# A Note on AIC and TIC for Model Selection

Yong Li
*Renmin University of China*

Zhou Wu
*Zhejiang University*

Jun Yu
*University of Macau*

Tao Zeng
*Zhejiang University*

November 9, 2024

**Abstract**

This note gives a rigorous justification to Akaike information criterion (AIC) and Takeuchi information criterion (TIC). The existing literature has shown that, when the candidate model is a good approximation of the true data generating process (DGP), AIC is an asymptotic unbiased estimator of the expected Kullback-Leibler divergence between the DGP and the plug-in predictive distribution. When the candidate model is misspecified, TIC can be regraded as a robust version of AIC with its justification following a similar line of argument. However, the justifications in current literature are predominantly confined to the iid scenario. In this note, we establish the asymptotic unbiasedness of AIC and TIC under certain regular conditions. These conditions are applicable in various scenarios, encompassing weakly dependent data.

*Keywords:* AIC; TIC; Expected loss function; Kullback-Leibler divergence; Model selection; Plug-in predictive distribution; weakly dependent data.

*JEL:* C11, C52, C25, C22, C32

# 1 Introduction

Arguably Akaike information criterion (AIC) of Akaike (1973) is the most well-known information criteria for model selection. AIC gives an answer of the following question: Given a set of candidate models, which model gives the best predictions of future observations generated by the same mechanism that gives the observed data?

This query can be viewed as a statistical decision problem following the selection of an appropriate loss and risk function. In 1973, Akaike opted for the Kullback-Leibler (KL) divergence between the data generating process (DGP) and the plug-in predictive distribution as the loss function. Consequently, the model showcasing the lowest AIC value is deemed the optimal choice.

The existing body of literature has established that AIC serves as an asymptotically unbiased estimator of predictive risk under the KL loss function. Akaike (1973) showed this property. Cavanaugh (1997) gave a unified justification of AIC and its invariant. Shi and Tsai (1998) showed the similar results for generalized AIC based on the M-estimator. Several prominent textbooks in model selection, such as Linhart and Zucchini (1986), Burnham and Anderson (2002), and Claeskens and Hjort (2008), also corroborate this outcome. The proofs in these works take the same approach. Initially, the KL divergence is decomposed into three components, with the first term representing the log-likelihood. Subsequently, the last two terms are approximated using the second-order Taylor expansion of the log-likelihood, disregarding the residual term when computing the expectation of a quadratic function of the maximum likelihood estimator (MLE). Finally, leveraging the asymptotic normality of the MLE, these expectations converge to the parameter's dimension.

To justify AIC, we need to assume all candidate models are good approximations of the true DGP. This assumption is often violated due to potential misspecification. In response to this challenge, Takeuchi Information Criterion (TIC), introduced by Takeuchi (1976), was formulated as a model-robust version of AIC, accommodating scenarios where candidate models are misspecified. The primary distinction between TIC and AIC lies in their respective penalties. Whereas AIC penalizes based on the model's dimension, TIC replaces this with the trace of the inverse Hessian matrix multiplied by the Jacobian matrix. This form of penalty in TIC was also noted by Stone (1977). When models are correctly specified, the information identity simplifies TIC back to AIC; see Claeskens and Hjort (2008), Cavanaugh and Neath (2019).

However, these justifications are only approximate for two main reasons. Firstly, although the remainder tends to zero in probability, it is a well-established fact that convergence in probability does not ensure the convergence of the expectation. This issue is elaborated on in Example 12.7 in Davidson (2021). Secondly, the asymptotic normality of MLE in general is insufficient to demonstrate the convergence of the expectation of the quadratic function of MLE. Additional regular conditions are needed for this purpose.

In this paper, we introduce slightly more robust regular conditions to guarantee the validity of these convergence results. We demonstrate that both AIC and TIC serve as asymptotically unbiased estimators of the KL divergence between the true Data Generating Process (DGP) and the plug-in predictive distribution. Finally, we discuss these regular conditions in various scenarios, encompassing iid data, stationary ergodic data, and weakly dependent data.

# 2 Model Selection as a Statistical Decision Problem

Before giving our justification of AIC and TIC, let us fix some notations. Let $\mathbf{y} = (y_1, y_2, \cdots, y_n)' \in \mathbf{Y}$ be observed data. Let a set of probabilistic models be $\{M_k\}_{k=1}^{K} =$

$\{p\left(\mathbf{y}|\boldsymbol{\theta}_k, M_k\right)\}_{k=1}^{K}$ where $n$ is the sample size, $\boldsymbol{\theta}_k$ (we simply write it as $\boldsymbol{\theta}$ when there is no confusion) is the set of parameters in candidate model $M_k$, and $p(\cdot)$ is its probability density function (pdf). Formally, the model selection problem is taken as a decision problem to select a model among $\{M_k\}_{k=1}^{K}$ where the action space has $K$ elements, namely, $\{d_k\}_{k=1}^{K}$, where $d_k$ means $M_k$ is selected.

For the decision problem, a loss function, $\ell(\mathbf{y}, d_k)$, which measures the loss of decision $d_k$ as a function of $\mathbf{y}$, must be specified. Given the loss function, the expected loss (or risk) can be defined as (Berger, 1985)

$$Risk(d_k) = E_{\mathbf{y}}\left[\ell(\mathbf{y}, d_k)\right] = \int \ell(\mathbf{y}, d_k) g(\mathbf{y}) d\mathbf{y}, \tag{1}$$

where $g(\mathbf{y})$ is the pdf of $\mathbf{y}$. Hence, the model selection problem is equivalent to optimizing the statistical decision,

$$k^* = \arg\min_{k} Risk(d_k).$$

Based on the set of candidate models $\{M_k\}_{k=1}^{K}$, model $M_{k^*}$ with decision $d_{k^*}$ is selected.

Let $\mathbf{y}_{rep} = (y_{1,rep}, \cdots, y_{n,rep})'$ be hypothetically replicate data, independently generated by the exact mechanism that gives $\mathbf{y}$. Assume the sample size in $\mathbf{y}_{rep}$ is also $n$. Consider the predictive density of this hypothetically replicate experiment for a candidate model $M_k$. The plug-in predictive density can be expressed as $p\left(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right)$ for $M_k$ where $\tilde{\boldsymbol{\theta}}_n(\mathbf{y})$ is an estimate of $\boldsymbol{\theta}$ based on $\mathbf{y}$ (we simply write $\tilde{\boldsymbol{\theta}}_n(\mathbf{y})$ as $\tilde{\boldsymbol{\theta}}_n$ when there is no confusion).

The quantity that has been used to measure the quality of the candidate model in terms of its ability to make predictions is the KL divergence between $g\left(\mathbf{y}_{rep}\right)$ and $p\left(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right)$ multiplied by 2, Naturally, the loss function associated with decision

$d_k$ is

$$\ell(\mathbf{y}, d_k) = 2\times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right)\right] = 2\int \log \frac{g\left(\mathbf{y}_{rep}\right)}{p\left(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right)} g\left(\mathbf{y}_{rep}\right) d\mathbf{y}_{rep}. \tag{2}$$

As a result, the model selection problem is,

$$\begin{aligned} k^* &= \arg\min_k Risk(d_k) = \arg\min_k E_{\mathbf{y}}\left[\ell(\mathbf{y}, d_k)\right] \\ &= \arg\min_k \left\{E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[2\log g\left(\mathbf{y}_{rep}\right)\right] + E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[-2\log p\left(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right)\right]\right\}. \end{aligned}$$

Since $g\left(\mathbf{y}_{rep}\right)$ is the DGP, $E_{\mathbf{y}_{rep}}\left[2\log g\left(\mathbf{y}_{rep}\right)\right]$ is the same across all candidate models, and hence, is dropped from the above equation. Consequently,

$$k^* = \arg\min_k Risk(d_k) = \arg\min_k E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[-2\log p\left(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right)\right]. \tag{3}$$

The smaller $Risk(d_k)$ is, the better the candidate model performs when using $p\left(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right)$ to predict $g\left(\mathbf{y}_{rep}\right)$. The optimal decision makes it necessary to evaluate the risk.

When there is no confusion, we simply write a generic candidate model $p\left(\mathbf{y}|\boldsymbol{\theta}, M_k\right)$ as $p\left(\mathbf{y}|\boldsymbol{\theta}\right)$ where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq R^P$ (i.e. the dimension of $\boldsymbol{\theta}$ is $P$). When the candidate model is different, the value of $P$ may be different. Note that we allow for model misspecification. We denote $\widehat{\boldsymbol{\theta}}_n(\mathbf{y})$ as the quasi-maximum likelihood estimator (see White (1982) and White (1996)) based on $\mathbf{y}$, which is the global maximum interior to $\boldsymbol{\Theta}$ defined by

$$\widehat{\boldsymbol{\theta}}_n(\mathbf{y}, M_k) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \log p\left(\mathbf{y}|\boldsymbol{\theta}, M_k\right). \tag{4}$$

Let $\boldsymbol{\theta}_n^p = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E\log p\left(\mathbf{y}|\boldsymbol{\theta}, M_k\right)$ denote the pseudo-true value of candidate model $M_k$. The expected Hessian matrix and the expected Jacobian matrix are defined as

$$\mathbf{H}_n = E\left[\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log p\left(\mathbf{y}|\boldsymbol{\theta}_n^p, M_k\right)\right],\ \mathbf{B}_n = Var\left[\frac{1}{\sqrt{n}}\frac{\partial}{\partial\boldsymbol{\theta}}\log p\left(\mathbf{y}|\boldsymbol{\theta}_n^p, M_k\right)\right].$$

AIC and TIC can be defined as

$$\text{AIC} = -2\log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right) + 2P, \tag{5}$$

$$\text{TIC} = -2\log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right) + 2\text{tr}(-\mathbf{B}_n\mathbf{H}_n^{-1}). \tag{6}$$

If the candidate models are good approximation of the true DGP, under a set of regularity conditions, it is well known (e.g. Burnham and Anderson (2002) and Claeskens and Hjort (2008)) that AIC is an asymptotically unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right)\right]$, that is, as $n \to \infty$,

$$E_{\mathbf{y}}(\text{AIC}) - E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right) \to 0.$$

If we allow for model misspecification, TIC should be used instead and

$$E_{\mathbf{y}}(\text{TIC}) - E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right) \to 0.$$

The decision-theoretic justification of AIC and TIC rests on a frequentist framework. Specifically, it requires a careful choice of the KL divergence, the use of QMLE, and a set of regularity conditions that ensure $\sqrt{n}$-consistency and the asymptotic normality of QMLE. In Burnham and Anderson (2002) and other standard references,

$E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[ -2 \log p \left( \mathbf{y}_{rep} | \widehat{\boldsymbol{\theta}}_n(\mathbf{y}), M_k \right) \right]$ is decomposed into three terms:

$$
\begin{aligned}
& E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left( -2 \log p \left( \mathbf{y}_{rep} | \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) \right) \right) \\
= & \underbrace{\left[ E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left( -2 \log p \left( \mathbf{y}_{rep} | \widehat{\boldsymbol{\theta}}_n(\mathbf{y}_{rep}) \right) \right) \right]}_{(T_1)} \\
+ & \underbrace{\left[ E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left( -2 \log p \left( \mathbf{y}_{rep} | \boldsymbol{\theta}_n^p \right) \right) - E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left( -2 \log p \left( \mathbf{y}_{rep} | \widehat{\boldsymbol{\theta}}_n(\mathbf{y}_{rep}) \right) \right) \right]}_{(T_2)} \\
+ & \underbrace{\left[ E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left( -2 \log p \left( \mathbf{y}_{rep} | \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) \right) \right) - E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left( -2 \log p \left( \mathbf{y}_{rep} | \boldsymbol{\theta}_n^p \right) \right) \right]}_{(T_3)}.
\end{aligned} \tag{7}
$$

Clearly, $T_1 = E_{\mathbf{y}} \left( -2 \log p \left( \mathbf{y} | \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) \right) \right)$. To justify AIC, it is suffice to show that $T_2 = T_3 + o(1) = P + o(1)$, where $P$ is the dimension of parameter $\boldsymbol{\theta}$. TIC can be justified in the same manner.

The existing literature, exemplified by Burnham and Anderson (2002), employs the Taylor expansion to derive a quadratic term of the centered MLE. Relying on the asymptotic normality of QMLE, they establish that the expectation of this quadratic term converges to $P$. Nevertheless, in general, additional conditions are essential to guarantee the validity of this convergence, as elaborated upon below.

# 3 Rigorous Justification of AIC and TIC

In this section, we provide some high level conditions, which can be satisfied in many cases. Then under these high level conditions, we rigorously justify that AIC and TIC are asymptotically unbiased estimators of $E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[ -2 \log p \left( \mathbf{y}_{rep} | \widehat{\boldsymbol{\theta}}_n(\mathbf{y}), M_k \right) \right]$. For simplification, we omit $M_k$ in this section.

First we fix some notations. Let $\mathbf{y}^t = (y_0, y_1, \ldots, y_t)$ for any $0 \leq t \leq n$ and $l_t(\mathbf{y}^t, \boldsymbol{\theta}) = \log p(\mathbf{y}^t | \boldsymbol{\theta}) - \log p(\mathbf{y}^{t-1} | \boldsymbol{\theta})$ be the conditional log-likelihood for the $t^{th}$ observation for any $1 \leq t \leq n$. When there is no confusion, we suppress $l_t(\mathbf{y}^t, \boldsymbol{\theta})$ as $l_t(\boldsymbol{\theta})$

so that the log-likelihood function $\log p(\mathbf{y}|\boldsymbol{\theta})$ is $\sum_{t=1}^{n} l_t(\boldsymbol{\theta})$.[1] Let $\nabla^j l_t(\boldsymbol{\theta})$ denote the $j^{th}$ derivative of $l_t(\boldsymbol{\theta})$ and $\nabla^j l_t(\boldsymbol{\theta}) = l_t(\boldsymbol{\theta})$ when $j = 0$. Furthermore, define

$$\begin{aligned}
&\mathbf{s}\left(\mathbf{y}^t, \boldsymbol{\theta}\right) = \frac{\partial \log p\left(\mathbf{y}^t|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{t} \nabla l_i(\boldsymbol{\theta}),\ \mathbf{h}\left(\mathbf{y}^t, \boldsymbol{\theta}\right) = \frac{\partial^2 \log p\left(\mathbf{y}^t|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^{t} \nabla^2 l_i(\boldsymbol{\theta}),\\
&\mathbf{s}_t(\boldsymbol{\theta}) = \nabla l_t(\boldsymbol{\theta}) = \mathbf{s}\left(\mathbf{y}^t, \boldsymbol{\theta}\right) - \mathbf{s}\left(\mathbf{y}^{t-1}, \boldsymbol{\theta}\right),\ \mathbf{h}_t(\boldsymbol{\theta}) = \nabla^2 l_t(\boldsymbol{\theta}) = \mathbf{h}\left(\mathbf{y}^t, \boldsymbol{\theta}\right) - \mathbf{h}\left(\mathbf{y}^{t-1}, \boldsymbol{\theta}\right),\\
&\mathbf{B}_n(\boldsymbol{\theta}) = Var\left[\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \triangledown l_t(\boldsymbol{\theta})\right], \bar{\mathbf{H}}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^{n} \mathbf{h}_t(\boldsymbol{\theta}),\\
&\mathbf{H}_n(\boldsymbol{\theta}) = \int \bar{\mathbf{H}}_n(\boldsymbol{\theta})\, g(\mathbf{y})\, d\mathbf{y},\ \mathbf{J}_n(\boldsymbol{\theta}) = \int \bar{\mathbf{J}}_n(\boldsymbol{\theta})\, g(\mathbf{y})\, d\mathbf{y}.
\end{aligned}$$

For simplification, we write $\mathbf{H}_n(\boldsymbol{\theta}_n^p)$ as $\mathbf{H}_n$, $\mathbf{B}_n(\boldsymbol{\theta}_n^p)$ as $\mathbf{B}_n$, and let $\mathbf{C}_n = \mathbf{H}_n^{-1} \mathbf{B}_n \mathbf{H}_n^{-1}$.

We impose the following high level conditions.

**Assumption 1 (Differentiable):** The log-likelihood $\log p(\mathbf{y}^t|\boldsymbol{\theta})$ is second-order differentiable for every $t$ so that the second-order Taylor's expansion is allowed.

**Assumption 2 (Lipschitz)**: For any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, $\|\mathbf{h}_t(\boldsymbol{\theta}) - \mathbf{h}_t(\boldsymbol{\theta}')\| \leq c_t(\mathbf{y}^t) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$, where $c_t(\mathbf{y}^t)$ is a positive random variable with $\sup_t E\|c_t(\mathbf{y}^t)\| < \infty$ and

$$E\left|\frac{1}{n} \sum_{t=1}^{n} c_t\left(\mathbf{y}^t\right)\right|^4 \leqslant C < \infty. \tag{8}$$

**Assumption 3 (Moment Conditions):** The following moment conditions are satisfied:

$$E\|\sqrt{n}\bar{\mathbf{s}}_n(\boldsymbol{\theta}_n^p)\|^2 \leqslant C < \infty, \tag{9}$$

$$E\|\sqrt{n}[\bar{\mathbf{H}}_n(\boldsymbol{\theta}_n^p) - \mathbf{H}_n(\boldsymbol{\theta}_n^p)]\|^2 \leqslant C < \infty. \tag{10}$$

[1] In the definition of log-likelihood, we ignore the initial condition $\log p(y_0)$. For weakly dependent data, the impact of ignoring the initial condition is asymptotically negligible.

**Assumption 4 (Quasi-Maximum Likelihood Estimator):** $\widehat{\boldsymbol{\theta}}_n(\mathbf{y})$ satisfies:

$$E\|\sqrt{n}(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p)\|^4 \leqslant C < \infty, \tag{11}$$

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p) = \bar{\mathbf{H}}_n^{-1}\left(\tilde{\boldsymbol{\theta}}_n(\mathbf{y})\right)\sqrt{n}\bar{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right) = \mathbf{H}_n^{-1}\left(\boldsymbol{\theta}_n^p\right)\sqrt{n}\bar{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right) + RT_n(\mathbf{y}), \tag{12}$$

where $\tilde{\boldsymbol{\theta}}_n(\mathbf{y})$ lies between $\widehat{\boldsymbol{\theta}}_n(\mathbf{y})$ and $\boldsymbol{\theta}_n^p$, $E\|RT_n(\mathbf{y})\|^2 = o(1)$.

These conditions are slightly stronger than the standard regular conditions of quasi-maximum likelihood estimator and are satisfied under quite general situations, including iid data, stationary ergodic data, and weakly dependent data. We will discuss these cases in the next section.

The following lemma states that these high level conditions ensure the expectation of remainders in the second order Taylor expansion converges to zero.

**Lemma 3.1** *Under Assumptions 1-4, we have the following expansions for the log likelihood*

$$\begin{aligned}\log p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right) = & \log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)\right) \\ & + \frac{1}{2}\left(\boldsymbol{\theta}_n^p-\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)'\frac{\partial^2\log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\boldsymbol{\theta}_n^p-\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right) + RT_n^1(\mathbf{y}),\end{aligned} \tag{13}$$

*where $E_y|RT_n^1(\mathbf{y})|$ is $o(1)$. And*

$$\begin{aligned}\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)\right) = & \log p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right) + \frac{\partial\log p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}'}\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)-\boldsymbol{\theta}_n^p\right) \\ & + \frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)-\boldsymbol{\theta}_n^p\right)'\frac{\partial^2\log p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)-\boldsymbol{\theta}_n^p\right) + RT_n^2(\mathbf{y},\mathbf{y}_{rep}),\end{aligned} \tag{14}$$

*where $E_yE_{y_{rep}}|RT_n^2(\mathbf{y},\mathbf{y_{rep}})|$ is $o(1)$.*

**Remark 3.1** *It is obvious that the remainders are $o_p(1)$. It seems reasonable to ignore them when taking the expectation if we can apply the dominant convergence theorem*

*(DCT). However, it is unreasonable to directly assume the remainder can be dominated by an integrable random variable because of its complex structure. The existing literature does not discuss this problem in detail. For example, in Lemma 1 of Cavanaugh (1997), Theorem 1 of Shi and Tsai (1998), Chapter 2.1-2.3 of Claeskens and Hjort (2008), Appendix A.1 and A.2 of Linhart and Zucchini (1986), Chapter 7.2 of Burnham and Anderson (2002), Theorem 28.4 of Hansen (2022) all directly assume uniformly integrability. Here, we try to directly bound the expectation of remainders to give a rigorous justification to it.*

**Lemma 3.2** *Under Assumptions 1-4, we have*

$$E_y\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)-\boldsymbol{\theta}_n^p\right)'(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)\right)-\mathbf{H}_n)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)-\boldsymbol{\theta}_n^p\right)\right]=o(1), \tag{15}$$

$$E_{\mathbf{y}}\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)-\boldsymbol{\theta}_n^p\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)-\boldsymbol{\theta}_n^p\right)'\right]=\mathbf{H}_n^{-1}\mathbf{B}_n\mathbf{H}_n^{-1}+o\left(1\right)=\mathbf{C}_n+o\left(1\right). \tag{16}$$

**Remark 3.2** *The left-hand side of (15) is the expectation of an $o_p(1)$ term. To get the convergence of the expectation in the first equation, a careful treatment is needed. For the left-hand side of (16), intuitively it should converge to the asymptotic covariance matrix of MLE. However, the convergence in distribution in general does ensure not this convergence. The existing literature does not address this issue, neither. Lemma 3.2 gives a rigorous justification to both convergence results.*

With the aid of the above two lemmas, we are now in the position to justify AIC as an asymptotically unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right]$, which is the expected KL loss (up to a constant, which is only dependent on the true DGP).

**Theorem 3.1** *Under Assumptions 1-4, we have, as $n\to\infty$,*

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right]=E_{\mathbf{y}}\left(TIC\right)+o(1), \tag{17}$$

*where* $TIC = -2\log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right) + 2tr(-\mathbf{B}_n\mathbf{H}_n^{-1})$. *If we further assume the candidate model is a good approximation of the true data generating process:* $\mathbf{B}_n + \mathbf{H}_n = o(1)$, *then*

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right] = E_{\mathbf{y}}\left(AIC\right) + o(1), \tag{18}$$

*where* $AIC = -2\log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right) + 2P$.

**Proof.** Recall (7), $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right] = T_1 + T_2 + T_3$.

Note that $T_1 = E_{\mathbf{y}}\left[-2\log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right]$. Lemma 3.1 and Lemma 3.2 imply

$$T_2 = T_3 + o(1) = \mathrm{tr}(-\mathbf{B}_n\mathbf{H}_n^{-1}) + o(1),$$

then we have (17). Furthermore, when candidate model is a good approximation of the true data generating process, $T_2 = T_3 + o(1) = P + o(1)$, then we have (18). ■

# 4 Discussion

In this section, we discuss our Assumptions 1-4 under different type data, including independent and identical distributed, stationary ergodic and weakly dependent cases.

For Assumption 1, the differentiability of log-likelihood is standard in the quasi-maximum likelihood estimator.

For Assumption 2, the Lipschitz condition for the Hessian and the score are widely used in literature. For example, Theorem 3.7 in Gallant and White (1988) used the Lipschitz condition to derive the uniform LLN, which is important for ensuing consistency of extreme estimators. The Lipschitz condition was also used in Li et al. (2020) to develop Deviance information criterion for latent variable models and misspecified models. Our Assumption 2 follows the literature but slightly strengthens the moment condition to ensure the remainder is negligible.

For Assumption 3, we have corresponding central limit theory for $\sqrt{n}\bar{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)$ and $\sqrt{n}(\bar{\mathbf{H}}_n\left(\boldsymbol{\theta}_n^p\right)-\mathbf{H}_n\left(\boldsymbol{\theta}_n^p\right))$ under independent and identical distributed, stationary ergodic and weakly dependent data, see Davidson (2021) for details. Then $\sqrt{n}\bar{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)$ and $\sqrt{n}(\bar{\mathbf{H}}_n\left(\boldsymbol{\theta}_n^p\right)-\mathbf{H}_n\left(\boldsymbol{\theta}_n^p\right))$ are both $O_p(1)$. The moment conditions are reasonable because of the appropriate order.

Assumption 4 is slightly stronger than that in the classical quasi-maximum likelihood theory, which provides

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p)=\mathbf{H}_n^{-1}\left(\boldsymbol{\theta}_n^p\right)\sqrt{n}\bar{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)+o_p(1)=O_p(1). \tag{19}$$

Assumption 4 in fact requires more moment conditions for $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p)$ and the remainder term $RT_n(\mathbf{y})$. Some moment inequality for the sum of a sequence of random variables will be useful to show the moment conditions hold. With some basic moment conditions, Assumption 4 can be verified under different data type, including independent and identical distributed, stationary ergodic and weakly dependent cases.

When $\mathbf{y}=(y_1,...,y_n)$ is independent and identical distributed or stationary ergodic, the Burkholder inequality for martingale will be useful to bound high order moment of $\sqrt{n}\bar{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)$, see Theorem 16.24 in Davidson (2021). For weakly dependent date, the concept of mixing and strong near epoch dependence can be assumed, see Lin (2004).

The following lemma may be useful to verify the moment conditions in Assumption 4.

**Lemma 4.1** *Suppose Assumptions 1-3 hold,* $\sup_{t\geq 1}E\|\mathbf{s}_t\left(\boldsymbol{\theta}_n^p\right)\|^8\leq C<\infty$*, and one of the following conditions hold:*

*(i)* $\{y_1,y_2,...\}$ *is independent and identical distributed or stationary ergodic, or*

*(ii)* $\{y_1,y_2,...\}$ *is a* $\phi$*-mixing sequence with* $\phi(m)=O\left((\log m)^{-4(1+\delta)}\right)$, $\{\mathbf{s}_t\left(\boldsymbol{\theta}_n^p\right),t\geq 1\}$ *is strong* $L_8$*-near epoch dependent sequence on* $\{y_1,y_2,...\}$ *with* $\{d_t\}$ *and* $\{v\left(m\right)\}$

*satisfying*

$$\limsup_{n^*\to\infty} \sup_{k\geq 0} \sum_{j=1}^{n^*} d_{k+j}^2/n^* = B < \infty,$$

*and* $v\left(m\right) = O\left(\left(\log m\right)^{-(1+\delta/2)}\right)$ *for some* $\delta > 0$,

*then* $E_{\mathbf{y}} \left\|\sqrt{n}\bar{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)\right\|^8$*=O(1). Further, if we have*

*(iii)* $E_{\mathbf{y}} \left\|\mathbf{H}_n - \bar{\mathbf{H}}_n\left(\tilde{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right\|^4 = o(1)$,

*then the moment condition in Assumption 4 holds.*

This lemma can be applied to weakly dependent data, which is quite general in application. $\phi$-mixing assumption is standard for weakly dependent data, and strong near epoch dependence property is proposed by Lin (2004), which is a strengthen of near epoch dependence property in Gallant and White (1988). The condition (iii) in Lemma 4.1 can also be verified by the same argument as the first part of this lemma under suitable moment condition and strong near epoch dependence order of the Hessian matrix sequence $\{\mathbf{h}_t\left(\boldsymbol{\theta}_n^p\right), t \geq 1\}$.

# Appendix

The appendix contains the proof details of Theorem 1 and Lemmas.

**Proof of Theorem 3.1.** We decompose $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right]$ into three terms:

$$\begin{aligned}
& E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right) \\
= & \left[\underbrace{E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y}_{rep})\right)\right)}_{(T_1)}\right] \\
+ & \left[\underbrace{E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\log p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)\right) - E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y}_{rep})\right)\right)}_{(T_2)}\right] \\
+ & \left[\underbrace{E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right) - E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\log p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)\right)}_{(T_3)}\right].
\end{aligned}$$

Note that $T_1 = E_{\mathbf{y}}\left(-2\log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right)$. Now let us analyze $T_2$ and $T_3$.

$$\begin{aligned}
T_2 &= \left[E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\log p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)\right) - E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y}_{rep})\right)\right)\right] \\
&= -2E_{\mathbf{y}}\left[\log p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right) - \log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right] \\
&= -2E_{\mathbf{y}}\left[\frac{1}{2}\left(\boldsymbol{\theta}_n^p - \widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)' \frac{\partial^2 \log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\boldsymbol{\theta}_n^p - \widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right) + RT_n^1(\mathbf{y})\right] \qquad (20) \\
&= E_{\mathbf{y}}\left[-\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)'\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right] + o(1) \\
&= E_{\mathbf{y}}\left[-\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)'\mathbf{H}_n\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right] + o(1).
\end{aligned}$$

The third equality and the fourth equality hold from (13) in Lemma 3.1. The last step comes from (15) in Lemma 3.2. Now we turn to $T_3$.

$$\begin{aligned}
T_3 &= \left[E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right) - E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\log p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)\right)\right] \\
&= -2E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right) - \log p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)\right] \\
&= -2E_{\mathbf{y}}\left[\frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)' E_{\mathbf{y}_{rep}}\frac{\partial^2 \log p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right) + RT_n^2(\mathbf{y}, \mathbf{y}_{rep})\right] \\
&= E_{\mathbf{y}}\left[-\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)'\mathbf{H}_n\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right] + o(1)
\end{aligned} \qquad (21)$$

The third equality and the fourth equality hold from (14) in Lemma 1.1 and the definition of $\mathbf{H}_n$. Then we can use (16) in Lemma 1.2

$$\begin{aligned}
&E_{\mathbf{y}}\left[-\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)'\mathbf{H}_n\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right] \\
=& -\mathbf{tr}\left[\mathbf{H}_n E_{\mathbf{y}}\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)'\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right]\right] \\
=& -\mathbf{tr}\left[\mathbf{H}_n\left(\mathbf{H}_n^{-1}\mathbf{B}_n\mathbf{H}_n^{-1} + o(1)\right)\right] \\
=& -\mathbf{tr}\left[\mathbf{B}_n\mathbf{H}_n^{-1}\right] + o(1).
\end{aligned}$$

This means $T_2 = T_3 + o(1) = -\mathbf{tr}\left[\mathbf{B}_n\mathbf{H}_n^{-1}\right] + o(1)$. For TIC, we have

$$\begin{aligned}
E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right] &= T_1 + T_2 + T_3 \\
&= E_{\mathbf{y}}\left(-2\log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right) + 2\mathbf{tr}\left[-\mathbf{B}_n\mathbf{H}_n^{-1}\right] + o(1) \\
&= E_{\mathbf{y}}\left(-2\log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right) + 2\mathbf{tr}\left[-\mathbf{B}_n\mathbf{H}_n^{-1}\right]\right) + o(1) \\
&= E_{\mathbf{y}}\left(TIC\right) + o(1).
\end{aligned}$$

If the model is a good approximation of the true DGP, i.e., $\mathbf{H}_n + \mathbf{B}_n = o(1)$, then

$$T_2 = T_3 + o(1) = -\mathbf{tr}\left[-\mathbf{H}_n\mathbf{H}_n^{-1}\right] + o(1) = P + o(1). \tag{22}$$

Then we finally justify AIC by

$$\begin{aligned}
E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\log p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right] &= T_1 + T_2 + T_3 \\
&= E_{\mathbf{y}}\left(-2\log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right) + 2P + o(1) \\
&= E_{\mathbf{y}}\left(-2\log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right) + 2P\right) + o(1) \\
&= E_{\mathbf{y}}\left(AIC\right) + o(1).
\end{aligned}$$

■

**Proof of Lemma 3.1.** For the first result of Lemma 3.1, consider the following Taylor expansion

$$\begin{aligned}
\log p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right) = \log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)\right) + \frac{\partial \log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)}{\partial \boldsymbol{\theta}'}\left(\boldsymbol{\theta}_n^p - \widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)\right) \\
+ \frac{1}{2}\left(\boldsymbol{\theta}_n^p - \widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)' \frac{\partial^2 \log p\left(\mathbf{y}|\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y})\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\left(\boldsymbol{\theta}_n^p - \widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right),
\end{aligned} \tag{23}$$

where $\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y})$ lies between $\widehat{\boldsymbol{\theta}}_n(\mathbf{y})$ and $\boldsymbol{\theta}_n^p$. Note that $\partial \log p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)/\partial\boldsymbol{\theta} = 0$, we have

$$\begin{aligned} RT_n^1(\mathbf{y}) &= \frac{1}{2}\left(\boldsymbol{\theta}_n^p - \widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)'\left[\frac{\partial^2 \log p\left(\mathbf{y}|\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y})\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{\partial^2 \log p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\left(\boldsymbol{\theta}_n^p - \widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right) \\ &= \frac{1}{2}\sqrt{n}\left(\boldsymbol{\theta}_n^p - \widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)'\left[\bar{\mathbf{H}}_n\left(\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y})\right) - \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right]\sqrt{n}\left(\boldsymbol{\theta}_n^p - \widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right). \end{aligned} \tag{24}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} E_y|RT_n^1(\mathbf{y})| &\leqslant E_y\left(\left\|\bar{\mathbf{H}}_n\left(\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y})\right) - \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right\|\left\|\sqrt{n}\left(\boldsymbol{\theta}_n^p - \widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right\|^2\right) \\ &\leqslant \left(E_y\left\|\bar{\mathbf{H}}_n\left(\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y})\right) - \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right\|^2\right)^{1/2}\left(E_y\left\|\sqrt{n}\left(\boldsymbol{\theta}_n^p - \widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right\|^4\right)^{1/2}. \end{aligned} \tag{25}$$

By the Lipschitz condition in **Assumption 2**, we have

$$\begin{aligned} \left\|\bar{\mathbf{H}}_n\left(\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y})\right) - \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right\|^2 &\leqslant \left|\frac{1}{n}\sum_{t=1}^n c_t\left(\mathbf{y}^t\right)\right|^2\left\|\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y}) - \widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right\|^2 \\ &\leqslant \frac{1}{n}\left|\frac{1}{n}\sum_{t=1}^n c_t\left(\mathbf{y}^t\right)\right|^2\left\|\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right\|^2. \end{aligned} \tag{26}$$

Use the Cauchy-Schwarz inequality and moment conditions (8) and (11), we have

$$\begin{aligned} E_y\left\|\bar{\mathbf{H}}_n\left(\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y})\right) - \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right\|^2 &\leqslant \frac{1}{n}\left(E_y\left|\frac{1}{n}\sum_{t=1}^n c_t\left(\mathbf{y}^t\right)\right|^4\right)^{1/2}\left(E_y\left\|\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right\|^4\right)^{1/2} \\ &\leqslant C_0^{1/2}C^{1/2}/n \to 0. \end{aligned} \tag{27}$$

Combining (11), (25) and (27), we have

$$E_y|RT_n^1(\mathbf{y})| \leqslant C^{1/2}\left(E_y\left\|\bar{\mathbf{H}}_n\left(\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y})\right)-\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right\|^2\right)^{1/2} \to 0.$$

Thus, we get (13), which is the first result of Lemma 3.1.

For the second result of Lemma 3.1 (equation (14)), note that

$$RT_n^2(\mathbf{y},\mathbf{y}_{rep}) = \frac{1}{2}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}_{rep})-\boldsymbol{\theta}_n^p\right)'\left[\bar{\mathbf{H}}_n\left(\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y})\right)-\bar{\mathbf{H}}_n\left(\boldsymbol{\theta}_n^p\right)\right]\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}_{rep})-\boldsymbol{\theta}_n^p\right), \tag{28}$$

where $\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y})$ lies between $\widehat{\boldsymbol{\theta}}_n(\mathbf{y})$ and $\boldsymbol{\theta}_n^p$. Use the same argument, we have

$$\begin{aligned}
&E_yE_{y_{rep}}|RT_n^2(\mathbf{y},\mathbf{y}_{rep})| \\
\leqslant&\left(E_yE_{y_{rep}}\left\|\bar{\mathbf{H}}_n\left(\tilde{\boldsymbol{\theta}}_n^*(\mathbf{y})\right)-\bar{\mathbf{H}}_n\left(\boldsymbol{\theta}_n^p\right)\right\|^2\right)^{1/2}\left(E_yE_{y_{rep}}\left\|\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\right\|^4\right)^{1/2} \\
\leqslant&(C_0^{1/2}C^{1/2}/n)^{1/2}C^{1/2} \to 0.
\end{aligned} \tag{29}$$

This is the second result in Lemma 3.1. ■

**Proof of Lemma 3.2.** For the left-hand side of (15), use the Cauchy-Schwarz inequality and the moment condition (11), we have

$$\begin{aligned}
&E_y\left|\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)-\boldsymbol{\theta}_n^p\right)'(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)\right)-\mathbf{H}_n)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)-\boldsymbol{\theta}_n^p\right)\right| \\
\leqslant&E_y\left[\left\|\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)\right)-\mathbf{H}_n\right\|\left\|\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)-\boldsymbol{\theta}_n^p\right)\right\|^2\right] \\
\leqslant&\left(E_y\left\|\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)\right)-\mathbf{H}_n\right\|^2\right)^{1/2}\left(E_y\left\|\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)-\boldsymbol{\theta}_n^p\right)\right\|^4\right)^{1/2} \\
\leqslant&C^{1/2}\left(E_y\left\|\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}\right)\right)-\mathbf{H}_n\right\|^2\right)^{1/2}.
\end{aligned} \tag{30}$$

Note that $(x+y)^2 \leqslant 2(x^2+y^2)$, we have

$$E_y \left\| \bar{\mathbf{H}}_n \left( \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) \right) - \mathbf{H}_n \right\|^2 \leqslant 2 \left( E_y \left\| \bar{\mathbf{H}}_n \left( \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) \right) - \bar{\mathbf{H}}_n \left( \boldsymbol{\theta}_n^p \right) \right\|^2 + E_y \left\| \bar{\mathbf{H}}_n \left( \boldsymbol{\theta}_n^p \right) - \mathbf{H}_n \right\|^2 \right). \tag{31}$$

By the Lipschitz condition and moment condition (8) in **Assumption 2**, we have

$$\begin{aligned} E_y \left\| \bar{\mathbf{H}}_n \left( \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) \right) - \bar{\mathbf{H}}_n \left( \boldsymbol{\theta}_n^p \right) \right\|^2 &\leqslant E_y \left| \frac{1}{n} \sum_{t=1}^{n} c_t \left( \mathbf{y}^t \right) \right|^2 \left\| \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p \right\|^2 \\ &\leqslant \frac{1}{n} \left( E_y \left| \frac{1}{n} \sum_{t=1}^{n} c_t \left( \mathbf{y}^t \right) \right|^4 \right)^{1/2} \left( E_y \left\| \sqrt{n} \left( \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p \right) \right\|^4 \right)^{1/2} \\ &\leqslant C_0^{1/2} C^{1/2}/n \to 0. \end{aligned} \tag{32}$$

By the moment conditions in **Assumption 4**, we have

$$E_y \left\| \bar{\mathbf{H}}_n \left( \boldsymbol{\theta}_n^p \right) - \mathbf{H}_n \right\|^2 \leqslant C/n \to 0. \tag{33}$$

Thus, $\lim_{n\to\infty} E_y \left\| \bar{\mathbf{H}}_n \left( \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) \right) - \mathbf{H}_n \right\|^2 = 0$ by (31), (32) and (33). Then we get the bound of (30):

$$\begin{aligned} & E_y \left| \sqrt{n} \left( \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p \right)' (\bar{\mathbf{H}}_n \left( \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) \right) - \mathbf{H}_n) \sqrt{n} \left( \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p \right) \right| \\ \leqslant & C^{1/2} \left( E_y \left\| \bar{\mathbf{H}}_n \left( \widehat{\boldsymbol{\theta}}_n(\mathbf{y}) \right) - \mathbf{H}_n \right\|^2 \right)^{1/2} \to 0. \end{aligned}$$

For the left-hand side of (16), use the MLE expression (12) in Assumption 3, we have

$$\begin{aligned}&E_{\mathbf{y}}\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)'\right]\\ =&E_y\left[\left(\mathbf{H}_n^{-1}\sqrt{n}\overline{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)+RT_n(\mathbf{y})\right)\left(\mathbf{H}_n^{-1}\sqrt{n}\overline{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)+RT_n(\mathbf{y})\right)'\right]\\ =&\mathbf{H}_n^{-1}E_y\left[\sqrt{n}\overline{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)\sqrt{n}\overline{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)'\right]\mathbf{H}_n^{-1}+E_y\left[RT_n(\mathbf{y})RT_n(\mathbf{y})'\right]\\ &+\mathbf{H}_n^{-1}E_y\left[\sqrt{n}\overline{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)RT_n(\mathbf{y})'\right]+E_y\left[RT_n(\mathbf{y})\sqrt{n}\overline{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)'\right]\mathbf{H}_n^{-1}.\end{aligned}\tag{34}$$

For the first term of (34), by the definition of $\mathbf{B}_n$, we have

$$\mathbf{H}_n^{-1}E_y\left[\sqrt{n}\overline{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)\sqrt{n}\overline{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)'\right]\mathbf{H}_n^{-1}=\mathbf{H}_n^{-1}\mathbf{B}_n\mathbf{H}_n^{-1}.$$

For the second term of (34), by the moment condition of $RT_n(\mathbf{y})$ in **Assumption 3**, we have

$$E_y\left\|RT_n(\mathbf{y})RT_n(\mathbf{y})'\right\|\leqslant E_y\left\|RT_n(\mathbf{y})\right\|^2\to 0.$$

For the third term and the fourth term of (34), using the Cauchy-Schwarz inequality and the moment conditions (9) in **Assumption 4**, we have

$$\begin{aligned}E_y\left\|\sqrt{n}\overline{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)RT_n(\mathbf{y})'\right\|&\leqslant\left(E_y\left\|\sqrt{n}\overline{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)\right\|^2\right)^{1/2}\left(E_y\left\|RT_n(\mathbf{y})\right\|^2\right)^{1/2}\\ &\leqslant C^{1/2}\left(E_y\left\|RT_n(\mathbf{y})\right\|^2\right)^{1/2}\to 0.\end{aligned}$$

Thus, we have $E_{\mathbf{y}}\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)'\right]=\mathbf{H}_n^{-1}\mathbf{B}_n\mathbf{H}_n^{-1}+o(1)$. ■

**Proof of Lemma 4.1.** Let $\mathbf{s}_{ni}\left(\boldsymbol{\theta}_n^p\right)$ denote the $i$th element of $\mathbf{s}_n\left(\boldsymbol{\theta}_n^p\right)$. When (i) holds, it can be shown that $\mathbf{s}_{ni}\left(\boldsymbol{\theta}_n^p\right)$ is an mds under the correctly specified model. The following Burkholder Inequality is useful; see Theorem 16.24 in Davidson (2021).

**Lemma 4.2** *(Burkholder Inequality) For a martingale $S_n$ with $S_0=0$, and increments*

*$X_t = S_t - S_{t-1}$, define the square function as*

$$Q\left(S_n\right) = \left(\sum_{t=1}^{n} X_t^2\right)^{1/2}.$$

*If $\{S_n, F_n\}_1^\infty$ is an $L_r$ bounded martingale for $r > 1$, then*

$$c_r \left\|Q\left(S_n\right)\right\|_r \leq \left\|S_n\right\|_r \leq C_r \left\|Q\left(S_n\right)\right\|_r,$$

*where $\left\|Q\left(S_n\right)\right\|_r = \left(E\left[\left|Q\left(S_n\right)\right|^r\right]\right)^{1/r}$, $c_r = \left(18r^{3/2}/\left(r-1\right)\right)^{-1}$ and $C_r = 18r^{3/2}/\left(r-1\right)^{1/2}$.*

**Remark 4.1** *Note that*

$$\begin{aligned}
\left\|Q\left(S_n\right)\right\|_r &= \left(E\left[\left|Q\left(S_n\right)\right|^r\right]\right)^{1/r} = \left(E\left[\left|\left(\sum_{t=1}^{n} X_t^2\right)^{1/2}\right|^r\right]\right)^{1/r} = \left(E\left[\left(\sum_{t=1}^{n} X_t^2\right)^{r/2}\right]\right)^{1/r} \\
&= \left[\left(E\left[\left(\sum_{t=1}^{n} \left|X_t\right|^2\right)^{r/2}\right]\right)^{2/r}\right]^{1/2} = \left\|\sum_{t=1}^{n} \left|X_t\right|^2\right\|_{L_{r/2}}^{1/2} \\
&\leq \left(\sum_{t=1}^{n} \left\|X_t^2\right\|_{L_{r/2}}\right)^{1/2} = \left[\sum_{t=1}^{n} \left(E\left|X_t\right|^r\right)^{2/r}\right]^{1/2} = O\left(n^{1/2}\right).
\end{aligned}$$

*If $E\left|X_t\right|^r$ is bounded for all $t$, we have*

$$\left\|S_n\right\|_r = \left(E\left[\left|S_n\right|^r\right]\right)^{1/r} \leq C_r \left\|Q\left(S_n\right)\right\|_r = O\left(n^{1/2}\right). \tag{35}$$

Thus, if we assume $E|\mathbf{s}_{ni}\left(\boldsymbol{\theta}_n^p\right)|^8$ is bounded for all $t$, by Lemma 4.2 and (35), we have

$$E\left[\left|\frac{1}{\sqrt{n}}\frac{\partial \log p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right)}{\partial \boldsymbol{\theta}_i}\right|^8\right] = \frac{1}{n^4} E\left[\left|\sum_{t=1}^{n} \mathbf{s}_{ni}\left(\boldsymbol{\theta}_n^p\right)\right|^8\right] \leq \frac{1}{n^4} O\left(n^4\right) = O(1).$$

Now we have $E_{\mathbf{y}}\left\|\sqrt{n}\bar{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)\right\|^8 = O(1)$.

When (ii) holds, let us first introduce the concept of strong near epoch dependence;

see Lin (2004).

**Definition 4.1** *Let $p > 0$, $\{X_t, t \geq 1\}$ is called a strong $L_p-$near epoch dependent sequence if there exist sequences $\{d_t\}$ and $\{v(m)\}$ of nonnegative constants, $v(m) \to 0$ as $m \to \infty$, such that for all $k \geq 0$, $n^* \geq 1$ and $m \geq 0$,*

$$\left[E\left(\left\|S_k(n^*) - E_{k+1-m}^{k+n^*+m} S_k(n^*)\right\|^p\right)\right]^{1/p} \leq v(m)\left(\sum_{j=1}^{n^*} d_{k+j}^2\right)^{1/2},$$

*where $S_k(n^*) = \sum_{j=k+1}^{k+n^*} X_j$.*

**Lemma 4.3** *Let $\{V_t, t \geq 1\}$ be a $\phi$-mixing sequence with $\varphi(m) = O\left((\log m)^{-p(1+\delta/2)}\right)$ and let $\{X_t, t \geq 1\}$ be a mean zero $L_p$ bounded and strong $L_p-$near epoch dependent sequence on $\{V_t, t \geq 1\}$, $p > 2$, with $\{d_t\}$ and $\{v(m)\}$ satisfying*

$$\lim_{n^* \to \infty} \sup \sup_{k \geq 0} \sum_{j=1}^{n^*} d_{k+j}^2 / n^* = B < \infty,$$

*and $v(m) = O\left((\log m)^{-(1+\delta/2)}\right)$ for some $\delta > 0$. Then there exists a finite constant $C$, depending only on $\{\varphi(\cdot)\}$ and $\{v(\cdot)\}$, such that for all positive integers $k$ and $n$,*

$$E\left(\max_{1 \leq i \leq n} |S_k(i)|^p\right) \leq C(Dn)^{p/2},$$

*where $D = \max\left\{B, \sup_n \left[E\left(|X_n|^p\right)\right]^{1/p}\right\}$.*

This lemma is proposed by Lin (2004). Use Lemma 4.3, we can show that

$$\begin{aligned} E\left[\left|\frac{1}{\sqrt{n}} \frac{\partial \log p\left(\mathbf{y}_{rep} | \boldsymbol{\theta}_n^p\right)}{\partial \boldsymbol{\theta}_i}\right|^8\right] &= \frac{1}{n^4} E\left[\max_{1 \leq i \leq n}\left|\sum_{t=1}^n \mathbf{s}_{ni}\left(\boldsymbol{\theta}_n^p\right)\right|^8\right] \leq \frac{1}{n^4} E\left[\left|\sum_{t=1}^n \mathbf{s}_{ni}\left(\boldsymbol{\theta}_n^p\right)\right|^8\right] \qquad (36) \\ &\leq \frac{1}{n^4} C(Dn)^4 = CD^4 < \infty, \end{aligned}$$

for some finite constant $C$, $D$. Then we have $E_{\mathbf{y}}\left\|\sqrt{n}\overline{\mathbf{s}}_n\left(\boldsymbol{\theta}_n^p\right)\right\|^8$ is $O(1)$.

Now we have proved that under (i) or (ii), we have $E_{\mathbf{y}} \left\| \sqrt{n}\bar{\mathbf{s}}_n \left(\boldsymbol{\theta}_n^p\right) \right\|^8 = O(1)$.

To obtain the moment conditions in Assumption 4, recall that the first order condition of quasi-maximum likelihood estimator is

$$0 = \bar{\mathbf{s}}_n \left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right) = \bar{\mathbf{s}}_n \left(\boldsymbol{\theta}_n^p\right) + \bar{\mathbf{H}}_n \left(\tilde{\boldsymbol{\theta}}_n(\mathbf{y})\right) (\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p). \tag{37}$$

Then we have the following expansion for quasi-maximum likelihood estimator

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p) = \bar{\mathbf{H}}_n^{-1} \left(\tilde{\boldsymbol{\theta}}_n(\mathbf{y})\right) \sqrt{n}\bar{\mathbf{s}}_n \left(\boldsymbol{\theta}_n^p\right) = \mathbf{H}_n^{-1} \sqrt{n}\bar{\mathbf{s}}_n \left(\boldsymbol{\theta}_n^p\right) + RT_n(\mathbf{y}).$$

To verify the moment conditions in Assumption 4, we need to show $E_{\mathbf{y}} \left\| \sqrt{n}(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p) \right\|^4 = O(1)$ and $E_{\mathbf{y}} \|RT_n(\mathbf{y})\|^2 = o(1)$.

Given the first part result and Cauchy-Schwartz inequality, we have

$$E_{\mathbf{y}} \left\| \sqrt{n}(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p) \right\|^4 \leqslant \left( E_{\mathbf{y}} \left\| \bar{\mathbf{H}}_n^{-1} \left(\tilde{\boldsymbol{\theta}}_n(\mathbf{y})\right) \right\|^8 \right)^{1/2} \left( E_{\mathbf{y}} \left\| \sqrt{n}\bar{\mathbf{s}}_n \left(\boldsymbol{\theta}_n^p\right) \right\|^8 \right)^{1/2} = O(1).$$

Now we turn to the remainder term. The remainder term can be expressed as

$$\begin{aligned} RT_n(\mathbf{y}) &= \bar{\mathbf{H}}_n^{-1} \left(\tilde{\boldsymbol{\theta}}_n(\mathbf{y})\right) \sqrt{n}\bar{\mathbf{s}}_n \left(\boldsymbol{\theta}_n^p\right) - \mathbf{H}_n^{-1} \sqrt{n}\bar{\mathbf{s}}_n \left(\boldsymbol{\theta}_n^p\right) \\ &= \mathbf{H}_n^{-1} \left[ \mathbf{H}_n - \bar{\mathbf{H}}_n \left(\tilde{\boldsymbol{\theta}}_n(\mathbf{y})\right) \right] \bar{\mathbf{H}}_n^{-1} \left(\tilde{\boldsymbol{\theta}}_n(\mathbf{y})\right) \bar{\mathbf{s}}_n \left(\boldsymbol{\theta}_n^p\right) \\ &= \mathbf{H}_n^{-1} \left[ \mathbf{H}_n - \bar{\mathbf{H}}_n \left(\tilde{\boldsymbol{\theta}}_n(\mathbf{y})\right) \right] \sqrt{n}(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p). \end{aligned}$$

then we have

$$\begin{aligned} E_{\mathbf{y}} \|RT_n(\mathbf{y})\|^2 &\leqslant \left\|\mathbf{H}_n^{-1}\right\|^2 E_{\mathbf{y}} \left\| \mathbf{H}_n - \bar{\mathbf{H}}_n \left(\tilde{\boldsymbol{\theta}}_n(\mathbf{y})\right) \right\|^2 \left\| \sqrt{n}(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p) \right\|^2 \\ &\leqslant \left\|\mathbf{H}_n^{-1}\right\|^2 \left( E_{\mathbf{y}} \left\| \mathbf{H}_n - \bar{\mathbf{H}}_n \left(\tilde{\boldsymbol{\theta}}_n(\mathbf{y})\right) \right\|^4 \right)^{1/2} \left( E_{\mathbf{y}} \left\| \sqrt{n}(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p) \right\|^4 \right)^{1/2}. \end{aligned}$$

Combined with $E_{\mathbf{y}} \left\| \sqrt{n}(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p) \right\|^4 = O(1)$, it is suffice to show $E_{\mathbf{y}} \left\| \mathbf{H}_n - \bar{\mathbf{H}}_n \left(\tilde{\boldsymbol{\theta}}_n(\mathbf{y})\right) \right\|^4 =$

$o(1)$, which is given as a condition.

■

# References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, 1:267–281.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer Series in Statistics. Springer New York, New York, NY.

Burnham, K. P. and Anderson, D. R., editors (2002). *Model Selection and Multimodel Inference.* Springer New York, New York, NY.

Cavanaugh, J. E. (1997). Unifying the derivations for the akaike and corrected akaike information criteria. *Statistics & Probability Letters*, 33(2):201–208.

Cavanaugh, J. E. and Neath, A. A. (2019). The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*, 11(3):e1460.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging.* Cambridge University Press, 1 edition.

Davidson, J. (2021). *Stochastic Limit Theory: An Introduction for Econometricians.* Oxford University PressOxford, 2 edition.

Gallant, A. R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models.* Blackwell, Oxford, UK, 1. publ edition.

Hansen, B. E. (2022). *Econometrics.* Princeton University Press, Princeton.

Li, Y., Yu, J., and Zeng, T. (2020). Deviance information criterion for latent variable models and misspecified models. *Journal of Econometrics*, 216(2):450–493.

Lin, Z. (2004). Strong near-epoch dependence. *Science in China Series A*, 47(4):497.

Linhart, H. and Zucchini, W. (1986). *Model Selection*. Wiley Series in Probability and Mathematical Statistics Applied Probability and Statistics. Wiley, New York.

Shi, P. and Tsai, C.-L. (1998). A note on the unification of the akaike information criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(3):551–558.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 39(1):44–47.

Takeuchi, K. (1976). Distribution of information statistics and a criterion of model fitting. *Mathematical Science*, 153:12–18.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

White, H. (1996). *Estimation, Inference and Specification Analysis*. Number 22 in Econometric Society Monographs. Cambridge Univ. Press, Cambridge, 1. paperback ed edition.