Model Selection in the Presence of Incidental Parameters

Yoonseok Lee*

University of Michigan

February 2012

Abstract

This paper considers model selection of nonlinear panel data models in the presence of incidental parameters. The main interest is in selecting the model that approximates best the structure with the common parameters after concentrating out the incidental parameters. New model selection information criteria are developed, either using the Kullback-Leibler information criterion based on the profile likelihood or using the Bayes factor based on the integrated likelihood with robust priors. These model selection criteria impose heavier penalties than those of the standard information criteria such as AIC and BIC. The additional penalty, which is data-dependent, properly reflects the model complexity from the incidental parameters. As a particular example, a lag order selection criterion is examined in the context of dynamic panel models with fixed individual effects, and it is illustrated how the over/under selection probabilities are controlled for.

Keywords: (Adaptive) model selection, incidental parameters, profile likelihood, Kullback-Leibler information, Bayes factor, integrated likelihood, robust prior, model complexity, fixed effects, lag order.

JEL Classifications: C23, C52

*University of Michigan. Address: Department of Economics, University of Michigan, 611 Tappan Street, Ann Arbor, MI 48109-1220. E-mail: yoolee@umich.edu.
1 Introduction

As available data get richer, it is possible to consider more sophisticated econometrics models, such as semiparametric models and large dimensional parametric models including heterogeneous effects in panel data models. In order to have valid inferences and policy implications, proper model selection is crucial, based on which the chosen model describes the data generating process correctly or most closely. For model selection, model specification tests and information-criterion-based model selection are two approaches. While the model specification test approach requires ad hoc null and alternative models, the model selection approach considers all the available candidate models and chooses what gives the minimal value of a proper information criterion. Examples of the model selection criteria are Akaike information criterion (AIC), Bayesian information criterion (BIC), posterior information criterion (PIC), Hannan-Quinn (HQ) criterion, Mellows’ $C_p$ criterion, bootstrap criteria and cross-validation approaches.

One of the important assumptions among these model selection procedures is that the number of parameters in each candidate model needs to be finite or at most growing very slowly comparing to the sample size. For example, Stone (1979) points that consistency of the standard BIC no longer holds when the number of parameters $k$ in the candidate model diverges with the sample size $N$. Such limitations in the standard model selection criteria is well understood and several approaches have been proposed for the model selection problem in the large dimensional models, particularly in the Bayesian statistics framework. Examples are Berger et al. (2003) and Chkrabarti and Ghosh (2006), who analyze the Laplace approximation for the exponential family under $k, N \to \infty$ to consistently estimate the Bayes factor.

In many cases, the large dimensional parameter problem is due to the many nuisance parameters, which are not of the main interest but required for a proper model specification or for handling omitted variables. Different from the aforementioned approaches in the Bayesian framework, which normally assumes the prior distribution of the nuisance parameters and uses the integration method to eliminate them (e.g., Berger et al. (1999)), this paper handles many (nuisance) parameter problem using the profile likelihood approach as a way to eliminate nuisance parameters. Most of the panel literature is based on the profile likelihood (e.g., the within-transformation approach in the linear fixed-effect panel regressions) and the main concern is to modify the profile likelihood as a way of bias reduction in the maximum likelihood estimator, which basically presumes that the parametric models are correctly specified (e.g., Hahn and Kuersteiner (2002, 2007); Hahn and Newey (2004); Arellano and Hahn (2005, 2006); Lee (2006, 2011, 2008); Bester and Hansen (2009)). This paper, on the other hand, focuses on the specification problem.

In particular, we consider the panel observations $z_{i,t}$ for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. 
whose unknown density (i.e., the model) is approximated by a parametric family \( f(z; \psi, \lambda_i) \), which does not need to include the true model. The parameter of interest is \( \psi \), which is common across \( i \), and the nuisance parameters are given by \( \lambda_1, \ldots, \lambda_n \), whose number increases at the same rate of the sample size (i.e., the incidental parameters problem; Neyman and Scott (1948)). Common examples of \( \lambda_i \) are unobserved heterogeneity (e.g., individual fixed effect) and heteroskedasticity. The main objective is to choose the model that fits best the data generating process, when only a subset of the parameters is of the main interest while the incidental parameters are concentrated out. Such approach is reasonable if we are interested in selecting the structure of the model in \( \psi \), while assuming the parameter space of \( \lambda_i \) is common across the candidate models.

The key method is to apply the profiling idea on the Kullback-Leibler information criterion (KLIC) in order to make a proper model selection criterion in the presence of incidental parameters. It is shown that the profile KLIC can be approximated by the standard KLIC based on the profile likelihoods, provided that a proper modification term is imposed. Such a result corresponds to the fact that the profile likelihood does not share the standard properties of the genuine likelihood function (e.g., the maximum likelihood estimator is biased and often inconsistent), which thus needs some modification (e.g., Sartori (2003)). It turns out that the new information criterion requires heavier penalty than those of the standard information criteria such as AIC so that the degrees of freedom in the model is properly counted. However, it is different from the total number of parameters (i.e., \( \dim(\psi) + n \dim(\lambda_i) \)). The additional penalty term depends on the model complexity measure (e.g., Rissanen (1986)) that reflects the level of difficulty of estimation and it is indeed closely related with the modification term in the modified profile likelihood function. The penalty term is data-dependent, so the new model selection rule is adaptive.

A Bayesian model selection criterion is also developed based on the Bayes factor, where the Bayes factor is obtained using the integrated likelihoods. An interesting finding is that these two approaches—profile likelihood based one and integrated likelihood based one—are closely related as in the standard AIC and BIC, provided that a proper prior of \( \lambda_i \) is used for the integration. It is found that the choice of the prior indeed corresponds to the robust prior of Arellano and Bonhomme (2009).

As an application, we present a lag order selection problem in the context of dynamic panel models with fixed effects. The standard lag order selection criteria do not work properly in this case and they reveal severe over-selection biases in particular. On the other hand, the new lag order selection criterion, whose additional penalty term is a positive function of the sample size and the degree of long-run autocorrelation in the fitted error, reduces the over-selection bias without much affecting the under-selection probability.

The remainder of this paper is organized as follows. Section 2 summarizes the incidental parameters problem in the quasi maximum likelihood setup. The modified profile likelihood
and the bias reduction in the panel data models are also discussed. Section 3 develops an AIC-type information criterion based on the profile likelihood. Section 4 obtains a BIC-type information criterion based on the integrated likelihood and finds connection between the AIC and BIC-type criteria by developing a robust prior. As a particular example, a lag order selection criterion for dynamic panel models is proposed in Section 5 and their statistical properties are examined. Section 6 concludes the paper with several remarks. All the technical proofs are provided in the Appendix.

2 Incidental Parameters Problem in QMLE

2.1 Misspecified models

We consider panel (or longitudinal, stratified) data observations \( z_{i,t} \in \mathcal{Z} \) for \( i = 1, 2, \cdots, n \) and \( t = 1, 2, \cdots, T \), which has an unknown distribution \( G_i(z) \) having probability density function \( g_i(z) \). As in the standard panel cases, \( z_{i,t} \) is allowed to have heterogeneous distributions across \( i \) but it is cross-sectionally independent. On the other hand, \( z_{i,t} \) could be serially correlated over \( t \) but it is stationary.

Since \( g_i(z) \) is unknown a priori, we instead consider a parametric family of densities \( \{ f(z; \theta_i) : \theta_i \in \Theta \} \) for each \( i \), which does not necessarily contain \( g_i(z) \). We assume that \( f(z; \theta_i) \) is continuous (and smooth enough as needed) in \( \theta_i \) for every \( z \in \mathcal{Z} \), the usual regularity conditions for \( f(z; \theta_i) \) hold (e.g., Severini (2000), Chapter 4), and that the parameters are all well identified. Note that we explicitly assume that the shape of marginal density is common for all \( i \) and \( t \) and the heterogeneity is solely controlled by the heterogenous parameter \( \theta_i \). The parameter vector \( \theta_i \) is decomposed as \( \theta_i = (\psi, \lambda_i)' \), where \( \psi \in \Psi \subset \mathbb{R}^r \) is the main parameter of interest that is common to all \( i \), whereas \( \lambda_i \in \Lambda \subset \mathbb{R} \) is the individual specific nuisance parameter. We could consider a multidimensional \( \lambda_i \) (e.g., Arellano and Hahn (2006)) but we focus on the scalar case for expositional simplicity. If we let \( Z = (z_{i,1}', \cdots, z_{i,T}')' \) for \( z_i = (z_{i,1}, \cdots, z_{i,T})' \) and \( \lambda = (\lambda_1, \cdots, \lambda_n)' \), then the joint (pseudo)likelihood can be written as

\[
 f(Z; \psi, \lambda) = \prod_{i=1}^n f(z_i; \psi, \lambda_i) = \prod_{i=1}^n \prod_{t=1}^T f(z_{i,t}; \psi, \lambda_i). \tag{1}
\]

Note that the joint likelihood (1) is separable with respect to nuisance parameters so that the nuisance parameter \( \lambda_i \) is only related to the \( i \)'s observations. Panel models with heterogenous parameters, such as fixed individual effects, (conditional) heteroskedasticity, or heterogenous slope coefficients, are good examples of \( f(\cdot; \psi, \lambda_i) \). More generally, semipara-

\[1\]When we consider dynamic models, \( f(z_{i,t}; \theta_i) \) should be understood as a conditional density on the lagged observations. For example, with \( z_{i,t} = (y_{i,t}, y_{i,t-1}, \cdots, y_{i,t-p}) \) for some \( p \geq 1 \), we define \( f(z_{i,t}; \theta_i) = f(y_{i,t} | y_{i,t-1}, \cdots, y_{i,t-p}; \theta_i) \).
metric models, whose nonparametric component is time-invariant, could be also understood in the similar context if we see $\lambda_i = \lambda(w_i)$ as a realization of the function $\lambda(\cdot)$ for a random variable $w_i$. We assume the following conditions as White (1982) though some stronger conditions are imposed for the later use.

**Assumption 1**

(i) $z_i$ is independent over $i$ with distribution $G_i$ on $Z$, a measurable Euclidean space, with measurable Radon-Nikodym density $g_i = dG_i/d\nu$ for each $i$. (ii) For each $i$, $f(z; \theta_i)$ is the Radon-Nikodym density of the distribution $F(z; \theta_i)$, where $f(z; \theta_i)$ is measurable in $z$ for every $\theta_i \in \Theta = \Psi \times \Lambda$, a compact subset of $\mathbb{R}^{r+1}$ and twice continuously differentiable in $\theta_i$ for every $z \in Z$. (iii) It can be decomposed as $\theta_i = (\psi', \lambda_i)'$, where $\lambda_i$ is related to the $i$-th observation only.

Since we are mainly interested in the parameter $\psi$, we first maximize out the nuisance parameter $\lambda_i$ to define the profile likelihood of $\psi$ as

$$f_{P_i}(z_{i,t}; \psi) = f(z_{i,t}; \psi, \hat{\lambda}_i(\psi))$$

for each $i$, where

$$\hat{\lambda}_i(\psi) = \arg \max_{\lambda_i \in \Lambda} \frac{1}{T} \sum_{t=1}^{T} \log f(z_{i,t}; \psi, \lambda_i)$$

is the quasi maximum likelihood estimator (QMLE) of $\lambda_i$ keeping $\psi$ fixed. Note that (3) is possible since the nuisance parameter is separable in $i$. From the separability, furthermore, we can consider the standard asymptotic results for $\hat{\lambda}_i(\psi)$ in powers of $T$. The quasi maximum profile likelihood estimator of $\psi$ is then obtained as

$$\hat{\psi} = \arg \max_{\psi \in \Psi} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \log f_{P_i}(z_{i,t}; \psi),$$

which indeed corresponds to the QMLE of $\psi$ because this is just taking the maximum in two steps instead of taking the maximum simultaneously. When $T$ is small as in the standard panel data cases, however, $f_{P_i}(\cdot; \psi)$ does not behave like the standard likelihood function due to the sampling variability of the estimator $\hat{\lambda}_i(\psi)$. For example, the expected score of the profile likelihood is nonzero and the standard information identity does not hold even when the true density is nested in $f(\cdot; \psi, \lambda_i)$. Intuitively, it is because the profile likelihood is itself a biased estimate of the original likelihood. Modification of the profile likelihoods in the form of

$$\log f_{Mi}(z_{i,t}; \psi) = \log f_{P_i}(z_{i,t}; \psi) - \frac{1}{T} M_i(\psi)$$

is widely studied, which makes the modified profile likelihood $f_{Mi}(\cdot; \psi)$ to behave more likely a proper likelihood (e.g., Barndorff-Nielsen (1983)). Typically, the modification term $M_i(\psi)$ corrects the leading sampling bias from $\hat{\lambda}_i(\psi)$ and it renders the expected score
of the modified profile likelihood to be closer to zero even with small $T$. A bias-reduced estimator for $\psi$ thus can be obtained by maximizing the modified profile likelihood (i.e., the quasi maximum modified profile likelihood estimation) as

$$\hat{\psi}_M = \arg \max_{\psi \in \Psi} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \log f_{M_i}(z_{i,t}; \psi).$$

Further discussions of the maximum modified profile likelihood estimator can be found in Barndorff-Nielsen (1983), Severini (1998, 2000) and Sartori (2003) to name a few, particularly for the proper choice of the the modification term $M_i(\psi)$. Closely related works are on the adjusted profile likelihood (e.g., McCullagh and Tibshirani (1990), DiCiccio et al. (1996)) and the conditional profile likelihood (e.g., Cox and Reid (1987)).

### 2.2 Incidental parameters problem

From the standard QMLE theory, we can show that the QML estimator (or the quasi maximum profile likelihood estimator) $\hat{\psi}$ in (4) is a consistent estimator for a nonrandom vector $\psi_T$, where

$$\psi_T = \arg \min_{\psi \in \Psi} \lim_{n \to \infty} \frac{1}{n} D(g \mid f(\psi; \hat{\Lambda}(\psi)))$$

with

$$D(g \mid f(\psi; \hat{\Lambda}(\psi))) = \int g(z) \log \left( \frac{g(z)}{f(z; \psi; \hat{\Lambda}(\psi))} \right) dz$$

being the Kullback-Leibler divergence (or the Kullback-Leibler information criterion; KLIC) of the true joint density $g(\cdot) = \prod_{i=1}^{n} g_i(\cdot)$ relative to a parametric joint profile likelihood $f(\cdot; \psi, \hat{\Lambda}(\psi))$, where the KLIC is well defined based on the following conditions.

**Assumption 2** For each $i$, (i) $\int \log g_i(z) dG_i$ exists and both $g_i(z)$ and $f(z; \theta_i)$ are bounded away from zero; (ii) $|\log f(z; \theta_i)|, |\partial^2 \log f(z; \theta_i)/\partial \theta_{i(j)} \partial \theta_{i(k)}|, |\partial \log f(z; \theta_i)/\partial \theta_{i(j)}|, |\partial f(z; \theta_i)/\partial \theta_{i(k)}|$, and $\partial f(z; \theta_i)/\partial \theta_{i(k)} dG_i$ is nonsingular, where $\theta_{i0} = (\psi_{i0}, \lambda_{i0})'$; (iv) $D(g \mid f(\psi, \hat{\Lambda}))$ has a unique minimum at $(\psi_0, \lambda_0) \in \Psi \times \Lambda^n$, where $(\psi_0, \lambda_0)$ is in the interior of the support.

More precisely, it holds that $\hat{\psi} = \psi_T + o_p(1)$ as $n \to \infty$ even with fixed $T$ under the Assump-

---

2Note that since $z_i$ is mutually independent and $f(\psi, \hat{\Lambda})$ is separable with respect to nuisance parameters so that the nuisance parameter $\lambda_i$ is only related to the $i$'s observations, we can indeed define $\psi_T$ as the minimizer of

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} D(g_i \mid f(\psi, \hat{\Lambda}_i(\psi))),$$

where $D(g_i \mid f(\psi, \hat{\Lambda}_i(\psi))) = \int g_i(z) \log \left( \frac{g_i(z)}{f(z; \psi, \hat{\Lambda}_i(\psi))} \right) dz$. 
tion 2 and the standard regularity conditions (e.g., White (1982)). It is also generally true that \( \sqrt{nT}(\hat{\psi} - \psi_T) \to_d \mathcal{N}(0, \Omega_{\psi}) \) as \( n, T \to \infty \) for some positive definite \( \Omega_{\psi} \). Particularly when the dimension of the nuisance parameter \( \Lambda \) is substantial relative to the sample size (e.g., when \( T \) is fixed and small in the panel data case), however, \( \psi_T \) is usually different from the standard KLIC minimizer \( \psi_0 \), which is defined as

\[
(\psi_0, \lambda_0) = \arg\min_{(\psi, \lambda) \in \Psi \times \Lambda} \lim_{n \to \infty} \frac{1}{n} D(g \| f(\psi, \lambda)),
\]

where \( (\psi_0, \lambda_0) = (\psi_0, \lambda_{10}, \cdots, \lambda_{n0}) \) is assumed to exist and to be unique. This is the incidental parameters problem (e.g., Neyman and Scott (1948)). In general, it can be shown that \( \hat{\psi} - \psi_0 = O_p\left(T^{-1}\right) \) for a smooth \( f \) under the standard conditions, and therefore, even when \( n, T \to \infty \) but \( n/T \to \gamma \in (0, \infty) \), we have

\[
\sqrt{nT}(\hat{\psi} - \psi_0) = \sqrt{nT}(\hat{\psi} - \psi_T) + \sqrt{T}B + O_p\left(\sqrt{T}\right) \to_d \mathcal{N}(\sqrt{T}B, \Omega_{\psi}) \tag{9}
\]

since

\[
\psi_T = \psi_0 + B/T + O(T^{-2}),
\]

where \( B/T \) is the \( O(T^{-1}) \) bias. The main source of the bias in (9) is that \( \hat{\lambda}_i(\psi) \) is still random and thus is not the same as

\[
\lambda_i(\psi) = \arg\min_{\lambda_i \in \Lambda} D(g_i \| f(\psi, \lambda_i)), \tag{10}
\]

where \( \lambda_i(\psi_0) = \lambda_{i0} \) for each \( i \). The estimation error of \( \hat{\lambda}_i(\psi) \) with finite \( T \) is not negligible even when \( n \to \infty \), based on which the expectation of the profile score is no longer zero for each \( i \) even under the sufficient regularity conditions.

More precisely, for each \( i \), we define the joint log likelihood functions as

\[
\ell_i(\psi, \lambda_i) = \sum_{t=1}^T \log f(z_{it}; \psi, \lambda_i),
\]

\[
\ell_{Pi}(\psi) = \sum_{t=1}^T \log f(z_{it}; \psi, \hat{\lambda}_i(\psi)),
\]

\[
\ell_{Mi}(\psi) = \ell_{Pi}(\psi) - M_i(\psi),
\]

and the pseudo-information matrix

\[
\mathcal{I}_i = \int \left. \frac{\partial \ell_i(\psi, \lambda_i)}{\partial \theta_i} \right|_{\theta_i = \theta_i(0)} \left. \frac{\partial \ell_i(\psi, \lambda_i)}{\partial \theta_i'} \right|_{\theta_i = \theta_i(0)} dG_i = \left( \begin{array}{cc} \mathcal{I}_{i,\psi\psi} & \mathcal{I}_{i,\psi\lambda} \\ \mathcal{I}_{i,\lambda\psi} & \mathcal{I}_{i,\lambda\lambda} \end{array} \right) \tag{11}
\]

\( ^3 \lambda_i(\psi) \) is normally referred to the least favorable curve.
conformable with \( \theta_i = (\psi', \lambda_i)' \in \mathbb{R}^{r+1} \), where \( \theta_{i0} = (\psi_0', \lambda_{i0})' \) in (8). \( \mathcal{I}_i, \mathcal{I}_{i,\psi} \) and \( \mathcal{I}_{i,\lambda} \) are all nonsingular from Assumption 2. We also define score functions:

\[
\begin{align*}
    u_i (\psi, \lambda_i) &= \partial \ell_i (\psi, \lambda_i) / \partial \psi, \\
v_i (\psi, \lambda_i) &= \partial \ell_i (\psi, \lambda_i) / \partial \lambda_i, \\
u^e_i (\psi, \lambda_i) &= u_i (\psi, \lambda_i) - \mathcal{I}_{i,\psi}^{-1} \mathcal{I}_{i,\psi,\lambda} v_i (\psi, \lambda_i).
\end{align*}
\]

For notational convenience, we suppress the arguments when expressions are evaluated at \( \theta_{0i} = (\psi_0', \lambda_{i0})' \) for each \( i \): \( u_i = u_i (\psi_0, \lambda_{i0}), v_i = v_i (\psi_0, \lambda_{i0}) \) and \( u^e_i = u^e_i (\psi_0, \lambda_{i0}) \). It can be shown that we have the following expansion (e.g., McCullagh and Tibshirani (1990) and Sartori (2003)):\(^4\)

\[
\frac{\partial \ell_{Fi} (\psi_0)}{\partial \psi} = u^e_i + b_i (\psi_0) + O_p (T^{-1/2})
\]

with \( u^e_i = O_p (T^{1/2}) \) and \( b_i (\psi_0) = O_p (1) \) for all \( i \). Though \( \int u^e_i dG_i = 0 \) by construction, \( \int b_i (\psi_0) dG_i \neq 0 \). As a consequence the bias of the profile score accumulates and an asymptotic bias appears as (9). The modification term \( M_i (\psi) \) in (5) can be found as a function in \( \psi \), provided that \( f (\cdot; \theta_i) \) be three-times differentiable in \( \theta_i \), satisfying

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \int \left[ \frac{dM_i (\psi_0)}{d\psi} - b_i (\psi_0) \right] dG_i = 0
\]

so that the expected score of the modified profile likelihood does not have the first order asymptotic bias from \( b_i (\psi_0) \). It is indeed possible to show that \( \int [\partial \ell_{M_i} (\psi_0) / \partial \psi] dG_i = O (T^{-1}) \) so that the modified profile log likelihood is asymptotically closer to the genuine log likelihood.

### 2.3 Bias reduction

The standard bias corrected estimators in nonlinear (dynamic) fixed effect regressions correspond to \( \hat{\psi}_M \) in (6), which is given by (e.g., Hahn and Newey (2004); Arellano and Hahn (2005))

\[
\hat{\psi}_M = \hat{\psi} - \frac{1}{T} \left( \frac{1}{n} \sum_{i=1}^{n} \hat{I}_i (\hat{\psi}_M) \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} \frac{d}{d\psi} M_i (\hat{\psi}_M) \right),
\]

\(^4u^e_i (\psi_0, \lambda_{i0}) \) is the efficient score for \( \psi \) at \( (\psi_0, \lambda_{i0}) \) and it can be understood as the orthogonal projection of the score function for \( \psi \) on the space spanned by the components of the nuisance score \( v_i (\psi_0, \lambda_{i0}) \) (e.g., Murphy and van der Vaart (2000)). It follows that \( \int u^e_i (\psi_0, \lambda_{i0}) / \partial \lambda_i \) \( dG_i = 0 \) since \( u^e_i (\psi, \lambda_i) \) and \( v_i (\psi, \lambda_i) \) are orthogonal at \( (\psi_0, \lambda_{i0}) \) by construction (e.g., Arellano and Hahn (2005)). Also note that the variance of \( u^e_i (\psi_0, \lambda_{i0}) \) is given by \( \hat{I}_i = \mathcal{I}_{i,\psi} - \mathcal{I}_{i,\lambda} \mathcal{I}_{i,\lambda,\psi}^{-1} \mathcal{I}_{i,\psi,\lambda} \equiv \mathcal{I}_{i,\psi | \lambda} \), which is called efficient information and it is nothing but the partial information of \( \psi \) (from not knowing \( \lambda_i \)).
where \( \hat{T}_i^e(\hat{\psi}_M) \) is some consistent estimator of \( T_i^e = T_{i,\psi\psi} - T_{i,\psi^I}T_{i,\psi^I}^{-1}T_{i,\lambda_i} \):

\[
\hat{T}_i^e(\hat{\psi}_M) = -\frac{1}{T} \sum_{t=1}^{T} \left\{ \frac{1}{T} \frac{\partial u_i(\hat{\theta}_{Mi})}{\partial \psi'} - \frac{1}{T} \frac{\partial u_i(\hat{\theta}_{Mi})}{\partial \lambda_i} \left[ \frac{1}{T} \frac{\partial u_i(\hat{\theta}_{Mi})}{\partial \lambda_i} \right]^{-1} \frac{\partial u_i(\hat{\theta}_{Mi})}{\partial \psi'} \right\}
\]

(14)

using the maximum modified profile likelihood estimator \( \hat{\theta}_{Mi} = (\hat{\psi}_M, \hat{\lambda}_i(\hat{\psi}_M))' \). Under the regularity conditions and Assumptions 1 and 2, a simple form of \( M_i(\psi) \) satisfying (13) can be found as follows. Note that, from the standard asymptotic result of the QML estimators, we have the first order stochastic expansion for an arbitrary fixed \( \psi \) as

\[
\sqrt{T}(\hat{\lambda}_i(\psi) - \lambda_i(\psi)) = \left\{ -\frac{1}{T} \frac{\partial^2 \ell_i(\psi, \lambda_i(\psi))}{\partial \lambda_i^2} \right\}^{-1} \frac{1}{\sqrt{T}} \frac{\partial \ell_i(\psi, \lambda_i(\psi))}{\partial \lambda_i} + O_p \left( \frac{1}{T^{1/2}} \right)
\]

(15)

for each \( i \). If we expand \( \ell_{P_i}(\psi) = \ell_i(\psi, \hat{\lambda}_i(\psi)) \) around \( \lambda_i(\psi) \) for given \( \psi \) as

\[
\ell_{P_i}(\psi) = \ell_i(\psi, \lambda_i(\psi)) + \frac{\partial \ell_i(\psi, \lambda_i(\psi))}{\partial \lambda_i}(\hat{\lambda}_i(\psi) - \lambda_i(\psi)) + \frac{1}{2} \frac{\partial^2 \ell_i(\psi, \lambda_i(\psi))}{\partial \lambda_i^2}(\hat{\lambda}_i(\psi) - \lambda_i(\psi))^2 + O_p \left( \frac{1}{T^{1/2}} \right),
\]

in which the first two terms are \( O_p(T^{1/2}) \) and the third term is \( O_p(1) \), then using (15) we can derive that (e.g., Severini (2000), Arellano and Hahn (2006))

\[
\mathbb{E}_{g_i} \left[ \ell_{P_i}(\psi) - \ell_i(\psi, \lambda_i(\psi)) - \frac{1}{2} \left\{ -\frac{\partial^2 \ell_i(\psi, \lambda_i(\psi))}{\partial \lambda_i^2} \right\}^{-1} \left( \frac{\partial \ell_i(\psi, \lambda_i(\psi))}{\partial \lambda_i} \right)^2 \right] = O \left( \frac{1}{T^{1/2}} \right).
\]

(16)

Here we simply denote \( \mathbb{E}_{g_i} [\cdot] = \int [\cdot] dG_i \). Since \( \lambda_i(\psi_0) = \lambda_{i0} \) and \( \mathbb{E}_{g_i} [\ell_i(\psi_0, \lambda_{i0})] = 0 \) by construction, comparing (12) and (16), this result suggests that a simple form of the modification function in \( M_{Mi}(\psi) \) could be found as

\[
M_i(\psi) = \frac{1}{2} \left\{ -\frac{\partial^2 \ell_i(\psi, \hat{\lambda}_i(\psi))}{\partial \lambda_i^2} \right\}^{-1} \left( \frac{\partial \ell_i(\psi, \hat{\lambda}_i(\psi))}{\partial \lambda_i} \right)^2,
\]

(17)

whose first derivative corrects the leading bias term in the profile score. For more general treatment of the modification on the profile likelihood, see Barndorff-Nielsen (1983) for the modified profile likelihood approach or McCullagh and Tibshirani (1990) for the adjusted profile likelihood approach.\(^5\)

In general, particularly for the panel data models, Arellano

\[^{5}\text{It can be also derived that } M_i(\psi) = \log |\mathbb{E}_{g_i} [v_i(\psi, \hat{\lambda}_i(\psi)) v_i]| - (1/2) \log | - \partial v_i(\psi, \hat{\lambda}_i(\psi))/\partial \lambda_i| \text{ using a higher order approximation instead of (15). (E.g., Barndorff-Nielsen (1983) and Severini (1998).)}\]
and Hahn (2006) and Bester and Hansen (2009) suggest the modification function given by

\[
M_i(\psi) = \frac{1}{2} \left\{ \frac{1}{T} \sum_{t=1}^{T} \frac{\partial^2 \log f(z_{i,t}; \psi, \hat{\lambda}_i(\psi))}{\partial \lambda_i^2} \right\}^{-1} \times \sum_{t=-m}^{m} K_\ell \frac{\partial \log f(z_{i,t}; \psi, \hat{\lambda}_i(\psi))}{\partial \lambda_i} \frac{\partial \log f(z_{i,t-\ell}; \psi, \hat{\lambda}_i(\psi))}{\partial \lambda_i},
\]

so that possible serial correlations in the score function could be accommodated. Note that the second term in (18) is a general expression of the variance estimator of the \(\sqrt{T}\)-

" restricted partial sum \(T^{-1/2} \sum_{t=1}^{T} \frac{\partial \log f(z_{i,t}; \psi, \hat{\lambda}_i(\psi))}{\partial \lambda_i} \) (i.e., the expectation of the second term in (17)), which allows for arbitrary serial correlations (e.g., heteroskedasticity and autocorrelation consistent estimator). The truncation parameter \(m \geq 0\) is chosen such that \(m/T^{1/2} \rightarrow 0\) as \(T \rightarrow \infty\), and the kernel function \(K_\ell\) guarantees positive definiteness of the variance estimate (e.g., the Bartlett kernel: \(K_\ell = 1 - \ell/(m + 1)\)).

Using a similar method, we could also expand \(\frac{\partial \ell_{Pi}(\psi_0)}{\partial \psi} = \frac{\partial \ell_i(\psi_0, \hat{\lambda}_i(\psi_0))}{\partial \psi} = u_i^e(\psi_0, \hat{\lambda}_i(\psi_0)) + I_{i,\psi\lambda_i} I_{i,\lambda_i,\lambda_i} v_i(\psi_0, \hat{\lambda}_i(\psi_0))\) in (12) around \(\lambda_i(\psi_0) = \lambda_{i0}\) and obtain (e.g., Hahn and Newey (2004) and Arellano and Hahn (2005))

\[
\frac{\partial \ell_{Pi}(\psi_0)}{\partial \psi} \approx u_i^e - v_i \left( \frac{\partial u_i^e}{\partial \lambda_i} \right)^{-1} \left\{ \frac{\partial u_i^e}{\partial \lambda_i} - \frac{1}{2} v_i \left( \frac{\partial v_i}{\partial \lambda_i} \right)^{-1} \left( \frac{\partial^2 u_i^e}{\partial \lambda_i^2} + I_{i,\psi\lambda_i} I_{i,\lambda_i,\lambda_i} \frac{\partial^2 v_i}{\partial \lambda_i^2} \right) \right\}
\]

with an approximation error \(O_p(T^{-1/2})\). From this expression, we can also define

\[
b_i(\psi_0) = -v_i \left( \frac{\partial u_i^e}{\partial \lambda_i} \right)^{-1} \left\{ \frac{\partial u_i^e}{\partial \lambda_i} - \frac{1}{2} v_i \left( \frac{\partial v_i}{\partial \lambda_i} \right)^{-1} \left( \frac{\partial^2 u_i^e}{\partial \lambda_i^2} + I_{i,\psi\lambda_i} I_{i,\lambda_i,\lambda_i} \frac{\partial^2 v_i}{\partial \lambda_i^2} \right) \right\}
\]

as the derivative of \(M_i(\psi_0)^6\), which can be estimated as

\[
\hat{b}_i(\theta_{Mi}) = -v_i(\theta_{Mi}) \left[ \frac{1}{T} \frac{\partial v_i(\theta_{Mi})}{\partial \lambda_i} \right]^{-1} \left\{ \frac{\partial u_i^e(\theta_{Mi})}{\partial \lambda_i} - \frac{1}{2} v_i(\theta_{Mi}) \left[ \frac{1}{T} \frac{\partial v_i(\theta_{Mi})}{\partial \lambda_i} \right]^{-1} \left( \frac{\partial^2 u_i^e(\theta_{Mi})}{\partial \lambda_i^2} + I_{i,\psi\lambda_i} I_{i,\lambda_i,\lambda_i} \frac{\partial^2 v_i(\theta_{Mi})}{\partial \lambda_i^2} \right) \right\}
\]

similarly as (14).

\[\text{It can be instead shown that (e.g., McCullagh and Tibshirani (1990) and Arellano and Hahn (2005))}
\]

\[
\mathbb{E}_{y_i} \left[ \frac{\partial \ell_{Pi}(\psi_0)}{\partial \psi} \right] = - (1/2) \mathbb{E}_{y_i} \left[ u_i^e \left\{ v_i^2 + (\partial v_i/\partial \lambda_i) \right\} (\mathbb{E}_{y_i} [v_i^2])^{-1} \right] + O(T^{-1}),
\]

where the first term corresponds to \(\mathbb{E}_{y_i}[b_i(\psi_0)]\).
3 Profile Likelihood and KLIC

3.1 Model selection and incidental parameters

Normally the panel data studies focus on reducing the first order bias (9) from the incidental parameters problem, which basically presume that the models considered are correctly specified. As discussed in Lee (2006, 2011), however, if the model is not correctly specified, the efforts of reducing bias from the incidental parameters could even exacerbate the bias. Therefore, the correct model specification is very important in this context particularly for dynamic or nonlinear panel models (e.g., choosing the lag order in ARMA models or the functional structure in the nonlinear models, respectively); the correct model specification should precede any bias corrections or bias reductions. This paper rather focuses on model specification. In particular, selecting a model \( f(\psi, \lambda) \), which is closest to the true one \( g(z) \), is the main interest.

In the standard setup, when the dimension of the entire parameter vector \( \theta \) is small and finite, we can conduct the standard model selection by comparing estimates of the KLIC:

\[
\min_{\theta} D(g \| f(\theta)) = D(g \| f(\hat{\theta})) = \int g(z) \log \left( \frac{g(z)}{f(z; \theta)} \right) dz, \tag{20}
\]

among different specifications \( f(\cdot; \theta) \), where \( \hat{\theta} \) is the QMLE for a given likelihood function \( f(\cdot; \theta) \). Note that \( \hat{\theta} \) is a consistent estimator of \( \theta_0 = \arg \min_{\theta} D(g \| f(\theta)) \) in this case. We select a model \( f(\cdot; \theta) \) whose KLIC in (20) is the minimum among the candidates. Equivalently, we select the model \( f(\cdot; \theta) \) minimizing the relative distance

\[
\Phi(\theta) = - \int \log f(z; \hat{\theta}) d\hat{G}(z),
\]

which can be estimated by \( \tilde{\Phi}(\theta) = - \int \log f(z; \hat{\theta}) d\hat{G}(z) \), where \( \hat{G} \) is the empirical distribution. As noted in Akaike (1973), however, \( \int \log f(z; \hat{\theta}) d\hat{G}(z) \) overestimates \( \int \log f(z; \hat{\theta}) dG(z) \) since \( \hat{G} \) corresponds more closely to \( \hat{\theta} \) than does the true \( G \). Therefore, it is suggested to minimize the bias-corrected version of \( \tilde{\Phi}(\theta) \) given by

\[
\tilde{\Phi}(\theta) = - \int \log f(z; \hat{\theta}) d\hat{G}(z) - B(\hat{G}) \tag{21}
\]

as an information criterion for model selection, where \( B(G) = \mathbb{E}_g[\tilde{\Phi}(\theta) - \Phi(\theta)] \). See, for example, Akaike (1973, 1974) for further details. Note that Akaike (1973) shows that \( B(G) \) is asymptotically the ratio of \( \text{dim}(\theta) \) to the sample size when \( \hat{\theta} \) is the QMLE and \( g \) is nested in \( f \).
Now we consider the case with incidental parameters $\lambda \in \mathbb{R}^n$, where $\theta = (\psi', \lambda')'$. Similarly as we discussed in the previous section, when the dimension of the parameter vector $\theta$ is substantial relative to the sample size, the incidental parameters problem still prevails and thus it is not straightforward to find a proper criterion similarly as (21). One possible solution is to reduce the dimension of the parameters by concentrating out the nuisance parameters. Particularly when we assume that the (finite-dimensional) parameter of main interest $\psi$ determines the main framework of the model and the specification does not change over $i$, it is then natural to concentrate out the nuisance parameters $\lambda_i$’s in the model selection problem. This is because the choice of a particular model does not depend on the realization of $\lambda_i$’s in this case. In other words, for model selection purposes, the family of candidate models is indexed by $\psi$ alone, while the parameter space of $\lambda_i$ remains the same across the candidate models. This is a similar idea of the profile likelihood approach when the main interest is in a subset of parameters $\psi$. Some examples are as follows.

**Example 1 (Variable selection in panel models)** We consider a parametric nonlinear fixed-effect model given by $y_{i,t} = m(x_{i,t}, u_{i,t}; \mu_i, \beta, \sigma_i^2)$ with known $m(\cdot; \cdot)$, where $u_{i,t}$ is independent over $i$ and $t$ with $u_{i,t} | x_{i,1}, \ldots, x_{i,T}, \mu_i \sim (0, \sigma_i^2)$, and $\beta$ is an $r$-dimensional parameter vector. The goal of this example is to select a set of regressors yielding the best fit in the presence of incidental parameters $(\mu_i, \sigma_i^2)$. Variable selection in a linear transformation model given by $\varphi_i(y_{i,t}) = x'_{i,t} \beta + u_{i,t}$ with strictly increasing incidental functions $\varphi_i(\cdot)$ is another example.

**Example 2 (Lag order selection in dynamic panel regressions)** We consider a panel $AR(p)$ model with fixed effect given by $y_{i,t} = \mu_i + \sum_{j=1}^{p} \alpha_{pj} y_{i,t-j} + \varepsilon_{i,t}$, where $\varepsilon_{i,t}$ is independent across $i$ and serially uncorrelated. The goal of this example is to choose the correct lag order $p$ in the presence of incidental parameters $\mu_i$. When $p = \infty$, this problem becomes to find the best approximation $AR(p)$ model.

**Example 3 (Number of support choice of random effects)** We consider a random-effect model given by $y_{i,t} = x'_{i,t} \beta + u_{i,t}$ with $u_{i,t} = \mu_i + \varepsilon_{i,t}$, where $\varepsilon_{i,t}$ is independent over $i$ and $t$ with $\varepsilon_{i,t} | x_{i,1}, \ldots, x_{i,T} \sim \mathcal{N}(0, \sigma^2_i)$, and $\mu_i$ is an i.i.d. unobserved random variable independent of $x_{i,t}$ and $\varepsilon_{i,t}$ with a common distribution over the finite support $\{q_1, \ldots, q_k\}$. The main interest in this example is to estimate the number of finite support $k$ in the presence of incidental parameters $\sigma^2_i$. In the context of mixed proportional hazard models, this problem is to choose the number of finite support of nonparametric frailty in the Heckman-Singer model (Heckman and Singer (1984)).
3.2 Profile likelihood information criterion

More precisely, for a proper model selection information criterion in the presence of incidental parameters, we consider the profile Kullback-Leibler divergence, in which the incidental parameters $\lambda_i$'s are concentrated out from the standard KLIC as follows.

**Definition (Profile KLIC)** The profile Kullback-Leibler divergence (or the profile KLIC) of $g(\cdot)$ relative to $f(\cdot; \psi, \lambda)$ is defined as

$$D_P(g \parallel f(\psi, \lambda); \psi) = \min_{\lambda \in \Lambda^n} D(g \parallel f(\psi, \lambda)) = D(g \parallel f(\psi, \lambda(\psi)))$$

(22)

for $\lambda(\psi) = (\lambda_1(\psi), \cdots, \lambda_n(\psi))'$ with $\lambda_i(\psi)$ given in (10), where the incidental parameters $\lambda_i$'s are concentrated out by minimizing the standard KLIC, $D(g \parallel f(\psi, \lambda))$, over $\lambda_i$'s.

Note that $D_P(g \parallel f(\psi, \lambda); \psi)$ indeed depends on $\psi$ only, not on $\lambda_i$'s. Since the profile KLIC ($D_P(g \parallel f(\psi, \lambda); \psi)$) is defined as the minimum of the standard KLIC ($D(g \parallel f(\psi, \lambda))$) in $\lambda$, it apparently satisfies the conditions that the standard KLIC has. For example, $D_P(g \parallel f(\psi, \lambda); \psi)$ is nonnegative and it is equal to zero when $g(\cdot)$ belongs to the parametric family of $f(\cdot; \psi, \lambda)$ (i.e., $g(\cdot) = f(\cdot; \psi, \lambda)$ for some $(\psi, \lambda) \in \Psi \times \Lambda^n$). Similarly as in the standard case, we select the model that has the smallest value of

$$\min_{\psi \in \Psi} \lim_{n \to \infty} \frac{1}{n} D_P(g \parallel f(\psi, \lambda); \psi).$$

(23)

But we have

$$\min_{\psi \in \Psi} D_P(g \parallel f(\psi, \lambda); \psi) = \min_{\psi \in \Psi} D(g \parallel f(\psi, \lambda(\psi)))$$

(24)

$$= \min_{(\psi, \lambda') \in \Psi \times \Lambda^n} D(g \parallel f(\psi, \lambda)),$$

so that the model with the smallest (23) indeed corresponds to the model with the smallest estimate of the standard KLIC, $D(g \parallel f(\psi, \lambda))$, over $\psi$ and $\lambda$. Note that, however, we cannot directly use (23) for the model selection procedure since (24) demonstrates that it contains infeasible components $\lambda_i(\psi)$'s, which are defined in (3). A natural candidate is then the standard KLIC based on the profile likelihoods given by

$$D(g \parallel f_P(\psi)) = D(g \parallel f(\psi, \tilde{\lambda}(\psi))),$$

(25)

which turns out to be equivalent to (7). Since $\tilde{\lambda}(\psi)$ is not a desirable estimator for $\lambda_i(\psi)$ when $T$ is small, however, the KLIC based on the profile likelihoods (25) would not be a good estimator for the profile KLIC in (22). The following lemma states the relation
between these two KLIC’s.

**Lemma 1** For a given \( \psi \in \Psi \), we have

\[
D_P (g \parallel f(\psi, \Delta); \psi) = D (g \parallel f_P(\psi)) + \delta(\psi; G),
\]

(26)

where the bias term is defined as

\[
\delta(\psi; G) = \sum_{i=1}^{n} \int \log \left( f(z_i; \psi, \lambda_i(\psi))/f(z_i; \psi, \lambda_i(\psi)) \right) dG_i
\]

with \( \lambda_i(\psi) \) and \( \lambda_i(\psi) \) being given in (3) and (10), respectively.

From (26), it can be seen that even when \( g \) is nested in \( f \), \( D (g \parallel f_P(\psi)) \) is not necessarily zero unless \( f(z; \psi, \lambda_i(\psi)) = f(z; \psi, \lambda_i(\psi)) \), which is very unlikely with small \( T \). It thus follows that model selection using \( D (g \parallel f_P(\psi)) \) itself is undesirable. An interesting finding is that, however, if we properly modify \( D (g \parallel f_P(\psi)) \) by correcting the bias \( \delta(\psi; G) \) as in Lemma 1, then we can conduct the model selection based on \( D (g \parallel f_P(\psi)) \) as long as we have a proper estimator for \( \delta(\psi; G) \). As well expected, the bias term in (26) is closely related with the modification term \( M_i(\psi) \) in constructing the modified profile likelihood in (5).

**Lemma 2** Let Assumptions 1 and 2 hold. For a given \( \psi \in \Psi \), under the regularity conditions, we have

\[
\delta(\psi; G) = \frac{1}{T} \sum_{i=1}^{n} \mathbb{E}_{g_i}[M_i(\psi)] + O \left( \frac{n}{T^{3/2}} \right),
\]

(27)

where \( M_i(\psi) \) is the modification term used for the modified profile likelihood function.

Similarly as (21), by letting

\[
\Phi_P(\hat{\psi}_M) = -\frac{1}{n} \sum_{i=1}^{n} \int \log f_i(z; \hat{\psi}_M, \hat{\lambda}_i(\hat{\psi}_M)) dG_i(z),
\]

we define an information criterion using the profile likelihood as

\[
\Phi_P(\hat{\psi}_M) = -\frac{1}{n} \sum_{i=1}^{n} \int \log f_i(z; \hat{\psi}_M, \hat{\lambda}_i(\hat{\psi}_M)) dG_i(z) - B_P(\hat{G}),
\]

(28)

where \( \hat{\psi}_M \) is the quasi maximum modified profile likelihood estimator of \( \psi_0 \) defined as (6) and \( B_P(\hat{G}) \) is an estimator of

\[
B_P(\hat{G}) = \mathbb{E}_{\hat{G}}[\Phi_P(\hat{\psi}_M) - \Phi_P(\hat{\psi}_M)] - \frac{1}{n} \delta(\hat{\psi}_M; G)
\]

\[
= -\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{g_i} \left[ \int \log f_i(z; \hat{\psi}_M, \hat{\lambda}_i(\hat{\psi}_M)) dG_i(z) - G_i(z) \right] - \frac{1}{n} \delta(\hat{\psi}_M; G)
\]

obtained by replacing the unknown distribution \( G \) by the empirical distribution \( \hat{G} \). We use the bias-corrected estimator (i.e., the quasi maximum modified profile likelihood estimator)
\[ \hat{\psi}_M = \arg\max_{\psi} (nT)^{-1} \sum_{i=1}^{n} \ell_{Mi}(z_i; \psi) \text{ for } \psi \in \Psi \subset \mathbb{R}^r \] when defining the information criterion. Note that the bias correction term \( B_P(\hat{G}) \) is somewhat different from the correction term \( B(\hat{G}) \) in (21). In particular, it includes an additional term \( n^{-1}\delta(\hat{\psi}_M; \hat{G}) \), which is needed because the information criterion is defined using \( D(g \parallel f_P(\psi)) \) instead of \( D_P(g \parallel f(\psi, \lambda); \psi) \). An approximated expression of \( B_P(G) \) and its consistent estimator are obtained in the following theorem.

**Theorem 3** Let Assumptions 1 and 2 hold. We suppose that there exists an \( r \)-dimensional regular function \( H \) such that \( \psi_0 = H(G) \) and \( \hat{\psi}_M = H(\hat{G}) \). \( H \) is assumed to be second order compact differentiable at \( G \). If \( n, T \to \infty \) satisfying \( n/T \to \gamma \in (0, \infty) \) and \( n/T^3 \to 0 \), under the regularity conditions (e.g., Hahn and Kuersteiner (2007)), we have

\[
B_P(G) = -\frac{1}{nT} \text{tr} \{ I(G)^{-1} J(G) \} - \frac{1}{n} \delta(\hat{\psi}_M; G),
\]

where \( \text{tr} \{ \cdot \} \) is the trace operator and

\[
I(G) = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{E}_{g_{it}} \left[ -\frac{\partial^2 \log f_i(z_{i,t}; \psi, \lambda_i(\psi))}{\partial \psi \partial \psi'} \bigg|_{\psi=H(G)} \right],
\]

\[
J(G) = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{s=1}^{T} \frac{\partial \log f_i(z_{i,t}; \psi, \lambda_i(\psi))}{\partial \psi} \bigg|_{\psi=H(G)} \frac{\partial \log f_i(z_{i,s}; \lambda_i(\psi), \hat{\lambda}_i(\psi))}{\partial \psi'} \bigg|_{\psi=H(G)}.
\]

Moreover, for some truncation parameter \( m \geq 0 \) such that \( m/T^{1/2} \to 0 \) as \( T \to \infty \), an consistent estimator for \( B_P(G) \) is obtained as

\[
B_P(\hat{G}) = -\frac{1}{nT} \text{tr} \{ I(\hat{G})^{-1} J(\hat{G}) \} - \frac{1}{nT} \sum_{i=1}^{n} M_i(\hat{\psi}_M), \tag{29}
\]

where\(^7\)

\[
I(\hat{G}) = -\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{\partial^2 \log f_{Mi}(z_{i,t}; \hat{\psi}_M)}{\partial \psi \partial \psi'} \text{ and}
\]

\[
J(\hat{G}) = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=-m}^{m} \min\{T, T+\ell\} \frac{\partial \log f_{Mi}(z_{i,t}; \hat{\psi}_M)}{\partial \psi} \sum_{t=-m}^{m} \frac{\partial \log f_{P_i}(z_{i,t-\ell}; \hat{\psi}_M)}{\partial \psi'}
\]

for \( \log f_{P_i}(z_{i,t}; \psi) = \log f(z_{i,t}; \psi, \hat{\lambda}_i(\psi)) \) and \( \log f_{Mi}(z_{i,t}; \psi) = \log f_{P_i}(z_{i,t}; \psi) - T^{-1} M_i(\psi) \).

From the equations (28) and (29), therefore, a general form of an information criterion

\(^7\)Note that \( J(\hat{G}) \) allows for possible serial correlations in the (modified) profile score functions similarly as \( M_i(\psi) \) does in (18).
for the model selection based on the bias-corrected profile likelihood (i.e., profile likelihood information criterion; PLIC) is defined as

\[
PLIC (f) = - \frac{2}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \log f_{Pi}(z_{i,t}; \hat{\psi}_M) - 2B_P(\hat{G})
\]

\[
= - \frac{2}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \log f(z_{i,t}; \hat{\psi}_M, \hat{\lambda}_i(\hat{\psi}_M)) \\
+ \frac{2}{nT} tr \{ I(\hat{G})^{-1} J(\hat{G}) \} + \frac{2}{nT} \sum_{i=1}^{n} M_i(\hat{\psi}_M),
\]

where \( M_i(\hat{\psi}_M) \) is given in (18) in general. Note that this new information criterion includes two penalty terms. The first penalty term corresponds to the standard finite sample adjustment as AIC whereas the second penalty term reflects bias correction from using of profile likelihood in the model selection problem. With further restricted conditions, we could derive the simpler form for the \( PLIC (f) \) using the following corollary.

**Corollary 4** Under the same conditions as Theorem 3, we have

\[
I (G) = \frac{1}{nT} \sum_{i=1}^{n} \mathbb{E}_{g_i} \left[ - \frac{\partial u_i^e}{\partial \psi} \right] \quad \text{and} \quad J (G) = \frac{1}{nT} \sum_{i=1}^{n} T_i^e + \frac{1}{nT} \sum_{i=1}^{n} \mathbb{E}_{g_i} \left[ u_i^e b_i(\psi_0)' \right] + o \left( \frac{1}{T^{3/2}} \right),
\]

where \( u_i^e = u_i - \mathcal{T}_i, \psi \mathcal{T}_i^{-1} \mathcal{V}_i \) is the efficient score of \( \psi \) at \( (\psi_0, \lambda_0) \) and \( T_i^e = \mathcal{T}_i, \psi \mathcal{T}_i^{-1} \mathcal{I}_i, \lambda_0 \) is its variance (i.e., efficient information) for each \( i \). Particularly when \( g \) is included in the family of \( f \), we have \( I (G) = (nT)^{-1} \sum_{i=1}^{n} I_i^e \), where \( r = \dim(\psi) \). Moreover, it could be approximated as \( \mathbb{E}_{g_i} [M_i(\hat{\psi}_M) - M_i(\psi_0)] = I (G)^{-1} \mathbb{E}_{g_i} [u_i^e b_i(\psi_0)'] + o(T^{-1/2}) \).

Using Corollary 4, when the true density \( g \) is nested in the family of \( f \), the penalty term \( B_P(G) \) can be approximated as

\[
B_P(G) \approx - \frac{r}{nT} + \frac{1}{nT} \sum_{i=1}^{n} I (G)^{-1} \mathbb{E}_{g_i} [u_i^e b_i(\psi_0)'] \\
- \frac{1}{nT} \sum_{i=1}^{n} \left\{ \mathbb{E}_{g_i} [M_i(\hat{\psi}_M)] - I (G)^{-1} \mathbb{E}_{g_i} [u_i^e b_i(\psi_0)'] \right\} \\
= - \frac{r}{nT} - \frac{1}{nT} \sum_{i=1}^{n} \mathbb{E}_{g_i} [M_i(\hat{\psi}_M)],
\]

which gives the simplified form of the information criterion (30) as

\[
PLIC (f) = - \frac{2}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \log f(z_{i,t}; \hat{\psi}_M, \hat{\lambda}_i(\hat{\psi}_M)) + \frac{2r}{nT} + \frac{2}{nT} \sum_{i=1}^{n} M_i(\hat{\psi}_M).
\]

Note that the goodness of fit is based on the maximized profile likelihood, which corresponds
to the standard maximized likelihood though it is evaluated at \( \hat{\psi}_M \) instead of the MLE. The additional penalty term \( (2/nT) \sum_{i=1}^{n} M_i(\hat{\psi}_M) \) is novel and it is not zero in the presence of incidental parameters. Since this additional penalty term is positive by construction, the new information criterion (30) or (31) has heavier penalty than the standard Akaike information criterion (AIC). Recall that in the standard AIC, the second part of the penalty term in (31) does not appear and the penalty term of the information criterion is simply given by \( 2r/nT \).

**Remark 1 (KLIC and modified profile likelihood)** In fact, \( PLIC(f) \) given in (31) is quite intuitive since it can be rewritten as

\[
PLIC(f) = -\frac{2}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \log f_{M_i}(z_{i,t}; \hat{\psi}_M) + \frac{2r}{nT}
\]  

using the modified profile likelihood function, where \( \log f_{M_i}(:, \psi) = \log f_{M_i}(\cdot; \psi) - T^{-1} M_i(\psi) \). Note that the modified profile likelihood function is closer to the genuine likelihood than is the profile likelihood function. (32) shows that such aspect extends even when we define the KLIC. More precisely, from Lemmas 1 and 2, we can derive that \( n^{-1} D_P (g \parallel f(\psi, \Delta); \psi) = n^{-1} D (g \parallel f_M(\psi)) + O(1/T^{3/2}) \), where \( \log f_M(\psi) = \sum_{i=1}^{n} \log f_{M_i}(\cdot; \psi) \).

**Remark 2 (Least favorable model)** The new information criterion is based on the profile likelihood, which reduces the high dimensional model to a finite dimensional, random submodel of the same dimension as \( \psi \). The family of distributions with parameter \((\psi, \Delta(\psi))\) is known as Stein’s least favorable family (Stein (1956)). More precisely, the submodel defined as the profile likelihood with reduced parameters \((\hat{\psi}, \hat{\Delta}) \rightarrow \psi)\) has the smallest information about \( \psi \) out of all possible submodels. Such loss of information can be expressed as the difference between \( I_{i,\hat{\psi}} \) and \( I_i = I_{i,\psi\hat{\lambda}} - I_{i,\psi\lambda} I_{i,\psi\hat{\lambda}}^{-1} I_{i,\psi\hat{\lambda}} \) for each \( i \), where \( I_i \) can be understood as the partial information of \( \psi \) with \( \lambda_i \) unknown whereas \( I_{i,\hat{\psi}} \) as the partial information of \( \psi \) with \( \lambda_i \) known. Therefore, it could be understood that the selected model by minimizing the profile Kullback-Leibler divergence \( D_P (g \parallel f(\psi, \Delta); \psi) \) is the least favorable at \((\psi_0, \Delta_0)\) (i.e., its information is minimal among the consistent model group; see Murphy and van der Vaart (2000), Severini (2000), Severini and Wong (1992)). In other words, the likelihood function \( f(\cdot; \psi, \lambda_i(\psi)) \) is a least favorable target likelihood in the sense that the expected information of \( \psi \) coincides with the partial expected information of \( \psi \). Based on such interpretation, we may expect that the new information criterion also tends to select over-parametrized models similarly as the standard AIC (i.e., choosing less efficient model).
4 Integrated Likelihood and Bayesian Approach

Instead of KLIC-based model selection criteria using the (modified) profile likelihood, we now consider the Bayesian approach using the integrated likelihood (e.g., Berger et al. (1999)). Interestingly, the result in this section shows that the difference between the integrated likelihood based approach and the profile likelihood based approach is merely their penalty terms, where the penalty terms are of the same form of the standard AIC and BIC cases.

We first assume a conditional prior of \( \pi_i(\lambda_i|\psi) \) for each \( i \). The integrated log-likelihood \( L_I(\psi) \) can be obtained as

\[
L_I(\psi) = \sum_{i=1}^{N} \ell_i(\psi) = \sum_{i=1}^{N} \log \left\{ \int f_i(\psi, \lambda_i) \pi_i(\lambda_i|\psi) d\lambda_i \right\},
\]

where \( f_i(\psi, \lambda_i) = \prod_{t=1}^T f(z_{it}; \psi, \lambda_i) \). We let \( \phi^k \) be the discrete prior over different models \( \mathcal{M}^1, \mathcal{M}^2, \ldots, \mathcal{M}^K \) and \( \eta(\psi^k|\mathcal{M}^k) \) be the prior on \( \psi^k \in \mathbb{R}^k \) given the model \( \mathcal{M}^k \). Then, the joint prior of \( \mathcal{M}^k \) and \( \psi^k \) is given as

\[
\phi^k \eta(\psi^k|\mathcal{M}^k) L(\psi^k|y) / g(y)
\]

where \( g(\cdot) \) is the marginal distribution of \( y \) and \( L(\psi^k|y) = \exp(L_I(\psi^k)) \) is the integrated likelihood function. In this case, the posterior is obtained as

\[
P(\mathcal{M}^k|y) = \frac{1}{g(y)} \phi^k \int \exp(L_I(\psi^k)) \eta(\psi^k|\mathcal{M}^k) d\psi^k
\]

and the Bayesian information criterion can be obtained based on \(-2 \log P(\mathcal{M}^k|y)\).

Note that, however, from Lemma 1 in Arellano and Bonhomme (2009), we can link the integrated and the (modified) profile likelihood as follows using the Laplace approximation:

\[
\ell_i(\psi^k) - \ell_{P_i}(\psi^k) = \frac{1}{2} \log (2\pi/T) - \frac{1}{2} \log \left( -\frac{\partial^2 \ell_i(\psi^k, \hat{\lambda}_i(\psi^k))}{\partial \lambda_i^2} \right) + \log \pi_i(\hat{\lambda}_i(\psi^k)|\psi^k) + O_p(T^{-1})
\]

or

\[
\ell_i(\psi^k) - \ell_{M_i}(\psi^k) = \frac{1}{2} \log (2\pi/T) - \frac{1}{2} \log \left( -\frac{\partial^2 \ell_i(\psi^k, \hat{\lambda}_i(\psi^k))}{\partial \lambda_i^2} \right) + \log \pi_i(\hat{\lambda}_i(\psi^k)|\psi^k) + M_i(\psi^k) + O_p(T^{-1})
\]

for each \( i \). These approximations imply that if we choose the conditional prior \( \pi_i(\lambda_i|\psi^k) \)
such that
\[
\log \pi_i(\tilde{\lambda}_i(\psi^k)|\psi^k) = \frac{1}{2} \log \left( -\frac{\partial^2 \ell_i(\psi^k, \tilde{\lambda}_i(\psi^k))}{\partial \lambda_i^2} \right) - M_i(\psi^k)
\]
then \( \ell_{II}(\psi^k) - \ell_{Mi}(\psi^k) = O_p(T^{-1}) \) up to a constant addition \((1/2) \log(2\pi/T)\). From (17), however, we obtain the explicit form of the conditional prior as
\[
\log \pi_i(\tilde{\lambda}_i(\psi^k)|\psi^k) = \frac{1}{2} \log \left( -\frac{\partial^2 \ell_i(\psi^k, \tilde{\lambda}_i(\psi^k))}{\partial \lambda_i^2} \right) - \frac{1}{2} \left\{ -\frac{\partial^2 \ell_i(\psi^k, \tilde{\lambda}_i(\psi^k))}{\partial \lambda_i^2} \right\}^{-1} \left( \frac{\partial \ell_i(\psi^k, \tilde{\lambda}_i(\psi^k))}{\partial \lambda_i} \right)^2,
\]
which indeed gives the same result of the robust prior in equation (14) of Arellano and Bonhomme (2009). Therefore, by choosing the robust prior as (33), we can show that
\[
\log \mathcal{P}\left(\mathcal{M}^k|y\right) = -\log g(y) + \log \phi^k + \log \int \exp(L_I(\psi^k)) \eta(\psi^k|\mathcal{M}^k) d\psi^k
\]
\[
= -\log g(y) + \log \phi^k + \log \int \exp \left( \sum_{i=1}^N \left\{ \ell_{Mi}(\psi^k) + O_p(T^{-1}) \right\} \right) \eta(\psi^k|\mathcal{M}^k) d\psi^k
\]
\[
= -\log g(y) + \log \phi^k + \sum_{i=1}^N \ell_{Mi}(\tilde{\psi}_M^k) + \frac{k}{2} \log 2\pi
\]
\[
- \frac{1}{2} \log NT - \frac{1}{2} \log \|T(\tilde{\psi}_M^k)\| + O_p\left(\frac{N}{T}\right),
\]
Therefore, using the uninformative flat prior \( \eta(\psi^k|\mathcal{M}^k) = 1 \), we have

\[
\log \int \exp(L_I(\psi^k)) d\psi^k = \log \int \exp \left( \sum_{i=1}^{N} \ell_{Mi}(\psi^k) + O_p \left( \frac{N}{T} \right) \right) d\psi^k \\
\approx \sum_{i=1}^{N} \ell_{Mi}(\hat{\psi}_M^k) + \frac{k}{2} \log 2\pi - \frac{k}{2} \log NT - \frac{1}{2} \log |\mathcal{T}(\hat{\psi}_M^k)|,
\]

which gives the result (34). But ignoring terms that do not depend on \( k \) (i.e., \(-\log g(y)\)) and terms that will be bounded as \( N, T \to \infty \) (i.e., they are of the smaller order of the other terms: \( \log \phi^k + (k/2) \log 2\pi - (1/2) \log |\mathcal{T}(\hat{\psi}_M^k)| + O_p(N/T) \)), we can define the integrated likelihood information criterion (ILIC) from \(-2/NT) \log P(\mathcal{M}^k|y)\) only using the relevant terms:

\[
ILIC(k) = -\frac{2}{NT} \sum_{i=1}^{N} \ell_{Mi}(\hat{\psi}_M^k) + \frac{k \log NT}{NT} \\
= -\frac{2}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \log f(z_{i,t}; \hat{\psi}_M, \hat{\lambda}_i(\hat{\psi}_M)) + \frac{k \log NT}{NT} + \frac{2}{nT} \sum_{i=1}^{n} M_i(\hat{\psi}_M). (35)
\]

Note that, comparing with \( PLIC \) in (31), the only difference in (35) is the second term (or the first penalty term), which corresponds to the standard penalty term in the BIC. This result implies that we also need to modify the BIC in the presence of the incidental parameters and the correction term (i.e., the additional penalty term) is the same as the KLIC-based (AIC-type) information criteria \( PLIC \) that we obtained in the previous section. Therefore, in general, we can construct the information criteria, which can be used in the presence of incidental parameters, given by

\[
IC^h(k) = -\frac{2}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \log f(z_{i,t}; \hat{\psi}_M, \hat{\lambda}_i(\hat{\psi}_M)) + \frac{h(N,T)}{NT} + \frac{2}{nT} \sum_{i=1}^{n} M_i(\hat{\psi}_M), \quad (36)
\]

where the choice of \( h(n, T) \) is 2 for AIC-type criteria, \( \log NT \) for BIC-type criteria. We can also conjecture that \( h(n, T) = 2 \log \log NT \) for HQ-type criteria. Note that the penalty term in \( IC^h \) is no longer deterministic; it is data-dependent and thus the model selection rule is adaptive.

\[\text{Note that the other terms diverged as } N, T \to \infty. \text{ Apparently, this condition requires that } N/T \text{ converges to some finite constant.}\]
5 Lag Order Selection in Dynamic Panel Models

5.1 Lag order selection criteria

We consider a specific example of the new model selection criterion in the context of dynamic panel regression. In particular, we consider a panel process \( \{y_{i,t}\} \) generated from the homogenous \( p_0 \)-th-order univariate autoregressive (AR(\( p_0 \))) model given by

\[
y_{i,t} = \mu_i + \sum_{j=1}^{p_0} \alpha_{p_0,j} y_{i,t-j} + \varepsilon_{i,t} \quad \text{for } i = 1, 2, \ldots, n \text{ and } t = 1, 2, \ldots, T,
\]

where \( p_0 \) is not necessarily finite.\(^9\) \( \varepsilon_{i,t} \) is serially uncorrelated and unobserved individual effects \( \mu_i \) are assumed fixed. For notational convenience we let the initial values \( (y_{i,0}, y_{i,-1}, \ldots, y_{i,-p_0+1}) \) be observed for all \( i \). We first assume the following conditions.

**Assumption A**

(i) \( \varepsilon_{i,t} \mid \{y_{i,s}\}_{s \leq t-1}, \mu_i \sim i.i.d. N(0, \sigma^2) \) for all \( i \) and \( t \), where \( 0 < \sigma^2 < \infty \). (ii) For given \( p_0 \), \( \sum_{j=1}^{p_0} |\alpha_{p_0,j}| < \infty \) and all roots of the characteristic equation \( 1 - \sum_{j=1}^{p_0} \alpha_{p_0,j} z^j = 0 \) lie outside the unit circle.

In Assumption A-(i), we assume that the higher order lags of \( y_{i,t} \) capture all the persistence and the error term does not have any serial correlation. We also exclude cross sectional dependence in \( \varepsilon_{i,t} \). Note that we assume the normality for analytical convenience, which is somewhat standard in model selection literature. We let the initial values remain unrestricted.

When \( p_0 \) is finite, the goal is to pick the correct lag order; when \( p_0 \) is infinite, the goal is to choose the lag order \( p \) among the nested models (i.e., with Gaussian distributions), which approximates the AR(\( p_0 \)) model (37) best. In this case, from (36) a new lag order selection criterion can be derived as

\[
IC^h (p) = \log \hat{\sigma}^2(p) + \frac{h(n,T)}{nT} p + \frac{2}{nT} \sum_{i=1}^{n} M_i(\hat{\alpha}(p), \hat{\sigma}^2(p))
\]

for some positive \( h(n,T) \), where \( (\hat{\alpha}(p), \hat{\sigma}^2(p)) = (\hat{\alpha}_{p1}, \ldots, \hat{\alpha}_{pp}, \hat{\sigma}_p^2) \) are the maximum modified profile likelihood estimators (i.e., bias corrected within group estimators) in a panel AR(\( p \)) regression. The lag order selection criterion (38) generalizes (31) by allowing for any positive \( h(n,T) \) instead of 2 in the original penalty term: examples are, \( h(n,T) = 2 \) for AIC type criteria as (31), \( h(n,T) = \log(nT) \) for BIC type criteria, and \( h(n,T) = 2 \log \log(nT) \)

---

\(^9\)When we are particularly interested in relatively short panels, it is reasonable to assume the true lag order \( p_0 \) to be finite. When the length of time series \( T \) is assumed to grow, however, we can consider an approximate AR(\( p_T \)) model with \( p_T \to \infty \) as \( T \to \infty \) but with further conditions (e.g., \( p_T^2/T \to 0 \)). Apparently, when we allow for an AR(\( \infty \)) process, the lag order selection problem becomes to choose the best AR(\( p \)) model that approximates the AR(\( \infty \)) process best.
for HQ type criteria. Note that from (18) it can easily derived that

$$
\sum_{i=1}^{n} M_i(\widehat{\alpha}(p), \widehat{\sigma}^2(p)) = \frac{1}{2\widehat{\sigma}^2(p)} \sum_{i=1}^{n} \sum_{t=-m}^{m} K_t \sum_{t=\max\{1, t+1\}}^{\min\{T, T+t\}} \widetilde{\varepsilon}_{i,t}^W(p)\widetilde{\varepsilon}_{i,t-t}^W(p),
$$

where \(\widetilde{\varepsilon}_{i,t}^W(p) = y_{i,t}^W - \sum_{j=1}^{p} \widehat{\alpha}_j y_{i,t-j}^W\) is the within-group estimation residual with \(y_{i,t}^W = y_{i,t} - T^{-1} \sum_{s=1}^{T} y_{i,s}\) indicating the within-transformation.\(^{10}\) Therefore, the penalty term in (38) is given by

$$
\frac{h(n, T)}{nT} p + \frac{1}{T} \widehat{R}_{n,T}(p),
$$

where \(\widehat{R}_{n,T}(p) = (2/n) \sum_{i=1}^{n} M_i(\widehat{\alpha}(p), \widehat{\sigma}^2(p))\) corresponds to the long-run autocorrelation estimator of \(\widetilde{\varepsilon}_{i,t}^W(p)\).

The interpretation of this new lag order selection criterion is quite intuitive. In (38), the first term indicates the goodness-of-fit as usual. The second term \(p \times h(n, T) / nT\), which is the first part of the penalty term, controls for the degrees of freedom of parameter of interest and thus tries to choose parsimonious models. Finally, the last term \((1/T)\widehat{R}_{n,T}(p)\), which is the second part of the penalty term, reflects the presence of nuisance parameters whose dimension is large. Particularly when \(h(n, T) = 2\), the entire penalty term can be rewritten as \((h(n, T) / nT)\{p + n \times (\widehat{R}_{n,T}(p)/2)\}\), which shows that the efficient number of parameters is not \(p + n\) in this case; the effect from the incidental parameters \(n\) is controlled by the size of \(\widehat{R}_{n,T}(p)/2\). Remark 3 below discusses more about the effective number of parameters in the context of the model complexity.

In fact, this last component of the penalty term tries to rule out any possible erroneous serial correlation in the regression error term. Since the within-transformation incurs serial correlation in the AR panel regression even when the original error \(\varepsilon_{i,t}\) is serially uncorrelated, \(\widehat{R}_{n,T}(p)\) will measure the degree of such pseudo serial correlation from the artificial transformation. Note that the serial correlation would be exacerbated if the lag order is not correctly chosen, particularly when it is under-selected.\(^{11}\) The additional penalty term controls for such aspect and thus it automatically controls for the under-selection probability. At the same time, this last term is positive and it adds heavier penalty, which also functions to control for the over-selection probability. Note that the last penalty term is \(O_p(1/T)\) and thus its roll becomes minor for large \(T\), which is indeed well expected since the incidental parameter problem gets attenuated by large \(T\).

---

\(^{10}\)The within transformation indeed corresponds to maximizing out the fixed effects \(\mu_i\)'s in MLE (i.e., forming the profile likelihood).

\(^{11}\)Note that the maximum modified profile likelihood estimators does not completely eliminate the within-group bias, which will give some pseudo serial correlation in the error term.
Remark 3 (Model complexity) The new penalty term of $IC^h(p)$ could be understood as a proper choice of the effective degrees of freedom (i.e., the model complexity). For example, Hodges and Sargent (2001) consider a one-way panel data model given by $y_{it} | \mu_i, \sigma^2 \sim iidN(\mu_i, \sigma^2)$ for all $i = 1, \ldots, n$ and $t = 1, \ldots, T$, where $\mu_i | \nu, \tau^2 \sim iidN(\nu, \tau^2)$ for all $i$. Under this specification, the number of parameters can be counted as either $n + 1$ if $\mu_i$ is considered as fixed effect (e.g., $\tau^2 = \infty$); or 3 if $\mu_i$ is considered as random effect. It is proposed that the model complexity can be measured by the degrees of freedom and it corresponds to the rank of the space into which $y_{i,t}$ is projected to give the fitted value $\hat{y}_{i,t}$. In this particular example, the degrees of freedom $\rho$ turns out to be

$$\rho = \frac{nT + (\sigma^2/\tau^2)}{T + (\sigma^2/\tau^2)} = \frac{(\sigma^2/\tau^2)}{1 + (\sigma^2/\tau^2)} + n(\sigma^2/\tau^2)/(T-1) \equiv \rho_1 + \rho_2.$$  

Notice that the first term $\rho_1$ corresponds to the “θ” value defined by Maddala (1971, eq.1.3 on p.343), which measures the weight given to the between-group variation in the standard random effect least squares estimator. Apparently, $\rho_1 \to 0$ if $T \to \infty$ or $\sigma^2/\tau^2 \to 0$, which reduces the random effect estimator to the standard within-group (or the fixed effect) estimator by ignoring between-group variations. The degrees of freedom $\rho$ also reflects such idea because for given $n$, $\rho \to n$ as the model gets closer to the fixed effect case (i.e., $T \to \infty$ or $\sigma^2/\tau^2 \to 0$ and thus the between-group variation is completely ignored) but $\rho$ will be close to one if $\sigma^2/\tau^2$ is large. The lag order selection example in this subsection corresponds to the case of fixed effect but the degrees of freedom in our case is different from $n$; it is instead given by $nR_{n,T}(p)/2$, which measures the model complexity somewhat differently. In a more general setup including nonlinear models, the model complexity is closely related with the Vapnik-Chervonenkis dimension (e.g., Cherkassky et al. (1999)).

In general, under the stationarity, the probability limit of the long-run autocorrelation estimator $R_{n,T}(p)$ in (39) is bounded and the entire penalty term multiplied by the sample size $nT$ (i.e., $h(n,T)p + nR_{n,T}(p)$) increases with the sample size. As noted in Shibata (1980), therefore, we can conjecture that the new lag order selection criterion is not asymptotically optimal when the true data generating model is $AR(\infty)$ with finite $\sigma^2$ (i.e., when $p_0 = \infty$, $\lim_{n,T \to \infty} [IC^h(p^*) / \inf_{p \geq 0} IC^h(p)] \neq 1$, where $p^*$ is the lag order estimator from $IC^h(p)$; see e.g., Li (1987)) even when $h(n,T)$ is fixed like $h(n,T) = 2$. If we assume that the true lag order $p_0$ exists and is finite, however, we can show that the new order selection criterion (38) is consistent under a certain condition, where we define a lag order estimator $p^*$ is consistent and thus the lag order selection criterion is consistent) if it satisfies $\lim_{n,T \to \infty} \mathbb{P}(p^* = p_0) = 1$.\(^{12}\)

\(^{12}\)This definition is somewhat different from the usual probability limit, but it is equivalent for integer valued random variables. $p^*$ is strongly consistent if $\mathbb{P}(\lim_{n,T \to \infty} p^* = p_0) = 1$. It is known that in the
Theorem 5 Under Assumption A, if we let \( n/T \to \gamma \in (0, \infty) \) and \( n/T^3 \to 0 \) as \( n, T \to \infty \), \( IC^h(p) \) is a consistent lag order selection criterion when the true lag order \( p_0 (\geq 1) \) is finite, provided that \( h(n, T) \) satisfies \( h(n, T)/nT \to 0 \) and \( h(n, T) \to \infty \) as \( n, T \to \infty \).

As discussed above, examples of \( h(n, T) \) for consistent criteria are \( \log(nT) \) and \( \omega \log \log(nT) \) for some \( \omega \geq 2 \), where the first one is the BIC type penalty term and the second one is the HQ type penalty term. We will see how the new lag order selection criteria perform by simulation studies in the following section.

It should be noted that Theorem 5 does not provide an analytical evidence why the new lag order selection criteria work better than the standard criteria, since the standard criteria based on the bias-corrected estimators (e.g., \( \log \hat{\sigma}^2(p) + p \times h(n, T)/nT \)) also satisfy the consistency with a suitable choice of \( h(n, T) \to \infty \). In fact, similarly as Guyon and Yao (1999), it can be shown that the under-selection probability vanishes exponentially fast for both cases (provided \( h(n, T)/nT \) is small), while the over-selection probability decreases at a slower rate depending on the magnitude of the penalty term. Therefore, the improvement of correct lag-order-selection probability mainly comes from the reduction of the over-selection probability of the new lag order selection criterion. Intuitively, since the new criterion includes additional positive penalty term, the lag order estimates cannot be larger than one from the conventional lag order selection criterion. The following theorem states that the over-selection probability reduced asymptotically by modifying the penalty term as in the new lag order selection criterion (39).

Theorem 6 Suppose the conditions in Theorem 5 hold. For some finite positive integer \( \overline{p} \), if we let \( p^* = \arg \min_{0 \leq p \leq \overline{p}} IC^h_0(p) \) with \( IC^h_0(p) = \log \hat{\sigma}^2(p) + p(h(n, T)/nT) \), then \( \lim_{n, T \to \infty} P(p^* > p_0) \geq \lim_{n, T \to \infty} P(p^* > p_0) \).

Finally note that Lee (2006) suggests a simplified form of the order selection criterion (38) as

\[
IC^h_c(p) = \log \hat{\sigma}^2(p) + \frac{1}{nT} \left\{ h(n, T) + c \left( \frac{n}{T} \right) \right\} \quad \text{(40)}
\]

for some positive constant \( c < \infty \), which uses deterministic penalty terms instead of data-dependent ones. The additional penalty term \( cp/T^2 \) in (40) is introduced to offset the higher order bias in the maximum profile likelihood estimator \( \hat{\sigma}^2(p) \) since it can be shown that \( \lim_{n \to \infty} \hat{\sigma}^2(p) - \sigma^2 = -cp/T^2 + O(T^{-3}) \), where the constant \( c \) depends on the parameter values \( \alpha_p \) and clearly on the stability of the system. Such bias is typically exacerbated when \( T \) is small and the system is less stable (i.e., close to unit root). For example, when \( p = 1 \), it can be derived that \( c = (1 + \alpha_1)/(1 - \alpha_1) \). If we ignore the asymptotic bias of \( \hat{\sigma}^2(p) \) and standard time series context, BIC and properly defined P1C are strongly consistent criteria; HQ is weakly consistent but not strongly; and other order selection criteria, such as final prediction error (FPE) and AIC are not consistent for finite \( p_0 \).
construct a model selection criterion without such adjustment, the total regression error is equal to the biases of the autoregressive coefficients plus the original \(i.i.d\). disturbance. As a consequence, the regression error has an erroneous serial correlation and behaves like an ARMA process, or an \(AR(\infty)\) process. Hence, the model selection is biased upward because it is prone to fit the model with \(p\) as large as possible to reflect the erroneous serial correlation. The second part of the penalty term in (40), or a heavier penalty overall, controls such phenomenon. Interestingly, this simplified order selection criterion can be obtained from the proof of Theorem 5, and thus it shares the same asymptotic properties as \(IC^h(p)\), like Theorems 5 and 6.

**Corollary 7** Under the same condition as Theorem 5, \(IC^h_c(p)\) shares the same asymptotic properties as \(IC^h(p)\).

### 5.2 Simulations

We compare the lag order selection criteria developed in the previous subsection with the conventional time series model selection methods. We first define the three most commonly used information criteria, which use the pooled information:

\[
\begin{align*}
AIC(p) &= \log \hat{\sigma}^2(p) + \frac{2}{nT}p, \\
BIC(p) &= \log \hat{\sigma}^2(p) + \frac{\log(nT)}{nT}p, \\
HQ(p) &= \log \hat{\sigma}^2(p) + \frac{2\log \log(nT)}{nT}p,
\end{align*}
\]

where \(\hat{\sigma}^2(p)\) is an bias-corrected estimate for \(\sigma^2\) in the panel AR(p) model. We only count the number of parameters as \(p\) instead of \(p + n\) (i.e., including fixed effect parameters). As well expected, the simulation results shows that constructing penalty terms using \(p + n\) too heavily penalize the criteria so that they yield high under-selection probabilities. The effective number of observations in each time series is adjusted to reflect the degrees of freedom by \(T - p\) (e.g., Ng and Perron (2005)). For the new criteria, we consider the following criteria that we suggested:

\[
\begin{align*}
IC^{AIC}(p) &= \log \hat{\sigma}^2(p) + \frac{2}{nT}p + \frac{1}{T}\tilde{R}_{n,T}(p), \\
IC^{BIC}(p) &= \log \hat{\sigma}^2(p) + \frac{\log(nT)}{nT}p + \frac{1}{T}\tilde{R}_{n,T}(p), \\
IC^{HQ}(p) &= \log \hat{\sigma}^2(p) + \frac{\log \log(nT)}{nT}p + \frac{1}{T}\tilde{R}_{n,T}(p),
\end{align*}
\]
as well as the simplified form as (40):

\[
    IC_c^{AIC} (p) = \log \hat{\sigma}^2 (p) + \frac{1}{nT} \left\{ 2 + \frac{n}{T} \right\} p,
\]

\[
    IC_c^{BIC} (p) = \log \hat{\sigma}^2 (p) + \frac{1}{nT} \left\{ \log (nT) + \frac{n}{T} \right\} p,
\]

\[
    IC_c^{HQ} (p) = \log \hat{\sigma}^2 (p) + \frac{1}{nT} \left\{ 2 \log \log (nT) + \frac{n}{T} \right\} p,
\]

in which \( c \) is simply chosen to one.

We generate \( AR (p_0) \) dynamic panel processes, with \( p_0 \) ranging from 1 to 4, of the form

\[
    y_{i,t} = \mu_i + \sum_{j=1}^{p_0} \alpha_{p_0j} y_{i,t-j} + \varepsilon_{i,t} \quad \text{for} \quad i = 1, 2, \ldots, n \quad \text{and} \quad t = 1, 2, \ldots, T,
\]

where \( \alpha_{p_0j} = 0.15 \) for all \( j = 1, \ldots, p_0 \). For each \( AR (p_0) \) model, all the autoregressive coefficients have the same value so that all the lagged terms are equally important. We consider nine different cases by combining different sample sizes of \( n = 20, 50, 100 \) and \( T = 12, 25, 50 \). Fixed effects \( \mu_i \) are randomly drawn from \( \mathcal{U} (-0.5, 0.5) \) and \( \varepsilon_{i,t} \) from \( \mathcal{N} (0, 1) \). We use the bias corrected within-group estimators (e.g., Lee (2011)) for \( \hat{\alpha}_{p_0j} \)'s and iterate the entire procedure 1000 times to compare the performance of different order selection criteria. For each case, we choose the optimal lag order \( p^* \) to minimize the criteria above, where we search the lag order from 1 to 10 (i.e., \( p = 10 \)). The simulation results are provided in Tables 1 to 4, which present the average values of \( p^* \) over 1000 iterations.

[TABLES 1 to 4 about here]

It is very promising that all the new lag order selection criteria, \( IC \) and \( IC_c \), perform much better than the two most commonly used criteria, \( AIC \), \( BIC \) and \( HQ \). In order to look at the distributional characteristics, we also provide Figures 1 to 4 for the case of \( (n, T) = (100, 50) \).

[FIGURES 1 to 4 about here]

One interesting finding is that \( BIC \) tends to overfit the panel models, which is contrary to the well known property that \( BIC \) normally underfits in the pure time series setup. On the other hand, the figures consistently show that the new order selection criteria significantly reduce the over-selection probabilities. Though we do not present this particular result of our simulation, heavier penalty of the new information criteria \( IC \) and \( IC_c \) slightly increases the under-selection probabilities. But the increment of the under-selection probability is very minor, so that the overall correct-selection probabilities increase notably.

[FIGURE 5 about here]
When $T$ is very small or $n$ is very large, so that the sample size ratio $n/T$ is large, the order selection performance is not much satisfactory or overall cases, which is somewhat expected due to the very limited number of time series observations and large number of nuisance parameters. However, as $T$ grows, the performances get better uniformly. (See Figure 5.) This is intuitively appealing because the dynamic structure is mainly determined by the time series dimension. But unlike the conventional time series information criteria, the new criteria tend to choose the correct lag orders even when $n$ is large, provided that $T$ is not so small. One remark is that though the simulation results look like $IC_c$ works better than $IC$, $IC_c$ cannot be always preferred to $IC$ empirically since it has larger under-selection probability than $IC$ has.

Lastly, one interesting finding is that, in general, the lag order selection is more accurate with $(n, T) = (50, 50)$ than $(n, T) = (100, 50)$. This implies that the sample size ratio $n/T$ matters in lag order selection: the smaller $n/T$, the better work the information criteria. As also stated in the proof of Theorem 5 in Appendix, it is because the under-selection probability becomes smaller as $T$ gets larger, whereas the over-selection probability becomes smaller as $n/T$ gets larger. Since the reduction of over-selection probability is the main source of improvement of the new information criterion, this simulation results confirms the analytical findings.

6 Concluding Remarks

It is not uncommon that only a sub-parameters are of the main interest among the entire set of the parameters. In such cases, the nuisance parameters account for the aspect of the model that are not of the main concern but they are still important for a realistic statistical modeling. Particularly when the dimension of the nuisance parameter is large, properly dealing with the nuisance parameters is important for valid inferences. As the current paper demonstrates, a proper model selection also should account for the nuisance parameters to obtain a meaningful specification. We deal with such nuisance parameters either using the profile likelihood (for the AIC-type approach) or the integrated likelihood (for the BIC-type approach) to develop a new model selection criterion that can be used in the presence of nuisance parameters. The penalty term is data-dependent and it properly controls for the model complexity.

For more general models, the semiparametric models could be handled in a similar context if we consider the nonparametric component as infinite dimensional parameters. In particular, using a similar approach as Severini and Wong (1992), for example, we can consider a model $f(z_i; \psi, \lambda_i(w_i))$ for given observations $\{z_i, w_i\}$, where $\lambda_i(w) = (\lambda_{1i}, \lambda_{2}(w))^T$ with $\lambda_2(\cdot)$ being an unknown (scalar) function. In this case, we could see that $\lambda_{2i} = \lambda_2(w_i)$ as the realization of $\lambda_2(\cdot)$ for each $i$th observation. Though it should be proved in the context of QML
estimation, we conjecture that for \( \tilde{\lambda}_{2,\psi}(\omega_i) = \arg \max_{\lambda} \sum_{t=1}^{T} \log f(z_{i,t}; \psi, \lambda_i, \lambda)K((\omega_i - w_{i,t})/h) \), where \( K \) and \( h \) are properly defined kernel function and the bandwidth parameter, respectively, we could derive a similar result as Theorem 3 under proper technical conditions. Note that, however, the conditions between for the incidental parameters and for the nonparametric components are different and thus their effects on the parametric component \( \psi \) need to be treated differently.\(^{13}\)

**Appendix: Mathematical Proofs**

**Proof of Lemma 1** First note that, since \( \lambda_i \)'s are separable across \( i \), we have

\[
D_P (g \parallel f(\psi, \lambda); \psi) = \min_{\lambda} D (g \parallel f(\psi, \lambda))
= \int g(z) \log g(z) \, dz - \sum_{i=1}^{n} \max_{\lambda_i \in \Lambda} \int g_i(z) \log f(z; \psi, \lambda_i) \, dz
= \int g(z) \log g(z) \, dz - \sum_{i=1}^{n} \int g_i(z) \log f(z; \psi, \lambda_i(\psi)) \, dz,
\]

where the last equality is from the definition:

\[
\lambda_i(\psi) = \arg \max_{\lambda_i \in \Lambda} D (g_i \parallel f(\psi, \lambda_i)) = \arg \max_{\lambda_i \in \Lambda} \int g_i(z) \log f(z; \psi, \lambda_i) \, dz.
\]

It thus follows that

\[
D_P (g \parallel f(\psi, \lambda); \psi) = \int g(z) \log g(z) \, dz - \sum_{i=1}^{n} \int g_i(z) \log f(z; \psi, \tilde{\lambda}_i(\psi)) \, dz
+ \sum_{i=1}^{n} \int g_i(z) \log \left( \frac{f(z; \psi, \tilde{\lambda}_i(\psi))}{f(z; \psi, \lambda_i(\psi))} \right) \, dz,
\]

in which the first line corresponds to \( D (g \parallel f_P) \) yielding the expression (26). \( \square \)

**Proof of Lemma 2** From (16) and (17), it can be derived that

\[
\int g_i(z) \log \left( \frac{f(z; \psi, \tilde{\lambda}_i(\psi))}{f(z; \psi, \lambda_i(\psi))} \right) \, dz = \mathbb{E}_{g_i} \left[ \log f(z_{i,t}; \psi, \tilde{\lambda}_i(\psi)) - \log f(z_{i,t}; \psi, \lambda_i(\psi)) \right]
= T^{-1} \mathbb{E}_{g_i} \left[ \ell_{P_i}(\psi) - \ell(\psi, \lambda_i(\psi)) \right]
= T^{-1} \mathbb{E}_{g_i} [M_i(\psi)] + O(T^{-3/2})
\]

for a given \( \psi \), where the last equality is because \( \tilde{\lambda}_i(\psi) - \lambda_i(\psi) = O_p(T^{-1/2}) \) under the regularity conditions. Note that the definition in (17) (and thus the last equality above) is indeed from the expansion (16), so as long as we define \( M_i(\psi) \) same as the leading bias term

\(^{13}\)In particular, the semiparametric component estimator does not affect the asymptotics of the parametric component estimator (e.g., two step estimation in semiparametric models) under the proper conditions (e.g., Andrews (1994) and Newey (1994)), whereas the nuisance parameters do.

27
in such expansion, we can conclude the result above independent of the particular form of $M_i(\psi)$. The result follows by summing each component over $i = 1, \cdots, n$. □

Proof of Theorem 3  For each $i$, we define $G_i(\cdot; \epsilon) = G_i(\cdot) + \epsilon(\hat{G}_i(\cdot) - G_i(\cdot))$ for some $\epsilon \in [0, 1]$. $G(\cdot; \epsilon)$, $G(\cdot)$ and $\hat{G}(\cdot)$ denote the collection of the marginal distributions (i.e., $G(Z; \epsilon) = (G_1(z_1; \epsilon), \cdots, G_n(z_n; \epsilon))$ and similarly for the others). We will also use notations $G_i$ and $\hat{G}_i$ instead of $G_i(\cdot)$ and $\hat{G}_i(\cdot)$ respectively if there is no confusion. For a fixed $\epsilon$, we let $\psi(\epsilon) = H(G(\cdot; \epsilon))$ be the solution of

$$
\frac{1}{n} \sum_{i=1}^{n} \int \frac{\partial}{\partial \psi} Q(z; \psi(\epsilon), G_i(z; \epsilon)) dG_i(z; \epsilon) \bigg|_{\psi(\epsilon)} = 0,
$$

(A.1)

where

$$
Q(z; \psi(\epsilon), G_i(z; \epsilon)) = \log f_i(z; \psi(\epsilon), \lambda_i(\psi(\epsilon); G_i(z; \epsilon))) - \mu_i(\psi(\epsilon); G_i(z; \epsilon))
$$

with

$$
\lambda_i(\psi(\epsilon); G_i(z; \epsilon)) = \begin{cases} 
\lambda_i(\psi(0); G_i) = \lambda_i(\psi(0)) & \text{if } \epsilon = 0; \\
\lambda_i(\psi(1); \hat{G}_i) = \hat{\lambda}_i(\psi(1)) & \text{if } \epsilon = 1,
\end{cases}
$$

and

$$
\mu_i(\psi(\epsilon); G_i(z; \epsilon)) = \begin{cases} 
\mu_i(\psi(0); G_i) = 0 & \text{if } \epsilon = 0; \\
\mu_i(\psi(1); \hat{G}_i) = M_i(\psi(1)) & \text{if } \epsilon = 1.
\end{cases}
$$

Recall that $\lambda_i(\psi)$, $\hat{\lambda}_i(\psi)$ and $M_i(\psi)$ are defined as (10), (3) and (5), respectively. Since

$$
Q(z; \psi(0), G_i(z; 0)) = \log f_i(z; \psi(1), \lambda_i(\psi(1))) \quad \text{and} \quad Q(z; \psi(1), G_i(z; 1)) = \log f_i(z; \psi(1), \hat{\lambda}_i(\psi(1))) - M_i(\psi(1)),
$$

it thus follows that $\psi(0) = H(G) = \psi_0$ and $\psi(1) = H(\hat{G}) = \psi_M$ by definition. Assuming $\hat{\psi}(\cdot)$ satisfies the usual assumption for a Taylor expansion to be valid (e.g., it is of the second order continuously differentiable and has a finite third order derivative over $[0, 1]$), the Taylor expansion for $\psi(\cdot)$ could be expressed as the following functional Taylor expansion for $H(\cdot)$: (e.g., Serfling (1980), Chapter 6.2)

$$
H(\hat{G}) - H(G) = d_1H(G; \hat{G} - G) + o_p((nT)^{-1/2}),
$$

(A.2)

where $d_1H(G; \hat{G} - G)$ is the standard first order Gâteaux differential of $H$ at $G$ in the direction of $\hat{G}$ and it is assumed to exist. The remainder term is negligible as shown in Hahn and Kuersteiner (2007) for $n, T \to \infty$ satisfying $n/T \to \gamma \in (0, \infty)$ and $n/T^3 \to 0$. It is well known that we can express $d_1H(G; \hat{G} - G)$ in the form of $V$-statistic as (e.g., Serfling (1980), Chapter 6.3; Withers (1983))

$$
d_1H(G; \hat{G} - G) = \int H^{(1)}(z; G)d[\hat{G}(z) - G(z)],
$$

where $H^{(1)}(z; G)$ is a symmetric function with $\int H^{(1)}(z; G)dG(z) = 0$. More precisely, by
differentiating (A.1) with respect to $\epsilon$, we have

$$0 = \frac{1}{n} \sum_{i=1}^{n} \int \frac{\partial^2}{\partial \psi \partial \psi'} Q(z; \psi(\epsilon), G_i(z; \epsilon)) dG_i(z; \epsilon) \times d_1 H(G; \hat{G} - G)$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \int \frac{\partial}{\partial \psi} Q(z; \psi(\epsilon), G_i(z; \epsilon)) d(\hat{G}_i(z) - G_i(z))$$

and by evaluating this result at $\epsilon = 0$ we derive (e.g., Hahn and Kuersteiner (2007))

$$d_1 H(G; \hat{G} - G) = \left( \frac{1}{n} \sum_{i=1}^{n} \int - \frac{\partial^2 \log f_i(z; \psi_0, \lambda_i)}{\partial \psi \partial \psi'} dG_i \right)^{-1} \times \frac{1}{n} \sum_{i=1}^{n} \int \frac{\partial \log f_i(z; \psi_0, \lambda_i)}{\partial \psi} dG_i$$

because $\lambda_i(\psi_0) = \lambda_i$ and $\int [\partial \log f_i(z; \psi_0, \lambda_i) / \partial \psi] dG_i = 0$. Note that this result implies that

$$H^{(1)}(z_{i,t}; G) = \left( \frac{1}{n} \sum_{i=1}^{n} \int - \frac{\partial^2 \log f_i(z; \psi, \lambda_i)}{\partial \psi \partial \psi'} dG_i \right)^{-1} \times \frac{\partial \log f_i(z_{i,t}; \psi, \lambda_i)}{\partial \psi} \bigg|_{\psi = \hat{\psi}(G)},$$

which indeed corresponds to the influence function of $\hat{\psi}_M = H(\hat{G})$ in the standard M-estimation theory (e.g., Huber (1981)). Therefore, from (A.2) and (A.3) we have the approximation of $\hat{\psi}_M$ given by

$$\hat{\psi}_M = H(\hat{G}) = H(G) + \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} H^{(1)}(z_{i,t}; G) + o_p((nT)^{-1/2}),$$

where $H^{(1)}(z_{i,t}; G)$ is given in (A.4). Apparently, this last results shows that $\hat{\psi}_M$ is $\sqrt{nT}$-consistent to $H(G)$ since $(nT)^{-1/2} \sum_{i=1}^{n} \sum_{t=1}^{T} H^{(1)}(z_{i,t}; G)$ is asymptotically normal with mean zero and variance

$$\frac{1}{nT} \sum_{i=1}^{n} \int \sum_{t=1}^{T} \sum_{s=1}^{T} H^{(1)}(z_{i,t}; G) H^{(1)}(z_{i,s}; G)' dG_i.$$
Now, for $\psi = (\psi_1, \cdots, \psi_r)'$, we let $H^{(1)}(z_{i,t}; G) = (H_r^{(1)}(z_{i,t}; G), \cdots, H_1^{(1)}(z_{i,t}; G))'$. Then, similarly as (Konishi and Kitagawa, 1996, Theorem 2.1), we have

$$
\mathbb{E}_g \left[ \phi_P(\widehat{\psi}_M) \right] = -\frac{1}{n} \sum_{i=1}^{n} \int \log f_i(z; \psi_0, \widehat{\lambda}_i(\psi_0)) dG_i \tag{A.5}
$$

$$
-\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{r} \int \frac{1}{nT} \sum_{j=1}^{n} \sum_{t=1}^{T} H_k^{(1)}(z_{j,t}; G) dG_j \int \frac{\partial \log f_i(z; \psi_0, \widehat{\lambda}_i(\psi_0))}{\partial \psi_k} dG_i
$$

$$
+ o((nT)^{-1/2})
$$

$$
= -\frac{1}{n} \sum_{i=1}^{n} \int \log f_i(z; \psi_0, \widehat{\lambda}_i(\psi_0)) dG_i + o((nT)^{-1/2})
$$

since $\int H_k^{(1)}(z; G) dG_i = 0$ for each $k = 1, \cdots, r$ by construction but

$$
\mathbb{E}_g \left[ \phi_P(\widehat{\psi}_M) \right] = -\frac{1}{n} \sum_{i=1}^{n} \int \log f_i(z; \psi_0, \widehat{\lambda}_i(\psi_0)) dG_i \tag{A.6}
$$

$$
-\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{r} \int \left[ \frac{1}{nT} \sum_{j=1}^{n} \sum_{t=1}^{T} \sum_{s=1}^{T} H_k^{(1)}(z_{j,t}; G) \frac{\partial \log f_i(z_{i,s}; \psi_0, \widehat{\lambda}_i(\psi_0))}{\partial \psi_k} \right] dG_i
$$

$$
+ o((nT)^{-1/2}).
$$

We thus derive that

$$
\mathbb{E}_g \left[ \phi_P(\widehat{\psi}_M) - \phi_P(\widehat{\psi}_M) \right]
$$

$$
= -\frac{1}{nT} \text{tr} \left\{ \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{g_i} \left[ \frac{1}{T} \sum_{t=1}^{T} \left. \frac{\partial^2 \log f_i(z_{i,t}; \psi, \lambda_i(\psi))}{\partial \psi \partial \psi'} \right|_{\psi = H(G)} \right] \right)^{-1}
$$

$$
\times \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{g_i} \left[ \left. \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \left. \frac{\partial \log f_i(z_{i,s}; \psi, \lambda_i(\psi))}{\partial \psi} \right|_{\psi = H(G)} \left. \frac{\partial \log f_i(z_{i,t}; \psi, \widehat{\lambda}_i(\psi))}{\partial \psi'} \right|_{\psi = H(G)} \right] \right\}
$$

$$
+ o \left( (nT)^{-1} \right)
$$

since the higher order terms in (A.5) and (A.6) are indeed the same and cancelled out to yield the remainder term of $B_P(G)$ is $o((nT)^{-1})$ as Konishi and Kitagawa (1996). The expression for $B_P(G)$ follows as $B_P(G) = \mathbb{E}_g[\phi_P(\widehat{\psi}_M) - \phi_P(\widehat{\psi}_M)] - n^{-1} \delta(\widehat{\psi}_M; G)$.

Finally, in order to derive an estimator for $B_P(G)$, we first note that

$$
\log f(z_{i,t}; \psi, \lambda_i(\psi)) = \log f_{P_i}(z_{i,t}; \psi) + \log \left( \frac{f(z_{i,t}; \psi, \lambda_i(\psi))}{f(z_{i,t}; \psi, \widehat{\lambda}_i(\psi))} \right)
$$

$$
= \log f_{M_i}(z_{i,t}; \psi) + \left\{ \frac{1}{T} M_i(\psi) + \log \left( \frac{f(z_{i,t}; \psi, \lambda_i(\psi))}{f(z_{i,t}; \psi, \widehat{\lambda}_i(\psi))} \right) \right\}
$$

$$
= \log f_{M_i}(z_{i,t}; \psi) + O_p(T^{-3/2})
$$
by construction, where the last equation is from Lemma 2. Loosely speaking, the modified profile likelihood \( f_{M_i}(z; \psi) \) behaves like the desirable likelihood \( f(z; \psi, \lambda_i(\psi)) \). It thus follows that the solution of

\[
\frac{1}{n} \sum_{i=1}^{n} \int \frac{\partial}{\partial \psi} \log f_{M_i}(z_i; \psi) \, dG_i = 0 \tag{A.7}
\]

is asymptotically equivalent to that of

\[
\frac{1}{n} \sum_{i=1}^{n} \int \frac{\partial}{\partial \psi} \log f(z_i; \psi, \lambda_i(\psi)) \, dG_i = 0. \tag{A.8}
\]

We denote \( \tilde{\psi}_M \) be the solution for (A.7). Note that \( \psi_0 = H(G) \) solves (A.8) since \( \lambda_i(\psi_0) = \lambda_{i0} \) for all \( i \). Since \( f \) is twice continuously differentiable at the neighborhood around \( \psi_0 \) with probability one, where the radius is smaller than \( O(T^{-3/2}) \), then we can claim that \( \tilde{\psi}_M - \psi_0 = O(T^{-3/2}) \). Therefore, for (moderately) large \( T \), we can define \( B_P(G) \) using \( \log f_{M_i}(z_i; \psi) \) instead of \( \log f(z_i; \psi, \lambda_i(\psi)) \). Since \( \tilde{\psi}_M \) is \( \sqrt{nT} \)-consistent to \( \psi_0 \) under the conditions in Theorem 3 (e.g., the Lyapunov’s theorem; Sartori (2003), Hahn and Kuersteiner (2007))\(^{14} \) and by the Envelope theorem, a consistent estimator for \( B_P(G) \) thus can be obtained as

\[
B_P(\tilde{G}) = -\frac{1}{nT} tr \left\{ \left( -\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{\partial^2 \log f_{M_i}(z_{i,t}; \tilde{\psi}_M)}{\partial \psi \partial \psi'} \right)^{-1} \right. \\
\times \left. \left( \frac{1}{nT} \sum_{i=1}^{n} \sum_{\ell=-m}^{m} \min\{T, T+\ell\} \sum_{t=1}^{\max\{1, \ell+1\}} \frac{\partial \log f_{M_i}(z_{i,t}; \tilde{\psi}_M)}{\partial \psi} \frac{\partial \log f_{P_i}(z_{i,t-\ell}; \tilde{\psi}_M)}{\partial \psi'} \right) \right\} \\
- \frac{1}{nT} \sum_{i=1}^{n} M_i(\tilde{\psi}_M),
\]

where the second line is using the HAC-type estimator for some truncation parameter \( m \geq 0 \) such that \( m/T^{1/2} \to 0 \) as \( T \to \infty \); the third line is from Lemma 2, where the remainder term is negligible provided that \( T \to \infty \).

**Proof of Corollary 4**  
First note that \( \partial \ell_i(\psi_0, \lambda_i(\psi_0)) / \partial \psi = u_i^e \) by construction and thus

\[
I(G) = \frac{1}{nT} \sum_{i=1}^{n} \int -\frac{\partial^2 \ell_i(\psi_0, \lambda_i(\psi_0))}{\partial \psi \partial \psi'} \, dG_i = \frac{1}{nT} \sum_{i=1}^{n} \mathbb{E}_{g_i} \left[ -\frac{\partial u_i^e}{\partial \psi'} \right].
\]

When \( g \) is nested in \( f \), moreover, the standard information matrix identity holds and thus \( \mathbb{E}_{g_i} \left[ -\partial u_i^e / \partial \psi' \right] = \mathbb{E}_{g_i} [u_i^e u_i^e'] = I_i^e \), which yields \( I(G) = (nT)^{-1} \sum_{i=1}^{n} I_i^e \). For \( J(G) \), since

\[^{14}\text{Recall that } \tilde{\psi}_M \text{ automatically correct the first order asymptotic bias in the asymptotic distribution of the standard QML estimator } \hat{\psi} \text{ and } \sqrt{nT}(\hat{\psi}_M - \psi_0) \to_d \mathcal{N}(0, \Omega_\psi) \text{ as } n, T \to \infty \text{ for } \Omega_\psi = \int f(z; G) \phi(z; G) \, dG(z) > 0 \text{ by construction, where } \phi(z; G) \text{ is the influence function of } \hat{\psi}_M \text{ defined above.} \]
Therefore, residual sum of squares does not increase as the number of regressors increases (e.g., Shibata limit order. We where \( \lambda_i(\psi_0) \) from (A.2) and (A.3) since 

\[
J(G) = \frac{1}{nT} \sum_{i=1}^{n} \int \left[ \frac{\partial \ell_i(\psi_0, \lambda_i(\psi_0))}{\partial \psi} \frac{\partial \ell_P(\psi_0)}{\partial \psi'} \right] dG_i 
\]

\[
\approx \frac{1}{nT} \sum_{i=1}^{n} \int \frac{\partial \ell_i(\psi_0, \lambda_i(\psi_0))}{\partial \psi} \frac{\partial \ell_i(\psi_0, \lambda_i(\psi_0))}{\partial \psi'} dG_i 
+ \frac{1}{nT} \sum_{i=1}^{n} \int \frac{\partial \ell_i(\psi_0, \lambda_i(\psi_0))}{\partial \psi} b_i(\psi_0)' dG_i
\]

with the approximation error is smaller than \( O(T^{-3/2}) \) since \( \int [\partial \ell_i(\psi_0, \lambda_i(\psi_0)) / \partial \psi] dG_i = 0 \). Therefore, we obtain \( J(G) = (nT)^{-1} \sum_{i=1}^{n} I_i^2 + (nT)^{-1} \sum_{i=1}^{n} \mathbb{E}_{\psi_i} [u_i^t b_i(\psi_0)'] \) by definition of the efficient information \( I_i^2 \).

For \( \mathbb{E}_{\psi_i}[M_i(\psi_0)] \), since we have \( M_i(H(\tilde{G})) = M_i(H(G)) + [\partial M_i(H(G))/\partial \psi]'[H(\tilde{G}) - H(G)]' + o_p(T^{-1/2}) \), using a similar argument as the proof of Theorem 3, it can be derived that

\[
\mathbb{E}_{\psi_i}[M_i(\tilde{\psi}_M) - M_i(\psi_0)] = I(G)^{-1} \mathbb{E}_{\psi_i} [u_i^t b_i(\psi_0)'] + o(T^{-1/2})
\]

from (A.2) and (A.3) since \( H(\tilde{G}) = \tilde{\psi}_M \) and \( H(G) = \psi_0 \). □

**Proof of Theorem 5** Recall that the selection rule is to choose \( p^* \) if \( IC^h(p^*) < IC^h(p) \), where \( 0 \leq p^*, p \leq \bar{p} \) for some finite positive integer \( \bar{p} \). We thus need to prove that \( \lim_{n,T \to \infty} \mathbb{P}[IC^h(p^*) < IC^h(p_0)] = 0 \) for all \( p^* \neq p_0 \), where \( p_0 \) is the (finite) true lag order. We first consider the case of under-selection, \( p^* < p_0 \). We write

\[
\mathbb{P}[IC^h(p^*) < IC^h(p_0)] = \mathbb{P}\left[ \log \left( \frac{\sigma^2(p^*)}{\sigma^2(p_0)} \right) < \frac{h(n, T)}{nT}(p_0 - p^*) + \frac{1}{T} \left( \bar{R}_{n,T}(p_0) - \bar{R}_{n,T}(p^*) \right) \right]. \tag{A.9}
\]

The left-hand-side of the inequality in (A.9) is nonnegative for any \( n \) and \( T \) because the residual sum of squares does not increase as the number of regressors increases (e.g., Shibata (1976)). On the other hand, the right-hand-side of the inequality in (A.9) converges to zero as \( n,T \to \infty \) since \( 0 < (p_0 - p^*) < \bar{p} < \infty \), \( |\bar{R}_{n,T}(p_0) - \bar{R}_{n,T}(p^*)| < \infty \) from the invertibility in Assumption A-(ii), and \( h(n, T)/nT \to 0 \) as \( n,T \to \infty \) by assumption. Therefore, \( \mathbb{P}[IC^h(p^*) < IC^h(p_0)] \to 0 \) as \( n,T \to \infty \). Now for the case of over-selection, \( p^* > p_0 \), we write

\[
\mathbb{P}[IC^h(p^*) < IC^h(p_0)] = \mathbb{P}\left[ nT \left( \log \bar{\sigma}^2(p^*) - \log \bar{\sigma}^2(p_0) \right) < h(n, T)(p_0 - p^*) + n \left( \bar{R}_{n,T}(p_0) - \bar{R}_{n,T}(p^*) \right) \right]. \tag{A.10}
\]

Similarly as Lee (2006, 2011), we can show that \( \lim_{n \to \infty} \bar{\sigma}^2(p) = \sigma^2 = O(T^{-2}) \) for any \( p \) and \( |\log \bar{\sigma}^2(p^*) - \log \bar{\sigma}^2(p_0)| = O_p(T^{-2}) \) for large \( n \). The left-hand-side of the inequality in (A.10) is thus \( O_p(1) \) for large \( n \) and \( T \) because it is assumed that \( n/T \to \gamma \in (0, \infty) \).
For the right-hand-side of the inequality in (A.10), we note that \( |\tilde{R}_{n,T}(p_0) - \tilde{R}_{n,T}(p^*)| = O_p(1/T) \) (e.g., Bhansali (1981)) and thus \( n(\tilde{R}_{n,T}(p_0) - \tilde{R}_{n,T}(p^*)) = O_p(n/T) = O_p(1) \) for \( n/T \to \gamma \in (0, \infty) \). It follows that the right-hand-side goes to negative infinity as \( n, T \to \infty \) since \( p_0 - p^* < 0 \) and \( h(n, T) \to \infty \). Therefore, \( \mathbb{P}[I_{C}^h(p^*) < I_{C}^h(p_0)] \to 0 \) as \( n, T \to \infty \) for \( p^* > p_0 \).\(^{15}\) □

**Proof of Theorem 6** We consider the case of over-selection, \( p^* > p_0 \) and \( p^{**} > p_0 \). We first define that

\[
\Delta I_{C}^h \equiv I_{C}^h(p^*) - I_{C}^h(p_0)
\]

\[
= \log \left( \frac{\hat{\sigma}^2(p^*)}{\sigma^2(p_0)} \right) + \frac{h(n, T)}{nT}(p^* - p_0) + \frac{1}{T} \left( \tilde{R}_{n,T}(p^*) - \tilde{R}_{n,T}(p_0) \right)
\]

and

\[
\Delta I_{C0}^h \equiv I_{C0}^h(p^{**}) - I_{C0}^h(p_0) = \log \left( \frac{\hat{\sigma}^2(p^{**})}{\sigma^2(p_0)} \right) + \frac{h(n, T)}{nT}(p^{**} - p_0).
\]

Then, similarly as in the proof of Theorem 5, we can show that

\[
\mathbb{P} \left[ \Delta I_{C}^h < \Delta I_{C0}^h \right] \quad \text{(A.11)}
\]

\[
= \mathbb{P} \left[ \log \left( \frac{\hat{\sigma}^2(p^*)}{\sigma^2(p^{**})} \right) < \frac{h(n, T)}{nT}(p^{**} - p^*) + \frac{1}{T} \left( \tilde{R}_{n,T}(p_0) - \tilde{R}_{n,T}(p^*) \right) \right] \to 0
\]

as \( n, T \to \infty \). Note that \( I_{C}^h(p) \) has the heavier penalty term than \( I_{C0}^h(p) \) and thus \( p^{**} \geq p^* \) by construction. Therefore, the left-hand-side of the last inequality in (A.11) is nonnegative for any \( n \) and \( T \), whereas the left-hand-side goes to zero as in (A.9). This result implies that \( \Delta I_{C}^h \) cannot be smaller than \( \Delta I_{C0}^h \) as probability approaches to one and thus \( \lim_{n,T \to \infty} \{ \mathbb{P}[\Delta I_{C}^h < 0] - \mathbb{P}[\Delta I_{C0}^h < 0] \} \leq 0 \). The desired result follows from the definition of \( p^* \) and \( p^{**} \). □

**Proof of Corollary 7** From Bhansali (1981), it can be shown that \( |\tilde{R}_{n,T}(p_0) - \tilde{R}_{n,T}(p^*)| = c|p_0 - p^*|/T + o_p(1/T) \). Therefore, \( \mathbb{P} \left[ I_{C}^h(p^*) < I_{C}^h(p_0) \right] \) is approximately the same as

\[
\mathbb{P} \left[ \log \hat{\sigma}^2(p^*) - \log \sigma^2(p_0) < \frac{h(n, T)}{nT}(p_0 - p^*) + \frac{c}{T}(p_0 - p^*) \right],
\]

which corresponds to \( \mathbb{P} \left[ I_{C}^h(p^*) < I_{C}^h(p_0) \right] \). From this relation, \( I_{C}^h(p) \) should share the same asymptotic properties as \( I_{C}^h(p) \). □

\(^{15}\)From the proof, the following properties are well expected: The under-selection probability becomes smaller as \( T \) gets larger since the relative speed between \( h(n, T)/nT \to 0 \) and \( 1/T \to 0 \) as \( n, T \to \infty \) determines how fast the probability in (A.9) goes to zero. On the other hand, the inequality for the over-selection probability in (A.10) can be more easily violated if the left-hand-side is larger (i.e., \( n/T \) is larger). The simulation results in Section 4.2 meet these expectations.
References


HAHN, J., AND G. KUERSTEINER (2007). Bias reduction for dynamic nonlinear panel models with fixed effects, unpublished manuscript.


\[ p_0 = 1 : \quad y_{i,t} = \mu_i + 0.15 y_{i,t-1} + \varepsilon_{i,t} \]

<table>
<thead>
<tr>
<th>( n )</th>
<th>( T )</th>
<th>AIC</th>
<th>BIC</th>
<th>HQ</th>
<th>IC_{AIC}</th>
<th>IC_{BIC}</th>
<th>IC_{HQ}</th>
<th>IC_{AIC}^{c}</th>
<th>IC_{BIC}^{c}</th>
<th>IC_{HQ}^{c}</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>12</td>
<td>10.0</td>
<td>9.71</td>
<td>10.0</td>
<td>9.82</td>
<td>8.43</td>
<td>9.57</td>
<td>1.98</td>
<td>1.05</td>
<td>2.39</td>
</tr>
<tr>
<td>20</td>
<td>25</td>
<td>6.41</td>
<td>2.63</td>
<td>4.77</td>
<td>4.67</td>
<td>1.27</td>
<td>2.69</td>
<td>3.02</td>
<td>1.15</td>
<td>3.33</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>4.81</td>
<td>1.69</td>
<td>3.19</td>
<td>3.24</td>
<td>1.10</td>
<td>1.65</td>
<td>4.08</td>
<td>1.22</td>
<td>4.27</td>
</tr>
<tr>
<td>50</td>
<td>12</td>
<td>10.0</td>
<td>9.99</td>
<td>10.0</td>
<td>9.99</td>
<td>9.72</td>
<td>9.95</td>
<td>1.51</td>
<td>1.00</td>
<td>1.73</td>
</tr>
<tr>
<td>50</td>
<td>25</td>
<td>8.43</td>
<td>4.25</td>
<td>6.47</td>
<td>7.12</td>
<td>1.76</td>
<td>4.18</td>
<td>2.13</td>
<td>1.06</td>
<td>2.40</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>6.00</td>
<td>2.12</td>
<td>4.23</td>
<td>4.50</td>
<td>1.11</td>
<td>2.20</td>
<td>3.51</td>
<td>1.10</td>
<td>3.51</td>
</tr>
<tr>
<td>100</td>
<td>12</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>1.15</td>
<td>1.00</td>
<td>1.29</td>
</tr>
<tr>
<td>100</td>
<td>25</td>
<td>9.08</td>
<td>6.10</td>
<td>8.22</td>
<td>8.58</td>
<td>3.42</td>
<td>6.95</td>
<td>2.25</td>
<td>1.03</td>
<td>2.29</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>6.29</td>
<td>3.23</td>
<td>4.76</td>
<td>5.10</td>
<td>1.42</td>
<td>3.03</td>
<td>3.24</td>
<td>1.13</td>
<td>3.11</td>
</tr>
</tbody>
</table>

**Table 1:** Average of lag order selections over 1000 iterations

\((p_0 = 1; \bar{p} = 10)\)

**Figure 1:** Order selection frequencies over 1000 iterations

\((p_0 = 1; (n, T) = (100, 50))\)
\[ p_0 = 2: \ y_{i,t} = \mu_t + 0.15y_{i,t-1} + 0.15y_{i,t-2} + \varepsilon_{i,t} \]

<table>
<thead>
<tr>
<th>( n )</th>
<th>( T )</th>
<th>AIC</th>
<th>BIC</th>
<th>HQ</th>
<th>IC(^{AIC} )</th>
<th>IC(^{BIC} )</th>
<th>IC(^{HQ} )</th>
<th>IC(^{c} )</th>
<th>IC(^{BC} )</th>
<th>IC(^{HC} )</th>
<th>IC(^{HQ} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>12</td>
<td>9.94</td>
<td>9.71</td>
<td>9.94</td>
<td>9.72</td>
<td>7.95</td>
<td>9.32</td>
<td>1.78</td>
<td>1.05</td>
<td>2.24</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>25</td>
<td>6.50</td>
<td>2.39</td>
<td>5.06</td>
<td>5.12</td>
<td>1.20</td>
<td>2.39</td>
<td>2.95</td>
<td>1.32</td>
<td>3.41</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>5.61</td>
<td>2.32</td>
<td>4.00</td>
<td>4.28</td>
<td>1.36</td>
<td>2.24</td>
<td>4.65</td>
<td>1.82</td>
<td>4.79</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>12</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>9.99</td>
<td>9.71</td>
<td>9.91</td>
<td>1.53</td>
<td>1.03</td>
<td>1.73</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>25</td>
<td>8.13</td>
<td>4.41</td>
<td>6.57</td>
<td>7.06</td>
<td>2.23</td>
<td>4.82</td>
<td>3.13</td>
<td>1.29</td>
<td>3.53</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>5.97</td>
<td>3.28</td>
<td>4.72</td>
<td>4.97</td>
<td>2.23</td>
<td>3.30</td>
<td>4.39</td>
<td>2.17</td>
<td>4.39</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>12</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>1.25</td>
<td>1.00</td>
<td>1.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>25</td>
<td>8.97</td>
<td>6.50</td>
<td>8.14</td>
<td>8.53</td>
<td>4.09</td>
<td>7.44</td>
<td>2.89</td>
<td>1.20</td>
<td>2.92</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>6.56</td>
<td>3.66</td>
<td>5.29</td>
<td>5.64</td>
<td>2.34</td>
<td>3.78</td>
<td>3.78</td>
<td>2.10</td>
<td>3.71</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Average of lag order selections over 1000 iterations
\((p_0 = 2; \overline{p} = 10)\)

Figure 2: Order selection frequencies over 1000 iterations
\((p_0 = 2; (n, T) = (100, 50))\)
\[ p_0 = 3 : \quad y_{i,t} = \mu_i + 0.15y_{i,t-1} + 0.15y_{i,t-2} + 0.15y_{i,t-3} + \varepsilon_{i,t} \]

<table>
<thead>
<tr>
<th>( n )</th>
<th>( T )</th>
<th>AIC</th>
<th>BIC</th>
<th>HQ</th>
<th>IC(_{AIC}^c)</th>
<th>IC(_{BIC}^c)</th>
<th>IC(_{HQ}^c)</th>
<th>IC(_{AIC}^c)</th>
<th>IC(_{BIC}^c)</th>
<th>IC(_{HQ}^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>12</td>
<td>10.0</td>
<td>9.75</td>
<td>9.98</td>
<td>9.68</td>
<td>7.59</td>
<td>9.20</td>
<td>1.95</td>
<td>1.13</td>
<td>2.66</td>
</tr>
<tr>
<td>20</td>
<td>25</td>
<td>6.97</td>
<td>2.91</td>
<td>5.04</td>
<td>5.15</td>
<td>1.16</td>
<td>2.65</td>
<td>3.51</td>
<td>1.47</td>
<td>3.80</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>6.19</td>
<td>3.53</td>
<td>4.78</td>
<td>5.15</td>
<td>1.63</td>
<td>3.07</td>
<td>5.08</td>
<td>2.75</td>
<td>5.28</td>
</tr>
<tr>
<td>50</td>
<td>12</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>9.99</td>
<td>9.81</td>
<td>9.97</td>
<td>1.56</td>
<td>1.05</td>
<td>1.76</td>
</tr>
<tr>
<td>50</td>
<td>25</td>
<td>8.20</td>
<td>5.24</td>
<td>7.19</td>
<td>7.71</td>
<td>2.71</td>
<td>5.94</td>
<td>3.74</td>
<td>1.67</td>
<td>3.84</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>6.93</td>
<td>4.05</td>
<td>5.91</td>
<td>5.99</td>
<td>3.08</td>
<td>4.34</td>
<td>5.08</td>
<td>3.03</td>
<td>5.17</td>
</tr>
<tr>
<td>100</td>
<td>12</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>1.12</td>
<td>1.00</td>
<td>1.29</td>
</tr>
<tr>
<td>100</td>
<td>25</td>
<td>9.23</td>
<td>7.19</td>
<td>8.76</td>
<td>8.89</td>
<td>5.10</td>
<td>8.05</td>
<td>3.54</td>
<td>1.45</td>
<td>3.72</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>7.24</td>
<td>4.79</td>
<td>6.40</td>
<td>6.70</td>
<td>3.32</td>
<td>4.98</td>
<td>4.75</td>
<td>3.11</td>
<td>4.68</td>
</tr>
</tbody>
</table>

**Table 3:** Average of lag order selections over 1000 iterations  
\((p_0 = 3; \bar{T} = 10)\)

**Figure 3:** Order selection frequencies over 1000 iterations  
\((p_0 = 3; (n, T) = (100, 50))\)
\[ p_0 = 4: \quad y_{i,t} = \mu_i + 0.15y_{i,t-1} + 0.15y_{i,t-2} + 0.15y_{i,t-3} + 0.15y_{i,t-4} + \epsilon_{i,t} \]

<table>
<thead>
<tr>
<th>( n )</th>
<th>( T )</th>
<th>AIC</th>
<th>BIC</th>
<th>HQ</th>
<th>( IC^{AIC} )</th>
<th>( IC^{BIC} )</th>
<th>( IC^{HQ} )</th>
<th>( IC^{AIC}_c )</th>
<th>( IC^{BIC}_c )</th>
<th>( IC^{HQ}_c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>12</td>
<td>9.90</td>
<td>9.52</td>
<td>9.82</td>
<td>9.42</td>
<td>7.09</td>
<td>8.91</td>
<td>2.10</td>
<td>1.06</td>
<td>2.48</td>
</tr>
<tr>
<td>20</td>
<td>25</td>
<td>7.23</td>
<td>3.65</td>
<td>5.64</td>
<td>5.65</td>
<td>1.36</td>
<td>2.89</td>
<td>4.36</td>
<td>1.76</td>
<td>4.92</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>6.56</td>
<td>4.41</td>
<td>5.40</td>
<td>5.48</td>
<td>2.20</td>
<td>3.72</td>
<td>5.57</td>
<td>3.62</td>
<td>5.58</td>
</tr>
<tr>
<td>50</td>
<td>12</td>
<td>10.0</td>
<td>9.99</td>
<td>10.0</td>
<td>9.90</td>
<td>9.40</td>
<td>9.88</td>
<td>1.63</td>
<td>1.03</td>
<td>1.70</td>
</tr>
<tr>
<td>50</td>
<td>25</td>
<td>8.31</td>
<td>5.86</td>
<td>7.62</td>
<td>7.94</td>
<td>3.09</td>
<td>6.21</td>
<td>4.61</td>
<td>1.76</td>
<td>4.89</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>7.11</td>
<td>5.17</td>
<td>6.26</td>
<td>6.60</td>
<td>3.93</td>
<td>5.51</td>
<td>6.14</td>
<td>4.07</td>
<td>6.19</td>
</tr>
<tr>
<td>100</td>
<td>12</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>1.39</td>
<td>1.01</td>
<td>1.46</td>
</tr>
<tr>
<td>100</td>
<td>25</td>
<td>9.11</td>
<td>7.56</td>
<td>8.54</td>
<td>8.89</td>
<td>6.46</td>
<td>8.07</td>
<td>4.38</td>
<td>1.80</td>
<td>4.47</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>7.73</td>
<td>5.56</td>
<td>6.94</td>
<td>7.22</td>
<td>4.37</td>
<td>6.05</td>
<td>5.54</td>
<td>4.04</td>
<td>5.54</td>
</tr>
</tbody>
</table>

**Table 4**: Average of lag order selections over 1000 iterations

\((p_0 = 4; \bar{p} = 10)\)

![Figure 4](image1)

**Figure 4**: Order selection frequencies over 1000 iterations

\((p_0 = 4; (n, T) = (100, 50))\)
Figure 5: Correct order selection frequencies over 1000 iterations when $p_0 = 3$