Model Selection in the Presence of Incidental Parameters*

Yoonseok Lee†
University of Michigan

October 2012

Abstract

This paper considers model selection of nonlinear panel data models in the presence of incidental parameters (i.e., large-dimensional nuisance 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 that use either the Kullback-Leibler information criterion based on the profile likelihood or the Bayes factor based on the integrated likelihood with the robust prior of Arellano and Bonhomme (2009, *Econometrica* 77: 489-536). 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

*The author acknowledges the valuable comments from Bruce Hansen, Chris Hansen, Yuichi Kitamura, Roger Moon, Peter Phillips, Peter Robinson and seminar participants at Yale, Columbia, Rochester, UVa, Boston College, Seoul National, MSU, Montreal, 2011 International Conference on Panel Data, and 2011 Midwest Econometrics Group. The author also thanks the Cowles Foundation for Research in Economics at Yale University, where he was a visiting fellow while working on this project.

†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 heterogenous 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, 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 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 of these model selection procedures is that the number of parameters in each candidate model is finite or at most growing very slowly comparing to the sample size. For example, Stone (1979) points out that consistency of the standard BIC no longer holds when the number of parameters in the candidate model diverges with the sample size.[1] 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 specifying heterogeneity or for handling omitted variables. This paper examines why the standard model selection criteria perform poorly for such cases (e.g., Figures 1 to 4 in Section 5.3) and proposes properly modified model selection criteria particularly when the candidate models include nuisance parameters whose number grows at the same rate of the sample size (i.e., incidental parameters; Neyman and Scott (1948)).

In particular, we study the specification problem of panel data models, where we focus on a subset of the parameters. We consider panel observations $z_{i,t}$ for $i = 1, \cdots, n$ and $t = 1, \cdots, T$, whose unknown density (i.e., the model) is approximated by a parametric family $f(z; \psi, \lambda_i)$ that does not need to include the true model. The parameter of interest is ψ, which is common across i, and the nuisance parameters are given by $\lambda_1, \cdots, \lambda_n$, whose number increases at the same rate of the sample size. Common examples of λ_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. Such approach is reasonable when we are interested in selecting the structure of the model in ψ, while assuming the parameter space of λ_i is common

[1] 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 large-dimensional exponential family to consistently estimate the Bayes factor.

across the candidate models. A similar approach can be found in Claeskens and Hjort (2003) in the context of cross sectional models with finite-dimensional nuisance parameters, though they consider the case with nested models via local misspecification. In comparison, we allow for infinite-dimensional nuisance parameters as well as nonnested cases.

Two different approaches are used to handle the incidental parameters and to obtain new model selection criteria. One method is to apply the profiling idea on the Kullback-Leibler information criterion (KLIC). 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 score has nonzero expectation or the information identity is violated), which thus needs proper 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) and Hodges and Sargent (2001)) that reflects the level of difficulty of estimation. The penalty term is data-dependent, so the new model selection rule is adaptive. As an alternative approach, a Bayesian model selection criterion is also developed based on the Bayes factor, where the posterior is obtained using the integrated likelihoods. These two approaches—the profile likelihood based one and the integrated likelihood based one—are closely related as in the standard AIC and BIC, provided that a proper prior of the incidental parameter is used for the integration. In the pseudo-likelihood setup, we obtain the prior such that it makes the integrated likelihood closer to the genuine likelihood (e.g., the robust prior of Arellano and Bonhomme (2009)), that depends on the data in general.

Note that the majority of the panel data studies focus on modifying the profile or integrated likelihood as a way of bias reduction in the maximum likelihood estimator, which basically presumes that the parametric models considered are correctly specified (e.g., Hahn and Kuersteiner (2002, 2011); Hahn and Newey (2004); Arellano and Hahn (2006, 2007); Lee (2006, 2010, 2012); Bester and Hansen (2009)). However, as discussed in Lee (2006, 2012), if the model is not correctly specified, the efforts to reduce bias from the incidental parameters could even exacerbate the bias and thus the correct model specification is very important in this context particularly for dynamic or nonlinear panel models; the correct model specification should precede any bias corrections or bias reductions. This paper focuses on the specification problem.

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. The profile KLIC is introduced

that is general enough to be applied for heterogenous panel data models. 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, Section 5 proposes a lag order selection criterion for dynamic panel models and examines their statistical properties with including some simulation studies. 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 data observations $\{z_{i,t}\}$ for $i = 1, 2, \cdots, n$ and $t = 1, 2, \cdots, T$, which has an unknown distribution $G_i(z)$ having probability density function $g_i(z)$. $z_{i,t}$ is allowed to have heterogenous 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 so that the marginal distribution of $z_{i,t}$ is invariant in t. T could vary over i (i.e., $T_i \neq T_j$) but we assume $T_i = T$ for all i for the sake of simplicity.

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 θ_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 the heterogeneity of the marginal distribution is solely controlled by the heterogenous parameter θ_i. We decompose the parameter vector 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 that 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 $f(\cdot;\psi,\lambda_i)$. We could consider a multidimensional λ_i (e.g., Arellano and Hahn (2006)) but we focus on the scalar case for expositional simplicity.

We denote the marginal (pseudo-)likelihood of $z_{i,t}$ as[2]

$$f_{it}(z_{i,t};\psi,\lambda_i) = f(z_{i,t};\psi,\lambda_i), \tag{1}$$

which leads to the expression for the scaled individual log-likelihood function given by

$$\ell_i(\psi,\lambda_i) = \frac{1}{T}\sum\nolimits_{t=1}^{T} \log f_{it}(z_{i,t};\psi,\lambda_i).$$

[2] When we consider dynamic models, $f_{it}(z_{i,t};\psi,\lambda_i)$ 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 $f_{it}(z_{i,t};\psi,\lambda_i) = f(y_{i,t}|y_{i,t-1}, \cdots, y_{i,t-p};\psi,\lambda_i)$.

We assume the following conditions as White (1982) though some stronger conditions are imposed for the later use.

Assumption 1 *(i) $z_{i,t}$ is independent over i with distribution G_i on $\mathcal{Z}$, a measurable Euclidean space, with measurable Radon-Nikodym density $g_i = dG_i/d\nu$ for each i and for all t. (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 θ_i for every $z \in \mathcal{Z}$. (iii) It can be decomposed as $\theta_i = (\psi', \lambda_i)'$, where λ_i is related to the i-th observations only.*

Since we are mainly interested in ψ, we first maximize out the nuisance parameter λ_i to define the *profile likelihood* of ψ as

$$f_{it}^P(z_{i,t};\psi) = f(z_{i,t};\psi,\widehat{\lambda}_i(\psi)) \text{ for each } i, \tag{2}$$

where

$$\widehat{\lambda}_i(\psi) = \arg\max_{\lambda_i \in \Lambda} \ell_i(\psi,\lambda_i) \tag{3}$$

is the quasi maximum likelihood estimator (QMLE) of λ_i keeping ψ 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 $\widehat{\lambda}_i(\psi)$ in powers of T. The quasi maximum profile likelihood estimator of ψ is then obtained as

$$\widehat{\psi} = \arg\max_{\psi \in \Psi} \frac{1}{n}\sum\nolimits_{i=1}^{n} \ell_i^P(\psi), \text{ where } \ell_i^P(\psi) = \frac{1}{T}\sum\nolimits_{t=1}^{T} \log f_{it}^P(z_{i,t};\psi), \tag{4}$$

which indeed corresponds to the QMLE of ψ because this is just taking the maximum in two steps instead of taking the maximum simultaneously. The existence of $\widehat{\psi}$ is obtained from Assumption 1 as White (1982). When T is small, however, $f_{it}^P(\cdot;\psi)$ does not behave like the standard likelihood function due to the sampling variability of the estimator $\widehat{\lambda}_i(\psi)$. 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

$$\log f_{it}^M(z_{i,t};\psi) = \log f_{it}^P(z_{i,t};\psi) - \frac{1}{T}M_i(\psi) \tag{5}$$

is widely studied, which makes the *modified profile likelihood* $f_{it}^M(\cdot;\psi)$ to behave more likely a genuine likelihood function (e.g., Barndorff-Nielsen (1983)). The modification term $M_i(\psi)$ is $O_p(1)$ and $M_i(\psi)/T$ corrects the leading $O_p(T^{-1})$ sampling bias from $\widehat{\lambda}_i(\psi)$ so that it

renders the expected score of the modified profile likelihood to be closer to zero even with small T. A bias-reduced estimator for ψ thus can be obtained by maximizing the modified profile likelihood (i.e., the quasi maximum modified profile likelihood estimation) as

$$\widehat{\psi}_M = \arg\max_{\psi\in\Psi} \frac{1}{n}\sum\nolimits_{i=1}^{n} \ell_i^M(\psi), \tag{6}$$

where

$$\ell_i^M(\psi) = \ell_i^P(\psi) - \frac{1}{T}M_i(\psi) = \frac{1}{T}\sum\nolimits_{t=1}^{T} \log f_{it}^M(z_{i,t};\psi).$$

Further discussions of the 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) $\widehat{\psi}$ in (4) is a consistent estimator for a nonrandom vector ψ_T for fixed T, where

$$\psi_T = \arg\min_{\psi\in\Psi} \lim_{n\to\infty} \overline{D}(g \parallel f(\psi, \widehat{\lambda}(\psi)))$$

with

$$\overline{D}(g \parallel f(\psi, \widehat{\lambda}(\psi))) = \frac{1}{nT}\sum\nolimits_{i=1}^{n}\sum\nolimits_{t=1}^{T} D(g_i \parallel f_{it}(\psi, \widehat{\lambda}_i(\psi))). \tag{7}$$

Note that

$$D(g_i \parallel f_{it}(\psi, \widehat{\lambda}_i(\psi))) = \mathbb{E}_{G_i}\left[\log\left(g_i(z_{i,t}) / f(z_{i,t};\psi,\widehat{\lambda}_i(\psi))\right)\right]$$

is the Kullback-Leibler divergence (or the Kullback-Leibler information criterion; KLIC) of the true marginal density $g_i(\cdot)$ relative to $f_{it}(\cdot;\psi,\widehat{\lambda}_i(\psi))) = f_{it}^P(\cdot;\psi)$, which is well defined by the conditions below.[3] Therefore, $\overline{D}(g \parallel f(\psi,\widehat{\lambda}(\psi)))$ is simply the averaged KLIC over i and t. We denote $\mathbb{E}_{G_i}[\cdot] = \int[\cdot]dG_i$ as the expectation taken with respect to the true distribution

[3] Note that we can understand the averaged KLIC (7) as the KLIC of $g_i(z_{i,t})$ relative to the scaled individual parametric profile likelihood $\overline{f}_i(\psi,\widehat{\lambda}_i(\psi)) = \exp[T^{-1}\sum_{t=1}^{T}\log f(z_{i,t};\psi,\widehat{\lambda}_i(\psi))]$ since

$$\frac{1}{T}\sum_{t=1}^{T} D(g_i \parallel f_{it}(\psi,\widehat{\lambda}_i(\psi)))) = \mathbb{E}_{G_i}\left[\log g_i(z_{i,t}) - \frac{1}{T}\sum_{t=1}^{T}\log f(z_{i,t};\psi,\widehat{\lambda}_i(\psi))\right] = D(g_i \parallel \overline{f}_i(\psi,\widehat{\lambda}_i(\psi)))$$

by the stationarity.

G_i for each i. We further let

$$\begin{aligned}(\psi_0, \lambda_0) &= \underset{(\psi,\lambda)\in\Psi\times\Lambda^n}{\arg\min} \lim_{T\to\infty} \overline{D}\left(g \parallel f(\psi,\lambda)\right) \\ &= \underset{(\psi,\lambda)\in\Psi\times\Lambda^n}{\arg\min} \lim_{T\to\infty} \frac{1}{nT}\sum\nolimits_{i=1}^{n}\sum\nolimits_{t=1}^{T} D\left(g_i \parallel f_{it}(\psi,\lambda_i)\right),\end{aligned} \tag{8}$$

where $D\left(g_i \parallel f_{it}(\psi,\lambda_i)\right) = \mathbb{E}_{G_i}\left[\log\left(g_i\left(z_{i,t}\right)/f(z_{i,t};\psi,\lambda_i)\right)\right]$ and $\lambda_0 = (\lambda_{10},\cdots,\lambda_{n0})'$.

Assumption 2 *For each i, (i) $\mathbb{E}_{G_i}[\log g_i(z)]$ exists and both $g_i(z)$ and $f(z;\theta_i)$ are bounded away from zero; (ii) $\partial \log f(z;\theta_i)/\partial\theta_{i(j)}$ for $j = 1,\cdots,r+1$ are measurable functions of z for each θ_i in Θ and continuously differentiable with respect to G_i for all z in $\mathcal{Z}$ and θ_i in Θ; (iii) $|\log f(z;\theta_i)|$, $|\partial^2 \log f(z;\theta_i)/\partial\theta_{i(j)}\partial\theta_{i(k)}|$ and $|\partial \log f(z;\theta_i)/\partial\theta_{i(j)}\cdot\partial \log f(z;\theta_i)/\partial\theta_{i(k)}|$ are all dominated by functions integrable with respect to G_i for all $j,k = 1,\cdots,r+1$, where $\theta_{i(j)}$ denotes the jth element of θ_i; and (iv) $\mathbb{E}_{G_i}[\partial^2 \log f(z;\theta_i)/\partial\theta_i\partial\theta_i']$ and $\mathbb{E}_{G_i}[\partial \log f\left(z;\theta_{i0}\right)/\partial\theta_i \cdot \partial \log f\left(z;\theta_{i0}\right)/\partial\theta_i']$ are both nonsingular, where $\theta_{i0} = \left(\psi_0', \lambda_{i0}\right)'$. (v) $(\psi_0,\lambda_0) \in \Psi\times\Lambda^n$ is the unique solution in (8), where (ψ_0,λ_0) is in the interior of the support.*

From White (1982) under Assumptions 1 and 2, it holds that $\widehat{\psi} = \psi_T + o_p(1)$ as $n\to\infty$ even with fixed T. When the dimension of the nuisance parameters $\lambda = (\lambda_1,\cdots,\lambda_n)'$ is substantial relative to the sample size (e.g., when T is small), however, ψ_T is usually different from the standard KLIC minimizer ψ_0 in (8). This is the incidental parameters problem (e.g., Neyman and Scott (1948)) in the context of the QMLE. In general, it can be shown that (e.g., Arellano and Hahn (2007)) $\widehat{\psi} - \psi_0 = O_p\left(T^{-1}\right)$, and even when $n,T\to\infty$ with $n/T\to\gamma\in(0,\infty)$ and $n/T^3\to 0$, we have

$$\sqrt{nT}(\widehat{\psi}-\psi_0) = \sqrt{nT}(\widehat{\psi}-\psi_T) + \sqrt{\frac{n}{T}}\mathcal{B} + O_p\left(\sqrt{\frac{n}{T^3}}\right) \to_d \mathcal{N}(\sqrt{\gamma}\mathcal{B},\Omega_\psi)$$

for some positive definite matrix Ω_ψ since

$$\psi_T = \psi_0 + \mathcal{B}/T + O(T^{-2}), \tag{9}$$

where $\mathcal{B}/T$ represents bias of $O(T^{-1})$. The main source of this bias is that $\widehat{\lambda}_i(\psi)$ is still random and thus is not the same as[4]

$$\lambda_i(\psi) = \underset{\lambda_i\in\Lambda}{\arg\min}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} D\left(g_i \parallel f_{it}(\psi,\lambda_i)\right) = \underset{\lambda_i\in\Lambda}{\arg\min} D\left(g_i \parallel f_{it}(\psi,\lambda_i)\right), \tag{10}$$

where $\lambda_i(\psi_0) = \lambda_{i0}$ for each i and the last equality is from the stationarity over t. The

[4] $\lambda_i(\psi)$ is normally referred to as the *least favorable curve*.

estimation error of $\widehat{\lambda}_i(\psi)$ with finite T in (3) is not negligible even when $n \to \infty$, and 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 (pseudo-)information matrix as

$$\mathcal{I}_i = \mathbb{E}_{G_i}\left[\frac{\partial \log f_{it}(z_{i,t};\psi_0,\lambda_{i0})}{\partial \theta_i} \cdot \frac{\partial \log f_{it}(z_{i,t};\psi_0,\lambda_{i0})}{\partial \theta_i'}\right] = \begin{pmatrix} \mathcal{I}_{i,\psi\psi} & \mathcal{I}_{i,\psi\lambda} \\ \mathcal{I}_{i,\lambda\psi} & \mathcal{I}_{i,\lambda\lambda} \end{pmatrix} \tag{11}$$

conformable with $\theta_i = \left(\psi', \lambda_i\right)' \in \mathbb{R}^{r+1}$, where (ψ_0, λ_{i0}) is defined in (8) for each i. $\mathcal{I}_i$, $\mathcal{I}_{i,\psi\psi}$ and $\mathcal{I}_{i,\lambda\lambda}$ are all nonsingular from Assumption 2. We also define the (scaled individual) score functions as

$$\begin{aligned} u_i(\psi,\lambda_i) &= \partial \ell_i(\psi,\lambda_i)/\partial\psi, \\ v_i(\psi,\lambda_i) &= \partial \ell_i(\psi,\lambda_i)/\partial\lambda_i, \\ u_i^e(\psi,\lambda_i) &= u_i(\psi,\lambda_i) - \mathcal{I}_{i,\psi\lambda}\mathcal{I}_{i,\lambda\lambda}^{-1} v_i(\psi,\lambda_i). \end{aligned}$$

Note that $u_i^e(\psi_0,\lambda_{i0})$ is the *efficient score* for ψ at (ψ_0,λ_{i0}) and it can be understood as the orthogonal projection of the score function for ψ 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)).[5] For notational convenience, we suppress the arguments when expressions are evaluated at $\theta_{0i} = \left(\psi_0', \lambda_{i0}\right)'$ for each i: $u_i = u_i(\psi_0,\lambda_{i0})$, $v_i = v_i(\psi_0,\lambda_{i0})$ and $u_i^e = u_i^e(\psi_0,\lambda_{i0})$. It can be shown that we have the following expansion (e.g., McCullagh and Tibshirani (1990), Severini (2000) and Sartori (2003)):

$$\frac{\partial \ell_i^P(\psi_0)}{\partial\psi} = u_i^e + b_i(\psi_0) + O_p\left(\frac{1}{T^{3/2}}\right) \tag{12}$$

with $u_i^e = O_p(T^{-1/2})$ and $b_i(\psi_0) = O_p\left(T^{-1}\right)$ for all i. Though $\mathbb{E}_{G_i}[u_i^e] = 0$ by construction, $\mathbb{E}_{G_i}[b_i(\psi_0)] \neq 0$, which incurs an asymptotic bias appears as (9). The modification term $M_i(\psi)$ in (5) can be found as a function in ψ, provided that $f(\cdot;\theta_i)$ be three-times differentiable in θ_i, satisfying

$$\mathbb{E}_{G_i}\left[\frac{1}{T}\frac{dM_i(\psi_0)}{d\psi} - b_i(\psi_0)\right] = O\left(\frac{1}{T^{3/2}}\right) \tag{13}$$

so that the expected score of the modified profile likelihood $\mathbb{E}_{G_i}[\partial \ell_i^M(\psi_0)/\partial\psi]$ does not have the first order asymptotic bias from $b_i(\psi_0)$.

[5] It follows that $\mathbb{E}_{G_i}\left[\partial u_i^e(\psi_0,\lambda_{i0})/\partial\lambda_i\right] = 0$ since $u_i^e(\psi,\lambda_i)$ and $v_i(\psi,\lambda_i)$ are orthogonal at (ψ_0,λ_{i0}) by construction (e.g., Arellano and Hahn (2007)).

2.3 Bias reduction

The standard bias corrected estimators in nonlinear (dynamic) fixed effect regressions correspond to $\widehat{\psi}_M$ in (6), which is given by (e.g., Hahn and Newey (2004); Arellano and Hahn (2007))

$$\widehat{\psi}_M = \widehat{\psi} - \frac{1}{T}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} \widehat{\mathcal{I}}_i^e(\widehat{\psi}_M)\right)^{-1}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{d}{d\psi} M_i(\widehat{\psi}_M)\right),$$

where $\widehat{\mathcal{I}}_i^e(\widehat{\psi}_M)$ is a consistent estimator of the efficient information $\mathcal{I}_i^e = \mathcal{I}_{i,\psi\psi} - \mathcal{I}_{i,\psi\lambda}\mathcal{I}_{i,\lambda\lambda}^{-1}\mathcal{I}_{i,\lambda\psi}$ for $T \to \infty$. In principle, $\widehat{\mathcal{I}}_i^e(\widehat{\psi}_M)$ can be driven as $-(1/T)\sum_{t=1}^{T} \partial^2 \log f_{it}^M(z_{i,t}; \widehat{\psi}_M)/\partial\psi\partial\psi'$, where the second derivative of $\log f_{it}^M(z;\psi)$ needs to be obtained numerically. Alternatively, we let $\widehat{\theta}_{Mi} = (\widehat{\psi}_M', \widehat{\lambda}_{Mi}) = (\widehat{\psi}_M', \widehat{\lambda}_i(\widehat{\psi}_M))$ be the maximum modified profile likelihood estimator and

$$\widehat{\mathcal{I}}_i(\widehat{\theta}_{Mi}) = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \log f_{it}(z_{i,t};\widehat{\theta}_{Mi})}{\partial\theta_i} \cdot \frac{\partial \log f_{it}(z_{i,t};\widehat{\theta}_{Mi})}{\partial\theta_i'} = \begin{pmatrix} \widehat{\mathcal{I}}_{i,\psi\psi}(\widehat{\theta}_{Mi}) & \widehat{\mathcal{I}}_{i,\psi\lambda}(\widehat{\theta}_{Mi}) \\ \widehat{\mathcal{I}}_{i,\lambda\psi}(\widehat{\theta}_{Mi}) & \widehat{\mathcal{I}}_{i,\lambda\lambda}(\widehat{\theta}_{Mi}) \end{pmatrix} \quad (14)$$

as a consistent estimator of $\mathcal{I}_i$ in (11). Then, $\widehat{\mathcal{I}}_i^e(\widehat{\theta}_{Mi})$, which indeed depends only on $\widehat{\psi}_M$, can be obtained using the elements in (14). The expression of $dM_i(\widehat{\psi}_M)/d\psi$ can be obtained similarly as the equation (12) in Arellano and Hahn (2007).

For later use, we can derive a simple form of $M_i(\psi)$ as follows under the regularity conditions and Assumptions 1 and 2. From the standard asymptotic result of the (Q)ML estimators, we have the first order stochastic expansion for an arbitrary fixed ψ as

$$\sqrt{T}(\widehat{\lambda}_i(\psi) - \lambda_i(\psi)) = \left\{-\frac{\partial^2 \ell_i(\psi, \lambda_i(\psi))}{\partial\lambda_i^2}\right\}^{-1} \cdot \sqrt{T}\frac{\partial \ell_i(\psi, \lambda_i(\psi))}{\partial\lambda_i} + O_p\left(\frac{1}{T^{1/2}}\right) \quad (15)$$

for each i. In addition, we can expand $\ell_i^P(\psi) = \ell_i(\psi, \widehat{\lambda}_i(\psi))$ around $\lambda_i(\psi)$ for given ψ as

$$\begin{aligned} \ell_i^P(\psi) - \ell_i(\psi, \lambda_i(\psi)) &= \frac{\partial \ell_i(\psi, \lambda_i(\psi))}{\partial\lambda_i}\left(\widehat{\lambda}_i(\psi) - \lambda_i(\psi)\right) \quad (16) \\ &\quad + \frac{1}{2}\frac{\partial^2 \ell_i(\psi, \lambda_i(\psi))}{\partial\lambda_i^2}\left(\widehat{\lambda}_i(\psi) - \lambda_i(\psi)\right)^2 + O_p\left(\frac{1}{T^{3/2}}\right) \\ &= \frac{1}{T}\cdot\frac{1}{2}\left\{-\frac{\partial^2 \ell_i(\psi, \lambda_i(\psi))}{\partial\lambda_i^2}\right\}^{-1}\left(\sqrt{T}\frac{\partial \ell_i(\psi, \lambda_i(\psi))}{\partial\lambda_i}\right)^2 + O_p\left(\frac{1}{T^{3/2}}\right) \end{aligned}$$

from (15), where the dominating term is $O_p(T^{-1})$ because $\partial^2 \ell_i(\psi, \lambda_i(\psi))/\partial\lambda_i^2 = O_p(1)$ and $\partial \ell_i(\psi, \lambda_i(\psi))/\partial\lambda_i = O_p(T^{-1/2})$. It follows that (e.g., Severini (2000), Arellano and Hahn

(2006))

$$\mathbb{E}_{G_i}\left[\frac{\partial \ell_i^P(\psi_0)}{\partial \psi}\right] = \mathbb{E}_{G_i}\left[\frac{\partial}{\partial \psi}\frac{1}{2}\left\{-\frac{\partial^2 \ell_i(\psi_0,\lambda_{i0})}{\partial \lambda_i^2}\right\}^{-1}\left(\frac{\partial \ell_i(\psi_0,\lambda_{i0})}{\partial \lambda_i}\right)^2\right] + O\left(\frac{1}{T^{3/2}}\right) \tag{17}$$

since $\lambda_i(\psi_0) = \lambda_{i0}$ and $\mathbb{E}_{G_i}[\partial \ell_i(\psi_0,\lambda_{i0})/\partial \psi] = 0$ by construction. Comparing (12) and (17), this result suggests that a simple form of the modification function in $\ell_i^M(\psi)$ can be obtained as

$$\frac{1}{T}M_i(\psi) = \frac{1}{2}\left\{-\left.\frac{\partial^2 \ell_i(\psi,\lambda_i)}{\partial \lambda_i^2}\right|_{\lambda_i=\widehat{\lambda}_i(\psi)}\right\}^{-1}\left(\left.\frac{\partial \ell_i(\psi,\lambda_i)}{\partial \lambda_i}\right|_{\lambda_i=\widehat{\lambda}_i(\psi)}\right)^2, \tag{18}$$

whose first derivative corrects the leading bias term $b_i(\psi)$ at $\psi = \psi_0$ in the profile score (12). 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. Furthermore, (18) can be generalized as

$$\begin{aligned}\frac{1}{T}M_i(\psi) &= \frac{1}{2}\left\{-\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 \log f_{it}(z_{i,t};\psi,\widehat{\lambda}_i(\psi))}{\partial \lambda_i^2}\right\}^{-1} \\ &\quad\times \sum_{j=-m}^{m}\frac{K_j}{T}\sum_{t=\max\{1,j+1\}}^{\min\{T,T+j\}}\frac{\partial \log f_{it}(z_{i,t};\psi,\widehat{\lambda}_i(\psi))}{\partial \lambda_i}\cdot\frac{\partial \log f_{it}(z_{i,t-j};\psi,\widehat{\lambda}_i(\psi))}{\partial \lambda_i},\end{aligned} \tag{19}$$

similarly as the modification functions suggested by Arellano and Hahn (2006) and Bester and Hansen (2009), which appears to be robust to arbitrary serial correlations in the score function. Note that the second component in (19) corresponds to the heteroskedasticity-robust variance estimator of $\sqrt{T}\partial \ell_i(\psi,\widehat{\lambda}_i(\psi))/\partial \lambda_i$. The truncation parameter $m \geq 0$ is chosen such that $m/T^{1/2} \to 0$ as $T \to \infty$, and the kernel function K_j guarantees positive definiteness of the variance estimate (e.g., the Bartlett kernel: $K_j = 1 - (j/(m+1))$).

3 Profile Likelihood and KLIC

3.1 Model selection

Normally the panel data studies focus on reducing the first order bias (9) from the incidental parameters problem, which basically presume that the models are correctly specified. As discussed in Lee (2006, 2012), however, if the model is not correctly specified, the efforts of reducing bias from the incidental parameters could even exacerbate the bias. Therefore, correct model specification is very important in this context particularly for dynamic or non-linear 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. We focus on model specification; in particular, we are interested in selecting a model $f(z|\psi, \lambda_i)$ that is closest to the true one $g_i(z)$ on average over i.

In the standard setup, when there are no nuisance parameters λ so that the dimension of the parameter vector $\theta = \psi$ is small and finite, we can conduct the standard model selection by comparing estimates of the averaged KLIC given by

$$\begin{aligned} \min_{\theta} \overline{D}(g \parallel f(\theta)) &= \min_{\theta} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} D(g_i \parallel f_{it}(\theta)) \\ &= \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \int \log g_i(z) \, dG_i(z) - \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \int \log f_{it}(z; \widehat{\theta}) dG_i(z), \end{aligned} \tag{20}$$

where $\widehat{\theta}$ is the QMLE, which is a consistent estimator of $\theta_0 = \arg\min_\theta \lim_{n,T\to\infty} \overline{D}(g \parallel f(\theta))$ in this case. Note that the averaged KLIC $\overline{D}(g \parallel f(\theta))$ is defined so that it could accommodate possibly heterogeneous panel data models. We select a model $f(\cdot; \theta)$ whose KLIC in (20) is the minimum among the candidates. Equivalently, since the first term in (20) does not depend on the model, we select the model $f(\cdot; \theta)$ minimizing the relative distance

$$\Phi(\widehat{\theta}) = -\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \int \log f_{it}(z; \widehat{\theta}) dG_i(z),$$

which can be estimated by

$$\widehat{\Phi}(\widehat{\theta}) = -\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \int \log f_{it}(z; \widehat{\theta}) d\widehat{G}_i(z),$$

where $\widehat{G}_i$ is the empirical distribution. As noted in Akaike (1973), however, $-\widehat{\Phi}(\widehat{\theta})$ overestimates $-\Phi(\widehat{\theta})$ since $\widehat{G}_i$ corresponds more closely to $\widehat{\theta}$ than does the true G_i. Therefore, it is suggested to minimize the bias-corrected version of $\widehat{\Phi}(\widehat{\theta})$ given by

$$\widetilde{\Phi}(\widehat{\theta}) = \widehat{\Phi}(\widehat{\theta}) - B(\widehat{G}) \tag{21}$$

as an information criterion for model selection, where $B(G) = \mathbb{E}[\widehat{\Phi}(\widehat{\theta}) - \Phi(\widehat{\theta})]$ and $\mathbb{E}[\cdot]$ is the expectation with respect to the joint distribution $G = (G_1, \cdots, G_n)'$. See, for example, Akaike (1973, 1974) for further details. Note that Akaike (1973) shows that $B(G)$ is asymptotically the ratio of $\dim(\theta)$ to the sample size when $\widehat{\theta}$ 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 θ is substantial relative to the sample size, the incidental parameters problem prevails and

thus it is not straightforward to use the standard criterion like (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 ψ determines the key structure of the model that does not change over i, it is then natural to concentrate out the nuisance parameters λ_i's in the model selection problem. The candidate models are indexed by ψ alone, while the parameter space of λ_i remains the same across them, and thus the choice of a particular model does not depend on the realization of λ_i's in this case. This is a similar idea to the profile likelihood approach when the main interest is in a subset of parameters. Some examples are as follows.

Example 1 (Variable or model selection in panel models) We consider a parametric nonlinear fixed-effect model given by $y_{i,t} = \xi(x_{i,t}, u_{i,t}; \mu_i, \beta, \sigma_i^2)$ with known $\xi(\cdot;\cdot)$, where $u_{i,t}$ is independent over i and t with $u_{i,t}|(x_{i,1}, \cdots, x_{i,T}, \mu_i) \sim (0, \sigma_i^2)$, and β is an r-dimensional parameter vector. The goal of this example is either to select a set of regressors or to choose a parametric function $\xi(\cdot;\cdot)$ yielding the best fit in the presence of incidental parameters (μ_i, σ_i^2). For $\xi(\cdot;\cdot)$, a possible example is choosing between Logit and Probit models. Variable selection in a linear transformation model given by $\varphi_i(y_{i,t}) = x_{i,t}'\beta + u_{i,t}$ with some strictly increasing incidental function $\varphi_i(\cdot)$ 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 μ_i. When $p = \infty$, this problem becomes to find the best approximation $AR(p)$ model.

Example 3 (Number of support choice of random effects or random coefficient) We consider a random-effect/coefficient model given by $y_{i,t} = x_{i,t}'\beta_i + \varepsilon_{i,t}$, where $\varepsilon_{i,t}$ is independent over i and t with $\varepsilon_{i,t}|(x_{i,1}, \cdots, x_{i,T}, \beta_i) \sim \mathcal{N}(0, \sigma_i^2)$, and β_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 $\{q_1, \cdots, q_k\}$. The main interest in this example is to determine the number of finite support k in the presence of incidental parameters σ_i^2. In the context of mixed proportional hazard models (or Cox partial likelihoods with unobserved heterogeneity), this problem is to choose the number of finite support of nonparametric frailty in the Heckman-Singer model (Heckman and Singer (1984)), if we see the Cox partial likelihood as a profile likelihood.

3.2 Profile likelihood information criterion

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 λ_i's are concentrated out from the standard KLIC as follows.

Definition (Profile KLIC) *The profile Kullback-Leibler divergence (or the profile KLIC) of $g_i(\cdot)$ relative to $f_{it}(\cdot;\psi,\lambda_i)$ is defined as*

$$D_P\left(g_i \parallel f_{it}(\psi,\lambda_i);\psi\right) = \min_{\lambda_i\in\Lambda} D\left(g_i \parallel f_{it}(\psi,\lambda_i)\right) = D\left(g_i \parallel f_{it}(\psi,\lambda_i(\psi))\right) \tag{22}$$

with $\lambda_i(\psi)$ given in (10).

Note that $D_P\left(g_i \parallel f_{it}(\psi,\lambda_i);\psi\right)$ depends on ψ only, not on λ_i. Since the profile KLIC is defined as the minimum of the standard KLIC ($D\left(g_i \parallel f_{it}(\psi,\lambda_i)\right)$) in λ_i, it apparently satisfies the conditions that the standard KLIC has. For example, $D_P\left(g_i \parallel f_{it}(\psi,\lambda_i);\psi\right)$ is nonnegative and it is equal to zero when $g_i(\cdot)$ belongs to the parametric family of $f(\cdot;\psi,\lambda_i)$ (i.e., $g_i(\cdot) = f(\cdot;\psi_*,\lambda_{i*})$ for some $(\psi_*',\lambda_{i*})' \in \Psi\times\Lambda$). Similarly as the standard case (20), we select the model that has the smallest value of the estimate of

$$\min_{\psi\in\Psi} \overline{D}_P\left(g \parallel f(\psi,\lambda);\psi\right) = \min_{\psi\in\Psi} \frac{1}{nT}\sum\nolimits_{i=1}^{n}\sum\nolimits_{t=1}^{T} D_P\left(g_i \parallel f_{it}(\psi,\lambda_i);\psi\right). \tag{23}$$

Since $\min_{\psi\in\Psi}\overline{D}_P\left(g \parallel f(\psi,\lambda);\psi\right) = \min_{\psi\in\Psi}\min_{\lambda_i\in\Lambda}(nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} D\left(g_i \parallel f_{it}(\psi,\lambda_i)\right) = \min_{(\psi,\lambda)\in\Psi\times\Lambda^n}\overline{D}\left(g \parallel f(\psi,\lambda)\right)$, the model with the smallest (23) corresponds to the model with the smallest estimate of the standard averaged KLIC, $\overline{D}\left(g \parallel f(\psi,\lambda)\right)$, over ψ and λ. Note that, however, we cannot directly use (23) for the model selection procedure since it contains infeasible components $\lambda_i(\psi)$'s. A natural candidate is then the averaged KLIC based on the profile likelihoods given by

$$\overline{D}\left(g \parallel f^P(\psi)\right) = \overline{D}(g \parallel f(\psi,\widehat{\lambda}(\psi))) = \frac{1}{nT}\sum\nolimits_{i=1}^{n}\sum\nolimits_{t=1}^{T} D(g_i \parallel f_{it}(\psi,\widehat{\lambda}_i(\psi))), \tag{24}$$

which turns out to be equivalent to (7). Since $\widehat{\lambda}_i(\psi)$ is a biased estimator of $\lambda_i(\psi)$ when T is small, the KLIC based on the profile likelihoods $D(g_i \parallel f_{it}^P(\psi)) = D(g_i \parallel f_{it}(\psi,\widehat{\lambda}_i(\psi)))$ in (24) is not the same as the profile KLIC $D_P\left(g_i \parallel f_{it}(\psi,\lambda_i);\psi\right) = D\left(g_i \parallel f_{it}(\psi,\lambda_i(\psi))\right)$ in (22). The following lemma states the relation between these two KLIC's.

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

$$D_P\left(g_i \parallel f_{it}(\psi,\lambda_i);\psi\right) = D(g_i \parallel f_{it}^P(\psi)) + \delta(\psi;G_i), \tag{25}$$

where the bias term is defined as $\delta(\psi;G_i) = \mathbb{E}_{G_i}\left[\log\left(f(z_{i,t};\psi,\widehat{\lambda}_i(\psi))/f(z_{i,t};\psi,\lambda_i(\psi))\right)\right]$ with $\widehat{\lambda}_i(\psi)$ and $\lambda_i(\psi)$ being given in (3) and (10), respectively. Furthermore, if Assumptions 1 and 2 hold, we have

$$\delta(\psi;G_i) = \mathbb{E}_{G_i}\left[M_i(\psi)/T\right] + O(T^{-3/2}) \tag{26}$$

under the regularity conditions, where $M_i(\psi)$ is the modification term used for the modified

profile likelihood function (5).

From (25), it can be seen that even when g_i is nested in f, $D(g_i \parallel f_{it}^P(\psi))$ is not necessarily zero unless $f(z;\psi,\lambda_i(\psi)) = f(z;\psi,\widehat{\lambda}_i(\psi))$, which is very unlikely with small T. It thus follows that model selection using $D(g_i \parallel f_{it}^P(\psi))$ is undesirable. However, Lemma 1 shows that if we modify $D(g_i \parallel f_{it}^P(\psi))$ by correcting the bias using some proper estimator of $\delta(\psi; G_i)$, then we can conduct the model selection based on the modified $D(g_i \parallel f_{it}^P(\psi))$. The result in (26) shows that the bias term in (25) is indeed closely related with the modification term $M_i(\psi)$.

Similarly as (21), by letting

$$\Phi_P(\widehat{\psi}_M) = -\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int \log f_{it}^P(z;\widehat{\psi}_M) dG_i(z),$$

where $f_{it}^P(z;\psi)) = f_{it}^P(z;\psi,\widehat{\lambda}_i(\psi))$, we define an information criterion using a bias-corrected estimator of $\Phi_P(\widehat{\psi}_M)$ given by

$$\widetilde{\Phi}_P(\widehat{\psi}_M) = -\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int \log f_{it}^P(z;\widehat{\psi}_M) d\widehat{G}_i(z) - B_P(\widehat{G}). \tag{27}$$

$\widehat{\psi}_M$ is the quasi maximum modified profile likelihood estimator (i.e., the bias-corrected estimator) of ψ_0 defined as (6) and $B_P(\widehat{G})$ is an estimator of

$$B_P(G) = \mathbb{E}\left[-\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int \log f_{it}^P(z;\widehat{\psi}_M) d(\widehat{G}_i(z) - G_i(z))\right] - \frac{1}{n}\sum_{i=1}^{n}\delta(\widehat{\psi}_M; G_i)$$

obtained by replacing the unknown distribution G_i by the empirical distribution $\widehat{G}_i$. Note that the bias correction term $B_P(\widehat{G})$ is somewhat different from the correction term $B(\widehat{G})$ in (21). In particular, it includes an additional correction term $n^{-1}\sum_{i=1}^{n}\delta(\widehat{\psi}_M;\widehat{G})$, which is needed because the feasible information criterion is defined using $D(g_i \parallel f_{it}^P(\widehat{\psi}_M))$ instead of $D_P(g_i \parallel f_{it}(\psi,\lambda_i);\widehat{\psi}_M)$. An approximated expression of $B_P(G)$ and its consistent estimator are obtained in the following theorem. We denote $z_i = (z_{i,1},\cdots,z_{i,T})'$.

Theorem 2 *Let Assumptions 1 and 2 hold. We suppose that there exists an r-dimensional regular function H such that $\psi_0 = H(G)$ and $\widehat{\psi}_M = H(\widehat{G})$, where G is the joint distribution of $(z_1,\cdots,z_n)$. 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 (2011)), we have*

$$B_P(G) = -\frac{1}{nT}tr\left\{I(G)^{-1}J(G)\right\} - \frac{1}{n}\sum_{i=1}^{n}\delta(\widehat{\psi}_M; G_i),$$

where $tr\{\cdot\}$ *is the trace operator and*

$$
\begin{aligned}
I(G) &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}_{G_i}\left[-\left.\frac{\partial^2 \log f_{it}\left(z_{i,t};\psi,\lambda_i(\psi)\right)}{\partial\psi\partial\psi'}\right|_{\psi=H(G)}\right], \\
J(G) &= \frac{1}{nT}\sum_{i=1}^{n}\mathbb{E}_{G_i}\left[\sum_{t=1}^{T}\sum_{s=1}^{T}\left.\frac{\partial \log f_{it}\left(z_{i,t};\psi,\lambda_i(\psi)\right)}{\partial\psi}\right|_{\psi=H(G)}\left.\frac{\partial \log f_{it}(z_{i,s};\psi,\widehat{\lambda}_i(\psi))}{\partial\psi'}\right|_{\psi=H(G)}\right].
\end{aligned}
$$

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

$$
B_P(\widehat{G}) = -\frac{1}{nT}tr\left\{I(\widehat{G})^{-1}J(\widehat{G})\right\} - \frac{1}{nT}\sum_{i=1}^{n}M_i(\widehat{\psi}_M), \tag{28}
$$

where[6]

$$
\begin{aligned}
I(\widehat{G}) &= -\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\partial^2 \log f_{it}^{M}(z_{i,t};\widehat{\psi}_M)}{\partial\psi\partial\psi'} \quad \textit{and} \\
J(\widehat{G}) &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\min\{T,T+j\}}\frac{\partial \log f_{it}^{M}(z_{i,t};\widehat{\psi}_M)}{\partial\psi}\frac{\partial \log f_{it}^{P}(z_{i,t-j};\widehat{\psi}_M)}{\partial\psi'}
\end{aligned}
$$

for $\log f_{it}^{P}\left(z_{i,t};\psi\right) = \log f_{it}(z_{i,t};\psi,\widehat{\lambda}_i(\psi))$ *and* $\log f_{it}^{M}\left(z_{i,t};\psi\right) = \log f_{it}^{P}\left(z_{i,t};\psi\right) - M_i(\psi)/T$*.*

From the equations (27) and (28), therefore, a general form of an information criterion for the model selection based on the bias-corrected profile likelihood (i.e., *profile likelihood information criterion*; PLIC) is defined as

$$
\begin{aligned}
PLIC\left(f\right) &= -\frac{2}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\log f_{it}^{P}(z_{i,t};\widehat{\psi}_M) - 2B_P(\widehat{G}) \qquad (29) \\
&= -\frac{2}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\log f_{it}(z_{i,t};\widehat{\psi}_M,\widehat{\lambda}_i(\widehat{\psi}_M)) \\
&\quad +\frac{2}{nT}tr\left\{I(\widehat{G})^{-1}J(\widehat{G})\right\} + \frac{2}{nT}\sum_{i=1}^{n}M_i(\widehat{\psi}_M),
\end{aligned}
$$

where $M_i(\widehat{\psi}_M)$ is given in (19) 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 the profile likelihood in the model selection problem. With further restricted conditions, we could derive

[6] $J(\widehat{G})$ allows for possible serial correlations in the (modified) profile score functions similarly as (19).

the simpler form for the $PLIC\,(f)$ as the following corollary.

Corollary 3 *Suppose that g is included in the family of f. Under the same conditions as Theorem 2, we then have*

$$PLIC\,(f) = -\frac{2}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\log f_{it}(z_{i,t};\widehat{\psi}_M,\widehat{\lambda}_i(\widehat{\psi}_M)) + \frac{2r}{nT} + \frac{2}{nT}\sum_{i=1}^{n} M_i(\widehat{\psi}_M), \qquad (30)$$

where $r = \dim(\psi)$.

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 $\widehat{\psi}_M$ instead of the MLE. The additional penalty term $(2/nT)\sum_{i=1}^{n} M_i(\widehat{\psi}_M)$ 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 (29) or (30) has heavier penalty than the standard Akaike information criterion (AIC). Recall that in the standard AIC, the second part of the penalty term in (30) does not appear and the penalty term of the information criterion is simply given by $2r/nT$.

Remark 1 $PLIC\,(f)$ in (30) can be rewritten as $-(2/nT)\sum_{i=1}^{n}\sum_{t=1}^{T}\log f_{it}^{M}(z_{i,t};\widehat{\psi}_M) + (2r/nT)$, where $\log f_{it}^{M}(\cdot;\psi) = \log f_{it}^{P}(\cdot;\psi) - T^{-1}M_i(\psi)$ is the modified profile likelihood function. Note that the modified profile likelihood function is closer to the genuine likelihood than is the profile likelihood function. It shows that such aspect extends even when we define the KLIC. More precisely, from Lemma 1, we can derive that $D_P\left(g_i \parallel f_{it}(\psi,\lambda_i);\psi\right) = D\left(g_i \parallel f_{it}^{M}(\psi)\right) + O(1/T^{3/2})$.

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)). 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 λ_i as $\pi_i(\lambda_i|\psi)$ for each i, which satisfies the following conditions as Arellano and Bonhomme (2009):

Assumption 3 (i) *The support of $\pi_i(\lambda_i|\psi)$ contains an open neighborhood of (ψ_0,λ_{i0}). (ii) When $T\to\infty$, $\log\pi_i(\lambda_i|\psi) = O(1)$ uniformly over i for all λ_i and ψ.*

Using $\pi_i(\lambda_i|\psi)$, the individual integrated log-likelihood $\ell_i^I(\psi)$ is defined as

$$\ell_i^I(\psi) = \frac{1}{T}\log\left\{\int f_i\left(\psi,\lambda_i\right)\pi_i(\lambda_i|\psi)d\lambda_i\right\}$$

for each i, where $f_i\left(\psi,\lambda_i\right) = \prod_{t=1}^T f_{it}\left(z_{i,t};\psi,\lambda_i\right) = \exp(T\ell_i(\psi,\lambda_i))$ is the joint density of $z_i = (z_{i,1},\cdots,z_{i,T})'$. We let ϕ^k be the discrete prior over different K models $\mathcal{M}^1,\mathcal{M}^2,\cdots,\mathcal{M}^K$ and $\eta(\psi^k|\mathcal{M}^k)$ be the prior on $\psi^k \in \mathbb{R}^{r_k}$ given the model $\mathcal{M}^k$. We further let $g\left(z\right) = \prod_{i=1}^n g_i(z_i)$ be the joint density of $(z_1,\cdots,z_n)$ and

$$L^I(\psi^k|z) = \exp\left(T\sum\nolimits_{i=1}^n \ell_i^I(\psi^k)\right)$$

be the *integrated (joint) likelihood function*. Then, the Bayes theorem yields the posterior probability of the model $\mathcal{M}^k$ as

$$\mathcal{P}\left(\mathcal{M}^k|z\right) = \frac{1}{g\left(z\right)}\phi^k\int L^I(\psi^k|z)\eta(\psi^k|\mathcal{M}^k)d\psi^k \tag{31}$$

and the Bayesian information criterion can be obtained based on $-2\log\mathcal{P}\left(\mathcal{M}^k|z\right)$. By choosing the candidate model corresponding to the minimum value of the Bayesian information criterion, one is attempting to select the candidate model corresponding to the highest Bayesian posterior probability; and this approach is approximately equivalent to model selection based on Bayes factors (e.g., Kass and Raftery (1995)).

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

$$\ell_i^I(\psi^k)-\ell_i^P(\psi^k) = \frac{1}{2T}\log\left(\frac{2\pi}{T}\right)-\frac{1}{2T}\log\left(-\frac{\partial^2\ell_i(\psi^k,\widehat{\lambda}_i(\psi^k))}{\partial\lambda_i^2}\right)+\frac{1}{T}\log\pi_i(\widehat{\lambda}_i(\psi^k)|\psi^k)+O_p\left(\frac{1}{T^2}\right)$$

or

$$\begin{aligned}\ell_i^I(\psi^k)-\ell_i^M(\psi^k) &= \frac{1}{2T}\log\left(\frac{2\pi}{T}\right)-\frac{1}{2T}\log\left(-\frac{\partial^2\ell_i(\psi^k,\widehat{\lambda}_i(\psi^k))}{\partial\lambda_i^2}\right)+\frac{1}{T}\log\pi_i(\widehat{\lambda}_i(\psi^k)|\psi^k)\\ &\quad+\frac{1}{T}M_i\left(\psi^k\right)+O_p\left(\frac{1}{T^2}\right)\end{aligned} \tag{32}$$

for each i. These approximations imply that if we choose the conditional prior $\pi_i(\lambda_i|\psi^k)$ such that it cancels out leading terms in (32), then we have better approximation as $\ell_i^I(\psi^k)-$

$\ell_i^M(\psi^k) = O_p(T^{-2})$. More precisely, from (18), we can obtain

$$\begin{aligned}\pi_i(\widehat{\lambda}_i(\psi^k)|\psi^k) &= C_\pi \left(\frac{2\pi}{T}\right)^{-1/2} \left(-\frac{\partial^2 \ell_i(\psi^k, \widehat{\lambda}_i(\psi^k))}{\partial \lambda_i^2}\right)^{1/2} \exp\left(-M_i\left(\psi^k\right)\right) \\ &= C_\pi \left(\frac{2\pi}{T}\right)^{-1/2} \left(-\frac{\partial^2 \ell_i(\psi^k, \widehat{\lambda}_i(\psi^k))}{\partial \lambda_i^2}\right)^{1/2} \\ &\quad \times \exp\left(-\frac{1}{2}\left\{-\frac{\partial^2 \ell_i(\psi^k, \widehat{\lambda}_i(\psi^k))}{\partial \lambda_i^2}\right\}^{-1} \left(\frac{\partial \ell_i(\psi^k, \widehat{\lambda}_i(\psi^k))}{\partial \lambda_i}\right)^2\right) \end{aligned} \tag{33}$$

for some nonzero finite constant C_π. Note that the explicit form of the conditional prior in (33) corresponds to the robust prior in equation (14) of Arellano and Bonhomme (2009) in the case of pseudo-likelihood. Arellano and Bonhomme (2009)'s robust prior is developed to obtain the first-order unbiased estimators in nonlinear panel models. This idea extends to our context since we find the conditional prior such that it better approximates the modified profile likelihood by the integrated likelihood, where the maximum modified profile likelihood estimator is first-order unbiased by construction (e.g., Section 2.3). Therefore, the general discussions in Arellano and Bonhomme (2009) also apply to the conditional prior $\pi_i(\widehat{\lambda}_i(\psi^k)|\psi^k)$ in (33): it generally depends on the data unless an orthogonal reparametrization (e.g., Lancaster (2002)) or some equivalent condition is available.

By choosing the conditional prior as (33), we can obtain the approximate posterior probability of the model $\mathcal{M}^k$ in (31) as follows.

Theorem 4 *Let Assumptions 1 and 2 hold and $n/T \to \gamma \in (0, \infty)$ as $n, T \to \infty$. If we suppose the conditional priors of λ_i as (33) and uninformative flat priors for ψ^k (i.e., $\eta(\psi^k|\mathcal{M}^k) = 1$ for all $k = 1, \cdots, K$) over the neighborhood of $\widehat{\psi}_M^k$ where $L^I(\psi^k|z)$ is dominant, we have an approximation*

$$\log \mathcal{P}\left(\mathcal{M}^k|z\right) = \sum_{i=1}^n \sum_{t=1}^T \log f_{it}^M(z_{i,t}; \widehat{\psi}_M^k) - \frac{r_k}{2} \log nT + c(z,k) + o_p(1), \tag{34}$$

where $\log f_{it}^M(z_{i,t}; \widehat{\psi}_M^k) = \log f_{it}(z_{i,t}; \widehat{\psi}_M^k, \widehat{\lambda}_i(\widehat{\psi}_M^k)) - M_i(\widehat{\psi}_M^k)/T$, $r_k = \dim(\psi^k)$, *and* $c(z,k) = O_p(1)$.

From (34), ignoring terms that do not depend on k and terms that are of the smaller order as $n, T \to \infty$, we can define the *integrated likelihood information criterion* ($ILIC$) from $-(2/nT)\log \mathcal{P}\left(\mathcal{M}^k|z\right)$ only using the relevant terms as follows:

$$ILIC\left(\mathcal{M}^k\right) = -\frac{2}{nT} \sum_{i=1}^n \sum_{t=1}^T \log f_{it}(z_{i,t}; \widehat{\psi}_M^k, \widehat{\lambda}_i(\widehat{\psi}_M^k)) + \frac{r_k \log nT}{nT} + \frac{2}{nT} \sum_{i=1}^n M_i(\widehat{\psi}_M^k). \tag{35}$$

Note that, comparing with $PLIC$ in (30), 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, where 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 an information criteria, which can be used in the presence of incidental parameters, given by

$$LIC^h\left(\mathcal{M}^k\right) = -\frac{2}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\log f_{it}(z_{i,t};\widehat{\psi}_M^k,\widehat{\lambda}_i(\widehat{\psi}_M^k)) + r_k\frac{h(n,T)}{nT} + \frac{2}{nT}\sum_{i=1}^{n}M_i(\widehat{\psi}_M^k) \quad (36)$$

for a candidate parametric model $\mathcal{M}^k$ whose parameter vector is given by $(\psi^k,\lambda_1,\cdots,\lambda_n)'$ with $\dim(\psi^k)=r_k$, 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 LIC^h is no longer deterministic; it is data-dependent and thus the model selection rule is adaptive.

5 Lag Order Selection in Dynamic Panel Models

5.1 Lag order selection criteria and model complexity

As an illustration, we consider model selection criteria 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

$$y_{i,t} = \mu_i + \sum\nolimits_{j=1}^{p_0}\alpha_{p_0 j}y_{i,t-j} + \varepsilon_{i,t} \quad \text{for } i=1,2,\cdots,n \text{ and } t=1,2,\cdots,T, \quad (37)$$

where p_0 is not necessarily finite.[7] $\varepsilon_{i,t}$ is serially uncorrelated and unobserved individual effects μ_i are assumed fixed. We also let the initial values $(y_{i,0},y_{i,-1},\cdots,y_{i,-p_0+1})$ be observed for all i, for notational convenience. We first assume the following conditions.

Assumption A *(i)* $\varepsilon_{i,t}|(\{y_{i,s}\}_{s\le t-1},\mu_i) \sim i.i.d.\mathcal{N}\left(0,\sigma^2\right)$ *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

[7] 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., $p_T^3/T\to 0$). 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.

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

$$LIC^h(p) = \log \widetilde{\sigma}^2(p) + \frac{h(n,T)}{nT}p + \frac{2}{nT}\sum_{i=1}^{n} M_i(\widetilde{\alpha}(p), \widetilde{\sigma}^2(p)) \tag{38}$$

for some positive $h(n,T)$. We let

$$\widetilde{\sigma}^2(p) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\widetilde{\varepsilon}_{i,t}^{W}(p)\right)^2, \tag{39}$$

where $\widetilde{\varepsilon}_{i,t}^{W}(p) = y_{i,t}^{W} - \sum_{j=1}^{p}\widetilde{\alpha}_{pj}y_{i,t-j}^{W}$ 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.[8] Note that $\widetilde{\alpha}(p) = (\widetilde{\alpha}_{p1}, \cdots, \widetilde{\alpha}_{pp})$ is the maximum modified profile likelihood estimators (i.e., bias corrected within group estimators) in a panel $AR(p)$ regression. From (19) it can be also derived that

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

Therefore, the penalty term in (38) is given by

$$\frac{h(n,T)}{nT}p + \frac{1}{T}\widetilde{R}_{n,T}(p), \tag{40}$$

where $\widetilde{R}_{n,T}(p) = (2/n)\sum_{i=1}^{n} M_i(\widetilde{\alpha}(p), \widetilde{\sigma}^2(p))$ 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. The last term $(1/T)\widetilde{R}_{n,T}(p)$, 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 $(2/nT)\{p + n \times (\widetilde{R}_{n,T}(p)/2)\}$, which shows that the efficient number of parameters is not

[8] The within transformation corresponds to maximizing out the fixed effects μ_i's in MLE (i.e., forming the profile likelihood).

$p+n$ in this case; the effect from the incidental parameters λ_i is smaller than n, where the degree is determined by the size of $\widetilde{R}_{n,T}(p)/2$. Remark 2 below discusses more about the effective number of parameters in the context of the model complexity.

In fact, the 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, $\widetilde{R}_{n,T}(p)$ 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.[9] 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.

Remark 2 (Model complexity) The new penalty term of $LIC^h(p)$ 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 $y_{i,t}|\mu_i,\sigma^2 \sim inid\mathcal{N}(\mu_i,\sigma^2)$ for all $i=1,\cdots,n$ and $t=1,\cdots,T$, where $\mu_i|\nu,\tau^2 \sim iid\mathcal{N}(\nu,\tau^2)$ for all i. Under this specification, the number of parameters can be counted as either $n+1$ if μ_i is considered as fixed effect (e.g., $\tau^2=\infty$); or 3 if μ_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 $\widehat{y}_{i,t}$. In this particular example, the degrees of freedom ρ turns out to be

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

Notice that the first term ρ_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 ρ 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 ρ will be close to one if σ^2/τ^2 is large. The lag order selection example in this section corresponds to the case of fixed effect but the degrees of freedom in our case is different from n; it is instead

[9] 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.

given by $n\widetilde{R}_{n,T}(p)/2$, 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)).

5.2 Statistical properties

In general, under the stationarity, the probability limit of the long-run autocorrelation estimator $\widetilde{R}_{n,T}(p)$ in (40) is bounded and the entire penalty term multiplied by the sample size nT (i.e., $h(n,T)\,p + n\widetilde{R}_{n,T}(p)$) increases with the sample size. As noted in Shibata (1980) and Yang (2005), therefore, we can conjecture that the new lag order selection criterion is not asymptotically optimal (i.e., when $p_0 = \infty$, $\text{plim}_{n,T\to\infty}[LIC^h(p^*) / \inf_{p\geq 0} LIC^h(p)] \neq 1$, where p^* is the lag order estimator from $LIC^h(p)$; e.g., Li (1987)) if the true data generating model is $AR(\infty)$ with finite σ^2 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. Note that we define a lag order estimator p^* is consistent (and thus the lag order selection criterion is consistent) if it satisfies $\liminf_{n,T\to\infty} \mathbb{P}(p^* = p_0) = 1$.[10]

Theorem 5 *Under Assumption A, if we let $n/T \to \gamma \in (0,\infty)$ and $n/T^3 \to 0$ as $n,T \to \infty$, then $LIC^h(p)$ 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 subsection.

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., $\log\widetilde{\sigma}^2(p) + p(h(n,T)/nT)$) also satisfy the consistency with a suitable choice of $h(n,T) \to \infty$. Similarly as Guyon and Yao (1999), however, it can be shown that the under-selection probability vanishes exponentially fast for both cases (provided both $h(n,T)/nT \to 0$ and $1/T \to 0$), while the over-selection probability decreases at a slower rate depending on the magnitude of the penalty term (provided $h(n,T) \to \infty$ and n/T does not diverge).[11] Therefore, the improvement of correct lag-order-selection proba-

[10] 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 standard time series context, BIC and properly defined PIC 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.

[11] In the proof of Theorem 5 in Appendix, we can find that controlling for the under-selection error probability requires $T \to \infty$ (as long as $h(n,T)/nT \to 0$), where the order of magnitude between n and T is

bility 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 corollary states that the over-selection probability reduced asymptotically by modifying the penalty term as in the new lag order selection criterion (40).

Corollary 6 *Suppose the conditions in Theorem 5 hold. For some finite positive integer* $\overline{p}$*, if we let* $p^{**} = \arg\min_{0\leq p\leq \overline{p}} LIC_0^h(p)$ *with* $LIC_0^h(p) = \log\widetilde{\sigma}^2(p) + p(h(n,T)/nT)$ *and* $p^* = \arg\min_{0\leq p\leq \overline{p}} LIC^h(p)$*, then* $\limsup_{n,T\to\infty}\mathbb{P}(p^{**} > p_0) \geq \limsup_{n,T\to\infty}\mathbb{P}(p^* > p_0)$.

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

$$LIC_c^h(p) = \log\widetilde{\sigma}^2(p) + \frac{p}{nT}\left\{h(n,T) + c\left(\frac{n}{T}\right)\right\} \tag{41}$$

for some positive constant $c < \infty$, which uses deterministic penalty terms instead of data-dependent ones. The additional penalty term cp/T^2 in (41) is introduced to offset the higher order bias in the maximum profile likelihood estimator $\widehat{\sigma}^2(p)$ since it can be shown that $\text{plim}_{n\to\infty}\widehat{\sigma}^2(p) - \sigma^2 = -cp/T^2 + O\left(T^{-3}\right)$, where the constant c depends on the parameter values α_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 $\widehat{\sigma}^2(p)$ and 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 (41), 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 $LIC^h(p)$, like Theorems 5 and 6.

Corollary 7 *Under the same condition as Theorem 5,* $LIC_c^h(p)$ *shares the same asymptotic properties as* $LIC^h(p)$.

5.3 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

not important. On the other hand, controlling for the over-selection error probability requires that n/T does not diverge (as long as $h(n,T) \to \infty$) so it needs much longer panel data set in practice comparing with the under-selection probability case.

used information criteria, which use the pooled information:

$$
\begin{aligned}
AIC\left(p\right) &= \log\widetilde{\sigma}^{2}(p) + \frac{2}{nT}p, \\
BIC\left(p\right) &= \log\widetilde{\sigma}^{2}(p) + \frac{\log\left(nT\right)}{nT}p, \\
HQ\left(p\right) &= \log\widetilde{\sigma}^{2}(p) + \frac{2\log\log\left(nT\right)}{nT}p,
\end{aligned}
$$

where $\widetilde{\sigma}^{2}(p)$ is defined as (39). As well expected, initial simulation results shows that constructing penalty terms using $p+n$ too heavily penalize the criteria so that they yield high under-selection probabilities. We thus only count the number of parameters as p instead of $p+n$ (i.e., including fixed effect parameters) in defining the information criteria above. 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 in the previous subsection:

$$
\begin{aligned}
LIC^{AIC}\left(p\right) &= \log\widetilde{\sigma}^{2}(p) + \frac{2}{nT}p + \frac{1}{T}\widetilde{R}_{n,T}(p), \\
LIC^{BIC}\left(p\right) &= \log\widetilde{\sigma}^{2}(p) + \frac{\log\left(nT\right)}{nT}p + \frac{1}{T}\widetilde{R}_{n,T}(p), \\
LIC^{HQ}\left(p\right) &= \log\widetilde{\sigma}^{2}(p) + \frac{2\log\log\left(nT\right)}{nT}p + \frac{1}{T}\widetilde{R}_{n,T}(p),
\end{aligned}
$$

as well as the simplified form as (41):

$$
\begin{aligned}
LIC_{c}^{AIC}\left(p\right) &= \log\widetilde{\sigma}^{2}(p) + \frac{p}{nT}\left\{2 + \frac{n}{T}\right\}, \\
LIC_{c}^{BIC}\left(p\right) &= \log\widetilde{\sigma}^{2}(p) + \frac{p}{nT}\left\{\log\left(nT\right) + \frac{n}{T}\right\}, \\
LIC_{c}^{HQ}\left(p\right) &= \log\widetilde{\sigma}^{2}(p) + \frac{p}{nT}\left\{2\log\log\left(nT\right) + \frac{n}{T}\right\},
\end{aligned}
$$

in which c is simply chosen to one.

We generate $AR\left(p_0\right)$ dynamic panel processes, with p_0 ranging from 1 to 4, of the form $y_{i,t} = \mu_i + \sum_{j=1}^{p_0}\alpha_{p_0 j}y_{i,t-j} + \varepsilon_{i,t}$ for $i = 1, 2, \cdots, n$ and $t = 1, 2, \cdots, T$, where $\alpha_{p_0 j} = 0.15$ for all $j = 1, \cdots, p_0$. For each $AR\left(p_0\right)$ 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 μ_i are randomly drawn from $\mathcal{U}\left(-0.5, 0.5\right)$ and $\varepsilon_{i,t}$ from $\mathcal{N}\left(0, 1\right)$. We use the bias corrected within-group estimators (e.g., Lee (2012)) for $\widetilde{\alpha}_{pj}$'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., $\overline{p} = 10$). 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, LIC and LIC_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 LIC and LIC_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 overall, 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 LIC_c works better than LIC, LIC_c cannot be always preferred to LIC empirically since it has larger under-selection probability than LIC 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. 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 we demonstrate, 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.

The incidental parameters can be understood as a subset of parameters whose estimators have slower rate of convergence than the rest of the parameter. Therefore, we could see this paper as a special case of a more general question: model selection problem on a sub-parameter set when the other (nuisance) parameter estimators potentially have slower rate of convergence than those of the parameter of interest. Semiparametric models thus 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 $f(z_i, w_i; \psi, \lambda_i(w_i))$ for given observations $\{z_i, w_i\}$, where $\lambda_i(w) = (\lambda_{1i}, \lambda_2(w))'$ 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 ith observation. Though it needs to be proved in the context of QML estimation, we conjecture that for $\widehat{\lambda}_{2,\psi}(\omega_i) = \arg\max_\lambda \sum_{t=1}^{T} \log f(z_{i,t}, w_{i,t}; \psi, \lambda_{1i}, \lambda) K((\omega_i - w_{i,t})/h)$, where $K(\cdot)$ and h are properly defined kernel function and the bandwidth parameter, respectively, we could derive a similar result as Theorem 2 under proper technical conditions. Note that, however, the conditions for the incidental parameters λ_{1i}'s and for the nonparametric components are different and thus their effects on the parametric component ψ need to be treated differently.[12]

[12] In fact, the semiparametric component estimator even does not affect the asymptotics of the parametric component estimator under the proper conditions (e.g., Andrews (1994) and Newey (1994)), whereas the nuisance parameters do without any information orthogonality with the parameter of interest.

Appendix: Mathematical Proofs

Proof of Lemma 1 The result follows immediately since

$$\begin{aligned} D_P\left(g_i \parallel f_{it}(\psi, \lambda_i); \psi\right) &= \int \log g_i\left(z\right) dG_i(z) - \int \log f_{it}(z; \psi, \lambda_i(\psi)) dG_i(z) \\ &= \int \log g_i\left(z\right) dG_i(z) - \int \log f_{it}(z; \psi, \widehat{\lambda}_i(\psi)) dG_i(z) \\ &\quad + \int \log\left(\frac{f_{it}(z; \psi, \widehat{\lambda}_i(\psi))}{f_{it}(z; \psi, \lambda_i(\psi))}\right) dG_i(z) = D\left(g_i \parallel f_{it}^P(\psi)\right) + \delta(\psi; G_i). \end{aligned}$$

Furthermore, from (16) and (18), it can be readily derived that

$$\delta(\psi; G_i) = \mathbb{E}_{G_i}\left[\ell_i^P(\psi) - \ell_i(\psi, \lambda_i(\psi))\right] = \mathbb{E}_{G_i}\left[M_i(\psi)/T\right] + O(T^{-3/2})$$

for a given ψ from the stationarity over t. □

Proof of Theorem 2 For each i, we define $G_i(\cdot; \epsilon) = G_i(\cdot) + \epsilon(\widehat{G}_i(\cdot) - G_i(\cdot))$ for some $\epsilon \in [0, 1]$. $G(\cdot; \epsilon)$, $G(\cdot)$ and $\widehat{G}(\cdot)$ denote the collection of the marginal distributions (i.e., $G(Z; \epsilon) = (G_1(z_1; \epsilon), \cdots, G_n(z_n; \epsilon))$ with $Z = (z_1, \cdots, z_n)'$ and similarly for the others). We will also use notations G_i and $\widehat{G}_i$ instead of $G_i(\cdot)$ and $\widehat{G}_i(\cdot)$ respectively if there is no confusion. For a fixed ϵ, we let $\psi(\epsilon) = H(G(\cdot; \epsilon))$ be the solution of

$$\frac{1}{n}\sum_{i=1}^{n} \int \frac{\partial}{\partial \psi} Q_{it}(z; \epsilon) dG_i\left(z; \epsilon\right) = 0, \tag{A.1}$$

where

$$Q_{it}(z; \epsilon) = \log f_{it}(z; \psi(\epsilon), \lambda_i(\epsilon)) - \frac{1}{T}\mu_i(\epsilon).$$

$\lambda_i(\epsilon)$ is the solution of $\int [\partial Q_{it}(z; \epsilon)/\partial \lambda_i] dG_i\left(z; \epsilon\right) = 0$ for each i so that

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

and $\mu_i(\epsilon) = \epsilon M_i(\psi(\epsilon))$ yielding

$$\mu_i(\epsilon) = \begin{cases} \mu_i(0) = 0 & \text{if } \epsilon = 0 \\ \mu_i(1) = M_i(\psi(1)) & \text{if } \epsilon = 1. \end{cases}$$

Recall that $\widehat{\lambda}_i(\psi)$, $\lambda_i(\psi)$ and $M_i(\psi)$ are defined as (3), (10) and (5), respectively. It then follows that $\psi(0) = H(G) = \psi_0$ and $\psi(1) = H(\widehat{G}) = \widehat{\psi}_M$ by construction. Therefore, the Taylor series expansion of $\widehat{\psi}_M$ about ψ_0 can be obtained as (e.g., Serfling (1980), Chapter 6.2)

$$\widehat{\psi}_M - \psi_0 = H(\widehat{G}) - H(G) = d_1 H(G; \widehat{G} - G) + o_p((nT)^{-1/2}), \tag{A.2}$$

where $d_1H(G;\widehat{G}-G)=\lim_{\epsilon\to 0+}\epsilon^{-1}\{H(G(\epsilon))-H(G)\}$ is the standard first order Gâteaux differential of H at G in the direction of $\widehat{G}$. The remainder term is negligible as shown in Hahn and Kuersteiner (2011) for $n,T\to\infty$ satisfying $n/T\to\gamma\in(0,\infty)$ and $n/T^3\to 0$.

Now similarly as Hahn and Kuersteiner (2011), by differentiating (A.1) with respect to ϵ, we have

$$\begin{aligned}0 &= \frac{1}{n}\sum_{i=1}^{n}\int\frac{\partial^2}{\partial\psi\partial\psi'}Q_{it}(z;\epsilon)dG_i(z;\epsilon)\times d_1H(G;\widehat{G}-G)\\ &+\frac{1}{n}\sum_{i=1}^{n}\int\frac{\partial^2}{\partial\psi\partial\lambda_i}Q_{it}(z;\epsilon)dG_i(z;\epsilon)\times\frac{\partial}{\partial\epsilon}\lambda_i(\psi(\epsilon);G_i(z;\epsilon))\\ &+\frac{1}{n}\sum_{i=1}^{n}\int\frac{\partial}{\partial\psi}Q_{it}(z;\epsilon)d(\widehat{G}_i(z)-G_i(z))\end{aligned}$$

and by evaluating this result at $\epsilon=0$ we derive

$$\begin{aligned}d_1H(G;\widehat{G}-G) &= \left(-\frac{1}{n}\sum_{i=1}^{n}\int\frac{\partial^2\log f_{it}(z;\psi_0,\lambda_{i0})}{\partial\psi\partial\psi'}dG_i(z)\right)^{-1}\\ &\times\frac{1}{n}\sum_{i=1}^{n}\int\frac{\partial\log f_{it}(z;\psi_0,\lambda_{i0})}{\partial\psi}d\widehat{G}_i(z).\end{aligned}\tag{A.3}$$

Note that $\lambda_i(\psi_0)=\lambda_{i0}$ and thus it holds that $\int[\partial\log f_{it}(z;\psi_0,\lambda_{i0})/\partial\psi]dG_i(z)=0$ and $\int[\partial^2\log f_{it}(z;\psi_0,\lambda_{i0})/\partial\psi\partial\lambda_i]dG_i(z)=\int[\partial^2\log f_{it}(z;\psi_0,\lambda_i(\psi_0))/\partial\psi\partial\lambda_i]dG_i(z)=0$. Therefore, from (A.2) and (A.3) we have the approximation of $\widehat{\psi}_M$ as (e.g., Withers (1983); Konishi and Kitagawa (1996))[13]

$$\widehat{\psi}_M-\psi_0=\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}H^{(1)}(z_{i,t};G)+o_p((nT)^{-1/2}),$$

where $H^{(1)}(z_{i,t};G)$ is given by

$$\begin{aligned}H^{(1)}(z_{i,t};G) &= \left(-\frac{1}{n}\sum_{i=1}^{n}\int\left.\frac{\partial^2\log f_{it}(z;\psi,\lambda_i(\psi))}{\partial\psi\partial\psi'}\right|_{\psi=\psi_0}dG_i(z)\right)^{-1}\\ &\times\left.\frac{\partial\log f_{it}(z_{i,t};\psi,\lambda_i(\psi))}{\partial\psi}\right|_{\psi=\psi_0}.\end{aligned}\tag{A.4}$$

Then, similarly as Theorem 2.1 of Konishi and Kitagawa (1996), by expanding $f_{it}^P(z;\widehat{\psi}_M)$ around ψ_0 for given i and t and combining the results above, we have stochastic expansions

[13] It also shows that $\widehat{\psi}_M$ is $\sqrt{nT}$-consistent to $H(G)=\psi_0$ since $(nT)^{-1/2}\sum_{i=1}^{n}\sum_{t=1}^{T}H^{(1)}(z_{i,t};G_i)$ is asymptotically normal with mean zero and variance $(nT)^{-1}\sum_{i=1}^{n}\int\sum_{t=1}^{T}\sum_{s=1}^{T}H^{(1)}(z_{i,t};G_i)H^{(1)}(z_{i,s};G_i)'dG_i$.

as

$$\begin{aligned}
&\int \log f_{it}^{P}(z;\widehat{\psi}_M)dG_i(z) \\
= &\int \log f_{it}(z;\psi_0,\widehat{\lambda}_i(\psi_0))dG_i(z) \\
&+\frac{1}{nT}\sum_{j=1}^{n}\sum_{s=1}^{T}\int \frac{\partial}{\partial\psi'}\log f_{it}(z;\psi_0,\widehat{\lambda}_i(\psi_0))dG_i(z)H^{(1)}(z_{j,s};G)+o_p\left(\frac{1}{\sqrt{nT}}\right)
\end{aligned}$$

and thus denoting $\mathbb{E}[\cdot]$ as the expectation with respect to the joint distribution $G=(G_1,\cdots,G_n)$,

$$\mathbb{E}\left[\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int \log f_{it}^{P}(z;\widehat{\psi}_M)dG_i(z)\right]=\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int \log f_{it}(z;\psi_0,\widehat{\lambda}_i(\psi_0))dG_i(z)+o\left(\frac{1}{\sqrt{nT}}\right)$$

since $\int H^{(1)}(z_{j,s};G)dG_j(z)=0$ for all j. Similarly,

$$\begin{aligned}
&\int \log f_{it}^{P}(z;\widehat{\psi}_M)d\widehat{G}_i(z) \\
= &\frac{1}{T}\sum_{t=1}^{T}\log f_{it}(z;\psi_0,\widehat{\lambda}_i(\psi_0)) \\
&+\frac{1}{nT^2}\sum_{j=1}^{n}\sum_{s=1}^{T}\sum_{t=1}^{T}\frac{\partial}{\partial\psi'}\log f_{it}(z;\psi_0,\widehat{\lambda}_i(\psi_0))H^{(1)}(z_{j,s};G)+o_p\left(\frac{1}{\sqrt{nT}}\right)
\end{aligned}$$

and by the stationarity over t

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int \log f_{it}^{P}(z;\widehat{\psi}_M)d\widehat{G}_i(z)\right] \\
= &\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int \log f_{it}(z;\psi_0,\widehat{\lambda}_i(\psi_0))dG_i(z) \\
&+\frac{1}{n^2T^2}\sum_{i=1}^{n}\int\sum_{s=1}^{T}\sum_{t=1}^{T}\frac{\partial}{\partial\psi'}\log f_{it}(z;\psi_0,\widehat{\lambda}_i(\psi_0))H^{(1)}(z_{i,s};G)dG_i(z)+o\left(\frac{1}{\sqrt{nT}}\right),
\end{aligned}$$

where the last term is nonzero only for the case of $i = j$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[-\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int \log f_{it}^{P}(z;\widehat{\psi}_M)d(\widehat{G}_i(z)-G_i(z))\right] \\
= \;& -tr\left\{\frac{1}{n^2T^2}\sum_{i=1}^{n}\int\sum_{s=1}^{T}\sum_{t=1}^{T}\frac{\partial}{\partial\psi'}\log f_{it}(z;\psi_0,\widehat{\lambda}_i(\psi_0))H^{(1)}(z_{i,s};G)dG_i(z)\right\}+o\left(\frac{1}{\sqrt{nT}}\right) \\
= \;& -\frac{1}{nT}tr\left\{\left(-\frac{1}{n}\sum_{i=1}^{n}\int\left.\frac{\partial^2\log f_{it}\left(z;\psi,\lambda_i(\psi)\right)}{\partial\psi\partial\psi'}\right|_{\psi=\psi_0}dG_i(z)\right)^{-1}\times\right. \\
& \left.\frac{1}{nT}\sum_{i=1}^{n}\int\left(\sum_{s=1}^{T}\sum_{t=1}^{T}\frac{\partial\log f_{it}(z;\psi_0,\widehat{\lambda}_i(\psi_0))}{\partial\psi}\frac{\partial\log f_{is}\left(z;\psi_0,\lambda_i(\psi_0)\right)}{\partial\psi'}\right)dG_i(z)\right\}+o\left(\frac{1}{\sqrt{nT}}\right)
\end{aligned}$$

by substituting (A.4), where the expression of $I(G)$ comes from the stationarity over t. This result gives the expression for $B_P(G)$.

Since $\widehat{\psi}_M$ is $\sqrt{nT}$-consistent to ψ_0 under the conditions in Theorem 2 (e.g., the Lyapounov's theorem; Sartori (2003), Hahn and Kuersteiner (2011))[14] and by the Envelope theorem, a consistent estimator for $B_P(G)$ under $n, T \to \infty$ can be obtained as (28) using the fact that

$$\begin{aligned}
\log f(z_{i,t};\psi,\lambda_i(\psi)) &= \log f_i^M(z_{i,t};\psi)+\left\{\frac{1}{T}M_i(\psi)-\log\left(\frac{f(z_{i,t};\psi,\widehat{\lambda}_i(\psi))}{f(z_{i,t};\psi,\lambda_i(\psi))}\right)\right\} \\
&= \log f_i^M(z_{i,t};\psi)+O_p(T^{-3/2})
\end{aligned}$$

and $\delta(\psi;G_i) = M_i(\psi)/T + O_p(T^{-3/2})$ from Lemma 1, where $J(\widehat{G})$ is defined using the HAC-type estimator like (19) for some truncation parameter $m \geq 0$ such that $m/T^{1/2} \to 0$ as $T \to \infty$. Note that the estimation error of $\delta(\psi;G_i)$ by $M_i(\widehat{\psi}_M)/T$ is at most $O_p(1/T^{3/2})$. This can be verified since we have

$$M_i(\widehat{\psi}_M)-M_i(\psi_0) = M_i(H(\widehat{G}))-M_i(H(G)) = \frac{dM_i(\psi_0)}{d\psi'}\times\frac{1}{nT}\sum_{j=1}^{n}\sum_{t=1}^{T}H^{(1)}(z_{j,t};G)+o_p\left(\frac{1}{(nT)^{1/2}}\right)$$

using a similar argument as above, and thus (13) yields

$$\begin{aligned}
\frac{1}{T}\mathbb{E}\left[M_i(\widehat{\psi}_M)-M_i(\psi_0)\right] &= \frac{1}{nT}\sum_{t=1}^{T}\int b_i(\psi_0)'I(G)^{-1}u_i^e dG_i(z)+o\left(\frac{1}{n^{1/2}T^{3/2}}\right) \\
&= \frac{1}{n}tr\left\{I(G)^{-1}\int u_i^e b_i(\psi_0)'dG_i(z)\right\}+o\left(\frac{1}{n^{1/2}T^{3/2}}\right)=O\left(\frac{1}{nT^{3/2}}\right).
\end{aligned}$$

Therefore, $\mathbb{E}_{G_i}[\delta(\psi;G_i) - M_i(\widehat{\psi}_M)/T] = \mathbb{E}_{G_i}[\delta(\psi;G_i) - M_i(\psi_0)/T] + T^{-1}\mathbb{E}_{G_i}[M_i(\widehat{\psi}_M) -$

[14] Recall that $\widehat{\psi}_M$ automatically correct the first order asymptotic bias in the asymptotic distribution of the standard QML estimator $\widehat{\psi}$ and $\sqrt{nT}(\widehat{\psi}_M - \psi_0) \to_d \mathcal{N}(0,\Omega_\psi)$ as $n,T \to \infty$ for $\Omega_\psi = \int \phi(z;G)\,\phi(z;G)'\,dG(z) > 0$ by construction, where $\phi(z;G)$ is the influence function of $\widehat{\psi}_M$ defined above.

$M_i(\psi_0)] = O(1/T^{3/2})$. □

Proof of Corollary 3 First note that $\partial \ell_i(\psi_0, \lambda_i(\psi_0))/\partial\psi = u_i^e$ by construction. Therefore, when g is nested in f, the standard information matrix identity gives

$$I(G) = \frac{1}{n}\sum_{i=1}^{n}\int -\frac{\partial^2 \ell_i(\psi_0, \lambda_i(\psi_0))}{\partial\psi\partial\psi'} dG_i = \frac{1}{n}\sum_{i=1}^{n} T\int u_i^e u_i^{e\prime} dG_i, \tag{A.5}$$

where the first equality is from the the stationarity over t. For $J(G)$, since $\partial \ell_{Pi}(\psi_0)/\partial\psi = u_i^e + b_i(\psi_0) + O_p(T^{-3/2})$ with $u_i^e = O_p(T^{-1/2})$ and $b_i(\psi_0) = O_p(T^{-1})$ from (12), we have

$$\begin{aligned} J(G) &= \frac{1}{n}\sum_{i=1}^{n} T\int \left[\frac{\partial \ell_i(\psi_0, \lambda_i(\psi_0))}{\partial\psi}\frac{\partial \ell_{Pi}(\psi_0)}{\partial\psi'}\right] dG_i \\ &= \frac{1}{n}\sum_{i=1}^{n} T\left\{\int u_i^e u_i^{e\prime} dG_i + \int u_i^e b_i(\psi_0)' dG_i + o_p(T^{-3/2})\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} T\left\{\int u_i^e u_i^{e\prime} dG_i + O_p(T^{-3/2})\right\}, \end{aligned} \tag{A.6}$$

where the remaining term in the second equality is $o_p(T^{-3/2})$ since $\int[\partial \ell_i(\psi_0, \lambda_i(\psi_0))/\partial\psi] dG_i = \int u_i^e dG_i = 0$. Therefore, by plugging (A.5) and (A.6) into $B_p(G)$, we have

$$B_p(G) = -\frac{r}{nT} - \frac{1}{n}\sum_{i=1}^{n}\delta(\psi; G_i) + O_p\left(\frac{1}{nT^{3/2}}\right),$$

which gives the expression of $B_p(\widehat{G})$. □

Proof of Theorem 4 By plugging the conditional prior (33) into the approximation (32), the log postertior probability of model $\mathcal{M}^k$ in (31) can be written as (we simply let $C_\pi = 1$)

$$\begin{aligned} \log \mathcal{P}\left(\mathcal{M}^k|z\right) &= -\log g(y) + \log\phi^k + \log\int \exp\left(T\sum_{i=1}^{n}\ell_i^I(\psi^k)\right)\eta(\psi^k|\mathcal{M}^k)d\psi^k \\ &= -\log g(y) + \log\phi^k \\ &\quad + \log\int \exp\left(\sum_{i=1}^{n} T\left\{\ell_i^M(\psi^k) + O_p\left(\frac{1}{T^2}\right)\right\}\right)\eta(\psi^k|\mathcal{M}^k)d\psi^k. \end{aligned}$$

But Taylor expansion yields

$$T\sum_{i=1}^{n}\ell_i^M(\psi^k) = T\sum_{i=1}^{n}\ell_i^M(\widehat{\psi}_M^k) - \frac{1}{2}\left(\widehat{\psi}_M^k - \psi^k\right)'\left[nT\widehat{\mathcal{I}}(\widehat{\psi}_M^k)\right]\left(\widehat{\psi}_M^k - \psi^k\right) + o_p(1),$$

where $\widehat{\psi}_M^k$ as the modified profile ML estimator of the model $\mathcal{M}^k$ and

$$\widehat{\mathcal{I}}(\widehat{\psi}_M^k) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathcal{I}}_i(\widehat{\psi}_M^k) = -\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\partial \log f_{it}(z_{i,t};\widehat{\psi}_M^k,\widehat{\lambda}_i(\widehat{\psi}_M^k))}{\partial\theta_i}\cdot\frac{\partial \log f_{it}(z_{i,t};\widehat{\psi}_M^k,\widehat{\lambda}_i(\widehat{\psi}_M^k))}{\partial\theta_i'}$$

is the averaged information matrix estimator in (14). Note that $\widehat{\psi}_M^k - \psi^k = O_p((nT)^{-1/2})$ when $n/T \to \gamma \in (0,\infty)$ and $\widehat{\mathcal{I}}(\widehat{\psi}_M^k) = O_p(1)$ from Assumptions 1 and 2. Therefore, using the uninformative flat prior $\eta(\psi^k|\mathcal{M}^k) = 1$, Laplace approximation (e.g., Tierney, Kass and Kadane (1989)) gives

$$\log\int \exp\left(T\textstyle\sum_{i=1}^{n}\ell_i^M(\psi^k)\right)d\psi^k = T\sum_{i=1}^{n}\ell_i^M(\widehat{\psi}_M^k) + \log\left\{(2\pi)^{r_k/2}\left|nT\widehat{\mathcal{I}}(\widehat{\psi}_M^k)\right|^{-1/2}\right\} + o_p(1),$$

and thus

$$\begin{aligned}\log\mathcal{P}\left(\mathcal{M}^k|z\right) &= -\log g(y) + \log\phi^k + O_p\left(\frac{n}{T}\right)\\ &\quad + T\sum_{i=1}^{n}\ell_i^M(\widehat{\psi}_M^k) + \frac{r_k}{2}\log 2\pi - \frac{r_k}{2}\log nT - \frac{1}{2}\log\left|\widehat{\mathcal{I}}(\widehat{\psi}_M^k)\right| + o_p(1),\end{aligned}$$

where $r_k = \dim(\psi^k)$. The result (34) follows by letting $c(z,k) = -\log g(y) + \log\phi^k + O_p(n/T) + (r_k/2)\log 2\pi - (1/2)\log|\widehat{\mathcal{I}}(\widehat{\psi}_M^k)|$, which is $O_p(1)$. □

Proof of Theorem 5 Recall that the selection rule is to choose p^* if $LIC^h(p^*) < LIC^h(p)$, where $0 \le p^*, p \le \overline{p}$ for some finite positive integer $\overline{p}$. We therefore need to prove that $\limsup_{n,T\to\infty}\mathbb{P}\left[LIC^h(p^*) < LIC^h(p_0)\right] = 0$ 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

$$\begin{aligned}&\mathbb{P}\left[LIC^h(p^*) < LIC^h(p_0)\right]\\ &= \mathbb{P}\left[\log\left(\frac{\widetilde{\sigma}^2(p^*)}{\widetilde{\sigma}^2(p_0)}\right) < \frac{h(n,T)}{nT}(p_0 - p^*) + \frac{1}{T}\left(\widetilde{R}_{n,T}(p_0) - \widetilde{R}_{n,T}(p^*)\right)\right].\end{aligned} \tag{A.7}$$

The left-hand-side of the inequality in (A.7) is nonnegative for any n and T because the residual sum of squares does not increase as the number of regressors increases. On the other hand, the right-hand-side of the inequality in (A.7) converges to zero as $n,T\to\infty$ since $0 < (p_0 - p^*) < \overline{p} < \infty$, $|\widetilde{R}_{n,T}(p_0) - \widetilde{R}_{n,T}(p^*)| < \infty$ from the invertibility in Assumption A-(ii), and $h(n,T)/nT \to 0$ as $n,T\to\infty$ by assumption. Therefore, $\limsup_{n,T\to\infty}\mathbb{P}[LIC^h(p^*) < LIC^h(p_0)] \le \mathbb{P}[\limsup_{n,T\to\infty}\{LIC^h(p^*) < LIC^h(p_0)\}] = \mathbb{P}[\varnothing] = 0$. Now for the case of over-selection, $p^* > p_0$, we write

$$\begin{aligned}&\mathbb{P}\left[LIC^h(p^*) < LIC^h(p_0)\right]\\ &= \mathbb{P}\left[nT\left(\log\widetilde{\sigma}^2(p^*) - \log\widetilde{\sigma}^2(p_0)\right) + n(\widetilde{R}_{n,T}(p^*) - \widetilde{R}_{n,T}(p_0)) < h(n,T)(p_0 - p^*)\right].\end{aligned} \tag{A.8}$$

Similarly as Lee (2006, 2012), we can show that $\text{plim}_{n\to\infty}\widetilde{\sigma}^2(p) - \sigma^2 = O\left(T^{-2}\right)$ for any p and thus

$$\log\widetilde{\sigma}^2(p) - \log\sigma^2 = \log\left(\frac{\widetilde{\sigma}^2(p) - \sigma^2}{\sigma^2} + 1\right) = \frac{\widetilde{\sigma}^2(p) - \sigma^2}{\sigma^2} + o_p\left(\left|\widetilde{\sigma}^2(p) - \sigma^2\right|\right) = O_p\left(T^{-2}\right)$$

for $0 < \sigma^2 < \infty$. It follows that $|\log\widetilde{\sigma}^2(p^*) - \log\widetilde{\sigma}^2(p_0)| \leq |\log\widetilde{\sigma}^2(p^*) - \log\widetilde{\sigma}^2(p_0)| + |\log\widetilde{\sigma}^2(p^*) - \log\widetilde{\sigma}^2(p_0)| = O_p\left(T^{-2}\right)$ for large n, and thus $nT(\log\widetilde{\sigma}^2(p^*) - \log\widetilde{\sigma}^2(p_0)) = O_p(n/T)$. We also note that (e.g., Bhansali (1981), Lee (2012)) $|\widetilde{R}_{n,T}(p_0) - \widetilde{R}_{n,T}(p^*)| = O_p(1/T)$ and $n(\widetilde{R}_{n,T}(p_0) - \widetilde{R}_{n,T}(p^*)) = O_p(n/T)$. The left-hand-side of the inequality in (A.8) is thus $O_p(1)$ for large n and T because it is assumed that $n/T \to \gamma \in (0,\infty)$. On the other hand, 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, it also holds that $\limsup_{n,T\to\infty}\mathbb{P}[LIC^h(p^*) < LIC^h(p_0)] = 0$ for $p^* > p_0$. □

Proof of Corollary 6 We consider the case of over-selection, $p^* > p_0$ and $p^{**} > p_0$. We first define that

$$\begin{aligned}\Delta LIC^h &\equiv LIC^h(p^*) - LIC^h(p_0) \\ &= \log\left(\frac{\widetilde{\sigma}^2(p^*)}{\widetilde{\sigma}^2(p_0)}\right) + \frac{h(n,T)}{nT}(p^* - p_0) + \frac{1}{T}\left(\widetilde{R}_{n,T}(p^*) - \widetilde{R}_{n,T}(p_0)\right)\end{aligned}$$

and

$$\Delta LIC_0^h \equiv LIC_0^h(p^{**}) - LIC_0^h(p_0) = \log\left(\frac{\widetilde{\sigma}^2(p^{**})}{\widetilde{\sigma}^2(p_0)}\right) + \frac{h(n,T)}{nT}(p^{**} - p_0).$$

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

$$\begin{aligned}&\limsup_{n,T\to\infty}\mathbb{P}\left[\Delta LIC^h < \Delta LIC_0^h\right] \qquad (A.9)\\ =\ &\limsup_{n,T\to\infty}\mathbb{P}\left[\log\left(\frac{\widetilde{\sigma}^2(p^*)}{\widetilde{\sigma}^2(p^{**})}\right) < \frac{h(n,T)}{nT}(p^{**} - p^*) + \frac{1}{T}\left(\widetilde{R}_{n,T}(p_0) - \widetilde{R}_{n,T}(p^*)\right)\right] = 0\end{aligned}$$

as $n, T \to \infty$. Note that $LIC^h(p)$ has the heavier penalty term than $LIC_0^h(p)$ and thus $p^{**} \geq p^*$ by construction. Therefore, the left-hand-side of the last inequality in (A.9) is nonnegative for any n and T, whereas the left-hand-side goes to zero as in (A.7). This result implies that ΔLIC^h cannot be smaller than ΔLIC_0^h as probability approaches to one and thus $\limsup_{n,T\to\infty}\{\mathbb{P}\left[\Delta LIC^h < 0\right] - \mathbb{P}\left[\Delta LIC_0^h < 0\right]\} \leq \limsup_{n,T\to\infty}\mathbb{P}\left[\Delta LIC^h - \Delta LIC_0^h < 0\right] \leq \mathbb{P}\left[\limsup_{n,T\to\infty}\{\Delta LIC^h - \Delta LIC_0^h < 0\}\right] = 0$. □

Proof of Corollary 7 From Bhansali (1981) and Lee (2012), it can be verified that $|\widetilde{R}_{n,T}(p_0) - \widetilde{R}_{n,T}(p^*)| = c|p_0 - p^*|/T + o_p(1/T)$. Therefore, $\mathbb{P}\left[LIC^h(p^*) < LIC^h(p_0)\right]$ is approximately the same as

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

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

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in B.N. Petrov and B.F. Csaki (Eds.), *2nd International Symposium on Information Theory*, 267–281, Budapest: Academia Kiado.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Andrews, D.W.K. (1994). Asymptotics for semi-parametric econometric models via stochastic equicontinuity, *Econometrica*, 62, 43–72.

Arellano, M., and S. Bonhomme (2009). Robust priors in nonlinear panel data models, *Econometrica*, 77, 489-536.

Arellano, M. and J. Hahn (2006). A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects, *CEMFI Working Paper*: No. 0613.

Arellano, M., and J. Hahn (2007). Understanding bias in nonlinear panel models: Some recent developments, R. Blundell, W.K. Newey, and T. Persson eds., *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Volume III*, Cambridge University Press.

Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator, *Biometrika*, 70, 343-365.

Berger, J.O., J.K. Ghosh, and N. Mukhopadhyay (2003). Approximations and consistency of the Bayes factors as model dimension grows, *Journal of Statistical Planning and Inference*, 112, 241-258.

Berger, J.O., B. Liseo, and R.L. Wolpert (1999). Integrated likelihood methods for eliminating nuisance parameters, *Statistical Science*, 14, 1-28.

Bester, C.A., and C. Hansen (2009). A Penalty Function Approach to Bias Reduction in Nonlinear Panel Models with Fixed Effects, *Journal of Business and Economic Statistics*, 27, 131-148.

Bhansali, R.J. (1981). Effects of not knowing the order of an autoregressive process on the mean squared error of prediction−I, *Journal of the American Statistical Association*, 76, 588-597.

Cherkassky, V., X. Shao, F.M. Mulier, and V.N. Vapnik (1999). Model complexity control for regression using VC generalization bounds, *IEEE Transactions on Neural Networks*, 10, 1075-1089.

Chkrabarti, A., and J.K. Ghosh (2006). A generalization of BIC for the general exponential family, *Journal of Statistical Planning and Inference*, 136, 2847-2872.

Claeskens, G. and N.L. Hjort (2003). The focused information criterion, *Journal of the American Statistical Association*, 98, 900-916.

Cox, D.R., and N. Reid (1987). Parameter orthogonality and approximate conditional inference (with Discussion), *Journal of the Royal Statistical Society*, B 49, 1-39.

DiCiccio, T.J., M.A. Martin, S.E. Stern, and G.A. Young (1996). Information bias and adjusted profile likelihoods, *Journal of the Royal Statistical Society*, B 58, 189-203.

Guyon, X., and J.-F. Yao (1999). On the underfitting and overfitting sets of models chosen by order selection criteria, *Journal of Multivariate Analysis*, 70, 221-249.

Hahn, J. and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects, *Econometrica*, 70, 1639-1657.

Hahn, J., and G. Kuersteiner (2011). Bias reduction for dynamic nonlinear panel models with fixed effects, *Econometric Theory*, 27, 1152-1191.

Hahn, J., and W. Newey (2004). Jackknife and analytical bias reduction for nonlinear panel models, *Econometrica*, 72, 1295-1319.

Heckman, J., and B.J. Singer (1984). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data, *Econometrica*, 52, 271-320.

Hodges, J.S. and D.J. Sargent (2001). Counting degrees of freedom in hierarchical and other richly-parametrised models, *Biometrika*, 88, 367-379.

Huber, P.J. (1981). *Robust Statistics*, New York: Wiley.

Kass, R. and A. Raftery (1995). Bayes Factors, *Journal of the American Statistical Association*, 90, 773-795.

Konishi, S. and G. Kitagawa (1996). Generalized information criteria in model selection, *Boimetrika*, 83, 875-890.

Lancaster, T. (2002). Orthogonal parameters and panel data, *Review of Economic Studies*, 69, 647-666.

Lee, Y. (2006). *Nonparametric Approaches to Dynamic Panel Modelling and Bias Correction*, Ph.D. dissertation, Yale.

Lee, Y. (2010). Nonparametric estimation of dynamic panel models with fixed effects, unpublished manuscript, University of Michigan.

Lee, Y. (2012). Bias in dynamic panel models under time series misspecification, *Journal of Econometrics*, 169, 54-60.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete Index Set, *Annals of Statistics*, 15, 958-975.

Maddala, G.S. (1971). The use of variance components models in pooling cross section and time series data, *Econometrica*, 39, 341-358.

McCullagh, P., and R. Tibshirani (1990). A simple method for the adjustment of profile likelihoods, *Journal of the Royal Statistical Society*, B 52, 325-344.

Murphy, S.A., and A.W. van der Vaart (2000). On Profile Likelihood. *Journal of the American Statistical Association*, 95, 449-465.

Newey, W.K. (1994). The asymptotic variance of semiparametric estimators, *Econometrica*, 62, 1349–1382.

Neyman, J. and E. Scott (1948). Consistent estimates based on partially consistent observations, *Econometrica*, 16, 1-32.

Ng, S., and P. Perron (2005). A note on the selection of time series models, *Oxford Bulletin of Economics and Statistics*, 67:1, 115-134.

Rissanen, J. (1986). Stochastic Complexity and Modeling, *Annals of Statistics*, 14, 1080-1100.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters, *Biometrika*, 90, 533-549.

Serfling, R. (1998). *Approximation Theorems of Mathematical Statistics*, Wiley.

Severini, T.A. (1998). An approximation to the modified profile likelihood function, *Boimetrika*, 85, 403-411.

Severini, T.A. (2000). *Likelihood Methods in Statistics*, New York: Oxford University Press.

Severini, T.A. and W.H. Wong (1992). Profile likelihood and conditionally parametric models, *Annals of Statistics*, 20, 1768-1802.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika*, 63, 117-126.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *Annals of Statistics*, 8, 147-164.

Stein, C. (1956). Efficient nonparametric testing and estimation, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 187-195.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwartz, *Journal of the Royal Statistical Society, Series B*, 41, 276-278.

Tierney, L., R.E. Kass and J.B. Kadane (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions, *Journal of the American Statistical Association*, 84, 710-716.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models, *Econometrica*, 50, 1-25.

Withers, C.S. (1983). Expansions for the distribution and quantiles of a regular functional of the empirical distribution with applications to nonparametric confidence intervals, *Annals of Statistics*, 11, 577-587.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation, *Biometrika*, 92, 937-950.

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

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

TABLE 1: Average of lag order selections over 1000 iterations
($p_0 = 1$; $\overline{p} = 10$)

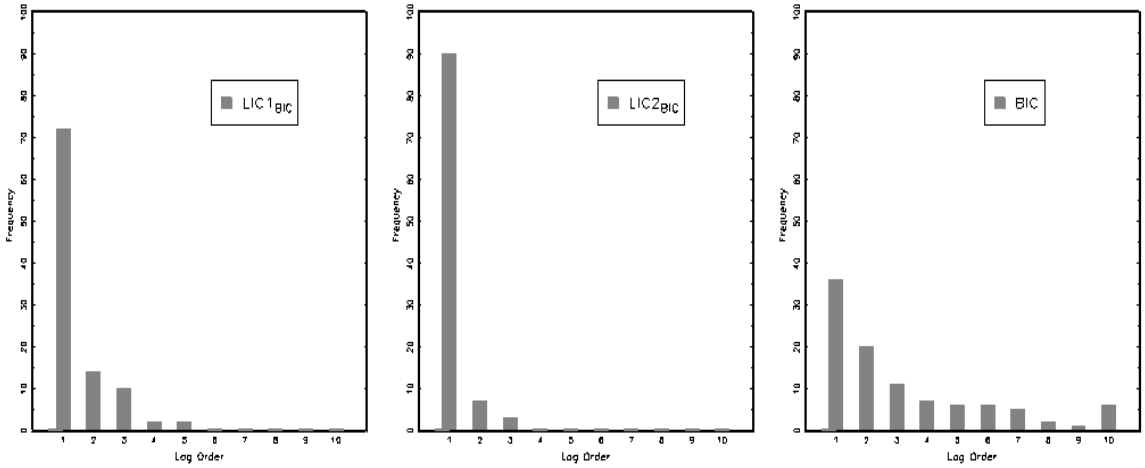

FIGURE 1: Order selection frequencies over 1000 iterations
($p_0 = 1$; $(n, T) = (100, 50)$)

$$p_0 = 2: \quad y_{i,t} = \mu_i + 0.15y_{i,t-1} + 0.15y_{i,t-2} + \varepsilon_{i,t}$$

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

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

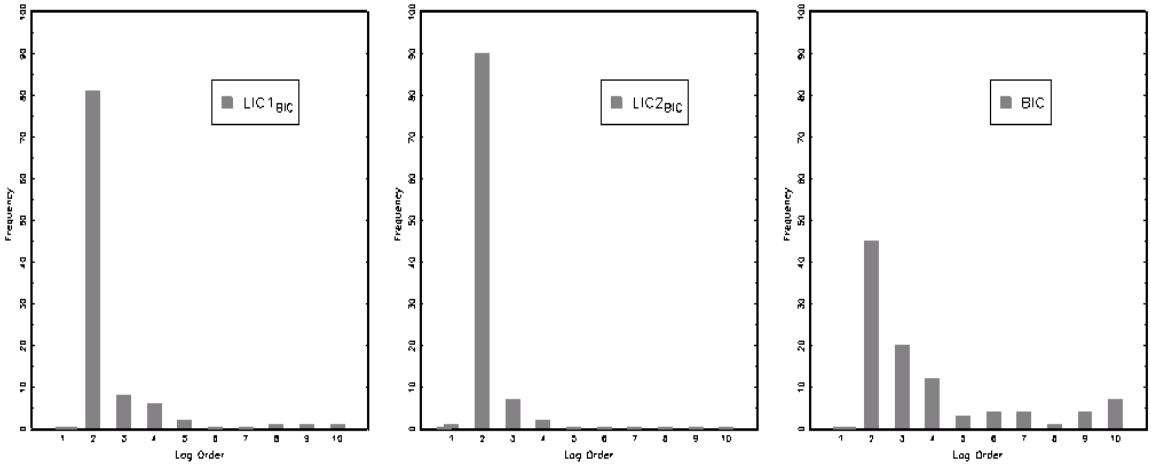

FIGURE 2: Order selection frequencies over 1000 iterations
$(p_0 = 2;\ (n, T) = (100, 50))$

$$p_0 = 3: \quad y_{i,t} = \mu_i + 0.15y_{i,t-1} + 0.15y_{i,t-2} + 0.15y_{i,t-3} + \varepsilon_{i,t}$$

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

TABLE 3: Average of lag order selections over 1000 iterations
($p_0 = 3$; $\overline{p} = 10$)

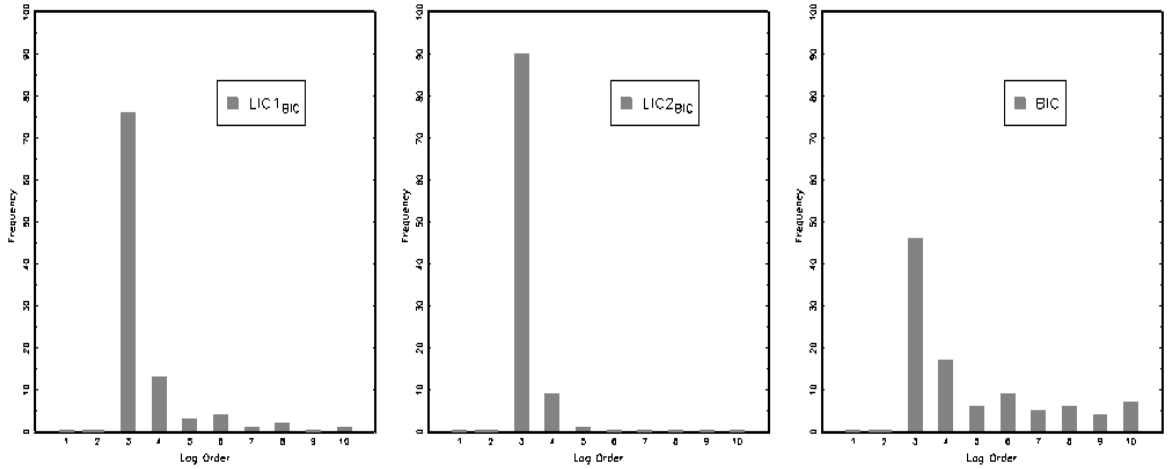

FIGURE 3: Order selection frequencies over 1000 iterations
($p_0 = 3$; $(n, T) = (100, 50)$)

$$p_0 = 4: \quad y_{i,t} = \mu_i + 0.15y_{i,t-1} + 0.15y_{i,t-2} + 0.15y_{i,t-3} + 0.15y_{i,t-4} + \varepsilon_{i,t}$$

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

TABLE 4: Average of lag order selections over 1000 iterations
($p_0 = 4$; $\overline{p} = 10$)

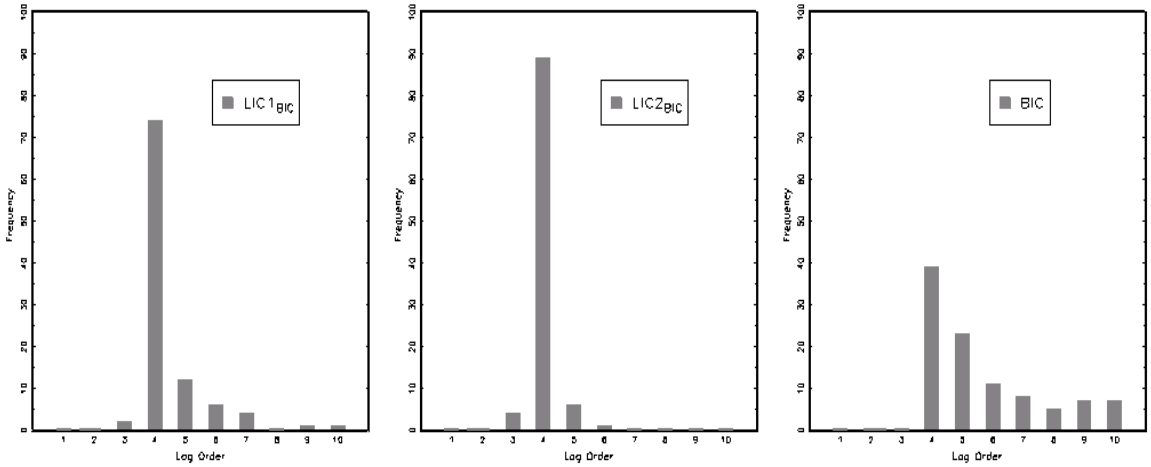

FIGURE 4: Order selection frequencies over 1000 iterations
($p_0 = 4$; $(n, T) = (100, 50)$)

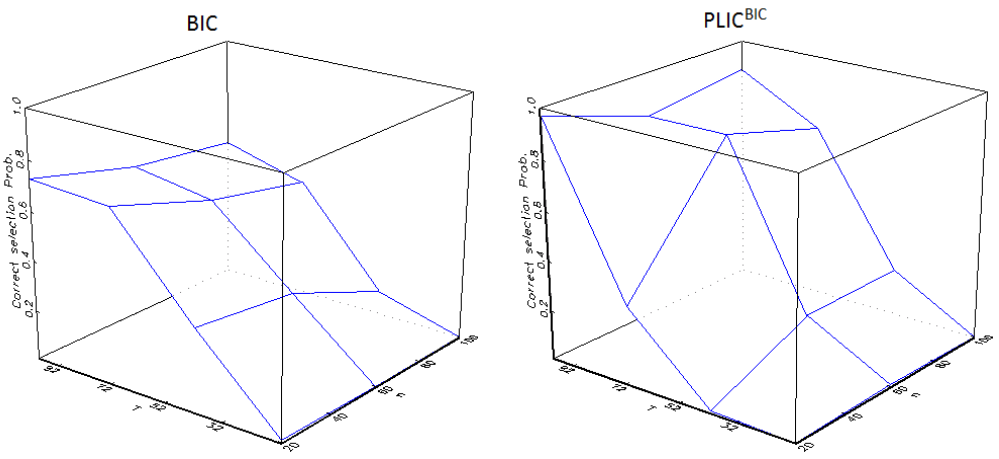

Figure 1:

FIGURE 5: Correct order selection frequencies over 1000 iterations when $p_0 = 3$