MODEL SELECTION IN THE PRESENCE OF INCIDENTAL PARAMETERS

By

Yoonseok Lee and Peter C.B. Phillips

October 2013

COWLES FOUNDATION DISCUSSION PAPER NO. 1919

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Model Selection in the Presence of Incidental Parameters[*]

Yoonseok Lee[†] and Peter C.B. Phillips[‡]

September 2013

Abstract

This paper considers model selection in nonlinear panel data models where incidental parameters or large-dimensional nuisance parameters are present. Primary interest typically centres on selecting a model that best approximates the underlying structure involving parameters that are common within the panel after concentrating out the incidental parameters. It is well known that conventional model selection procedures are often inconsistent in panel models and this can be so even without nuisance parameters (Han et al, 2012). Modifications are then needed to achieve consistency. New model selection information criteria are developed here 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). These model selection criteria impose heavier penalties than those associated with standard information criteria such as AIC and BIC. The additional penalty, which is data-dependent, properly reflects the model complexity arising from the presence of incidental parameters. A particular example is studied in detail involving lag order selection in dynamic panel models with fixed individual effects. The new criteria are shown to control for over/under-selection probabilities in these models and lead to consistent order selection criteria.

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 authors acknowledge helpful comments from Bruce Hansen, Chris Hansen, Yuichi Kitamura, Roger Moon, Peter Robinson and seminar participants at Yale, Columbia, Rochester, UVa, Boston College, Seoul National, MSU, Montreal, Pittsburgh, 2011 International Conference on Panel Data, and 2011 Midwest Econometrics Group. Lee thanks the Cowles Foundation for Research in Economics at Yale University, where he was a visiting fellow while working on this project. Phillips acknowledges support from the NSF under Grants No. SES 0956687 and SES 1258258.

[†] Syracuse University. *Address*: Department of Economics and Center for Policy Research, Syracuse University, 426 Eggers Hall, Syracuse, NY 13244. *E-mail*: `ylee41@maxwell.syr.edu`.

[‡] Yale University, University of Auckland, University of Southampton and Singapore Management University. *Address*: Department of Economics, Yale University, Box 208281, New Haven, CT 06520-8281. *E-mail*: `peter.phillips@yale.edu`.

1 Introduction

As datasets grow richer, more sophisticated models are being used in empirical econometric work, including semiparametric models, large dimensional parametric models, and panel systems with manifold heterogenous effects that lead to a proliferation of nuisance parameters. Good model selection procedures are an important element in empirical work to avoid bias, to help in validating inference, and to assist in ensuring sound policy implications. They are particularly important in more sophisticated systems where multi-index asymptotics and high dimensional nuisance parameters can affect the properties of estimators, inference and model selection.

Some of these panel modeling issues were considered in the pioneering work by Anderson and Hsiao (1981), which examined the use of multi-index asymptotics, dynamic panel estimation inconsistency, and the possible use of instrumental variable (IV) methods to avoid inconsistencies in dynamic panel regression with short wide panels. Following that paper, there was a massive flowering of research on dynamic panel modeling, efficient IV estimation techniques and semiparametric methods, to all of which Cheng Hsiao has made significant contributions. Much of this work is overviewed in Hsiao (2003).

One topic that is still relatively unexplored in this field is model selection in dynamic panels. Specification tests and information-criteria provide two standard approaches to model selection and are available for use in dynamic panels. The specification test approach requires an *ad hoc* null, a set of alternative models, and a test sequence to evaluate the alternatives. On the other hand, the model selection approach considers all the candidate models jointly and chooses one that optimizes an information criterion. Examples include the Akaike information criterion (AIC), Bayesian information criterion (BIC), posterior information criterion (PIC), Hannan-Quinn (HQ) criterion, the Mellows' C_p criterion, bootstrap criteria and cross-validation approaches.

An important assumption in most model selection approaches is that the number of parameters in each candidate model is finite or at most grows slowly compared to the sample size. For example, Stone (1979) showed that consistency of the standard BIC order selector breaks down when the number of parameters in the candidate model diverges with the sample size.[1] In many cases, large dimensional parameter spaces arise from the proliferation of nuisance parameters which, though they are not of primary interest, are required for specifying heterogeneity or for handling omitted variables. The present paper examines why standard model selection criteria perform poorly for such cases and proposes modified selection criteria that are effective when the candidate models have nuisance parameters whose dimension grows

[1] This limitation in standard criteria is now well understood and several approaches have been proposed for model selection in large dimensional models, particularly in the Bayesian framework. Examples are Berger et al. (2003) and Chakrabarti and Ghosh (2006), who analyze the use of the Laplace approximation in large-dimensional exponential families to compute the Bayes factor and achieve a consistent selector.

with the sample size, analogous to incidental parameters (Neyman and Scott (1948)).

In particular, we study the specification of panel data models in which the focus of interest is 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 effects) and heteroskedastic variances. The main objective is to choose the model that fits best the data generating process when only a subset of the parameters is of central interest. Such an 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 across the candidate models. A similar approach can be found in Claeskens and Hjort (2003) in the context of cross section 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 incidental parameters and to obtain new model selection criteria. One method applies profiling to 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. This 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 therefore needs appropriate modification (e.g., Sartori (2003)). It turns out that the new information criterion requires a heavier penalty than that of standard information criteria such as AIC so that the degrees of freedom in the model are properly counted. However, the penalty is different from the total number of parameters (i.e., $\dim(\psi) + n \dim(\lambda_i)$). The additional penalty depends on a 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 a second approach, we develop a Bayesian model selection criterion that is based on the Bayes factor, in which the posterior is obtained using the integrated likelihoods. These two approaches – one based on the profile likelihood and the other based on the integrated likelihood – are closely related, as in the standard AIC and BIC, provided that a proper prior for the incidental parameter is used in performing the integration. In the pseudo-likelihood setup, we obtain the prior so that the integrated likelihood is close to the genuine likelihood (e.g., the robust prior of Arellano and Bonhomme (2009)) and that depends on the data in general.

The majority of panel data studies focus on modifying the profile or integrated likelihood as a means of bias reduction in maximum likelihood estimation, which 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, 2013, 2012); Bester and Hansen (2009)). However, as discussed in Lee (2006, 2012), if the model is not correctly specified, effort to reduce bias stemming from incidental parameters may exacerbate bias. Hence, correct model specification is very important, particularly for dynamic or nonlinear panel models where bias occurs naturally in estimation. Correct model specification should ideally precede the use of bias correction or bias reduction procedures. The focus of the present paper is on mechanisms to address the specification problem.

The remainder of the paper is organized as follows. Section 2 summarizes the incidental parameter problem in the quasi maximum likelihood setup. The modified profile likelihood and bias reduction in panel data models are also discussed. Section 3 develops an AIC-type information criterion based on the profile likelihood. A profile KLIC is introduced that is general enough to be applied in heterogenous panel data models. Section 4 obtains a BIC-type information criterion based on the integrated likelihood and explores connections between AIC-type and BIC-type criteria by developing a robust prior. In Section 5, the methodology is mobilized in the particular example of lag order selection for dynamic panel models. This Section also reports simulations that examine the statistical performance properties of the procedures. Section 6 concludes. Proofs are given in the Appendix.

2 Incidental Parameter Problems 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 have an unknown distribution $G_i(z)$ with probability density $g_i(z)$. The components $z_{i,t}$ are allowed to have heterogenous distributions across i but are cross-section independent. On the other hand, $z_{i,t}$ may be serially correlated over t but is assumed to be 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 simplicity in what follows.

Since $g_i(z)$ is unknown a priori, we 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 common to all i, whereas the $\lambda_i \in \Lambda \subset \mathbb{R}$ are individual nuisance parameter that are specific to i. 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 may consider multidimensional λ_i (e.g., Arellano and Hahn (2006)) but focus on the scalar case for expositional simplicity.

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

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

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

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

We assume the following conditions as in 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 observation 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

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

where

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

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. By separability, furthermore, we can consider the standard asymptotic results for $\widehat{\lambda}_i\left(\psi\right)$ in powers of T. The quasi maximum profile likelihood estimator of ψ is then obtained as

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

which indeed corresponds to the QMLE of ψ because the maximum is obtained in two suc-

[2] When we consider dynamic models, $f_{it}\left(z_{i,t};\psi,\lambda_i\right)$ is understood as a conditional density given the lagged observations. For example, with $z_{i,t}=\left(y_{i,t},y_{i,t-1},\cdots,y_{i,t-p}\right)$ for some $p\geq 1$, we define $f_{it}\left(z_{i,t};\psi,\lambda_i\right)=f\left(y_{i,t}|y_{i,t-1},\cdots,y_{i,t-p};\psi,\lambda_i\right)$.

cessive steps rather than simultaneously. The existence of $\widehat{\psi}$ follows from Assumption 1 as in 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 $\{f(\cdot;\psi,\lambda_i)\}$. The intuitive explanation is that the profile likelihood is itself a biased estimate of the original likelihood. Modification of the profile likelihoods in the form of

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

is widely studied, where

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

Such modification makes the *modified profile likelihood* $f_{it}^{M}(\cdot;\psi)$ behave more like 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 for small T. A bias-reduced estimator for ψ can therefore be obtained by maximizing the modified profile likelihood (i.e., the quasi maximum modified profile likelihood estimation) as

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

Further discussion of the maximum modified profile likelihood estimator can be found in Barndorff-Nielsen (1983), Severini (1998, 2000) and Sartori (2003) among others, particularly regarding appropriate choices of the modification term $M_i(\psi)$. Closely related works consider 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 parameter problem

From 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

$$\psi_T = \arg\min_{\psi\in\Psi}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{G_i}\left[\frac{1}{T}\sum_{t=1}^{T}\log\left(\frac{g_i(z_{i,t})}{f(z_{i,t};\psi,\widehat{\lambda}_i(\psi))}\right)\right].$$

We denote by $\mathbb{E}_{G_i}[\cdot] = \int[\cdot]dG_i$ the expectation taken with respect to the true distribution G_i for each i. From the stationarity assumption over t, ψ_T can be rewritten as $\psi_T = \arg\min_{\psi\in\Psi}\lim_{n\to\infty}\overline{D}(g \parallel f(\psi, \widehat{\lambda}(\psi)))$ with

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

Note that

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

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] $\overline{D}(g \parallel f(\psi, \widehat{\lambda}(\psi)))$ is thus simply the averaged KLIC over i and t. We further let[4]

$$\lambda_i(\psi) = \arg\min_{\lambda_i\in\Lambda}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} D(g_i \parallel f_{it}(\psi, \lambda_i)) \tag{8}$$

for each i and

$$\begin{aligned} \psi_0 &= \arg\min_{\psi\in\Psi}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{G_i}\left[\frac{1}{T}\sum_{t=1}^{T}\log\left(\frac{g_i(z_{i,t})}{f(z_{i,t}; \psi, \lambda_i(\psi))}\right)\right] \\ &= \arg\min_{\psi\in\Psi}\lim_{n\to\infty}\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} D(g_i \parallel f_{it}(\psi, \lambda_i(\psi))) \end{aligned} \tag{9}$$

by stationarity. The KLIC minimizers ψ_0 and $\lambda_0 = (\lambda_{10}, \cdots, \lambda_{n0})'$ are obtained from (9) and $\lambda_{i0} = \lambda_i(\psi_0)$ for each i.

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 $\mathbb{E}_{G_i}[\partial\log f(z; \theta_{i0})/\partial\theta_i \cdot$*

[3] We may interpret 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 stationarity.

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

$\partial \log f(z;\theta_{i0})/\partial\theta_i']$ *are both nonsingular, where* $\theta_{i0} = (\psi_0', \lambda_{i0})'$. *(v)* $(\psi_0', \lambda_0')' \in \Psi \times \Lambda^n$ *is the unique solution from (8) and (9), where* $(\psi_0', \lambda_0')'$ *lies in the interior of the support.*

From White (1982) under Assumptions 1 and 2, we have 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 (9). This inconsistency is a manifestation of the incidental parameter problem (e.g., Neyman and Scott (1948)) in the context of the QMLE. In general, it can often be shown that (e.g., Arellano and Hahn (2007), Bester and Hansen (2009))

$$\psi_T - \psi_0 = \frac{\Upsilon}{T} + O\left(\frac{1}{T^2}\right) \tag{10}$$

where Υ/T represents bias of $O(T^{-1})$, and when $n, T \to \infty$ with $n/T \to \gamma \in (0,\infty)$ and $n/T^3 \to 0$, we have

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

for some positive definite matrix Ω_ψ. The main source of this bias is that $\widehat{\lambda}_i(\psi)$ in (3) is still random and thus is not the same as $\lambda_i(\psi)$ in (8). The estimation error of $\widehat{\lambda}_i(\psi)$ with finite T is not negligible even when $n \to \infty$, and the expectation of the profile score is no longer zero for each i even under sufficient regularity conditions.

More precisely, for each i, we define the (pseudo-)information matrix as

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

where the partition is conformable with $\theta_i = (\psi', \lambda_i)' \in \mathbb{R}^{r+1}$. The matrices $\mathcal{I}_i$, $\mathcal{I}_{i,\psi\psi}$ and $\mathcal{I}_{i,\lambda\lambda}$ are all nonsingular from Assumption 2. We also define the (scaled individual) score functions as

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

Note that $u_i^e(\psi_0,\lambda_{i0})$ is the *efficient score* for ψ at (ψ_0,λ_{i0}) and 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 leads to an asymptotic bias that appears in (10). The modification term $M_i(\psi)$ in (5) can be found as a function in ψ, provided that $f(\cdot; \theta_i)$ is thrice 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 $b_i(\psi_0)$.

2.3 Bias reduction

The standard bias corrected estimators in nonlinear (dynamic) fixed effect regressions correspond to $\widehat{\psi}_M$ in (6) and are given by (e.g., Hahn and Newey (2004); Arellano and Hahn (2007); Hahn and Kuersteiner (2011))

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

where $\widehat{\mathcal{I}}_i^e(\widehat{\psi}_M)$ is a consistent estimator of the efficient information $\mathcal{I}_i^e = \mathcal{I}_{i,\psi\psi} - \mathcal{I}_{i,\psi\lambda}\mathcal{I}_{i,\lambda\lambda}^{-1}\mathcal{I}_{i,\lambda\psi}$ as $T \to \infty$. In principle, $\widehat{\mathcal{I}}_i^e(\widehat{\psi}_M)$ can be derived 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 may 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 use

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

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 in the same way as equation (12) in Arellano and Hahn (2007).

[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)).

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 standard asymptotic results for (Q)ML estimators, we have the first order stochastic expansion for an arbitrary fixed ψ as

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

for each i, where $\overline{\mathcal{H}}_i(\psi) = \lim_{T\to\infty} \mathbb{E}_{G_i}(-\partial^2 \ell_i(\psi, \lambda_i(\psi))/\partial\lambda_i^2)$. Similarly we can expand $\ell_i^P(\psi) = \ell_i(\psi, \widehat{\lambda}_i(\psi))$ around $\lambda_i(\psi)$ for given ψ as

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

from (15), where the dominating term is $O_p(T^{-1})$ because $\overline{\mathcal{H}}_i(\psi) = O(1)$ 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))

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

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), (13) and (17), this result suggests that a simple form of the modification function in $\ell_i^M(\psi)$ can be obtained as

$$\begin{aligned} \frac{1}{T}M_i(\psi) &= \frac{1}{2T}\left(-\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 \log f_{it}(z_{i,t};\psi, \widehat{\lambda}_i(\psi))}{\partial \lambda_i^2}\right)^{-1} \qquad (18) \\ &\quad \times \sum_{j=-m}^{m}\frac{K_j}{T}\sum_{t=\max\{1,j+1\}}^{\min\{T,T+j\}}\frac{\partial \log f_{it}(z_{i,t};\psi, \widehat{\lambda}_i(\psi))}{\partial \lambda_i}\cdot\frac{\partial \log f_{it}(z_{i,t-j};\psi, \widehat{\lambda}_i(\psi))}{\partial \lambda_i}, \end{aligned}$$

whose first derivative corrects the leading bias term $b_i(\psi_0)$ at $\psi = \psi_0$ in the profile score (12) with probability approaching to one. The second component in (18) corresponds to the robust variance estimator of $\sqrt{T}\partial \ell_i(\psi, \widehat{\lambda}_i(\psi))/\partial\lambda_i$. For a more general treatment of the modification to the profile likelihood, see Barndorff-Nielsen (1983) for the modified profile likelihood approach or McCullagh and Tibshirani (1990) for the adjusted profile likelihood approach. Note that $M_i(\psi)/T$ in (18) is similar to the modification functions suggested by Arellano and Hahn (2006) and Bester and Hansen (2009), which appears to be robust to arbitrary serial correlation in $\partial \log f_{it}(z_{i,t};\psi, \widehat{\lambda}_i(\psi))/\partial\lambda_i$. The truncation parameter $m \geq 0$ is chosen so that $m/T^{1/2} \to 0$ as $T \to \infty$, and the lag kernel function K_j generally guarantees positive definiteness of the variance estimate (e.g., by use of the Bartlett kernel: $K_j = 1-(j/(m+1))$).

3 Profile Likelihood and KLIC

3.1 Model selection

Panel data studies conventionally focus on reducing the first order bias (10) arising from the presence of incidental parameters under a presumption that the models are correctly specified. As discussed in Lee (2006, 2012), however, if the model is not correctly specified effort to reduce bias due to incidental parameters may be counterproductive and even exacerbate bias. Achieving correct model specification is therefore an important component in successful bias reduction, particularly for dynamic and nonlinear panel models. Examples include the choice of lag order in panel $ARMA$ models or the functional structure in nonlinear panel models. Importantly, correct model specification should precede the use of any bias corrections. We focus here on model specification – in particular, we are interested in selecting a model $f\left(z|\psi,\lambda_i\right)$ that is closest to the true model $g_i\left(z\right)$ 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 standard model selection by comparing estimates of the averaged KLIC given by

$$\begin{aligned}\min_{\theta}\overline{D}\left(g\parallel f(\theta)\right) &= \min_{\theta}\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}D\left(g_i\parallel f_{it}(\theta)\right) \qquad (19)\\ &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int\log g_i\left(z\right)dG_i(z)-\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int\log f_{it}(z;\widehat{\theta})dG_i(z),\end{aligned}$$

where $\widehat{\theta}$ is the QMLE, which is a consistent estimator of $\theta_0=\arg\min_{\theta}\lim_{n,T\to\infty}\overline{D}\left(g\parallel f(\theta)\right)$ in this case. Note that averaged KLIC $\overline{D}\left(g\parallel f(\theta)\right)$ is defined so that it could accommodate possibly heterogeneous panel data models. We select a model $f\left(\cdot;\theta\right)$ whose KLIC in (19) is the minimum among the candidates. Equivalently, since the first term in (19) does not depend on the model, we select the model $f\left(\cdot;\theta\right)$ 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

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

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

$$\widetilde{\Phi}(\widehat{\theta}) = \widehat{\Phi}(\widehat{\theta}) - B(\widehat{G}) \tag{20}$$

as an information criterion for model selection, where $B\left(G\right) = \mathbb{E}[\widehat{\Phi}(\widehat{\theta}) - \Phi(\widehat{\theta})]$ 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 $B\left(G\right)$ is asymptotically the ratio of $\dim(\theta)$ to the sample size when $\widehat{\theta}$ is the QMLE and g is nested in f.

Now consider the case with incidental parameters $\lambda \in \mathbb{R}^n$, where $\theta = \left(\psi', \lambda'\right)'$. Similar to the discussion of the previous section, when the dimension of the parameter vector θ is substantial relative to the sample size, the incidental parameter problem prevails and it is not straightforward to use a standard criterion like (20). One possible solution is to reduce the dimension of the parameters by concentrating out the nuisance parameters. Particularly when it is assumed that the (finite-dimensional) parameter of central interest ψ governs the key structure of the model that is unchanging over i it is natural to concentrate out the nuisance parameters λ_i in conducting model selection. 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 idea is similar to the profile likelihood approach when interest lies in a subset of parameters. Some examples are as follows.

Example 1 (Variable or model selection in panel models) Consider a parametric nonlinear fixed-effect model given by $y_{i,t} = \xi(x_{i,t}, u_{i,t}; \mu_i, \beta, \sigma_i^2)$ where $\xi(\cdot;\cdot)$ is some known specified function, $u_{i,t}$ 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 in this case 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 common choice would be 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) Consider a panel $AR(p)$ model with fixed effects 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 here is to choose the correct lag order p, allowing for the presence of incidental parameters μ_i. When $p = \infty$, the problem becomes one of finding a best approximation in the finite order $AR(p)$ class.

Example 3 (Number of support choice of random effects or random coefficient) 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 finite support number 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), the problem is to choose the finite support number of nonparametric frailty in the Heckman-Singer model (Heckman and Singer (1984)), if the Cox partial likelihood is viewed as a profile likelihood.

3.2 Profile likelihood information criterion

For model selection using an information criterion in the presence of incidental parameters we consider the *profile Kullback-Leibler divergence*, in which the incidental parameters λ_i are concentrated out of 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*

$$D_P\left(g_i \parallel f_{it}(\psi, \lambda_i); \psi\right) = \min_{\lambda_i \in \Lambda} D\left(g_i \parallel f_{it}(\psi, \lambda_i)\right). \tag{21}$$

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 same conditions as standard KLIC. For example, $D_P\left(g_i \parallel f_{it}(\psi, \lambda_i); \psi\right)$ is nonnegative and equals 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 $(\psi'_*, \lambda_{i*})' \in \Psi \times \Lambda$).

Similar to the standard case (19), we select the model that has the smallest value of the estimate of

$$\min_{\psi \in \Psi} \overline{D}_P\left(g \parallel f(\psi, \lambda); \psi\right) = \min_{\psi \in \Psi} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} D_P\left(g_i \parallel f_{it}(\psi, \lambda_i); \psi\right). \tag{22}$$

Under stationarity over t, however, it holds that $D_P\left(g_i \parallel f_{it}(\psi, \lambda_i); \psi\right) = D\left(g_i \parallel f_{it}(\psi, \lambda_i(\psi))\right)$, where $\lambda_i(\psi)$ given in (8), and the minimization problem in (22) can be rewritten as

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

Therefore, the model with the smallest (22) corresponds to the model with the smallest estimate of the standard averaged KLIC, $\overline{D}\left(g \parallel f(\psi,\lambda)\right) = (nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}D\left(g_i \parallel f_{it}(\psi,\lambda_i)\right)$, over ψ and λ.

In practice, we cannot directly use (22) for model selection since it contains the infeasible components $\lambda_i(\psi)$. A natural candidate is then the averaged KLIC based on the profile likelihoods given by

$$\overline{D}\left(g \parallel f^P(\psi)\right) = \overline{D}(g \parallel f(\psi,\widehat{\lambda}(\psi))) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}D(g_i \parallel f_{it}(\psi,\widehat{\lambda}_i(\psi))), \qquad (23)$$

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, however, 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 (23) is not the same as the profile KLIC $D_P\left(g_i \parallel f_{it}(\psi,\lambda_i);\psi\right) = D(g_i \parallel f_{it}(\psi,\lambda_i(\psi)))$ in (21). The following lemma states the relation between these two KLIC's.

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

$$\frac{1}{T}\sum_{t=1}^{T}D_P\left(g_i \parallel f_{it}(\psi,\lambda_i);\psi\right) = \frac{1}{T}\sum_{t=1}^{T}D(g_i \parallel f_{it}^P(\psi)) + \delta(\psi;G_i), \qquad (24)$$

where the bias term is defined as $\delta(\psi;G_i) = \mathbb{E}_{G_i}\left[\ell_i^P(\psi) - \ell_i(\psi,\lambda_i(\psi))\right]$. Furthermore, if Assumptions 1 and 2 hold, $\delta(\psi;G_i)$ satisfies

$$\mathbb{E}_{G_i}\left[\delta(\psi;G_i) - \left(\frac{M_i(\psi)}{T}\right)\right] = O\left(\frac{1}{T^{3/2}}\right) \qquad (25)$$

under the regularity conditions, where $M_i(\psi)$ is the modification term used for the modified profile likelihood function (5).

From (24), 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 unlikely with small T. It 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 suitable estimator of $\delta(\psi;G_i)$, then we can conduct model selection based on the modified $D(g_i \parallel f_{it}^P(\psi))$. The result in (25) shows that the bias term in (24) is indeed closely related with the modification term $M_i(\psi)$.

Similar to (20) by letting

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

we define an information criterion using a bias-corrected estimator of $\Phi_P(\widehat{\psi}_M)$ given by

$$\widetilde{\Phi}_P(\widehat{\psi}_M) = -\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int \log f_{it}^P(z;\widehat{\psi}_M)d\widehat{G}_i\left(z\right) - \left\{B_P(\widehat{G}) - \frac{1}{nT}\sum_{i=1}^{n}M_i(\widehat{\psi}_M)\right\}. \qquad (26)$$

Here $\widehat{\psi}_M$ 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

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

obtained by replacing the unknown distribution G_i by the empirical distribution $\widehat{G}_i$. Note that (26) includes two bias correction terms: From Lemma 1, the additional correction term $(nT)^{-1}\sum_{i=1}^{n}M_i(\widehat{\psi}_M)$ is introduced because the feasible information criterion is defined using $D(g_i \parallel f_{it}^P(\widehat{\psi}_M))$ instead of $D_P(g_i \parallel f_{it}(\psi,\lambda_i);\widehat{\psi}_M)$. The following theorem derives an approximate expression for $B_P(G)$ based on which the information criterion is to be developed. 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 regularity conditions (given for example in Hahn and Kuersteiner (2011)), we have*

$$B_P(G) = -\frac{1}{nT}tr\left\{I(G)^{-1}J(G)\right\} + o\left(\frac{1}{nT}\right),$$

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

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

Similarly as $M_i(\psi)$, for some truncation parameter $m\geq 0$ such that $m/T^{1/2}\to 0$ as $T\to\infty$

and a properly chosen lag kernel function K_j, a consistent estimator for $B_P(G)$ can be obtained as

$$B_P(\widehat{G}) = -\frac{1}{nT} tr\left\{I(\widehat{G})^{-1} J(\widehat{G})\right\}, \tag{27}$$

where

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

From equations (26) and (27), therefore, a general form of information criterion for model selection based on the bias-corrected profile likelihood (i.e., a *profile likelihood information criterion*; PLIC) may be defined as

$$\begin{aligned} PLIC\,(f) &= -\frac{2}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\log f_{it}^P(z_{i,t};\widehat{\psi}_M) - 2\left\{B_P(\widehat{G}) - \frac{1}{nT}\sum_{i=1}^{n} M_i(\widehat{\psi}_M)\right\} \qquad (28) \\ &= -\frac{2}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\log f_{it}(z_{i,t};\widehat{\psi}_M,\widehat{\lambda}_i(\widehat{\psi}_M)) \\ &\quad + \frac{2}{nT}tr\left\{I(\widehat{G})^{-1}J(\widehat{G})\right\} + \frac{2}{nT}\sum_{i=1}^{n}M_i(\widehat{\psi}_M), \end{aligned}$$

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

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

$$PLIC\,(f) = -\frac{2}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\log f_{it}(z_{i,t};\widehat{\psi}_M,\widehat{\lambda}_i(\widehat{\psi}_M)) + \frac{2r}{nT} + \frac{2}{nT}\sum_{i=1}^{n}M_i(\widehat{\psi}_M), \tag{29}$$

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 at the MLE. The additional penalty term $(2/nT)\sum_{i=1}^{n} M_i(\widehat{\psi}_M)$ is novel and is nonzero in the presence of incidental parameters. Since this additional penalty term is positive by construction, the new information criterion (28) or (29) has heavier penalty than the standard Akaike information criterion (AIC). Since $(2/nT)\sum_{i=1}^{n} M_i(\widehat{\psi}_M) = O_p(T^{-1})$, the second penalty term can dominate the first one by a big margin when n is quite large. Recall that in the standard AIC, this additional penalty term does not appear and the penalty term of the information criterion is simply given by $2r/nT$ via a standardized parameter count.

Remark 1 *$PLIC(f)$ in (29) 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 this feature applies when we define the KLIC.*

4 Integrated Likelihood and Bayesian Approach

Instead of a KLIC-based model selection criteria using the (modified) profile likelihood, we next consider a 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 lies in their penalty terms, where the penalty terms are of the same form as 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 in 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

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

for each i, where $f_i(\psi,\lambda_i) = \prod_{t=1}^{T} f_{it}(z_{i,t};\psi,\lambda_i) = \exp(T\ell_i(\psi,\lambda_i))$ is the joint density of $z_i = (z_{i,1},\cdots,z_{i,T})'$. 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 $\mathcal{M}^k$. Further, let $g(z) = \prod_{i=1}^{n} g_i(z_i)$ be the joint density of $(z_1,\cdots,z_n)$ and

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

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

$$\mathcal{P}\left(\mathcal{M}^k|z\right) = \frac{1}{g\left(z\right)}\phi^k \int L^I(\psi^k|z)\eta(\psi^k|\mathcal{M}^k)d\psi^k, \tag{30}$$

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, the goal is to select the candidate model corresponding to the highest Bayesian posterior probability. This approach is approximately equivalent to model selection based on Bayes factors (e.g., Kass and Raftery (1995)).

Note 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

$$\begin{aligned}\ell_i^I(\psi^k) - \ell_i^M(\psi^k) \quad &= \quad \frac{1}{2T}\log\left(\frac{2\pi}{T}\right) - \frac{1}{2T}\log\left(-\frac{\partial^2 \ell_i(\psi^k, \widehat{\lambda}_i(\psi^k))}{\partial \lambda_i^2}\right) + \frac{1}{T}\log \pi_i(\widehat{\lambda}_i(\psi^k)|\psi^k) \\ &\quad + \frac{1}{T}M_i\left(\psi^k\right) + O_p\left(\frac{1}{T^2}\right) \end{aligned} \tag{31}$$

for each i. These expansions imply that if we choose the conditional prior $\pi_i(\lambda_i|\psi^k)$ such that it cancels out $O_p(T^{-1})$ leading terms in (31) at $\lambda_i = \widehat{\lambda}_i(\psi^k)$, then we have an improved approximation. More precisely, from (16) and (18), we obtain

$$\begin{aligned}\pi_i(\lambda_i|\psi^k) \quad &= \quad C_\pi \left(\mathbb{E}_{\widehat{G}_i}\left[-\frac{\partial^2 \ell_i(\psi^k, \lambda_i)}{\partial \lambda_i^2}\right]\right)^{1/2} \\ &\quad \times \exp\left\{-\frac{T}{2}\left(\mathbb{E}_{\widehat{G}_i}\left[-\frac{\partial^2 \ell_i(\psi^k, \lambda_i)}{\partial \lambda_i^2}\right]\right)^{-1}\left(\mathbb{E}_{\widehat{G}_i}\left[\frac{\partial \ell_i(\psi^k, \lambda_i)}{\partial \lambda_i}\right]\right)^2\right\}\end{aligned} \tag{32}$$

for some finite positive constant C_π, where $\mathbb{E}_{\widehat{G}_i}[\cdot]$ denotes the empirical expectation for each i. Note that the explicit form of the conditional prior in (32) corresponds to the robust (bias-reducing) prior in equation (14) of Arellano and Bonhomme (2009) in the case of a pseudo-likelihood. Arellano and Bonhomme (2009)'s robust prior is developed to obtain 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 discussion in Arellano and Bonhomme (2009) also applies to the conditional prior $\pi_i(\lambda_i|\psi^k)$ in (32): unlike the Jeffreys' prior, 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 (32), we obtain the approximate posterior probability of the model $\mathcal{M}^k$ in (30) as follows.

Theorem 4 *Let Assumptions 1 to 3 hold and $n/T \to \gamma \in (0,\infty)$ as $n,T \to \infty$. If we suppose conditional priors of λ_i as in (32) 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 the approximation*

$$\log \mathcal{P}\left(\mathcal{M}^k|z\right) = \sum_{i=1}^{n}\sum_{t=1}^{T} \log f_{it}^M(z_{i,t};\widehat{\psi}_M^k) - \frac{r_k}{2}\log nT + c(z,k) + o_p(1), \tag{33}$$

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 (33), 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 $-(2/nT)\log \mathcal{P}\left(\mathcal{M}^k|z\right)$ retaining only the relevant terms as follows:

$$ILIC\left(\mathcal{M}^k\right) = -\frac{2}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\log f_{it}(z_{i,t};\widehat{\psi}_M^k,\widehat{\lambda}_i(\widehat{\psi}_M^k)) + \frac{r_k \log nT}{nT} + \frac{2}{nT}\sum_{i=1}^{n} M_i(\widehat{\psi}_M^k). \tag{34}$$

Comparing with $PLIC$ in (29), the only difference in (34) is the second term (or the first penalty term), which corresponds to the standard penalty term in BIC. This result implies that we also need to modify 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 $PLIC$ obtained in the previous section. Therefore, in general, we can construct the following information criteria, which can be used in the presence of incidental parameters,

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

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 $h(nT)$ is some nondecreasing positive function of the sample size nT. The choice of $h(nT)$ is 2 for AIC-type criteria and $\log nT$ for BIC-type criteria. We conjecture that $h(nT) = 2\log\log nT$ for HQ-type criteria, although this formulation is not derived here. Note that the penalty term in LIC is no longer deterministic. It is data-dependent. So this model selection 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 homogeneous p_0'th order univariate autoregressive ($AR\,(p_0)$) model given by

$$y_{i,t} = \mu_i + \sum_{j=1}^{p_0} \alpha_{p_0 j} y_{i,t-j} + \varepsilon_{i,t} \quad \text{for } i = 1, 2, \cdots, n \text{ and } t = 1, 2, \cdots, T, \tag{36}$$

where p_0 is not necessarily finite.[6] The errors $\varepsilon_{i,t}$ are serially uncorrelated and the unobserved individual effects μ_i are assumed fixed. Let the initial values $(y_{i,0}, y_{i,-1}, \cdots, y_{i,-p_0+1})$ be observed for all i and assume the following conditions.

Assumption A *(i)* $\varepsilon_{i,t}|(\{y_{i,s}\}_{s\leq t-1}, \mu_i) \sim i.i.d.\mathcal{N}\left(0, \sigma^2\right)$ *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, the error term is serially uncorrelated, and there is no cross sectional dependence in $\varepsilon_{i,t}$. Normality is assumed for analytic convenience, which is common in the 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 (with Gaussian distributions) that best approximates the $AR(p_0)$ model (36). To develop a lag order selection criterion, we first obtain the maximum modified profile likelihood estimators in a Gaussian panel $AR(p)$ regression, $\widetilde{\alpha}(p) = (\widetilde{\alpha}_{p1}, \cdots, \widetilde{\alpha}_{pp})$ and $\widetilde{\sigma}^2(p)$, using the truncated sample $(y_{i,\overline{p}+1}, \cdots, y_{i,T})$ for each i, where $\bar{p} \geq p_0$ is the maximum AR lag considered. We define $y_{i,t}^W = y_{i,t} - \overline{T}^{-1} \sum_{s=\overline{p}+1}^{T} y_{i,s}$ as the within-transformed observation and $X_{i,t}^W(p) = (y_{i,t-1}^W, \cdots, y_{i,t-j}^W)'$, where $\overline{T} = T - \overline{p}$ is the number of truncated time series observations. Note that within-transformation corresponds to maximizing out the fixed effects μ_i's in MLE (i.e., forming the

[6] 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 time series sample T is longer and we allow $T \to \infty$, we can consider an approximate $AR\,(p_T)$ model with $p_T \to \infty$ as $T \to \infty$ with further rate conditions (e.g., $p_T^3/T \to 0$). Apparently, when we allow for an underlying $AR(\infty)$ process, the lag order selection problem becomes one of choosing the best $AR(p)$ model to approximate the $AR(\infty)$ process.

profile likelihood). Using the expression of $M_i(\cdot)$ in (18), it can be derived that

$$\widetilde{\alpha}(p) = \left[\sum_{i=1}^{n} \sum_{t=\overline{p}+1}^{T} X_{i,t}^{W}(p) X_{i,t}^{W}(p)' + \sum_{i=1}^{n} \sum_{j=-m}^{m} \frac{K_j}{\overline{T}} \sum_{t=\max\{\overline{p}+1, \overline{p}+j+1\}}^{\min\{T, T+j\}} X_{i,t}^{W}(p) X_{i,t-j}^{W}(p)' \right]^{-1} \times \left[\sum_{i=1}^{n} \sum_{t=\overline{p}+1}^{T} X_{i,t}^{W}(p) y_{i,t}^{W} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=-m}^{m} \frac{K_j}{\overline{T}} \sum_{t=\max\{\overline{p}+1, \overline{p}+j+1\}}^{\min\{T, T+j\}} \left\{ X_{i,t}^{W}(p) y_{i,t-j}^{W} + X_{i,t-j}^{W}(p) y_{i,t}^{W} \right\} \right] \tag{37}$$

and

$$\widetilde{\sigma}^2(p) = \frac{1}{n\overline{T}} \sum_{i=1}^{n} \sum_{t=\overline{p}+1}^{T} \left(\widetilde{\varepsilon}_{i,t}^{W}(p) \right)^2 + \frac{1}{n\overline{T}} \sum_{i=1}^{n} \sum_{j=-m}^{m} \frac{K_j}{\overline{T}} \sum_{t=\max\{\overline{p}+1, \overline{p}+j+1\}}^{\min\{T, T+j\}} \widetilde{\varepsilon}_{i,t}^{W}(p) \widetilde{\varepsilon}_{i,t-j}^{W}(p), \tag{38}$$

where $\widetilde{\varepsilon}_{i,t}^{W}(p) = y_{i,t}^{W} - \sum_{j=1}^{p} \widetilde{\alpha}_{pj} y_{i,t-j}^{W}$. As discussed in Section 2.3, $\widetilde{\alpha}(p)$ in (37) corresponds to the bias-corrected within-group estimator and other bias-corrected estimators can be used instead. The bias corrected variance estimator $\widetilde{\sigma}^2(p)$ in (38) is novel in the literature; instead of $\widetilde{\sigma}^2(p)$, it is normally used that

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

where the difference between $\widetilde{\sigma}^2(p)$ and $\widehat{\sigma}^2(p)$ is of $O_p(\overline{T}^{-1})$. If we denote

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

then

$$\widetilde{\sigma}^2(p) = \widehat{\sigma}^2(p) \left\{ 1 + \frac{R(p)}{\overline{T}} \right\}$$

and

$$-\frac{2}{n\overline{T}} \sum_{i=1}^{n} \sum_{t=\overline{p}+1}^{T} \log f_{it}^{P}(\widetilde{\alpha}(p), \widetilde{\sigma}^2(p)) = \log \widetilde{\sigma}^2(p) + \frac{\widehat{\sigma}^2(p)}{\widetilde{\sigma}^2(p)}.$$

In this case, therefore, from (35), a new lag order selection criterion can be obtained as

$$\begin{aligned} LIC\,(p) &= -\frac{2}{n\overline{T}}\sum_{i=1}^{n}\sum_{t=\overline{p}+1}^{T}\log f_{it}^{P}(\widetilde{\alpha}(p),\widetilde{\sigma}^{2}(p)) + \frac{2}{n\overline{T}}\sum_{i=1}^{n}M_i(\widetilde{\alpha}(p),\widetilde{\sigma}^{2}(p)) + \frac{h\left(n\overline{T}\right)}{n\overline{T}}p \\ &= \left\{\log\widetilde{\sigma}^{2}(p) + \frac{\widehat{\sigma}^{2}(p)}{\widetilde{\sigma}^{2}(p)}\right\} + \left(\frac{\widehat{\sigma}^{2}(p)}{\widetilde{\sigma}^{2}(p)}\right)\frac{R(p)}{\overline{T}} + \frac{h\left(n\overline{T}\right)}{n\overline{T}}p \\ &= \log\left(\widehat{\sigma}^{2}(p)\left\{1+\frac{R(p)}{\overline{T}}\right\}\right) + \left(\frac{\widehat{\sigma}^{2}(p)}{\widehat{\sigma}^{2}(p)\left\{1+(R(p)/\overline{T}\right\}}\right)\left\{1+\frac{R(p)}{\overline{T}}\right\} + \frac{h\left(n\overline{T}\right)}{n\overline{T}}p. \end{aligned}$$

Using an expansion of $\log(1+(R(p)/\overline{T}))$, whose remainder term is expected to depend on p in general, and by retaining only the relevant terms above, we can define the new lag order selection criterion as

$$LIC\,(p) = \log\widehat{\sigma}^{2}(p) + \frac{p}{n\overline{T}}\left(h\left(n\overline{T}\right) + c\frac{n}{\overline{T}}\right) + \frac{1}{\overline{T}}R(p) \tag{40}$$

for some positive $h\,(\cdot)$ and positive constant c.

The first term in (40) indicates goodness-of-fit, which resembles the standard lag order selection case. As suggested in Han et al. (2012) we utilize a homogeneous time series sample in the construction of the residual variance estimates $\widehat{\sigma}^{2}(p)$ as (39). The adjustment to employ a homogeneous time series sample in the residual variance estimates $\widehat{\sigma}^{2}(p)$ is important in controlling the probability of lag order overestimation and applies even in cases where there are no fixed effects, as shown in Han et al. (2012).

This new lag order selection criterion (40) has the penalty term given by

$$\frac{p}{n\overline{T}}\left(h\left(n\overline{T}\right) + c\frac{n}{\overline{T}}\right) + \frac{1}{\overline{T}}R(p) = \frac{p}{n\overline{T}}h\left(n\overline{T}\right) + c\frac{p}{n\overline{T}}\left(\frac{n}{\overline{T}}\right) + \frac{1}{\overline{T}}R(p), \tag{41}$$

where $R(p)$ corresponds to the long-run autocorrelation estimator of $\widetilde{\varepsilon}_{i,t}^{W}(p)$. The first penalty term in (41), which is quite standard in the model selection criteria, controls for degrees of freedom of the parameter of interest and therefore favors parsimonious models. The second and third penalty terms reflect the presence of nuisance parameters whose dimension is large. They are positive and add a heavier penalty to the information criterion, which will control for the over-selection probability. They are at most $O_p(T^{-1})$ and their role becomes minor for large T, which is well expected since the incidental parameter problem is attenuated with large T. However, they can be quite important compared to the first penalty term particularly when $\overline{T}$ is small and n is large.

The last element in the penalty term (41) deserves more explanation. Intuitively this term tries to rule out erroneous serial correlation in the regression residuals. Since the within-transformation incurs serial correlation in the AR panel regression even when the original error $\varepsilon_{i,t}$ is serially uncorrelated, $R(p)$ measures the degree of such pseudo serial correlation induced by the transformation. The maximum modified profile likelihood estimators may not completely eliminate the within-group bias and thus the pseudo serial correlation still remains in the residual. Since serial correlation will generally be exacerbated if the lag order is not correctly chosen – particularly when it is under-selected – the additional penalty term controls for this aspect and automatically controls for the under-selection probability. At the same time, this last term is positive and adds a heavier penalty, which also functions to control for the over-selection probability.

Remark 2 (Model complexity) The new penalty term in $LIC(p)$ can be understood as an appropriate choice of the effective degrees of freedom (i.e., the model complexity). For example, when $h(n\overline{T}) = 2$, the entire penalty term can be rewritten as $(2/n\overline{T})\{p + (n/2)(cp/\overline{T} + R(p))\}$, which shows that the efficient number of parameters is not $p + n$ in this case; the effect from the incidental parameters λ_i is smaller than n, where the degree is determined by the size of $(cp/\overline{T} + R(p))/2$.

Hodges and Sargent (2001) also 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 the μ_i are considered as fixed effects (e.g., $\tau^2 = \infty$); or 3 if the μ_i are considered as random effects. It is proposed that model complexity can be measured by the degrees of freedom and so 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

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

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 fixed effect) estimator by ignoring between-group variations. The degrees of freedom ρ also reflects this 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 effects but the degrees of freedom in our case is different from n; it is instead given by $(n/2)(cp/\overline{T} + R(p))$, which measures the model complexity somewhat differently. In a more general setup including nonlinear models, model complexity is closely related to the Vapnik-Chervonenkis dimension (e.g., Cherkassky et al. (1999)).

5.2 Statistical properties

Under stationarity the probability limit of the long-run autocorrelation estimator $R(p)$ in (41) is bounded and the penalty times $n\overline{T}$ (i.e., $p\left(h\left(n\overline{T}\right)+c(n/\overline{T})\right)+nR(p))$ increases with the sample size. As noted in Shibata (1980) and Yang (2005), we therefore conjecture that the new lag order selection criterion is not asymptotically optimal (i.e., $\operatorname{plim}_{n,\overline{T}\to\infty}[LIC(p^*)/\inf_{p\geq 0} LIC(p)] \neq 1$, where p^* is the lag order estimator from $LIC\,(p)$, e.g., Li (1987)) if the true data generating model is $AR(\infty)$ with finite σ^2 even when $h\left(n\overline{T}\right)$ is fixed like $h\left(n\overline{T}\right)=2$. When the true lag order p_0 exists and is finite, however, the new order selection criterion (40) is consistent under a certain side condition, as shown in the following result. We define a lag order estimator p^* to be consistent (and so the corresponding selection criterion is consistent) if it satisfies $\liminf_{n,\overline{T}\to\infty}\mathbb{P}\left(p^*=p_0\right)=1$.[7]

Theorem 5 *Under Assumption A, if we let $n/\overline{T}\to\gamma\in(0,\infty)$ and $n/\overline{T}^3\to 0$ as $n,\overline{T}\to\infty$, then $LIC\,(p)$ is a consistent lag order selection criterion when $p_0\leq\overline{p}<\infty$, provided that $h\left(n\overline{T}\right)$ satisfies $h\left(n\overline{T}\right)/n\overline{T}\to 0$ and $h\left(n\overline{T}\right)\to\infty$ as $n\overline{T}\to\infty$.*

As discussed above, examples of $h\left(n\overline{T}\right)$ for consistent criteria are $\log\left(n\overline{T}\right)$ and $\omega\log\log\left(n\overline{T}\right)$ for some $\omega\geq 2$, where the first is a BIC type penalty term and the second is a HQ type penalty term. Performance of the new lag order selection criteria is studied in simulations reported in the following subsection.

Theorem 5 does not provide analytical evidence explaining why the new lag order selection criteria work better than standard criteria such as $LIC_0(p)=\log\widehat{\sigma}^2(p)+p(h\left(n\overline{T}\right)/n\overline{T})$. Note that this standard criteria $LIC_0(p)$ is based on the truncated sample as suggested by Han et al. (2013), so it is also expected to be consistent with a suitable choice of $h\left(n\overline{T}\right)\to\infty$. It can be conjectured that the under-selection probability vanishes exponentially fast for both cases (provided that $h\left(n\overline{T}\right)/n\overline{T}\to 0$ and $1/\overline{T}\to 0$) similarly as Guyon and Yao (1999), while their over-selection probabilities decrease at different rates depending on the magnitude of the penalty term. Therefore, the observed improvement in correct selection probability of the new lag order selection criterion comes from reduction in the over-selection probability. Intuitively, since the new criterion includes an additional positive penalty term, the lag order estimates cannot be larger than those obtained by conventional lag order selection criteria. The following corollary states that the over-selection probability is reduced asymptotically by modifying the penalty term as in the new lag order selection criterion given in (40) and (41).

[7] This definition is somewhat different from the usual defintion of consistency but is equivalent for integer valued random variables. The lag estimator p^* is strongly consistent if $\mathbb{P}\left(\lim_{n,\overline{T}\to\infty}p^*=p_0\right)=1$. 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 consistent; and other order selection criteria, such as the final prediction error (FPE) and AIC are not consistent for finite p_0.

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

5.3 Simulations

We study the finite sample performance of the lag order selection criteria developed in the previous subsection and compare it with conventional time series model selection methods. We first define the two most commonly used information criteria, which use the pooled information as $LIC_0(p)$ in Corollary 6:

$$AIC\left(p\right) = \log \widehat{\sigma}^2(p) + \frac{2}{n\overline{T}} p,$$
$$BIC\left(p\right) = \log \widehat{\sigma}^2(p) + \frac{\log\left(n\overline{T}\right)}{n\overline{T}} p,$$

where $\widehat{\sigma}^2(p)$ is defined as (39) using the truncated uniform time series sample following Han et al. (2013). Preliminary simulation results show that constructing penalty terms using the parameter count $p + n$ too heavily penalizes 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. For the new criteria, we consider the following forms suggested in (40):

$$LIC^{AIC}\left(p\right) = \log \widehat{\sigma}^2(p) + \frac{p}{n\overline{T}} \left(2 + \frac{n}{\overline{T}}\right) + \frac{1}{\overline{T}} R(p),$$
$$LIC^{BIC}\left(p\right) = \log \widehat{\sigma}^2(p) + \frac{p}{n\overline{T}} \left(\log\left(n\overline{T}\right) + \frac{n}{\overline{T}}\right) + \frac{1}{\overline{T}} R(p),$$

in which c is simply set to unity.

We generate $AR\left(3\right)$ dynamic panel processes of the form $y_{i,t} = \mu_i + \sum_{j=1}^{3} \alpha_{3j} y_{i,t-j} + \varepsilon_{i,t}$ for $i = 1, 2, \cdots, n$ and $t = 1, 2, \cdots, T$, where $\alpha_{3j} = 0.15$ for all $j = 1, 2, 3$. This design is analogous to the one used in the simulation study of Han et al. (2013). All the autoregressive coefficients have the same value so that the lagged terms are equally important. We consider 64 different cases by combining different sample sizes of $n = 100, 200, 300, \cdots, 800$ and $T = 25, 30, 35, \cdots, 60$. 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 the $\widetilde{\alpha}_{pj}$ and replicate 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 over lag orders from 1 to 7 (i.e., $\overline{p} = 7$). The simulation results are provided in Figures 1 to 3, which present the correct-selection, over-selection, and under-selection probabilities of each case, respectively.

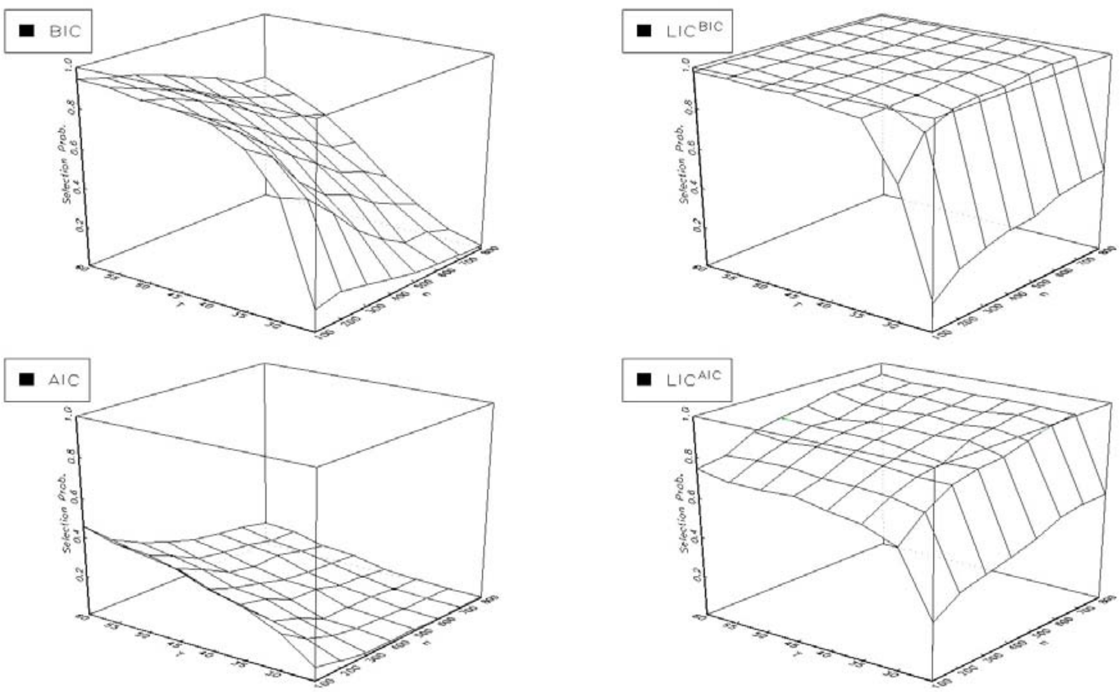

Figure 1: Correct order selection frequencies over 1000 iterations when $p_0 = 3$

Figure 1 shows clearly that the new lag order selection criteria LIC^{AIC} and LIC^{BIC} perform much better than the common criteria AIC and BIC. With the new criteria the correct-selection probability improves quite fast with T and does so uniformly over n. From Figures 2 and 3 it is evident that the improvement comes from the reduction in the over-selection probability. Since we impose a heavier penalty, however, the under-selection probability is high for very small T, which corresponds to the well-known property of BIC in a pure time series setup.

By comparison Figure 1 shows that the common criteria perform poorly with large n, and consistency seems to hold only with very large T and small n. From Figures 2 and 3, such poor performance is due to the high over-selection probability discussed in the previous subsection. Even BIC tends to overfit the order in dynamic panel models, where the over-selection probability increases quite fast with n. This finding is contrary to the well known property that BIC normally underfits lag order in a pure time series setup. In addition, since BIC is formulated here in the modified form developed by Han et al. (2013) with a uniformly truncated sample (to ensure consistency), it is apparent that this modified criterion seems to require large T to perform well when the dynamic panel model includes individual fixed effects.

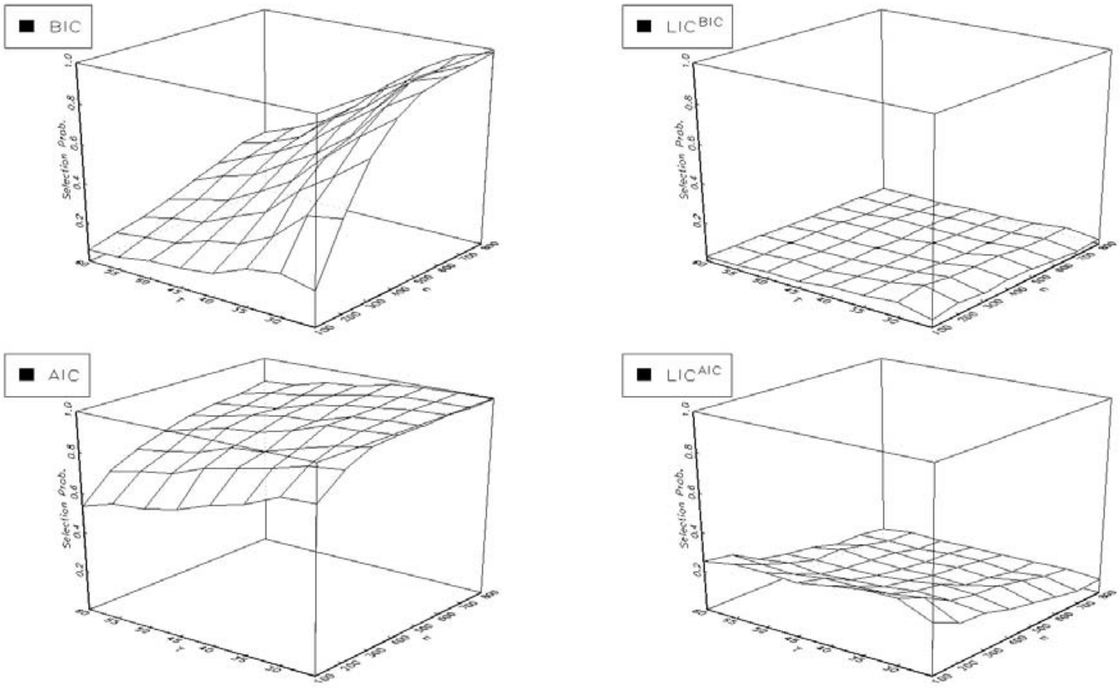

Figure 2: Over selection frequencies over 1000 iterations when $p_0 = 3$

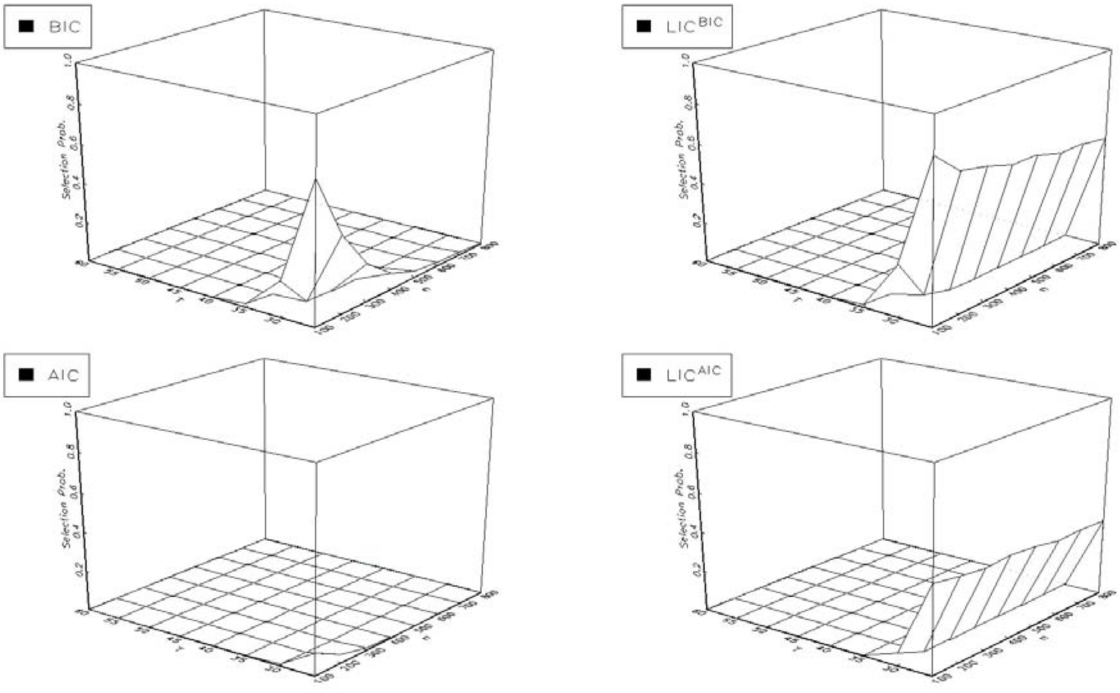

Figure 3: Under selection frequencies over 1000 iterations when $p_0 = 3$

6 Concluding Remarks

It is not uncommon in empirical work for a subset of parameters to be the central interest. In such cases, the nuisance parameters account for aspects of the model that are not of immediate concern but are nonetheless needed for realistic statistical modeling. Particularly when the dimension of the nuisance parameter space is large, dealing adequately with nuisance parameters is important for valid inference. As we demonstrate, model selection also needs to account for the presence of nuisance parameters to obtain correct model specification. The approach adopted in the present paper is to deal with nuisance parameters using either the profile likelihood (for AIC-type selectors) or integrated likelihood (for BIC-type selectors). The result is a new model selection criterion that can be used in the presence of nuisance parameters. The new penalty term in the selector is data-dependent and properly controls for model complexity.

Incidental parameters form a subset of parameters whose estimators typically have slower rates of convergence than those of the primary parameters under dual index asymptotics. We may therefore view the present paper as addressing a special case of a more general question: model selection involving a sub-set of parameters when the remaining parameters are estimable only at a slower rate of convergence than the primary parameters. Semiparametric models come within the same framework when we consider the nonparametric component as an infinite dimensional nuisance parameter. For example, using a similar approach to Severini and Wong (1992), consider a model with density $f(z_i, w_i; \psi, \lambda_i(w_i))$ for given observations $\{z_i, w_i\}$, where $\lambda_i(w) = (\lambda_{1i}, \lambda_2(w))'$ in which $\lambda_2(\cdot)$ is an unknown (scalar) function. In this case, we can regard $\lambda_{2i} = \lambda_2(w_i)$ as the realization of $\lambda_2(\cdot)$ at the ith observation. In the context of QML estimation, we conjecture that for $\widehat{\lambda}_{2,\psi}(\omega_i) = \arg\max_\lambda \sum_{t=1}^T \log f(z_{i,t}, w_{i,t}; \psi, \lambda_{1i}, \lambda)K((\omega_i - w_{i,t})/a)$, where $K(\cdot)$ and a are a kernel function and bandwidth, a similar result to Theorem 2 can be derived under suitable technical conditions. Note however that the conditions on the incidental parameters λ_{1i} and those on the nonparametric components λ_{2i} are different and their effects on the parametric component ψ need to be treated differently.[8]

For the particular problem of lag order selection in panel autoregression, Han et al. (2012) recently showed that the conventional BIC selector is inconsistent even in a panel model without fixed effects. The analysis in Han et al. (2012) reveals that dual index asymptotics typically induce order overestimation (with an asymptotic probability as high as 50%) in lag order selectors. The heuristic reason for the overestimation is that residual variance estimates in panel models with lag orders that exceed the true value will involve fewer innovations

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

than the residual variance estimate obtained from the true dynamic specification. Cross section averaging then produces $O(n)$ such differences (which after normalization contribute $O\left(\frac{1}{\sqrt{nT}}\right)$ rather than $O\left(\frac{1}{nT}\right)$ to the fit component of the selector) and these components end up dominating the standard BIC penalty, thereby blinding BIC to the overspecification. Modifications to BIC that are explored in Han et al. (2012) involve increasing the penalty, as we have done in the present paper to attenuate overspecification, and truncating the time series sample so that a common sample is used for the residual variance calculation. With these modifications, the BIC criterion is a consistent lag order selector in panel autoregression with fixed effects.

Appendix: Proofs

Proof of Lemma 1 The result follows immediately since

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} D_P\left(g_i \parallel f_{it}(\psi,\lambda_i);\psi\right) &= \int \log g_i(z)\,dG_i(z) - \int \frac{1}{T}\sum_{t=1}^{T}\log f_{it}(z;\psi,\lambda_i(\psi))dG_i(z) \\
&= \int \log g_i(z)\,dG_i(z) - \int \frac{1}{T}\sum_{t=1}^{T}\log f_{it}(z;\psi,\widehat{\lambda}_i(\psi))dG_i(z) \\
&\quad + \int \log\left(\frac{T^{-1}\sum_{t=1}^{T} f_{it}(z;\psi,\widehat{\lambda}_i(\psi))}{T^{-1}\sum_{t=1}^{T} f_{it}(z;\psi,\lambda_i(\psi))}\right) dG_i(z) \\
&= \frac{1}{T}\sum_{t=1}^{T} D\left(g_i \parallel f_{it}^{P}(\psi)\right) + \delta(\psi;G_i)
\end{aligned}$$

by stationarity. Furthermore, from (16) and (18), it can be seen that

$$\mathbb{E}_{G_i}\left[\delta(\psi;G_i) - \frac{M_i(\psi)}{T}\right] = \mathbb{E}_{G_i}\left[\ell_i^P(\psi) - \ell_i(\psi,\lambda_i(\psi)) - \frac{M_i(\psi)}{T}\right] = O\left(\frac{1}{T^{3/2}}\right)$$

for a given ψ. ■

Proof of Theorem 2 For each i, 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 also use notations G_i and $\widehat{G}_i$ instead of $G_i(\cdot)$ and $\widehat{G}_i(\cdot)$ when there is no risk of confusion. For a fixed ϵ, we let $\psi(\epsilon) = H(G(\cdot;\epsilon))$ be the solution of

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

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 $\int[\partial Q_{it}(z;\epsilon)/\partial\lambda_i]dG_i\,(z;\epsilon) = 0$ 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), (8) 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, for $n, T \to \infty$ satisfying $n/T \to \gamma \in (0,\infty)$ and $n/T^3 \to 0$, the Taylor series expansion of $\widehat{\psi}_M$ about ψ_0 can be obtained as (e.g., Chapter 6.2 in Serfling (1980), Konishi and Kitagawa (1996))

$$\begin{aligned} \widehat{\psi}_M - \psi_0 &= H(\widehat{G}) - H(G) \qquad \text{(A.2)} \\ &= d_1 H(G; \widehat{G} - G) + \frac{1}{2} d_2 H(G; \widehat{G} - G) + o_p\left(\frac{1}{nT}\right) \\ &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} H^{(1)}(z_{i,t}; G) + \frac{1}{2n^2T^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T} H^{(2)}(z_{i,t}, z_{j,s}; G) + o_p\left(\frac{1}{nT}\right), \end{aligned}$$

where $d_1 H(G; \widehat{G} - G) = \lim_{\epsilon\to 0+} \epsilon^{-1}\{H(G(\epsilon)) - H(G)\}$ is the standard first order Gâteaux differential of H at G in the direction of $\widehat{G}$ and $d_2 H(G; \widehat{G} - G) = d^2 H(G(\epsilon))/d\epsilon^2\big|_{\epsilon=0+}$ provided limit exists. $H^{(k)}$ are defined as $d^k H(G(\epsilon))/d\epsilon^k = \int\cdots\int H^{(k)}(z^1,\cdots,z^k;G)\prod_{a=1}^{k} d(\widehat{G}(z^a) - G(z^a))$ at $\epsilon = 0$ and $\int H^{(k)}(z^1,\cdots,z^k;G)dG(z^a) = 0$ for $1 \le a \le k$ and $k = 1, 2$.

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

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial^2}{\partial\psi\partial\psi'} Q_{it}(z;\epsilon) dG_i\,(z;\epsilon) \times d_1 H(G; \widehat{G} - G) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial^2}{\partial\psi\partial\lambda_i} Q_{it}(z;\epsilon) dG_i\,(z;\epsilon) \times \frac{\partial}{\partial\epsilon}\lambda_i(\psi(\epsilon); G_i(z;\epsilon)) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial}{\partial\psi} Q_{it}(z;\epsilon) d(\widehat{G}_i(z) - G_i(z)), \end{aligned}$$

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

$$\begin{aligned} d_1 H(G; \widehat{G} - G) &= \left(-\frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial^2 \log f_{it}\,(z;\psi_0,\lambda_{i0})}{\partial\psi\partial\psi'} dG_i(z)\right)^{-1} \qquad \text{(A.3)} \\ &\quad \times \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \log f_{it}\,(z;\psi_0,\lambda_{i0})}{\partial\psi} d\widehat{G}_i(z). \end{aligned}$$

Note that $\lambda_i(\psi_0) = \lambda_{i0}$ and thus $\int [\partial \log f_{it}(z; \psi_0, \lambda_{i0}) / \partial\psi] dG_i(z) = 0$ 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 explicit expression of $H^{(1)}(z_{i,t}; G)$ as (e.g., Withers (1983), Konishi and Kitagawa (1996))[9]

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

Since we only need an expression of $H^{(1)}(z_{i,t}; G)$ to derive the main result, we do not find $H^{(2)}$ in details as well as any terms associated with $H^{(2)}$ below.

Similar to 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 then have stochastic expansions as

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

for some $V_{it}(z; H^{(1)}, H^{(2)}, G) = O_p(1)$. Using $\mathbb{E}[\cdot]$ to signify expectation with respect to the joint distribution of $(G_1, \cdots, G_n)$, we then have

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

since $\int H^{(1)}(z_{j,s}; G) dG_j(z) = 0$ for all j, where $V = \mathbb{E}\left[(nT)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} V_{it}(z; H^{(2)}, G) \right] = O(1)$. Similarly,

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

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

and by stationarity over t

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

where the second term is nonzero only for the case $i = j$. It can be also verified that $\mathbb{E}\left[(nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}V_{it}(z;H^{(1)},H^{(2)},\widehat{G})\right] = \mathbb{E}\left[(nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}V_{it}(z;H^{(1)},H^{(2)},G)\right] = V$. Therefore,

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

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

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

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

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

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

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

$$
B_p(G) = -\frac{r}{nT} + O\left(\frac{1}{nT^{3/2}}\right) + o\left(\frac{1}{nT}\right) = -\frac{r}{nT} + o\left(\frac{1}{nT}\right),
$$

from which the information criterion (29) is obtained. ■

Proof of Theorem 4 By plugging the conditional prior (32) into the approximation (31), the log posterior probability of model $\mathcal{M}^k$ in (30) can be written as (we simply let $C_\pi = 1$)

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

But Taylor expansion yields

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

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 et al. (1989))

gives

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

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 (33) 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 $LIC(p^*) < LIC(p)$, where $0 \leq p^*, p \leq \overline{p}$ for some finite positive integer $\overline{p}$. We therefore need to prove that $\limsup_{n,\overline{T}\to\infty} \mathbb{P}[LIC(p^*) < LIC(p_0)] = 0$ for all $p^* \neq p_0$, where p_0 is the (finite) true lag order.

First consider the case of under-selection, $p^* < p_0$. We write

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

The left-hand-side of the inequality in (A.7) is positive in the limit as $n, \overline{T} \to \infty$ because $\widehat{\sigma}^2(p_0) = \sigma^2 + o_p(1)$ and $\widehat{\sigma}^2(p^*) = \sigma^2 + A + o_p(1)$ for some $A > 0$ (due to the under-specification) whenever $p^* < p_0$, as shown in Lemma 1 of Han et al. (2012). On the other hand, the right-hand-side of the inequality in (A.7) converges to zero as $n, \overline{T} \to \infty$ since $0 < (p_0 - p^*) \leq \overline{p} < \infty$, $|R(p_0) - R(p^*)| < \infty$ from the invertibility in Assumption A-(ii), and $h(n\overline{T})/n\overline{T} \to 0$ as $n\overline{T} \to \infty$ by assumption. Therefore, $\limsup_{n,\overline{T}\to\infty} \mathbb{P}[LIC(p^*) < LIC(p_0)] \leq \mathbb{P}[\limsup_{n,\overline{T}\to\infty}\{LIC(p^*) < LIC(p_0)\}] = \mathbb{P}[\varnothing] = 0$.

For the case of over-selection, $p^* > p_0$, we consider

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

As in the proof of Theorem 2 of Han et al. (2012) we have $n\overline{T}(\log\widehat{\sigma}^2(p^*) - \log\widehat{\sigma}^2(p_0)) = O_p(1)$. Further, as in Bhansali (1981) and Lee (2012), it can be verified that $|R(p_0) - R(p^*)| = a|p_0 - p^*|/\overline{T} + o_p(1/\overline{T})$ for some finite constant $a > 0$, which yields $n(R(p_0) - R(p^*)) = O_p(n/\overline{T})$. The left-hand-side of the inequality in the expression (A.8) is thus $O_p(1)$ for large n and T because it is assumed that $n/\overline{T} \to \gamma \in (0,\infty)$. On the other hand, the right-hand-side goes to negative infinity as $n\overline{T} \to \infty$ since $p_0 - p^* < 0$ and $h(n\overline{T}) \to \infty$. It follows that

$\limsup_{n,\overline{T}\to\infty} \mathbb{P}[LIC(p^*) < LIC(p_0)] = 0$ 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

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

and

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

Then, similar to the proof of Theorem 5, we write

$$\begin{aligned} &\mathbb{P}\left[\Delta LIC < \Delta LIC_0\right] \qquad\qquad (A.9) \\ =\;& \mathbb{P}\left[\log\left(\frac{\widehat{\sigma}^2(p^*)}{\widehat{\sigma}^2(p^{**})}\right) < \frac{h\left(n\overline{T}\right)}{n\overline{T}}(p^{**} - p^*) + \frac{c}{\overline{\overline{T}}^2}(p_0 - p^*) + \frac{1}{\overline{\overline{T}}}\left(R\left(p_0\right) - R\left(p^*\right)\right)\right]. \end{aligned}$$

Since $LIC\left(p\right)$ has the heavier penalty than $LIC_0\left(p\right)$, $p^{**} \geq p^*$ by construction and thus the left-hand-side of the last inequality in (A.9) is nonnegative for any n and $\overline{T}$, whereas the right-hand-side goes to zero with $n, \overline{T} \to \infty$ as in (A.7). Therefore, $\limsup_{n,\overline{T}\to\infty}\{\mathbb{P}\left[\Delta LIC < 0\right] - \mathbb{P}\left[\Delta LIC_0 < 0\right]\} \leq \limsup_{n,\overline{T}\to\infty} \mathbb{P}\left[\Delta LIC - \Delta LIC_0 < 0\right] \leq \mathbb{P}[\limsup_{n,\overline{T}\to\infty}\{\Delta LIC - \Delta LIC_0 < 0\}] = 0$, which implies $\limsup_{n,\overline{T}\to\infty} \mathbb{P}\left(p^{**} > p_0\right) \geq \limsup_{n,\overline{T}\to\infty} \mathbb{P}\left(p^* > p_0\right)$. ■

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.

Anderson, T. W. and C. Hsiao (1981). Estimation of Dynamic Models with error components, *Journal of the American Statistical Association*, 76, 598-606.

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.

Chakrabarti, 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.

Han, C., P.C.B. Phillips, and D. Sul (2012). Lag length selection in panel autoregression, unpublished manuscript.

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.

Hsiao, C. (2003). *Analysis of Panel Data*, 2nd edition, Cambridge University Press.

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 University.

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

Lee, Y. (2013). Nonparametric estimation of dynamic panel models with fixed effects, *Econometric Theory*, forthcoming.

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.

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 identification and regression estimation, *Biometrika*, 92, 937-950.