n} \sum_{j=1}^{n} \partial^2 b_m (X_j' \hat{\theta}_{m,n}) X_j X_j' \right) < (p_m - p_{m_0}) \log n \Bigg] \\
&= P\Bigg[\frac{-2}{n} (\mathbb{H}_{m_0,n}(\hat{\theta}_{m_0,n}) - \mathbb{H}_{m,n}(\hat{\theta}_{m,n})) + \frac{1}{n} \log\det \left(\frac{1}{n} \sum_{j=1}^{n} \partial^2 b_m (X_j' \hat{\theta}_{m_0,n}) X_j X_j' \right) \\
&\qquad - \frac{1}{n} \log\det \left(\frac{1}{n} \sum_{j=1}^{n} \partial^2 b_m (X_j' \hat{\theta}_{m,n}) X_j X_j' \right) < (p_m - p_{m_0}) \frac{\log n}{n} \Bigg] \\
&\to P\Big[-2\big(\mathbb{H}_{m_0,0}(\theta_{m_0,0}) - \mathbb{H}_{m,0}(\theta_{m_0,0})\big) < 0 \Big] = 1
\end{aligned}$$

as $n \to \infty$.

Acknowledgements

The author wishes to thank Professor H. Masuda for several valuable suggestions and comments.

References

[1] Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csàki (Eds.), *2nd Int. Symp. Information Theory*, pp. 267-281. Budapest: Academiai Kiado.
[2] Akaike, H. (1974), A new look at the statistical model identification. *IEEE Trans. Autom. Control.* **19**, 716-723.
[3] Antonio, K. and Beirlant, J. (2007). Actuarial statistics with generalized linear mixed models. *Insur. Math. Econ.*, **40**, 58–76.
[4] Berg, D. (2007), Bankruptcy prediction by generalized additive models. *Appl. Stochastic Models Bus. Ind.*, **23**, 129–143.
[5] Bozdogan, H. (1987), Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.
[6] Burnham, K. P. and Anderson, D. R. (2002), *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York.
[7] Casella, G., Girón, F. J., Martínez, M. L. and Moreno, E. (2009), Consistency of Bayesian procedures for variable selection. *Ann. Statist.* **37**, 1207–1228.
[8] Cavanaugh, J. E. and Neath, A. A. (1999), Generalizing the derivation of the Schwarz information criterion. *Comm. Statist. Theory Methods*, **28**, 49–66.
[9] Chen, J. and Chen, Z. (2008), Extended Bayesian information criteria for model selection with large model spaces. *Biometrika.* **95**, 759–771.
[10] Davydov, Y. A. (1973), Mixing conditions for Markov chains. *Theory of Probability & Its Applications*, **18**, 312–328.
[11] Eguchi, S. and Masuda, H. (2015), Quasi-Bayesian comparison for LAQ models. MI Preprint series 2015–7.
[12] Fahrmeir, L. and Kaufmann, H. (1985), Consistency and asymptotic normality of the maximum likelihood estimator in generalized liner models. *Ann. Statist.*, **13**, 342–368.
[13] Fasen, V. and Kimmig, S. (2015). Information Criteria for Multivariate CARMA Processes. arXiv preprint arXiv:1505.00901.
[14] Haberman, S. and Renshaw, A. E. (1996), Generalized linear models and actuarial science. *The Statistician*, 407–436.
[15] Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall, London.
[16] Herrndorf, N. (1984), A functional central limit theorem for weakly dependent sequences of random variables. *Ann. Prob.*, 141–153.
[17] Imhof, J. P. (1961), Computing the distribution of quadratic forms in normal variables. *Biometrika*, 419–426.
[18] Konishi, S. and Kitagawa, G. (1996), Generalised information criteria in model selection. *Biometrika.* **83**, 875–890.
[19] Liebscher, E. (2005), Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *Journal of Time Series Analysis*, **26**, 669–689.
[20] Lv, J. and Liu, J. S. (2014), Model selection principles in misspecified models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **76**, 141–167.
[21] Masuda, H. (2013), Convergence of Gaussian quasi-likelihood random fields for ergodic Lévy driven SDE observed at high frequency. *Ann. Statist.*, **41**, 1593–1641.
[22] McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman & Hall, London.
[23] McNeil, A. J. and Wendin, J. P. (2007), Bayesian inference for generalized linear mixed models of portfolio credit risk. *Journal of Empirical Finance*, **14**, 131–149.
[24] Schwarz, G. (1978), Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
[25] Sei, T. and Komaki, F. (2007), Bayesian prediction and model selection for locally asymptotically mixed normal models. *Journal of Statistical Planning and Inference* **137**, 2523–2534.
[26] Uchida, M. (2010), Contrast-based information criterion for ergodic diffusion processes from discrete observations. *Ann. Inst. Statist. Math.* **62**, 161–187.
[27] van der Vaart, A. W. (1998), *Asymptotic Statistics*. Cambridge University press.
[28] Yoshida, N. (2011), Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. *Ann. Inst. Statist. Math.*, **63**, 431–479.

Graduate School of Mathematics, Kyushu University. 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan.
E-mail address: s-eguchi@math.kyushu-u.ac.jp