Likelihood Based Model Selection in the Presence of Incidental Parameters

Yoonseok Lee*
University of Michigan

July 2009

Abstract

This paper considers model selection problem in the presence of incidental parameters. The main interest is in selecting the structure of the model in the common parameters after concentrating out the incidental parameters. Using KLIC based on the profile likelihood, a new model selection information criterion is developed, which impose heavier penalties than those of the standard information criteria. As a particular example, a lag order selection criterion is examined in the context of dynamic panel models with fixed individual effects.

Keywords: Model selection, incidental parameters, profile likelihood, lag order, dynamic panel, fixed effects.
JEL Classifications: C23

1 Introduction

[To be added]

2 Incidental Parameters Problem in QMLE

We consider the panel data (or stratified) observations $z_{i,t}$ for $i = 1, 2, \cdots, n$ and $t = 1, 2, \cdots, T$. We suppose that $z_i = (z_{i,1}, \cdots, z_{i,T})'$ is independent across i with an unknown distribution $G_i(z)$ having probability density function $g_i(z)$. Since $g_i(z)$ is unknown, we instead consider a parametric family of densities $\{f_i(z; \theta_i) : \theta_i \in \Theta\}$ for each i, which does

Preliminary; comments are welcome. Valuable comments are obtained from seminar participants at Columbia, Rochester, Seoul National, Kyunghee and 2009 Korean Econometric Society Summer Meeting. The usual disclaimer applies. *Address*: Department of Economics, University of Michigan, 611 Tappan Street, 365C Lorch Hall, Ann Arbor, MI 48109-1220. *E-mail*: yoolee@umich.edu.

not necessarily contain $g_i(z)$, where we assume that the shape of density is common for all i and t:[1]

$$f_i(z_i;\theta_i) = f(z_i;\theta_i) \quad \text{and} \quad f(z_i;\theta_i) = \prod_{t=1}^{T} f(z_{i,t};\theta_i). \tag{1}$$

The parameter vector θ_i is decomposed as $\theta_i = \left(\psi', \lambda_i\right)'$. $\psi \in \Psi \subset \mathbb{R}^r$ is the main parameter of interest, which is common to all i, whereas $\lambda_i \in \Lambda \subset \mathbb{R}$ is the individual specific nuisance parameter. If we let $Z_{n,T} = (z_1', \cdots, z_n')'$ and $\underline{\lambda} = (\lambda_1, \cdots, \lambda_n)'$, then the joint likelihood can be written as

$$f(Z_{n,T};\psi,\underline{\lambda}) = \prod_{i=1}^{n} f(z_i;\psi,\lambda_i). \tag{2}$$

It is assumed that the joint likelihood (2) is separable with respect to nuisance parameters so that the nuisance parameter λ_i is only related to the i's observations. Panel models with heterogenous parameters, such as fixed individual effects, (conditional) heteroskedasticity, or heterogenous slope coefficients, are good examples of $f(\cdot;\psi,\lambda_i)$. More generally, semiparametric models, whose nonparametric component is time-invariant, could be also understood in the similar context if we see $\lambda_i = \lambda(w_i)$ as a realization of the function $\lambda(\cdot)$ for a random variable w_i. Since we are mainly interested in the parameter ψ, we can concentrate out the nuisance parameter λ_i to define the *profile likelihood* of ψ as

$$f_P(z_i;\psi) = f(z_i;\psi,\widehat{\lambda}_i(\psi)), \tag{3}$$

where

$$\widehat{\lambda}_i(\psi) = \arg\max_{\lambda_i} \log f(z_i;\psi,\lambda_i)$$

is the quasi maximum likelihood estimator (QMLE) of λ_i keeping ψ fixed. The quasi maximum profile likelihood estimator of ψ is then obtained as

$$\widehat{\psi} = \arg\max_{\psi} \sum_{i=1}^{n} \log f_P(z_i;\psi), \tag{4}$$

which is indeed the QMLE of ψ.[2] In general, however, $f_P(z_i;\psi)$ does not behave like the standard likelihood function due to the sampling variability of the estimator $\widehat{\lambda}_i(\psi)$, particularly when T is small as in the standard panel cases. 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. Intuitively, the profile likelihood is itself a biased estimate of the original likelihood. Modification of the profile likelihoods in the form of

$$\log f_M(z_i;\psi) = \log f_P(z_i;\psi) - M_i(\psi) \tag{5}$$

[1] When $z_{i,t}$ is serially correlated, $f(z_{i,t};\theta_i)$ should be understood as a conditional density on the lagged values $z_{i,t-s}$ for some $s \geq 1$.

[2] This is just taking the maximum in two steps instead of taking the maximum simultaneously.

is widely studied, which makes the *modified profile likelihood* $f_M(z;\psi)$ to behave like a proper likelihood (e.g., Barndorff-Nielsen (1983)). The modification term $M_i(\psi)$ corrects the sampling bias and it renders the expected score of the modified profile likelihood to be zero and thus an asymptotically unbiased estimator for ψ (more precisely, bias-reduced estimator for ψ) can be obtained by maximizing the modified profile likelihood (i.e., the quasi maximum modified profile likelihood estimation):

$$\widehat{\psi}_M = \arg\max_{\psi} \sum\nolimits_{i=1}^{n} \log f_M(z_i;\psi). \tag{6}$$

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

From the standard QMLE theory, we can show that $\widehat{\psi} = \psi_T + o_p(1)$ as $n \to \infty$ and T fixed under the regularity conditions (e.g., White (1982)), where

$$\psi_T = \arg\min_{\psi} KL\left(g(\cdot) \parallel f(\cdot;\psi,\widehat{\underline{\lambda}}(\psi))\right)$$

with $\widehat{\underline{\lambda}}(\psi) = (\widehat{\lambda}_1(\psi),\cdots,\widehat{\lambda}_n(\psi))'$ and $KL(g(\cdot) \parallel f(\cdot;\psi,\widehat{\underline{\lambda}}(\psi)))$ is the Kullback-Leibler divergence of $g(\cdot)$ relative to $f(\cdot;\psi,\widehat{\underline{\lambda}}(\psi))$ defined as

$$KL\left(g(\cdot) \parallel f(\cdot;\psi,\widehat{\underline{\lambda}}(\psi))\right) = \int g(z) \log\left(\frac{g(z)}{f\left(z;\psi,\widehat{\underline{\lambda}}(\psi)\right)}\right) dz, \tag{7}$$

in which $g = \prod_{i=1}^{n} g_i$ is the true joint density of $Z_{n,T}$. If we let

$$(\psi_0, \underline{\lambda}_0) = \arg\min_{\psi,\underline{\lambda}} KL\left(g \parallel f(\cdot;\psi,\underline{\lambda})\right), \tag{8}$$

where $(\psi_0, \underline{\lambda}_0)$ is assumed to be unique, however, ψ_T is normally different from ψ_0 particularly when the dimension of the nuisance parameter $\underline{\lambda}$ is substantial relative to the sample size (or when T is fixed and small in the panel data case), which is the incidental parameters problem (e.g., Neyman and Scott (1948)). More precisely, it can be shown that

$$\widehat{\psi} - \psi_0 = O_p\left(T^{-1}\right) \tag{9}$$

in general for smooth f under the standard conditions. The main source of the bias in (9)

is that $\widehat{\lambda}_i(\psi)$ is not the same as

$$\lambda_i(\psi) = \arg\min_{\lambda_i} KL\left(g_i(\cdot) \parallel f(\cdot;\psi,\lambda_i)\right) \tag{10}$$

and thus the estimation error of $\widehat{\lambda}_i(\psi)$ with finite T is not negligible even when $n \to \infty$, based on which the expectation of the profile score is normally not zero for each i.[3]

To show this, we first let

$$\begin{aligned}
\ell_i(\psi,\lambda_i) &= \log f(z_i;\psi,\lambda_i) = \sum\nolimits_{t=1}^{T} \log f(z_{i,t};\psi,\lambda_i), \\
\ell_{Pi}(\psi) &= \log f_P(z_i;\psi,\lambda_i) = \sum\nolimits_{t=1}^{T} \log f\left(z_{i,t};\psi,\widehat{\lambda}_i(\psi)\right), \\
\ell_{Mi}(\psi) &= \log f_M(z_i;\psi,\lambda_i) = \ell_{Pi}(\psi) - M_i(\psi),
\end{aligned}$$

and the pseudo-information matrix

$$\mathcal{I}_i = \mathbb{E}_{g_i}\left[-\left.\frac{\partial^2 \ell_i(\psi,\lambda_i)}{\partial\theta_i \partial\theta_i'}\right|_{\theta_i=\theta_{i0}}\right] = \begin{pmatrix} \mathcal{I}_{i,\psi\psi} & \mathcal{I}_{i,\psi\lambda_i} \\ \mathcal{I}_{i,\lambda_i\psi} & \mathcal{I}_{i,\lambda_i\lambda_i} \end{pmatrix} \tag{11}$$

conformable with $\theta_i = \left(\psi', \lambda_i\right)' \in \mathbb{R}^{r+1}$, where $\theta_{i0} = \left(\psi_0', \lambda_{i0}\right)'$ in (8). We also define scores $u_i(\psi,\lambda_i) = \partial\ell_i(\psi,\lambda_i)/\partial\psi$, $v_i(\psi,\lambda_i) = \partial\ell_i(\psi,\lambda_i)/\partial\lambda_i$, and $u_i^e(\psi,\lambda_i) = u_i(\psi,\lambda_i) - \mathcal{I}_{i,\psi\lambda_i}\mathcal{I}_{i,\lambda_i\lambda_i}^{-1} v_i(\psi,\lambda_i)$.[4] For notational convenience, we suppress the arguments when expressions are evaluated at $\theta_{0i} = \left(\psi_0', \lambda_{i0}\right)'$ for each i: for example, $u_i = u_i(\psi_0,\lambda_{i0})$, $v_i = v_i(\psi_0,\lambda_{i0})$ and $u_i^e = u_i^e(\psi_0,\lambda_{i0})$. We can derive the following expansion (e.g., Sartori (2003)):

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

with $u_i^e = O_p(T^{1/2})$ and $b_i(\psi_0) = O_p(1)$ for all i even for the QML estimators $\widehat{\lambda}_i(\psi)$. Though $\mathbb{E}_{g_i}[b_i(\psi_0)] = 0$ by construction, $\mathbb{E}_{g_i}[b_i(\psi_0)] \neq 0$. As a consequence the bias of the profile score accumulates and an asymptotic bias appears as (9). The modification term (or the penalizing term) $M_i(\psi)$ in (5) is normally found as a smooth function in ψ, provided that $f(\cdot;\theta_i)$ be three-times differentiable in θ_i, satisfying $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} \mathbb{E}_{g_i}[dM_i(\psi_0)/d\psi - b_i(\psi_0)] = 0$ so that the expected score of the modified profile likelihood does not have the

[3] Even when $T \to \infty$, if $n/T \to \gamma \in (0,\infty)$, $\widehat{\psi} - \psi_0 \neq o_p(1)$ as $n, T \to \infty$ (e.g., Sartori (2003), Hahn and Kuersteiner (2002)).

[4] $u_i^e(\psi_0,\lambda_{i0})$ is the *efficient score* for ψ at (ψ_0,λ_{i0}) and it can be understood as the orthogonal projection of the score function for ψ on the space spanned by the components of the nuisance score $v_i(\psi_0,\lambda_{i0})$ (e.g., Murphy and van der Vaart (2000)). Therefore, it is essentially the score function for ψ when the nuisance parameters are known. It also follows that $\mathbb{E}_{g_i}[\partial u_i^e(\psi_0,\lambda_{i0})/\partial\lambda_i] = 0$ 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 (2005)). Also note that the variance of $u_i^e(\psi_0,\lambda_{i0})$ is $\mathcal{I}_i^e = \mathcal{I}_{i,\psi\psi} - \mathcal{I}_{i,\psi\lambda_i}\mathcal{I}_{i,\lambda_i\lambda_i}^{-1}\mathcal{I}_{i,\lambda_i\psi} = \mathcal{I}_{i,\psi\psi|\lambda_i}$, which is called *efficient information* and it is nothing but the partial information of ψ.

first order asymptotic bias from $b_i(\psi_0)$'s.

In particular, note that since $\partial \ell_i(\psi, \lambda_i)/\partial \psi = u_i^e(\psi, \lambda_i) + \mathcal{I}_{i,\psi\lambda_i}\mathcal{I}_{i,\lambda_i\lambda_i}^{-1} v_i(\psi, \lambda_i)$ from the definition of $u_i^e(\psi, \lambda_i)$, we can write

$$\frac{\partial \ell_{Pi}(\psi_0)}{\partial \psi} = \frac{\partial \ell_i(\psi_0, \widehat{\lambda}_i(\psi_0))}{\partial \psi} = u_i^e(\psi_0, \widehat{\lambda}_i(\psi_0)) + \mathcal{I}_{i,\psi\lambda_i}\mathcal{I}_{i,\lambda_i\lambda_i}^{-1} v_i(\psi_0, \widehat{\lambda}_i(\psi_0)).$$

If we expand $\partial \ell_{Pi}(\psi_0)/\partial \psi$ around (ψ_0, λ_{i0}), we have

$$\begin{aligned}
\frac{\partial \ell_{Pi}(\psi_0)}{\partial \psi} &= u_i^e + \mathcal{I}_{i,\psi\lambda_i}\mathcal{I}_{i,\lambda_i\lambda_i}^{-1} v_i \\
&\quad + \left\{ \frac{\partial u_i^e}{\partial \lambda_i} + \mathcal{I}_{i,\psi\lambda_i}\mathcal{I}_{i,\lambda_i\lambda_i}^{-1} \frac{\partial v_i}{\partial \lambda_i} \right\} \left(\widehat{\lambda}_i(\psi_0) - \lambda_{i0} \right) \\
&\quad + \frac{1}{2} \left\{ \frac{\partial^2 u_i^e}{\partial \lambda_i^2} + \mathcal{I}_{i,\psi\lambda_i}\mathcal{I}_{i,\lambda_i\lambda_i}^{-1} \frac{\partial^2 v_i}{\partial \lambda_i^2} \right\} \left(\widehat{\lambda}_i(\psi_0) - \lambda_{i0} \right)^2 + \cdots,
\end{aligned}$$

where the remainder terms are $O_p\left(T^{-1/2}\right)$ under the usual regularity conditions in which the QML estimator is consistent (e.g., McCullagh and Tibshirani (1990)). However, since $\widehat{\lambda}_i(\psi) = \arg\max_{\lambda_i} \ell_i(\psi, \lambda_i)$ satisfies $\partial \ell_i(\psi, \widehat{\lambda}_i(\psi))/\partial \lambda_i = 0$ for any given value of ψ, we can obtain $\widehat{\lambda}_i(\psi_0) - \lambda_{i0} \approx -v_i \left(\partial v_i/\partial \lambda_i\right)^{-1}$, which is from expanding the first order condition $\partial \ell_i(\psi_0, \widehat{\lambda}_i(\psi_0))/\partial \lambda_i = 0$ around (ψ_0, λ_{i0}). Therefore, we have[5]

$$\frac{\partial \ell_{Pi}(\psi_0)}{\partial \psi} \approx u_i^e - v_i \left(\frac{\partial v_i}{\partial \lambda_i} \right)^{-1} \left\{ \frac{\partial u_i^e}{\partial \lambda_i} - \frac{1}{2} v_i \left(\frac{\partial v_i}{\partial \lambda_i} \right)^{-1} \left(\frac{\partial^2 u_i^e}{\partial \lambda_i^2} + \mathcal{I}_{i,\psi\lambda_i}\mathcal{I}_{i,\lambda_i\lambda_i}^{-1} \frac{\partial^2 v_i}{\partial \lambda_i^2} \right) \right\}$$

with an approximation error $O_p(T^{-1/2})$. From this expression, it can be derived that

$$b_i(\psi_0) = -v_i \left(\frac{\partial v_i}{\partial \lambda_i} \right)^{-1} \left\{ \frac{\partial u_i^e}{\partial \lambda_i} - \frac{1}{2} v_i \left(\frac{\partial v_i}{\partial \lambda_i} \right)^{-1} \left(\frac{\partial^2 u_i^e}{\partial \lambda_i^2} + \mathcal{I}_{i,\psi\lambda_i}\mathcal{I}_{i,\lambda_i\lambda_i}^{-1} \frac{\partial^2 v_i}{\partial \lambda_i^2} \right) \right\}. \tag{13}$$

Furthermore, since subtracting $b_i(\psi_0)$ from $\partial \ell_{Pi}(\psi_0)/\partial \psi$ corrects the leading bias term in score, we can define $\widehat{\psi}_M$ in (6) as a bias corrected estimator given by

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

where $\widehat{\mathcal{I}}_i^e(\widehat{\psi}_M)$ and $\widehat{b}_i(\widehat{\psi}_M)$ are some estimates of $\mathcal{I}_i^e = \mathcal{I}_{i,\psi\psi} - \mathcal{I}_{i,\psi\lambda_i}\mathcal{I}_{i,\lambda_i\lambda_i}^{-1}\mathcal{I}_{i,\lambda_i\psi}$ and $b_i(\psi_0)$, respectively. See the following section for the explicit forms of them. For more discussions on the bias reduction, see Arellano and Hahn (2005) for the survey of bias corrections in nonlinear panel regression models. In a similar vein of our approach, Arellano and Hahn

[5] Similar expression can be also found in Hahn and Newey (2004) and Arellano and Hahn (2005).

(2006) and Bester and Hansen (2007) study the penalized likelihood approach to reduce the bias in nonlinear panel models with fixed effects, whose modification formula $M_i(\psi)$ is given as

$$\begin{aligned} M_i(\psi) &= \frac{1}{2} tr \left\{ \left(- \lim_{T\to\infty} \mathbb{E}_{g_i} \left(\frac{1}{T} \sum_{t=1}^{T} \frac{\partial v_{i,t}(\psi, \lambda_i(\psi))}{\partial \lambda_i} \right) \right)^{-1} \right. \\ &\quad \left. \times \sum_{\ell=-\infty}^{\infty} \lim_{T\to\infty} \mathbb{E}_{g_i} \left(\frac{1}{T} \sum_{t=\ell+1}^{T} v_{i,t}(\psi, \lambda_i(\psi)) v_{i,t-\ell}(\psi, \lambda_i(\psi)) \right) \right\}, \end{aligned}$$

which is using the infeasible value $\lambda_i(\psi)$ in (10), where $v_{i,t}(\psi, \lambda_i) = \partial \log f\left(z_{i,t}; \psi, \lambda_i\right) / \partial \lambda_i$. Sometimes it is more convenient to work with $M_i(\psi)$ instead of $b_i\left(\psi\right)$..

3 Profile Likelihood and Kullback-Leibler Divergence

In the previous section, we overview the incidental parameters problem in panel data models when the model has individual specific parameters (e.g., fixed effects) and thus the model has an increasing number of nuisance parameters. Normally the panel data studies focus on reducing the first order bias (9) from the incidental parameters problem, which basically presume that the models considered are correctly specified. However, as discussed in Lee (2006) and Lee (2009), if the model is not correctly specified, the efforts of reducing bias from the incidental parameters could even exacerbate the bias and thus the correct model specification is very important in this context particularly for dynamic and/or nonlinear panel models (e.g., choosing the lag order in $ARMA$ models or the functional structure in the nonlinear models); the correct model specification should precede any bias corrections or bias reductions. This paper, therefore, rather focuses on model specification. In particular, selecting a model $f\left(z|\psi, \underline{\lambda}\right)$, which is closest to the true one $g\left(z\right)$, is the main interest.

When the dimension of the entire parameter vector θ is small and finite, the standard model selection is based on comparing estimates of the Kullback-Leibler divergence

$$KL\left(g \parallel f(\cdot;\widehat{\theta})\right) = \int g\left(z\right) \log \left(\frac{g\left(z\right)}{f(z;\widehat{\theta})} \right) dz \tag{14}$$

among different specifications $f\left(\cdot;\theta\right)$, where $\widehat{\theta}$ is the QMLE (i.e., a consistent estimator of $\theta_0 = \arg\min_\theta KL\left(g \parallel f\left(\cdot;\theta\right)\right)$). Equivalently, we select the model $f\left(\cdot;\theta\right)$ minimizing the relative distance $\Phi(\widehat{\theta}) = -\mathbb{E}_g \log f(\cdot;\widehat{\theta}) = -\int g\left(z\right) \log f(z;\widehat{\theta}) dz$, which can be estimated by

$$\widehat{\Phi}(\widehat{\theta}) = -\int \log f(z;\widehat{\theta}) d\widehat{G}\left(z\right),$$

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

$$\widetilde{\Phi}(\widehat{\theta}) = -\int \log f(z;\widehat{\theta}) d\widehat{G}(z) + B(\widehat{G}) \tag{15}$$

as an information criterion for model selection, where $B(G) = \mathbb{E}_g[\widehat{\Phi}(\widehat{\theta}) - \Phi(\widehat{\theta})]$. See, for example, Akaike (1973) and Akaike (1974) for further details. Recall that Akaike (1973) showed that $B(G)$ is asymptotically the ratio of $\dim(\theta)$ to the sample size when $\widehat{\theta}$ is the QMLE and g is nested in f.

Now we consider the original case that $\theta = \left(\psi', \underline{\lambda}'\right)'$, where $\underline{\lambda}$ is the n-dimensional nuisance parameter. When the dimension of the parameter vector θ is substantial relative to the sample size, it is not straightforward to find a proper criterion similarly as (15). The main reason is that a consistent estimator for a large dimensional parameter is hard to obtain as we discussed in the previous section (e.g., $n \to \infty$ but T fixed). Even when both n and T tend to infinity, under some circumstances (e.g., $n, T \to \infty$ with $n/T \to \gamma \in (0, \infty)$), the standard QML yields an asymptotically biased estimator for the minimizer of $KL(g \parallel f(\cdot;\theta))$. (See, e.g., Sartori (2003).) Apparently, one possible solution is to reduce the dimension of the parameter set by concentrating out the nuisance parameters. Particularly when we assume that the parameter of main interest ψ, which is of small and finite dimension, determines the main framework of the model and the specification does not change over i, it is then natural to concentrate out the nuisance parameters λ_i's from the likelihood in the model selection problem. In other words, we consider λ_i's as pure nuisance parameters and we assume that the choice of a particular model does not depends on the realization of λ_i's. Of course, this is a similar idea of the profile likelihood approach when the main interest is in a subset of parameters ψ. A good example is the lag order selection problem in a homogenous autoregressive dynamic panel regression with fixed effects (i.e., $y_{i,t} = \mu_i + \sum_{j=1}^{k} \alpha_{kj} y_{i,t-j} + \varepsilon_{i,t}$; e.g., Lee (2009)).

More precisely, we consider the Kullback-Leibler divergence based on the profile likelihoods as a measure for the model selection problem:

$$KL_P(g(\cdot) \parallel f_P(\cdot;\psi)) = \int g(z) \log\left(\frac{g(z)}{f_P(z;\psi)}\right) dz, \tag{16}$$

which is indeed equivalent to (7). Note that since $f_P(z;\psi) = f(z;\psi,\widehat{\underline{\lambda}}(\psi))$, the profile likelihood is still the original likelihood $f(z;\psi,\underline{\lambda})$ including all the parameters but by letting the incidental parameters be known functions of ψ. Though $\widehat{\underline{\lambda}}(\psi)$ is different from $\underline{\lambda}(\psi)$ in

(10) so $f_P(z;\psi)$ cannot be the exact proxy for $f(z;\psi,\underline{\lambda})$, it is still reasonable to compare the profile likelihood with the true density, where the true density could be heterogenous across i.[6] Furthermore, it is easy to verify that $KL_P\left(g \parallel f_P(\cdot;\psi)\right) = \min_{\underline{\lambda}} KL\left(g \parallel f(\cdot;\psi,\underline{\lambda})\right)$ since $\underline{\lambda}$ appears only in $f(\cdot;\psi,\underline{\lambda})$ and $f(\cdot;\psi,\underline{\lambda})$ is separable with respect to nuisance parameters λ_i that is only related to the i's observations. Therefore, $KL_P\left(g \parallel f_P(\cdot;\psi)\right)$ can be also understood as the *profile Kullback-Leibler divergence*. So selecting a model minimizing (16) indeed corresponds to selecting a model minimizing the standard Kullback-Leibler divergence based on the original candidate likelihood $f(z;\psi,\underline{\lambda})$. See Remark 3.2 for further explanation.

Similarly as (15), we can define an information criterion as

$$\widetilde{\Phi}_P(\widehat{\psi}_M) = -\int \log f_P(z;\widehat{\psi}_M) d\widehat{G}\left(z\right) + B_P(\widehat{G}), \tag{17}$$

where $\widehat{\psi}_M$ is the quasi maximum modified profile likelihood estimator of ψ_0 defined as (6) and $B_P(\widehat{G})$ is an estimate of $B_P\left(G\right) = \mathbb{E}_g[\widehat{\Phi}_P(\widehat{\psi}_M) - \Phi_P(\widehat{\psi}_M)]$ obtained by replacing the unknown distribution G by the empirical distribution $\widehat{G}$. Note that $\widehat{\psi}_M$ has desired properties as in the standard information criteria. An explicit expression of $B_P(G)$ can be obtained using the following theorem, particularly when both n and T tend to infinity.

Theorem 3.1 *We let $\widehat{\psi}_M = H(\widehat{G})$ with $H(\cdot)$ being a suitably defined r-dimensional regular function, which is second order compact differentiable at G. It also satisfies that $H(F(\theta)) = \psi$ for all $\theta \in \Theta$, where $F\left(\theta\right) = F\left(\psi,\underline{\lambda}\right)$ is the distribution function of the specified joint density $f\left(\cdot;\psi,\underline{\lambda}\right)$. If $n,T \to \infty$ satisfying $n/T \to \gamma \in (0,\infty)$ and $n/T^3 \to 0$, under the regularity conditions,[7] we can derive*

$$\frac{1}{T}B_P(G) = \frac{1}{nT} tr\left\{I_M\left(G\right)^{-1} J_{MP}\left(G\right)\right\} + o\left(\frac{1}{nT}\right),$$

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

$$\begin{aligned} I_M\left(G\right) &= \int -\left.\frac{\partial^2 \ell_M\left(\psi\right)}{\partial\psi\partial\psi'}\right|_{\psi=H(G)} dG, \\ J_{MP}\left(G\right) &= \int \left.\frac{\partial \ell_M\left(\psi\right)}{\partial\psi}\right|_{\psi=H(G)} \left.\frac{\partial \ell_P\left(\psi\right)}{\partial\psi'}\right|_{\psi=H(G)} dG. \end{aligned}$$

[6] In comparison, one could consider the Kullback-Leibler divergence based on the marginal likelihood $f(\cdot;\psi)$, in which the incidental parameters were integrated out with assuming that its marginal distribution is known. However, once the heterogeneity in the original f is integrated out, it is meaningless to measure the distance between this marginal likelihood and the truth, since the true density still includes the heterogeneity in it. Unless we could marginalize the heterogeneity in the true density, therefore, this method is indeed infeasible.

[7] For example, see the technical conditions in Hahn and Kuersteiner (2004).

Furthermore, it can be approximated that[8] $I_M(G) \approx -\sum_{i=1}^n \mathbb{E}_{g_i}\left[\partial u_i^e/\partial\psi'\right]$ *and* $J_{MP}(G) \approx \sum_{i=1}^n \mathcal{I}_i^e + \sum_{i=1}^n \mathbb{E}_{g_i}\left[u_i^e b_i(\psi_0)'\right]$*, where* $u_i^e = u_i - \mathcal{I}_{i,\psi\lambda_i}\mathcal{I}_{i,\lambda_i\lambda_i}^{-1} v_i$ *is the efficient score of* ψ *at* (ψ_0, λ_{i0}) *and* $\mathcal{I}_i^e = \mathcal{I}_{i,\psi\psi} - \mathcal{I}_{i,\psi\lambda_i}\mathcal{I}_{i,\lambda_i\lambda_i}^{-1}\mathcal{I}_{i,\lambda_i\psi}$ *is its variance (i.e., efficient information) for each* i*. Particularly when* g *is included in the family of* f*, we have* $I_M(G) \approx \mathcal{I}_i^e$*.*

Since $\widehat{\psi}_M$ is $\sqrt{nT}$-consistent under the conditions in Theorem 3.1 (e.g., Sartori (2003), Hahn and Kuersteiner (2004)), the estimate for $B(G)/T$ is simply given by

$$\frac{1}{T}B_P(\widehat{G}) = \frac{1}{nT} tr\left\{I_M^{-1}(\widehat{G})^{-1} J_{MP}(\widehat{G})\right\}, \tag{18}$$

where

$$\begin{aligned} I_M(\widehat{G}) &= -T^{-1}\sum\nolimits_{i=1}^n \partial^2 \ell_{Mi}(\widehat{\psi}_M)/\partial\psi\partial\psi' \text{ and} \\ J_{MP}(\widehat{G}) &= T^{-1}\sum\nolimits_{i=1}^n (\partial\ell_{Mi}(\widehat{\psi}_M)/\partial\psi)(\partial\ell_{Pi}(\widehat{\psi}_M)/\partial\psi'). \end{aligned}$$

Note that for $\ell_{Mi}(\psi) = \sum_{t=1}^T \log f(z_{i,t};\psi,\widehat{\lambda}_i(\psi)) - M_i(\psi)$, we can further derive that

$$I_M(\widehat{G}) = -\frac{1}{T}\sum\nolimits_{i=1}^n \sum\nolimits_{t=1}^T \frac{\partial u_{i,t}^P(\psi,\widehat{\lambda}_i(\psi))}{\partial\psi'} + \frac{1}{T}\sum\nolimits_{i=1}^n \frac{M_i(\widehat{\psi}_M)}{\partial\psi\partial\psi'},$$

and

$$\begin{aligned} J_{MP}(\widehat{G}) &= \sum\nolimits_{i=1}^n \sum\nolimits_{\ell=-m}^m \frac{1}{T}\sum\nolimits_{t=\max(1,\ell+1)}^{\min(T,T+\ell)} u_{i,t}^P(\psi,\widehat{\lambda}_i(\psi)) u_{i,t-\ell}^P(\psi,\widehat{\lambda}_i(\psi))' \\ &\quad -\sum\nolimits_{i=1}^n \frac{M_i(\widehat{\psi}_M)}{\partial\psi}\frac{1}{T}\sum\nolimits_{t=1}^T u_{i,t}^P(\psi,\widehat{\lambda}_i(\psi))', \end{aligned} \tag{19}$$

where $u_{i,t}^P(\psi,\widehat{\lambda}_i(\psi)) = \partial \log f(z_{i,t};\psi,\widehat{\lambda}_i(\psi))/\partial\psi$ for some truncation parameter m. Notice that (19) allows for possible serial correlations in the profile score functions. Using (18) and (17), therefore, we can develop an information criterion for the model selection based on the bias corrected profile likelihood (i.e., *profile likelihood information criterion*; PLIC) as $(2/T)\widetilde{\Phi}_P(\widehat{\psi}_M)$:

$$PLIC(f) = -\frac{2}{nT}\sum_{i=1}^n\sum_{t=1}^T \log f\left(z_{i,t};\widehat{\psi}_M,\widehat{\lambda}_i\left(\widehat{\psi}_M\right)\right) + \frac{2}{nT} tr\left\{I_M(\widehat{G})^{-1} J_{MP}(\widehat{G})\right\}. \tag{20}$$

[8] It is important to note that the modified profile likelihood behaves like the standard likelihoods in that the score and the information matrix of $f_M(\cdot|\psi)$ is approximately the same as the efficient score and the efficient information, respectively. More precisely, (5) implies that $\mathbb{E}_g\left[\partial\ell_M(z_i|\psi_0)/\partial\psi\right] \approx \mathbb{E}_g\left[u_i^e\right] = 0$ and $\mathbb{E}_g\left[(\partial\ell_M(z_i|\psi_0)/\partial\psi)(\partial\ell_M(z_i|\psi_0)/\partial\psi')\right] \approx \mathbb{E}_g\left[u_i^e u_i^{e\prime}\right] = \mathcal{I}_i^e$.

From Theorem 3.1, $B_P(G)$ can be further approximated as

$$\begin{aligned}\frac{1}{T}B_P(G) &\approx \frac{1}{nT}tr\left\{\left(-\sum_{i=1}^{n}\mathbb{E}_{g_i}\left[\frac{\partial u_i^e}{\partial \psi'}\right]\right)^{-1}\sum_{i=1}^{n}\mathcal{I}_i^e\right\} \\ &\quad +\frac{1}{nT}tr\left\{\left(-\sum_{i=1}^{n}\mathbb{E}_{g_i}\left[\frac{\partial u_i^e}{\partial \psi'}\right]\right)^{-1}\sum_{i=1}^{n}\mathbb{E}_{g_i}\left[u_i^e b_i\left(\psi_0\right)'\right]\right\}. \end{aligned} \tag{21}$$

The first term in (21) is simply r/nT under the condition that g is nested in the family of f since $-\mathbb{E}_g\left[\partial u_i^e/\partial\psi'\right]=\mathcal{I}_i^e$. In this particular case, the penalty term $B_P(\widehat{G})/T$ is reduced to

$$\frac{1}{T}B_P(\widehat{G})=\frac{r}{nT}+\frac{1}{nT}tr\left\{\widehat{R}(\widehat{\psi}_M)\right\}, \tag{22}$$

where $\widehat{R}(\widehat{\psi}_M)$ is some consistent estimate of

$$R\left(\psi_0\right)=\left(\sum_{i=1}^{n}\mathcal{I}_i^e\right)^{-1}\sum_{i=1}^{n}\mathbb{E}_{g_i}\left[u_i^e b_i\left(\psi_0\right)'\right]. \tag{23}$$

Therefore, the penalty term of the information criterion $PLIC\left(f\right)$ in (20) can be written as

$$PLIC\left(f\right)=\frac{2r}{nT}+\frac{2}{nT}tr\{\widehat{R}(\widehat{\psi}_M)\}. \tag{24}$$

Note that the additional term $(2/nT)\,tr\{\widehat{R}(\widehat{\psi}_M)\}$ of the penalty function is novel and it is not zero in the presence of incidental parameters. If $tr\{\widehat{R}(\widehat{\psi}_M)\}>0$, then the new information criterion (20) (with the penalty term written as (22)) has heavier penalty than the standard Akaike information criterion (AIC). Recall that in the standard AIC, the second term in (22) does not appear and the penalty term of the information criterion is simply given by $2r/nT$. Finally note that, using the maximum modified profile likelihood estimator $\widehat{\theta}_{Mi}=(\widehat{\psi}'_M,\widehat{\lambda}_i(\widehat{\psi}_M))'$ or using some bias corrected estimators, we can estimate $R\left(\psi_0\right)$ as

$$\widehat{R}(\widehat{\psi}_M)=\left(-\frac{1}{T}\sum_{i=1}^{n}\left\{\frac{\partial u_i(\widehat{\theta}_{Mi})}{\partial\psi'}-\frac{\partial u_i(\widehat{\theta}_{Mi})}{\partial\lambda_i}\left[\frac{\partial v_i(\widehat{\theta}_{Mi})}{\partial\lambda_i}\right]^{-1}\frac{\partial v_i(\widehat{\theta}_{Mi})}{\partial\psi'}\right\}\right)^{-1}\left(\frac{1}{T}\sum_{i=1}^{n}\widehat{u}_i^e\widehat{b}_i'\right),$$

where

$$\widehat{u}_i^e=u_i(\widehat{\theta}_{Mi})-\frac{1}{T}\frac{\partial u_i(\widehat{\theta}_{Mi})}{\partial\lambda_i}\left[\frac{1}{T}\frac{\partial v_i(\widehat{\theta}_{Mi})}{\partial\lambda_i}\right]^{-1}v_i(\widehat{\theta}_{Mi})$$

and

$$\widehat{b}_i = -v_i(\widehat{\theta}_{Mi}) \left[\frac{1}{T} \frac{\partial v_i(\widehat{\theta}_{Mi})}{\partial \lambda_i} \right]^{-1} \left\{ \frac{\partial u_i^e(\widehat{\theta}_{Mi})}{\partial \lambda_i} - \frac{1}{2} v_i(\widehat{\theta}_{Mi}) \left[\frac{1}{T} \frac{\partial v_i(\widehat{\theta}_{Mi})}{\partial \lambda_i} \right]^{-1} \right.$$
$$\left. \times \left(\frac{1}{T} \frac{\partial^2 u_i^e(\widehat{\theta}_{Mi})}{\partial \lambda_i^2} + \frac{1}{T} \frac{\partial u_i(\widehat{\theta}_{Mi})}{\partial \lambda_i} \left[\frac{1}{T} \frac{\partial v_i(\widehat{\theta}_{Mi})}{\partial \lambda_i} \right]^{-1} \frac{\partial^2 v_i(\widehat{\theta}_{Mi})}{\partial \lambda_i^2} \right) \right\}$$

from (13).

Remark 3.2 The information criteria (17) or (20) are based on the profile likelihood, which reduces the high dimensional model to a finite dimensional, random submodel of the same dimension as ψ. The family of distributions with parameter $(\psi, \widehat{\underline{\lambda}}(\psi))$ is known as Stein's least favorable family (e.g., Stein (1956)). More precisely, the submodel defined as the profile likelihood with reduced parameters $((\psi, \underline{\lambda}) \to \psi)$ has the smallest information about ψ out of all possible submodels. Such loss of information can be expressed as the difference between $\mathcal{I}_{i,\psi\psi}$ and $\mathcal{I}_i^e = \mathcal{I}_{i,\psi\psi} - \mathcal{I}_{i,\psi\lambda_i} \mathcal{I}_{i,\lambda_i\lambda_i}^{-1} \mathcal{I}_{i,\lambda_i\psi}$, where $\mathcal{I}_i^e$ can be understood as the partial information of ψ with λ_i unknown whereas $\mathcal{I}_{i,\psi\psi}$ as the partial information of ψ with λ_i known. Therefore, it could be understood that the selected model by minimizing $KL_P\left(g(\cdot) \parallel f_P(\cdot;\psi)\right)$ is the least favorable at $(\psi_0, \underline{\lambda}_0)$ (i.e., its information is minimal among the consistent model group; see Murphy and van der Vaart (2000), Severini (2000)). Based on such interpretation, we may expect that the new information criterion also tends to select over-parametrized models similarly as the standard AIC (i.e., choosing less efficient model).

Remark 3.3 As noted in the previous section, the semiparametric models could be handled in a similar context if we consider the nonparametric component as infinite dimensional parameter models (i.e., incidental parameters). In particular, using the results by Severini and Wong (1992), for example, we can consider a model $f(z_i;\psi,\lambda(w_i))$ for given observations (z_i, w_i), where $\lambda(\cdot)$ is an unknown function. Apparently, we could see that $\lambda_i = \lambda(w_i)$ as the realization of $\lambda(\cdot)$ for each observation. Though it should be proved in the context of QML estimation, we could conjecture that for

$$\widehat{\lambda}_\psi(w) = \arg\max_\lambda \sum\nolimits_{i=1}^n \log f(z_i;\psi,\lambda(w_i)) K\left(\frac{w - w_i}{h}\right),$$

where K and h are properly defined kernel function and the bandwidth parameter, respectively, we could derive a similar result as Theorem 3.1 under proper technical conditions. Notice that, for the partially linear regression case, the QMLE for ψ is nothing but the

estimator considered by Robinson (1988). [To be added further]

4 Lag Order Selection in Dynamic Panel Models

We now consider a specific example of the new model selection criterion (20) in the context of dynamic panel regression. In particular, we consider a panel process $\{y_{i,t}\}$ generated from the homogenous pth-order univariate autoregressive ($AR(k)$) model given by

$$y_{i,t} = \mu_i + \sum_{j=1}^{k} \alpha_{kj} y_{i,t-j} + \varepsilon_{i,t} \quad \text{for } i = 1, 2, \cdots, N \text{ and } t = 1, 2, \cdots, T, \tag{25}$$

where k is not necessarily finite.[9] $\varepsilon_{i,t}$ is serially uncorrelated and unobserved individual effects μ_i are assumed fixed. For notational convenience we let the initial values $(y_{i,0}, y_{i,-1}, \cdots, y_{i,-k+1})$ be observed for all i. We first 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* k*,* $\sum_{j=1}^{k} |\alpha_{kj}| < \infty$ *and all roots of the characteristic equation* $1 - \sum_{j=1}^{k} \alpha_{kj} z^j = 0$ *lie outside the unit circle.*

In Assumption A-(i), we assume that the higher order lags of $y_{i,t}$ capture all the persistence and the error term does not have any serial correlation. We also exclude cross sectional dependence in $\varepsilon_{i,t}$. Note that we assume the normality for analytical convenience, which is somewhat standard in model selection literature. Assumption A-(ii) is standard condition for stationary autoregressive processes. We do not impose any initial conditions on $(y_{i,0}, \cdots, y_{i,-k+1})$ and the initial values remain unrestricted.

We let $y_i = (y_{i,1}, \cdots, y_{i,T})'$, $\varepsilon_i = (\varepsilon_{i,1}, \cdots, \varepsilon_{i,T})'$, $x_i = (x_{i,1}, \cdots, x_{i,T})'$ with $x_{i,t} = (y_{i,t-1}, \cdots, y_{i,t-k})'$, $\alpha_k = (\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{kk})'$, and ι be the $T \times 1$ vector of ones. Then we can rewrite (25) in a vector form as

$$y_i = \mu_i \iota + x_i \alpha_k + \varepsilon_i,$$

whose log-likelihood conditional on μ_i and $x_{i,1} = (y_{i,0}, \cdots, y_{i,-k+1})'$ is given by

$$\ell\left(y_i | x_{i,1}, \mu_i, \alpha_k, \sigma^2\right) = -\frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \left(y_i - \mu_i \iota - x_i \alpha_k\right)' \left(y_i - \mu_i \iota - x_i \alpha_k\right) \tag{26}$$

with ignoring constant terms. Apparently, the (Gaussian) maximum likelihood estimation

[9] When we are particularly interested in relatively short panels, it is reasonable to assume the true lag order k to be finite. When the length of time series T is assumed to grow, however, we can consider an approximate $AR(k_T)$ model with $k_T \to \infty$ as $T \to \infty$ but with further conditions (e.g., $p_T^3/T \to 0$).

of $\sum_{i=1}^{N} \ell\left(y_i|x_{i,1},\mu_i,\alpha_k,\sigma^2\right)$ yields the within group (WG) estimator $\widehat{\alpha}_k$ defined as

$$\widehat{\alpha}_k = \left(\sum\nolimits_{i=1}^{N} x_i^{0\prime} x_i^0\right)^{-1} \sum\nolimits_{i=1}^{N} x_i^{0\prime} y_i^0, \tag{27}$$

where all the elements in x_i^0 and y_i^0 are within-transformed: $y_{i,t}^0 = y_{i,t} - (1/T)\sum_{s=1}^{T} y_{i,s}$ for all i and t. It is easy to verify that the within transformation is concentrating out the nuisance parameters μ_i based on its MLE with other parameters fixed. Therefore, the WG estimator $\widehat{\alpha}_k$ in (27) is nothing but the maximum profile likelihood estimator. More precisely,

$$\widehat{\mu}_i\left(\alpha_k,\sigma^2\right) = \arg\max_{\mu_i} \ell\left(y_i|x_{i,1},\mu_i,\alpha_k,\sigma^2\right) = \frac{1}{T}\left(y_i - x_i\alpha_k\right)'\iota$$

and the profile likelihood is given by (also ignoring constant terms)

$$\ell_P\left(y_i|x_{i,1},\mu_i,\alpha_k,\sigma^2\right) = -\frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(y_i^0 - x_i^0\alpha_k\right)'\left(y_i^0 - x_i^0\alpha_k\right).$$

When N is large but T is small, it is well known that the WG estimator $\widehat{\alpha}_k$ is inconsistent (e.g., Nickell (1981)), which is a particular example of the incidental parameters problem or the problem of the profile likelihood as discussed in the previous sections. See Hahn and Kuersteiner (2002), Alvarez and Arellano (2003) and Arellano and Hahn (2005) for further discussions.

When we are interested in the correctly specified model and its reliable estimates for policy analysis, choosing the true lag order k could be crucial. Particularly for the WG estimation case, which requires bias correction in general, correct lag order selection is more important since any bias correction methods presumes correct model specification. As shown in Lee (2009), any attempts to correct the bias in the WG estimators of dynamic panel models under incorrect model specification could exacerbate the bias.[10] In the standard time series context, lag order selection is normally conducted by optimizing a penalized goodness-of-fit criterion. Commonly used order selection criteria include the final prediction error (FPE) criterion, Akaike information criterion (AIC), Bayesian information criterion (BIC), posterior information criterion (PIC), Hannan-Quinn criterion (HQ), Mallows' C_p and Rissanen's minimum description length. However, as we will see in the simulation results in the following section, such conventional lag order selection criteria do not work properly for panel models in the presence of fixed effects.

[10] Alternatively, one can consider the GMM estimation using the first-differenced regression model, which is known to be consistent even for small T. However, this approach also requires correct specification particularly on the serial correlation structure of the error term, which can be affected by the dynamic misspecification of the level $y_{i,t}$. (Recall that in the standard dynamic panel models, the valid instrumental variables are constructed based on the white-noise error term.) Therefore, such an approach could not be an alternative answer (or an robust approach) to the dynamic specification.

Instead of using the general form of information criterion $PLIC$ in (20), we want to develop an automatic lag order selection criterion particularly tailored for the dynamic panel case. To this end, we first need to obtain an explicit expression of $tr\left\{R\left(\alpha_k, \sigma^2\right)\right\}$ in (23) in this particular case of the nested models with Gaussian distribution so that we can use the simplified information criterion in (24). For notational convenience, we define A as a $k \times k$ matrix given by

$$A = \begin{pmatrix} \alpha_k' \\ I_{k-1},\ 0 \end{pmatrix},$$

where I_{k-1} is the identity matrix of rank $k-1$.

Lemma 4.1 *We let Assumption A hold. For fixed T,*

$$tr\left\{R\left(\alpha_k, \sigma^2\right)\right\} = k + 2tr\left\{A\left(I_k - A\right)^{-1}\right\} + 1, \tag{28}$$

where

$$0 < \frac{tr\left(I_k + A'A\right)}{\Lambda_{\max}\left(\left(I_k - A\right)'\left(I_k - A\right)\right)} \leq k + 2tr\left\{A\left(I_k - A\right)^{-1}\right\} \leq \frac{tr\left(I_k + A'A\right)}{\Lambda_{\min}\left(\left(I_k - A\right)'\left(I_k - A\right)\right)}$$

for all $k \geq 1$. $\Lambda_{\max}(Q)$ and $\Lambda_{\min}(Q)$ denote the minimum and maximum eigenvalue of a square matrix Q, respectively.

As discussed in the previous section, the additional penalty term is positive and thus we expect that the new information criterion has heavier penalty than the standard lag order selection criteria, which will reduce the over-selection probability. More precisely, based on (20) and Lemma 4.1, we now propose a general form of the information criterion for lag order selection, which is suitable for dynamic panels involving fixed effects as follows.

Theorem 4.2 *We let Assumption A hold. We also assume that $n/T \to \gamma \in (0, \infty)$ and $n/T^3 \to 0$ as $n, T \to \infty$. Then a lag order selection criterion for dynamic panel models in (25) can be obtained as $k^* = \arg\min_{0 \leq \kappa \leq \bar{k}} PLIC_1(\kappa)$, where*

$$PLIC_1(\kappa) = \log \widetilde{\sigma}_\kappa^2 + \frac{h(n,T)}{nT}\left\{(2\kappa + 1) + \frac{2\sum_{j=1}^{\kappa} j\widetilde{\alpha}_{\kappa j}}{1 - \sum_{j=1}^{\kappa} \widetilde{\alpha}_{\kappa j}}\right\}, \tag{29}$$

in which $\overline{k}$ *is some predetermined upper bound,*[11] $h(n,T)$ *is some constant depending on* n *or* T*,*

$$\widetilde{\sigma}_{\kappa}^{2}=\frac{1}{nT}\sum\nolimits_{i=1}^{n}\left(y_{i}^{0}-x_{i}^{0}\widetilde{\alpha}_{\kappa}\right)'\left(y_{i}^{0}-x_{i}^{0}\widetilde{\alpha}_{\kappa}\right), \tag{30}$$

and $\widetilde{\alpha}_{\kappa}=(\widetilde{\alpha}_{\kappa 1},\cdots,\widetilde{\alpha}_{\kappa\kappa})'$ *is some bias corrected estimate for the panel* $AR(\kappa)$ *regression coefficients* α_{κ}*. More simply, if* $k\geq 1$*, a lower bound of* $PLIC_{1}(\kappa)$ *can be derived as*

$$PLIC_{1L}(\kappa)=\log\widetilde{\sigma}_{\kappa}^{2}+\frac{h(n,T)}{nT}\left\{(2\kappa+1)-\frac{2}{\kappa+1}\right\}. \tag{31}$$

If we let $h(n,T)=2$, then (29) becomes the model selection criterion derived in (20). >From (29), we can see that the additional penalty term reflects the stability of the system: if $\{y_{i,t}\}$ is close to unit root, then the additional penalty term explodes. The simpler form (31) provides the lower bound of (29), so theoretically it would select more over-parametrized models than (29) does. However, it is free from the nuisance parameters α_{κ} and thus it is more robust to the choice of bias correction method for them. An interesting finding is that, roughly speaking, the penalty term is doubled comparing to the standard penalty terms in the lag order selection.

In comparison, we can also suggest a slightly different lag order selection criterion given by[12]

$$PLIC_{2}(\kappa)=\log\widetilde{\sigma}_{\kappa}^{2}+\frac{1}{nT}\left\{h(n,T)\,\kappa+c\kappa\left(\frac{n}{T}\right)\right\} \tag{32}$$

for some positive constant c. It is based on the information criterion suggested by Lee (2006): $\log\widehat{\sigma}_{\kappa}^{2}+(nT)^{-1}\{h(n,T)(\kappa+n)+c\kappa(n/T)\}$. The original intuition of such modification is that the additional penalty term $c\kappa/T^{2}$ in (32) is introduced to offset the bias in $\widehat{\sigma}_{\kappa}^{2}=(nT)^{-1}\sum_{i=1}^{n}\left(y_{i}^{0}-x_{i}^{0}\widehat{\alpha}_{\kappa}\right)'\left(y_{i}^{0}-x_{i}^{0}\widehat{\alpha}_{\kappa}\right)$, where it is obtained using the standard (and thus biased) WG estimator $\widehat{\alpha}_{k}$ for α_{κ}. Specifically, using the results in Lee (2009), we have $\operatorname{plim}_{n\to\infty}\widehat{\sigma}_{\kappa}^{2}-\sigma_{\kappa}^{2}=-c\kappa/T^{2}+O\left(T^{-3}\right)$, where the constant c depends on the parameter values α_{κ} and clearly on the stability of the system similarly as the case of $PLIC_{1}(\kappa)$ in (29). Such bias is typically exacerbated when T is small and the system is less stable (i.e.,

[11] In practice, it is impossible to calculate $PLIC_{1}(k)$ for all orders. Usually a positive integer $\overline{k}$ is preassigned as an upper bound of possible orders, and the searching process is conducted up to this order. We simply assume that $\overline{k}$ is large enough to contain the true order. More discussion on choosing $\overline{k}$ can be found in Hannan and Deistler (1988), Choi (1992), and references therein.

[12] Note that the penalty in $PLIC_{2}$ is larger than $PLIC_{1L}$ in general since

$$\frac{h(n,T)\left\{(2\kappa+1)-\frac{2}{\kappa+1}\right\}}{h(n,T)\,\kappa+c\kappa\left(\frac{n}{T}\right)}=\frac{2\kappa^{2}+3\kappa-1}{\kappa^{2}+\kappa}\times\frac{1}{1+\frac{c}{h(n,T)}\left(\frac{n}{T}\right)}<1,$$

where $2\leq\frac{2\kappa^{2}+3\kappa-1}{\kappa^{2}+\kappa}\leq\frac{13}{6}<\infty$ and $\frac{n}{T}\times\frac{c}{h(n,T)}$ is large enough.

close to unit root). For example, when $\kappa = 1$, it can be derived that $c = (1 + \alpha_1) / (1 - \alpha_1)$. If we ignore the asymptotic bias of $\widehat{\sigma}^2_\kappa$ and construct a model selection criterion without such adjustment, the total regression error is equal to the biases of the autoregressive coefficients plus the original *i.i.d.* disturbance. As a consequence, the regression error has an erroneous serial correlation and behaves like an $ARMA$ process, or an $AR(\infty)$ process. Hence, the model selection is biased upward because it is prone to fit the model with κ as large as possible to reflect the erroneous serial correlation. The second part of the penalty term (32), or a heavier penalty overall, controls such phenomenon. The first part of the penalty term (32), on the other hand, is the conventional penalty function reflecting the total number of parameters.[13]

If we assume that the true lag order k exists and is finite, we can define a lag order estimator k^* is consistent if it satisfies $\lim_{n,T\to\infty} \mathbb{P}(k^* = k) = 1$. This definition is somewhat different from the usual probability limit, but it is equivalent for integer valued random variables. k^* is strongly consistent if $\mathbb{P}(\lim_{n,T\to\infty} k^* = k) = 1$. It is known that in the standard time series context, BIC and the properly defined PIC are strongly consistent criteria; HQ is weakly consistent but not strongly; and other order selection criteria, such as FPE and AIC, are not consistent for finite k. The constant $h(n,T)$ can be chosen so that it guarantees the consistency of the new order selection criterion.

Theorem 4.3 *Let $h(n,T)$ satisfy $h(n,T)/nT \to 0$ and $h(n,T) \to \infty$ as $n,T \to \infty$. Under Assumption A, if we let $n/T \to \gamma \in (0,\infty)$ and $n/T^3 \to 0$ as $n,T \to \infty$, $PLIC_{1L}(\kappa)$ in (31) is consistent lag order selection criterion, provided that the true lag order $k\,(\geq 1)$ is finite.*

Examples of $h(n,T)$ for consistent criteria are $\log(nT)$ and $\omega \log\log(nT)$ for some $\omega \geq 2$, where the first one is the BIC type penalty term and the second one is the HQ type penalty term. In fact, $\log(nT)$ renders strongly consistent criterion, whereas $\omega \log\log(nT)$ is so only when $\omega > 2$. We will see how the new lag order selection criteria perform by simulation studies in the following section.

Remark 4.4 Even when we want to consider the original likelihood in (26), instead of the profile likelihood, the standard AIC needs to be adjusted. In this case, the dimension of the parameter is $(n + k + 1)$ and the sample size is nT. As the dimension of the parameter

[13] More precisely, for the standard MLE as in (26), the total number of parameters is $(\kappa + n + 1)$ instead of κ, which includes the incidental parameters. Adding $n + 1$ does not seem to affect the lag order selection since the additional $(n+1)\,h(n,T)/nT$ term does not depend on the lag order κ. In practice, however, when we adjust the degrees of freedom using $(T - \kappa)$ instead of T, this term still depends on k though the effect is minor for large T. Ng and Perron (2005) discuss about choosing the effective number of observations in lag order selection.

space increases in comparison to the sample size, the standard AIC is known to become a strongly negatively biased estimate of the information (e.g., Hurvich and Tsai (1989)). This second-order bias (in addition to $B_P(\widehat{G})$ in (17)) can lead to overfitting but a proper modification can correct this small-sample-size bias. For example, similarly as Hurvich and Tsai (1989)'s AIC_C, we can consider a small-sample-size-bias-modified information criterion

$$PLIC_C(\kappa) = \log \widetilde{\sigma}_\kappa^2 + \frac{2(k+n+1)}{nT}\left\{1 + \frac{k+n+2}{n(T-1)-k}\right\},$$

which is also free from the nuisance parameters α_κ as (31).

5 Simulations

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

$$\begin{aligned} AIC(\kappa) &= \log \widetilde{\sigma}_\kappa^2 + \frac{2}{nT}(\kappa+n), \\ BIC(\kappa) &= \log \widetilde{\sigma}_\kappa^2 + \frac{\log(nT)}{nT}(\kappa+n) \\ HQ(\kappa) &= \log \widetilde{\sigma}_\kappa^2 + \frac{2\log\log(nT)}{nT}(\kappa+n), \end{aligned}$$

where $\widetilde{\sigma}_\kappa^2$ is an estimate for σ^2 in the panel $AR(\kappa)$ model as defined in (30). The number of parameters is $\kappa + n$ including fixed effect parameters. In practice, the effective number of observations in each time series is adjusted to reflect the degrees of freedom by $T - \kappa$, so including n for the total number of parameters is meaningful in finite samples. We consider the following criteria that we suggested in the previous section:

$$\begin{aligned} PLIC_1^{AIC}(\kappa) &= \log \widetilde{\sigma}_\kappa^2 + \frac{4}{nT}\left\{\kappa + \frac{\sum_{j=1}^{\kappa} j\widetilde{\alpha}_{\kappa j}}{1-\sum_{j=1}^{\kappa}\widetilde{\alpha}_{\kappa j}}\right\} \\ PLIC_1^{BIC}(\kappa) &= \log \widetilde{\sigma}_\kappa^2 + \frac{2\log(nT)}{nT}\left\{\kappa + \frac{\sum_{j=1}^{\kappa} j\widetilde{\alpha}_{\kappa j}}{1-\sum_{j=1}^{\kappa}\widetilde{\alpha}_{\kappa j}}\right\}, \\ PLIC_1^{HQ}(\kappa) &= \log \widetilde{\sigma}_\kappa^2 + \frac{2\log\log(nT)}{nT}\left\{\kappa + \frac{\sum_{j=1}^{\kappa} j\widetilde{\alpha}_{\kappa j}}{1-\sum_{j=1}^{\kappa}\widetilde{\alpha}_{\kappa j}}\right\}, \end{aligned}$$

For comparison purposes, we also consider the criteria by Lee (2006):

$$PLIC_2^{AIC}(\kappa) = \log \widetilde{\sigma}_\kappa^2 + \frac{1}{nT}\left\{2(\kappa+n) + \frac{n}{T}\kappa\right\},$$
$$PLIC_2^{BIC}(\kappa) = \log \widetilde{\sigma}_\kappa^2 + \frac{1}{nT}\left\{\log(nT)(\kappa+n) + \frac{n}{T}\kappa\right\},$$
$$PLIC_2^{HQ}(\kappa) = \log \widetilde{\sigma}_\kappa^2 + \frac{1}{nT}\left\{2\log\log(nT)(\kappa+n) + \frac{n}{T}\kappa\right\},$$

where c is simply chosen to one.

We generate $AR(k)$ dynamic panel processes, with k ranging from 1 to 4, of the form $y_{i,t} = \mu_i + \sum_{j=1}^{k} \alpha_{kj} y_{i,t-j} + \varepsilon_{i,t}$ for $i = 1, 2, \cdots, n$ and $t = 1, 2, \cdots, T$, where $\alpha_{kj} = 0.15$ for all $j = 1, \cdots, k$. For each $AR(k)$ model, all the autoregressive coefficients have the same value so that all the lagged terms are equally important. We consider nine different cases by combining different sample sizes of $n = 20, 50, 100$ and $T = 12, 25, 50$. Fixed effects μ_i are randomly drawn from $\mathcal{U}(-0.5, 0.5)$ and $\varepsilon_{i,t}$ from $\mathcal{N}(0, 1)$. We use the bias corrected WG estimators (e.g., Lee (2009)) for $\widetilde{\alpha}_{\kappa j}$'s and iterate the entire procedure 1000 times to compare the performance of different order selection criteria. For each case, we choose the optimal lag order k^* to minimize the criteria above, where we search the lag order from 1 to 10. (i.e., $\overline{k} = 10$) The simulation results are provided in Appendix A2. Tables A.1 to A.4 present the average values of k^* over 1000 iterations. It is very promising that all the new lag order selection criteria, $PLIC_1$ and $PLIC_2$, perform much better than the two most commonly used criteria, $AIC(\kappa)$, $BIC(\kappa)$ and $HQ(\kappa)$. In order to look at the distributional characteristics, we also provide Figures A.1 to A.4 for the case of $(n, T) = (100, 50)$. One interesting finding is that $BIC(\kappa)$ tends to overfit the panel models, which is contrary to the well known property that $BIC(\kappa)$ normally underfits in the pure time series setup. On the other hand, the figures consistently show that the new order selection criteria significantly reduce the over-selection probabilities. Though we do not present this particular result of our simulation, heavier penalty of the new information criteria $PLIC_1(\kappa)$ and $PLIC_2(\kappa)$ slightly increases the under-selection probabilities. But the increment of the under-selection probability is very minor, so that the overall correct-selection probabilities increase notably.

When T is very small and/or n is very large, so that the sample size ratio n/T is large, the order selection performance is not much satisfactory or overall cases, which is somewhat expected due to the very limited number of time series observations and large number of nuisance parameters. However, as T grows, the performances get better uniformly. This is intuitively appealing because the dynamic structure is mainly determined by the time series dimension. But unlike the conventional time series information criteria, the new criteria tend to choose the correct lag orders even when n is large, provided that T is not

so small. One remark is that though the simulation results look like $PLIC_2$ works better than $PLIC_1$, $PLIC_1$ cannot be always preferred to $PLIC_2$ empirically since $PLIC_2$ has larger under-selection probability than $PLIC_2$ does..

Lastly, one interesting finding is that, in general, the lag order selection is more accurate with $(n,T) = (50,50)$ than $(n,T) = (100,50)$. This implies that the sample size ratio n/T matters in lag order selection: the smaller n/T, the better work the information criteria. More simulation works need to be done to investigate such phenomenon.

6 Concluding Remarks

This paper considers model selection problem in the presence of incidental parameters. The main interest is in selecting the structure of the model in the common parameters after concentrating out the incidental parameters. As a particular example, a lag order selection criterion is examined in the context of dynamic panel models with fixed individual effects. An application of the new lag order selection criteria can be found in Lee (2006), which analyzes habit formation in consumption preferences and the specification of the habit function can be converted into the lag order selection problem.

For dynamic panel models, particularly for forecasting, it may not necessary to get the correct lag order as long as we can properly control for the serial correlation in the system. As an alternative approach to the lag order selection in dynamic panels, one of the authors is developing serial correlation robust GMM estimation for dynamic panel models with fixed effects. As noted in Remark 3.3, model selection in the context of semiparametric models with incidental parameters is also an interesting topic to investigate. In such context, it is convenient to use the kernel based local estimation for the nonparametric component. Alternatively, we could consider series approximation (or sieve approach) to estimate the nonparametric component and the coefficients of series functionals can be understood as the incidental parameters; but in this case, the number of basis functionals should be restricted in order to have a consistent nonparametric estimator.

7 Appendix

A.1 Mathematical Proofs

Proof of Theorem 3.1 Since the quasi maximum modified profile likelihood estimator is given by $\widehat{\psi}_M = \arg\max_\psi \log f_M(Z_{n,T};\psi)$ for $\psi \in \mathbb{R}^r$, we can define r-dimensional function H as a solution of the implicit equation:

$$\int \left. \frac{\partial \log f_M(z;\psi)}{\partial \psi} \right|_{\psi = H(G)} dG(z) = 0$$

and thus $\widehat{\psi}_M = H(\widehat{G})$. Note that using the standard M-estimator theory, the influence function of $\widehat{\psi}_M = H(\widehat{G})$ can be derived as (e.g., Huber (1981))

$$\phi(z;G) = \left(-\int \left. \frac{\partial^2 \log f_M(z;\psi)}{\partial\psi\partial\psi'} \right|_{\psi=H(G)} dG(z) \right)^{-1} \left. \frac{\partial \log f_M(z;\psi)}{\partial\psi'} \right|_{\psi=H(G)}.$$

From the discussions in Section 2, it can be shown that (e.g., Sartori (2003), Hahn and Kuersteiner (2004)), under the certain conditions such as Lyapounov's theorem, $\widehat{\psi}_M$ is $\sqrt{nT}$-consistent to $\psi_0 = H(G)$ when $n, T \to \infty$ satisfying $n/T \to \gamma \in (0,\infty)$ and $n/T^3 \to 0$.[14] Therefore, by Theorem 2.1 in Konishi and Kitagawa (1996),

$$\begin{aligned}
\frac{1}{T}B_P(G) &= \frac{1}{nT} tr\left\{ \int \phi(z;G) \left. \frac{\partial \log f_P(z;\psi)}{\partial\psi} \right|_{\psi=H(G)} dG(z) \right\} + o\left(\frac{1}{nT}\right) \\
&= \frac{1}{nT} tr\left\{ \left(-\int \left. \frac{\partial^2 \log f_M(z;\psi)}{\partial\psi\partial\psi'} \right|_{\psi=H(G)} dG(z) \right)^{-1} \times \right. \\
&\quad \left. \int \left. \frac{\partial \log f_M(z;\psi)}{\partial\psi} \right|_{\psi=H(G)} \left. \frac{\partial \log f_P(z;\psi)}{\partial\psi'} \right|_{\psi=H(G)} dG(z) \right\} + o\left(\frac{1}{nT}\right).
\end{aligned}$$

Now, similarly as Sartori (2003), it can be shown that $\partial\ell_{Mi}(\psi_0)/\partial\psi = u_i^e + O_p\left(T^{-1/2}\right)$. For $I_M(G)$, therefore, we have

$$I_M(G) = \sum\nolimits_{i=1}^n \mathbb{E}_{g_i}\left[-\frac{\partial^2 \ell_{Mi}(\psi_0)}{\partial\psi\partial\psi'} \right] \approx -\sum\nolimits_{i=1}^n \mathbb{E}_{g_i}\left[\frac{\partial u_i^e}{\partial\psi'} \right]$$

using the independence across i. Particularly when g is included in the family of f, however, the standard information identity holds and thus $-\mathbb{E}_{g_i}\left[\partial u_i^e/\partial\psi'\right] = \mathbb{E}_{g_i}\left[u_i^e u_i^{e\prime}\right] = \mathcal{I}_i^e$, which gives $I_M(G) \approx \sum_{i=1}^n \mathcal{I}_i^e$. For $J_{MP}(G)$, since $\partial\ell_{Pi}(\psi)/\partial\psi = \partial\ell_{Mi}(\psi)/\partial\psi + \partial M_i(\psi)/\partial\psi$ by construction, we similarly have

$$\begin{aligned}
J_{MP}(G) &= \sum\nolimits_{i=1}^n \mathbb{E}_{g_i}\left[\frac{\partial\ell_{Mi}(\psi_0)}{\partial\psi} \frac{\partial\ell_{Mi}(\psi_0)}{\partial\psi'} \right] + \sum\nolimits_{i=1}^n \mathbb{E}_{g_i}\left[\frac{\partial\ell_{Mi}(\psi_0)}{\partial\psi} \frac{\partial M_i(\psi_0)}{\partial\psi'} \right] \\
&\approx \sum\nolimits_{i=1}^n \mathbb{E}_{g_i}\left[u_i^e u_i^{e\prime}\right] + \sum\nolimits_{i=1}^n \mathbb{E}_{g_i}\left[u_i^e b_i(\psi_0)'\right]
\end{aligned}$$

by simply letting $\partial M_i(\psi_0)/\partial\psi = b_i(\psi_0)$. *Q.E.D.*

To prove Lemma 4.1, we first state useful results in the following lemma.

Lemma A.1 *For the model* (25) *satisfying Assumption A,*

(a) $\mathbb{E}\left(\sum_{t=1}^T \varepsilon_{i,t}\right)^2 = \mathbb{E}\sum_{t=1}^T \varepsilon_{i,t}^2 = T\sigma^2$

[14] Recall that $\widehat{\psi}_M$ automatically correct the first order asymptotic bias in the asymptotic distribution of the standard QML estimator $\widehat{\psi}$ (i.e., $\sqrt{nT}(\widehat{\psi}-\psi_0) \to_d N(0,V)$ as $n,T \to \infty$ for some $V > 0$) by construction.

(b) $\mathbb{E}\left(\sum_{t=1}^{T}\varepsilon_{i,t}\right)^2\sum_{t=1}^{T}\varepsilon_{i,t}^2 = T(T+2)\sigma^4$

(c) $\mathbb{E}\left(\sum_{t=1}^{T}\left(x_{i,t}x_{i,t}' - \mathbb{E}x_{i,t}\mathbb{E}x_{i,t}'\right)\right) = T\sigma^2 D$

(d) $\mathbb{E}\left(\sum_{t=1}^{T}(x_{i,t}-\mathbb{E}x_{i,t})\varepsilon_{i,t}\right)\left(\sum_{t=1}^{T}\varepsilon_{i,t}\right)\left(\sum_{t=1}^{T}(x_{i,t}-\mathbb{E}x_{i,t})'\right)$
$= T\sigma^4\left\{D + (I_k - A')^{-1}A'D + DA(I_k-A)^{-1}\right\} + o(T)$

where $D = \sum_{j=0}^{\infty}A^j ee' A'^j$ *and* e *as the* $k\times 1$ *column vector with one in the first element and zeros elsewhere.*

Proof of Lemma A.1 Since $\varepsilon_{i,t} \sim i.i.d.\left(0,\sigma^2\right)$, (a) is straightforward. (b) also follows easily since $\mathbb{E}\left(\sum_{t=1}^{T}\varepsilon_{i,t}\right)^2\sum_{t=1}^{T}\varepsilon_{i,t}^2 = T\mathbb{E}\varepsilon_{i,t}^4 + T(T-1)\left(\mathbb{E}\varepsilon_{i,t}^2\right)^2$. For (c), similarly as Lee (2009), we first write $x_{i,t}-\mathbb{E}x_{i,t} = \sum_{j=0}^{\infty}A^j e\varepsilon_{i,t-j-1}$. Note that, Assumption A-(ii) is equivalent to $\det[I_p - Az] \neq 0$ for all $|z|\leq 1$, or that each eigenvalue of A has modulus less than one. It thus guarantees that the sequence $\left\{A^j : j = 0,1,2,\cdots\right\}$ is absolutely summable and its infinite sum exists. Hence, the vector linear process $\sum_{j=0}^{\infty}A^j e\varepsilon_{i,t-j-1}$ is well defined in the mean square sense. It then follows that since $\mathbb{E}\left(\sum_{t=1}^{T}(x_{i,t}-\mathbb{E}x_{i,t})\right) = 0$ by construction, we have $\mathbb{E}\left(\sum_{t=1}^{T}\left(x_{i,t}x_{i,t}' - \mathbb{E}x_{i,t}\mathbb{E}x_{i,t}'\right)\right) = \mathbb{E}\left(\sum_{t=1}^{T}(x_{i,t}-\mathbb{E}x_{i,t})(x_{i,t}-\mathbb{E}x_{i,t})'\right) = \mathbb{E}\left(\sum_{t=1}^{T}\sum_{j=0}^{\infty}A^j ee'A'^j\varepsilon_{i,t-j-1}^2\right) = T\sigma^2\sum_{j=0}^{\infty}A^j ee'A'^j$. Finally, for (d), it can be derived that

$$
\begin{aligned}
&\mathbb{E}\left(\sum_{t=1}^{T}(x_{i,t}-\mathbb{E}x_{i,t})\varepsilon_{i,t}\right)\left(\sum_{t=1}^{T}\varepsilon_{i,t}\right)\left(\sum_{t=1}^{T}(x_{i,t}-\mathbb{E}x_{i,t})'\right)\\
=\;&\mathbb{E}\left(\sum_{t=1}^{T}\sum_{j=0}^{\infty}\varepsilon_{i,t}^2\{\varepsilon_{i,t-j-1}^2 A^j ee'A'^j\right.\\
&\left.+\sum_{s=1}^{t-1}\varepsilon_{i,s-j-1}^2 A^j ee'A'^{j+s} + \sum_{s=1}^{T-t}\varepsilon_{i,s-j-1}^2 A^{j+s}ee'A'^j\}\right)\\
=\;&T\sigma^4\left(\sum_{j=0}^{\infty}A^j ee'A'^j\right)\\
&+\sigma^4\left(TI_k - \left(I_k - A'^T\right)\left(I_k-A'\right)^{-1}\right)\left(I_k-A'\right)^{-1}A'\left(\sum_{j=0}^{\infty}A^j ee'A'^j\right)\\
&+\sigma^4\left(\sum_{j=0}^{\infty}A^j ee'A'^j\right)A(I_k-A)^{-1}\left(TI_k - (I_k-A)^{-1}\left(I_k-A^T\right)\right)
\end{aligned}
$$

similarly as Lemma A1 in Lee (2009). Note that Assumption A-(ii) guarantees $\sum_{j=0}^{\infty}A^j = (I_k-A)^{-1}$ exists and $\left\|A^T\right\| \to 0$ as $T\to\infty$. Therefore, we can approximate (f) as

$$
\begin{aligned}
&\mathbb{E}\left(\sum_{t=1}^{T}(x_{i,t}-\mathbb{E}x_{i,t})\varepsilon_{i,t}\right)\left(\sum_{t=1}^{T}\varepsilon_{i,t}\right)\left(\sum_{t=1}^{T}(x_{i,t}-\mathbb{E}x_{i,t})'\right)\\
=\;&T\sigma^4\left(\sum_{j=0}^{\infty}A^j ee'A'^j\right)\\
&+T\sigma^4\left(I_k-A'\right)^{-1}A'\left(\sum_{j=0}^{\infty}A^j ee'A'^j\right) + T\sigma^4\left(\sum_{j=0}^{\infty}A^j ee'A'^j\right)A(I_k-A)^{-1} + o(T).
\end{aligned}
$$

Q.E.D.

Proof of Lemma 4.1 We first let $\ell_i\left(\mu_i,\alpha_k,\sigma^2\right)=\ell\left(y_i|x_{i,1},\mu_i,\alpha_k,\sigma^2\right)$ and $\ell_{Pi}\left(\alpha_k,\sigma^2\right)=\ell\left(y_i|x_{i,1},\widehat{\mu}_i\left(\alpha_k,\sigma^2\right),\alpha_k,\sigma^2\right)$. Note that, in this particular case, we can obtain that

$$v_i\left(\mu_i,\alpha_k,\sigma^2\right)=\iota'\varepsilon_i/\sigma^2,\;\; u_i\left(\mu_i,\alpha_k,\sigma^2\right)=\left(\begin{array}{c} x_i'\varepsilon_i/\sigma^2 \\ -T/2\sigma^2+\varepsilon_i'\varepsilon_i/2\sigma^4 \end{array}\right),$$

and

$$\mathcal{I}_i=\left(\begin{array}{ccc} \mathbb{E}(x_i'x_i)/\sigma^2 & 0 & T\mathbb{E}(\overline{x}_i)/\sigma^2 \\ 0 & T/2\sigma^4 & 0 \\ T\mathbb{E}(\overline{x}_i)/\sigma^2 & 0 & T/\sigma^2 \end{array}\right),$$

where $\mathbb{E}(\overline{x}_i)=\mathbb{E}(T^{-1}\sum_{t=1}^{T}x_{i,t})$. Therefore, we can derive that

$$\begin{aligned} u_i^e\left(\mu_i,\alpha_k,\sigma^2\right) &= \left(\begin{array}{c} \left(x_i-\iota\mathbb{E}(\overline{x}_i)\right)'\varepsilon_i/\sigma^2 \\ -T/2\sigma^2+\varepsilon_i'\varepsilon_i/2\sigma^4 \end{array}\right) \\ b_i\left(\alpha_k,\sigma^2\right) &= \left(\begin{array}{c} \left(\sum_{t=1}^{T}\varepsilon_{i,t}\right)\left(\sum_{t=1}^{T}\left(x_{i,t}-\mathbb{E}(\overline{x}_i)\right)\right)/T\sigma^2 \\ \left(\sum_{t=1}^{T}\varepsilon_{i,t}\right)^2/2T\sigma^4 \end{array}\right) \\ \mathcal{I}_i^e &= \left(\begin{array}{cc} \mathbb{E}\left(\sum_{t=1}^{T}\left(x_{i,t}x_{i,t}'-\mathbb{E}(\overline{x}_i)\mathbb{E}(\overline{x}_i)'\right)\right)/\sigma^2 & 0 \\ 0 & T/2\sigma^4 \end{array}\right). \end{aligned}$$

Now we let

$$cov\left(u_i^e,b_i\left(\alpha_k,\sigma^2\right)\right)\equiv\left(\begin{array}{cc} C_{i,11} & C_{i,12} \\ C_{i,21} & C_{i,22} \end{array}\right).$$

Since we want to find $tr\{[\sum_{i=1}^{n}\mathcal{I}_i^e]^{-1}\sum_{i=1}^{n}cov\left(u_i^e,b_i\left(\alpha_k,\sigma^2\right)\right)\}$ and $\mathcal{I}_i^e$ is block-diagonal for all i, we only need to derive the expression for $C_{i,11}$ and $C_{i,22}$. Using Lemma A.1, where we simply let $\mathbb{E}(\overline{x}_i)=\mathbb{E}x_{i,t}$ from the mean stationarity of $y_{i,t}$, we can derive

$$C_{i,11}=D+\left(I_k-A'\right)^{-1}A'D+DA\left(I_k-A\right)^{-1}\;\;\text{and}\;\;C_{i,22}=1/2\sigma^4.$$

Therefore,

$$\begin{aligned} tr\left(\left[\sum_{i=1}^{n}\mathcal{I}_i^e\right]^{-1}\sum_{i=1}^{n}cov\left(u_i^e,b_i\left(\alpha_k,\sigma^2\right)\right)\right) &= tr\left(D^{-1}C_{i,11}\right)+2\sigma^4C_{i,22} \qquad \text{(A.1)} \\ &= k+tr\left\{A\left(I_k-A\right)^{-1}+\left(I_k-A'\right)^{-1}A'\right\}+1. \end{aligned}$$

The result follows since D is symmetric and $tr\left(Q\right)=tr\left(Q'\right)$ for a square matrix Q.

To show that $k + tr\left\{A\left(I_k - A\right)^{-1} + \left(I_k - A'\right)^{-1} A'\right\}$ is nonnegative, we define

$$B = \left(I_k + A'A\right)^{1/2} \left(I_k - A\right)^{-1},$$

in which $I_k + A'A = I_k + A'I_kA$ is positive definite by construction, then

$$\begin{aligned} tr\left\{B'B\right\} &= tr\left\{\left(I_k - A'\right)^{-1}\left(I_k + A'A\right)\left(I_k - A\right)^{-1}\right\} \\ &= tr\left\{\left(I_k - A'\right)^{-1}\left(A' + \left(I_k - A'\right)\left(I_k - A\right) + A\right)\left(I_k - A\right)^{-1}\right\} \\ &= k + tr\left\{A\left(I_k - A\right)^{-1} + \left(I_k - A'\right)^{-1} A'\right\}. \end{aligned}$$

Since $B'B = B'I_kB$ is positive definite for $k \geq 1$, which is equivalent that all the eigenvalues of $B'B$ are positive, and the trace is sum of eigenvalues, it follows that $tr\left\{B'B\right\} > 0$. Note that $tr\left\{B'B\right\} = 0$ when $k = 0$..

Finally, the bounds for $k + tr\left\{A\left(I_k - A\right)^{-1} + \left(I_k - A'\right)^{-1} A'\right\}$ can be obtained easily since

$$\begin{aligned} & k + tr\left\{A\left(I_k - A\right)^{-1} + \left(I_k - A'\right)^{-1} A'\right\} \\ = \; & tr\left\{\left(I_k - A'\right)^{-1}\left(I_k + A'A\right)\left(I_k - A\right)^{-1}\right\} \\ = \; & tr\left\{\left[\left(I_k - A\right)'\left(I_k - A\right)\right]^{-1}\left(I_k + A'A\right)\right\} \end{aligned}$$

and

$$\frac{tr\left(Q_2\right)}{\Lambda_{\max}\left(Q_1\right)} = \Lambda_{\min}\left(Q_1^{-1}\right) tr\left(Q_2\right) \leq tr\left(Q_1^{-1}Q_2\right) \leq \Lambda_{\max}\left(Q_1^{-1}\right) tr\left(Q_2\right) = \frac{tr\left(Q_2\right)}{\Lambda_{\min}\left(Q_1\right)}$$

for some $k \times k$ positive definite matrices Q_1 and Q_2. *Q.E.D.*

Proof of Theorem 4.2 First note that, thanks to the specific structure of A, we can derive $tr\left(A\left(I_k - A\right)^{-1}\right) = \left(\sum_{j=1}^{k} j\alpha_{kj}\right) / \left(1 - \sum_{j=1}^{k} \alpha_{kj}\right)$. From (20), (22) and Lemma 4.1, by replacing k with κ, it thus follows that the information criterion (20) can be simplified as

$$\log \widetilde{\sigma}_\kappa^2 + \frac{2\kappa}{nT} + \frac{2}{nT}\left(\kappa + \frac{2\sum_{j=1}^{\kappa} j\widetilde{\alpha}_{\kappa j}}{1 - \sum_{j=1}^{\kappa} \widetilde{\alpha}_{\kappa j}} + 1\right) = \log \widetilde{\sigma}_\kappa^2 + \frac{2}{nT}\left((2\kappa + 1) + \frac{2\sum_{j=1}^{\kappa} j\widetilde{\alpha}_{\kappa j}}{1 - \sum_{j=1}^{\kappa} \widetilde{\alpha}_{\kappa j}}\right)$$

for some consistent estimators $\widetilde{\sigma}_\kappa^2$ and $\widetilde{\alpha}_{\kappa 1}, \cdots, \widetilde{\alpha}_{\kappa\kappa}$. $PLIC_1\left(\kappa\right)$ is simply obtained using $h\left(n, T\right)$ instead of 2 in front of the penalty function.

For deriving $PLIC_{1L}\left(\kappa\right)$, we have $tr\left(A\left(I_\kappa - A\right)^{-1}\right) \geq tr\left(A\right)\Lambda_{\min}\left(\left(I_\kappa - A\right)^{-1}\right) \geq tr\left(A\right)/\Lambda_{\max}\left(I_\kappa - A\right) \geq tr\left(A\right)/tr\left(I_\kappa - A\right)$. Since $tr\left(A\right) = \alpha_{\kappa 1}$ and $tr\left(I_\kappa - A\right) = \kappa - \alpha_{\kappa 1}$,

it thus follows that

$$tr\left(A\left(I_\kappa - A\right)^{-1}\right) \geq \frac{\alpha_{\kappa 1}}{\kappa - \alpha_{\kappa 1}}.$$

Furthermore, since $|\alpha_{\kappa 1}| < 1$ for any κ from Assumption A-(ii), we have $\alpha_{\kappa 1}/\left(\kappa - \alpha_{\kappa 1}\right) \geq -1/\left(\kappa + 1\right)$. From $(A.1)$, therefore,

$$\kappa + 2tr\left(A\left(I_\kappa - A\right)^{-1}\right) + 1 \geq \left(\kappa + 1\right) - \frac{2}{\kappa + 1}.$$

Q.E.D.

Proof of Theorem 4.3 Since the selection rule is to choose $k^* \left(\neq k\right)$ if $PLIC_1\left(k^*\right) < PLIC_1\left(k\right)$, we shall prove that $\lim_{n,T\to\infty} \mathbb{P}\left[PLIC_1\left(k^*\right) < PLIC_1\left(k\right)\right] = 0$ for all $k^* \neq k$ and $k^* \leq \overline{k}$, where k is the (finite) true lag order. We first consider the case $k^* < k$. We write

$$\begin{aligned} & \mathbb{P}\left[PLIC_1\left(k^*\right) < PLIC_1\left(k\right)\right] \\ = \; & \mathbb{P}\left[\left(\log\widetilde{\sigma}_{k^*}^2 - \log\widetilde{\sigma}_k^2\right) < \frac{2h\left(n,T\right)}{nT}\left\{\left(k - k^*\right) + \left(\xi_{k^*} - \xi_k\right)\right\}\right], \end{aligned} \tag{A.2}$$

where $\xi_k = \left(\sum_{j=1}^k j\alpha_{kj}\right) / \left(1 - \sum_{j=1}^k \alpha_{kj}\right)$. The left-hand-side in $(A.2)$, $\left(\log\widetilde{\sigma}_{k^*}^2 - \log\widetilde{\sigma}_k^2\right)$, is nonnegative for any n and T because the residual sum of squares does not increase (mostly decreases) as the number of regressors increases. On the other hand, the right-hand-side in $(A.2)$, $\frac{2h(n,T)}{nT}\left\{\left(k - k^*\right) + \left(\xi_{k^*} - \xi_k\right)\right\}$, converges to zero as $n, T \to \infty$ since $0 < \left(k - k^*\right) < \overline{k} < \infty$, $\left|\xi_{k^*} - \xi_k\right| < \left|\xi_{k^*}\right| + \left|\xi_k\right| < \infty$ from Assumption A-(ii) and $h\left(n,T\right)/nT \to 0$ as $n, T \to \infty$ by assumption. Therefore, $\mathbb{P}\left[PLIC_1\left(k^*\right) < PLIC_1\left(k\right)\right] \to 0$ as $n, T \to \infty$.

Now we consider the case $k^* > k$. By the Kronecker's lemma and Assumption A-(ii), $(1/k)\sum_{j=1}^k j\alpha_{kj}$ should be small if k is large. Therefore, we can write $\xi_k = k\epsilon_k$ for some finite ϵ_k. Note that Assumption A-(ii) also implies $0 < \left|1 - \sum_{j=1}^k \alpha_{kj}\right| < \infty$. It follows that

$$\begin{aligned} & \mathbb{P}\left[PLIC_1\left(k^*\right) < PLIC_1\left(k\right)\right] \\ = \; & \mathbb{P}\left[nT\left(\log\widetilde{\sigma}_{k^*}^2 - \log\widetilde{\sigma}_k^2\right) < 2h\left(n,T\right)\left\{\left(k - k^*\right) + \left(\xi_{k^*} - \xi_k\right)\right\}\right] \\ \leq \; & \mathbb{P}\left[nT\left(\log\widetilde{\sigma}_{k^*}^2 - \log\widetilde{\sigma}_k^2\right) < 2h\left(n,T\right)\left\{\left(k - k^*\right) + \left|\xi_{k^*} - \xi_k\right|\right\}\right] \\ \leq \; & \mathbb{P}\left[nT\left(\log\widetilde{\sigma}_{k^*}^2 - \log\widetilde{\sigma}_k^2\right) < 2h\left(n,T\right)\left(k - k^*\right)\left\{1 + \max\left\{\left|\epsilon_k\right|, \left|\epsilon_{k^*}\right|\right\}\right\}\right]. \end{aligned} \tag{A.3}$$

Similarly as Lee (2009), when $k^* > k$, we can simply let $\alpha_{k^*j} = 0$ for $j > k$ and we can show that (e.g., Lee (2006), Lee (2009)) $\text{plim}_{n\to\infty}\widetilde{\sigma}_k^2 - \sigma_k^2 = O\left(T^{-2}\right)$ and $\text{plim}_{n\to\infty}\widetilde{\sigma}_{k^*}^2 - \sigma^2\left(k\right) = O\left(T^{-2}\right)$. It follows that $\left|\log\widetilde{\sigma}_{k^*}^2 - \log\widetilde{\sigma}_k^2\right| = O_p\left(T^{-2}\right)$ for large n. Therefore, the left-hand-side in $(A.3)$ is $O_p\left(1\right)$ for large n and T because $n/T \to \gamma \in (0, \infty)$. However, the right-hand-side in $(A.3)$ goes to negative infinity because $\left(k - k^*\right) < 0$ and $h\left(n,T\right) \to \infty$ as $N, T \to \infty$. Therefore, $\mathbb{P}\left[PLIC_1\left(k^*\right) < PLIC_1\left(k\right)\right] \to 0$ in this case, too. In sum, $\lim_{n,T\to\infty} \mathbb{P}\left[PLIC_1\left(k^*\right) < PLIC_1\left(k\right)\right] = 0$ for any $k^* \leq \overline{k}$, or the $PLIC_1\left(k\right)$ is a consistent order selection criterion. *Q.E.D.*

A.2 Simulation Results

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

n	T	AIC	BIC	HQ	$\text{PLIC}_1^{\text{AIC}}$	$\text{PLIC}_1^{\text{BIC}}$	$\text{PLIC}_1^{\text{HQ}}$	$\text{PLIC}_2^{\text{AIC}}$	$\text{PLIC}_2^{\text{BIC}}$	$\text{PLIC}_2^{\text{HQ}}$
20	12	10.0	9.71	10.0	9.82	8.43	9.57	1.98	1.05	2.39
20	25	6.41	2.63	4.77	4.67	1.27	2.69	3.02	1.15	3.33
20	50	4.81	1.69	3.19	3.24	1.10	1.65	4.08	1.22	4.27
50	12	10.0	9.99	10.0	9.99	9.72	9.95	1.51	1.00	1.73
50	25	8.43	4.25	6.47	7.12	1.76	4.18	2.13	1.06	2.40
50	50	6.00	2.12	4.23	4.50	1.11	2.20	3.51	1.10	3.51
100	12	10.0	10.0	10.0	10.0	10.0	10.0	1.15	1.00	1.29
100	25	9.08	6.10	8.22	8.58	3.42	6.95	2.25	1.03	2.29
100	50	6.29	3.23	4.76	5.10	1.42	3.03	3.24	1.13	3.11

Table A.1: Average of lag order selections over 1000 iterations
$\big(p = 1;\ \overline{k} = 10\big)$

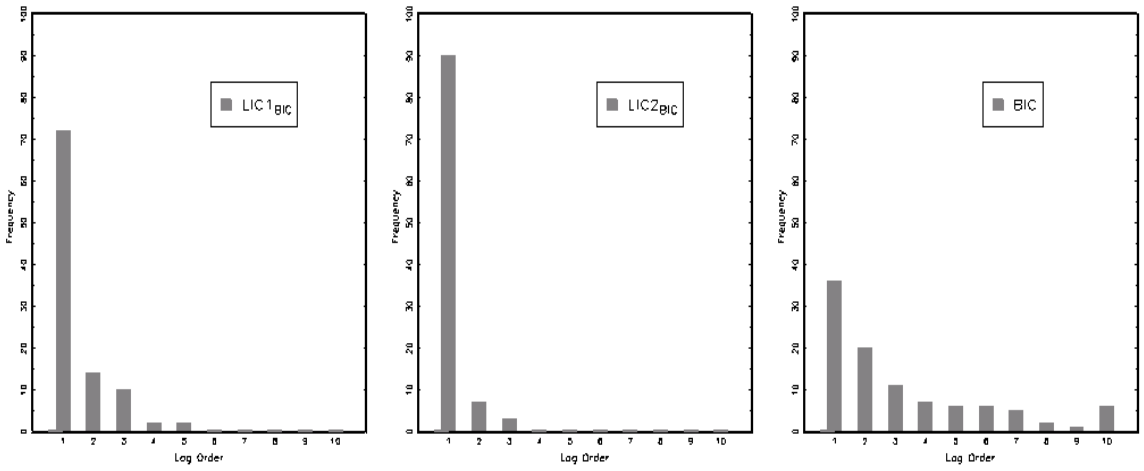

Figure A.1: Order selection frequencies over 1000 iterations
$(p = 1;\ (n, T) = (100, 50))$

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

n	T	AIC	BIC	HQ	$\text{PLIC}_1^{\text{AIC}}$	$\text{PLIC}_1^{\text{BIC}}$	$\text{PLIC}_1^{\text{HQ}}$	$\text{PLIC}_2^{\text{AIC}}$	$\text{PLIC}_2^{\text{BIC}}$	$\text{PLIC}_2^{\text{HQ}}$
20	12	9.94	9.71	9.94	9.72	7.95	9.32	1.78	1.05	2.24
20	25	6.50	2.39	5.06	5.12	1.20	2.39	2.95	1.32	3.41
20	50	5.61	2.32	4.00	4.28	1.36	2.24	4.65	1.82	4.79
50	12	10.0	10.0	10.0	9.99	9.71	9.91	1.53	1.03	1.73
50	25	8.13	4.41	6.57	7.06	2.23	4.82	3.13	1.29	3.53
50	50	5.97	3.28	4.72	4.97	2.23	3.30	4.39	2.17	4.39
100	12	10.0	10.0	10.0	10.0	10.0	10.0	1.25	1.00	1.37
100	25	8.97	6.50	8.14	8.53	4.09	7.44	2.89	1.20	2.92
100	50	6.56	3.66	5.29	5.64	2.34	3.78	3.78	2.10	3.71

Table A.2: Average of lag order selections over 1000 iterations
$\big(p = 2;\ \overline{k} = 10\big)$

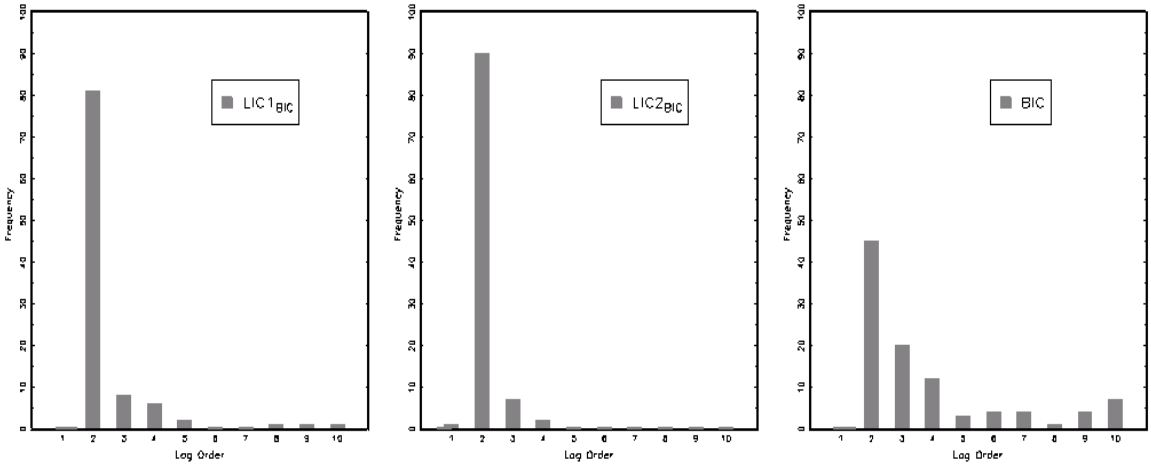

Figure A.2: Order selection frequencies over 1000 iterations
$(p = 2;\ (n, T) = (100, 50))$

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

n	T	AIC	BIC	HQ	$\text{PLIC}_1^{\text{AIC}}$	$\text{PLIC}_1^{\text{BIC}}$	$\text{PLIC}_1^{\text{HQ}}$	$\text{PLIC}_2^{\text{AIC}}$	$\text{PLIC}_2^{\text{BIC}}$	$\text{PLIC}_2^{\text{HQ}}$
20	12	10.0	9.75	9.98	9.68	7.59	9.20	1.95	1.13	2.66
20	25	6.97	2.91	5.04	5.15	1.16	2.65	3.51	1.47	3.80
20	50	6.19	3.53	4.78	5.15	1.63	3.07	5.08	2.75	5.28
50	12	10.0	10.0	10.0	9.99	9.81	9.97	1.56	1.05	1.76
50	25	8.20	5.24	7.19	7.71	2.71	5.94	3.74	1.67	3.84
50	50	6.93	4.05	5.91	5.99	3.08	4.34	5.08	3.03	5.17
100	12	10.0	10.0	10.0	10.0	10.0	10.0	1.12	1.00	1.29
100	25	9.23	7.19	8.76	8.89	5.10	8.05	3.54	1.45	3.72
100	50	7.24	4.79	6.40	6.70	3.32	4.98	4.75	3.11	4.68

Table A.3: Average of lag order selections over 1000 iterations
$\big(p = 3;\ \overline{k} = 10\big)$

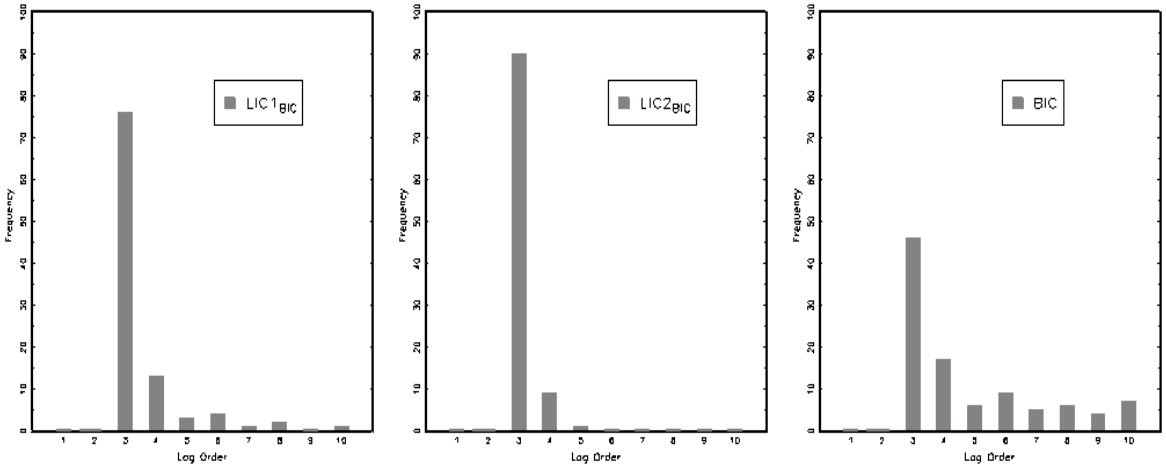

Figure A.3: Order selection frequencies over 1000 iterations
$(p = 3;\ (n, T) = (100, 50))$

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

n	T	AIC	BIC	HQ	$\text{PLIC}_1^{\text{AIC}}$	$\text{PLIC}_1^{\text{BIC}}$	$\text{PLIC}_1^{\text{HQ}}$	$\text{PLIC}_2^{\text{AIC}}$	$\text{PLIC}_2^{\text{BIC}}$	$\text{PLIC}_2^{\text{HQ}}$
20	12	9.90	9.52	9.82	9.42	7.09	8.91	2.10	1.06	2.48
20	25	7.23	3.65	5.64	5.65	1.36	2.89	4.36	1.76	4.92
20	50	6.56	4.41	5.40	5.48	2.20	3.72	5.57	3.62	5.58
50	12	10.0	9.99	10.0	9.90	9.40	9.88	1.63	1.03	1.70
50	25	8.31	5.86	7.62	7.94	3.09	6.21	4.61	1.76	4.89
50	50	7.11	5.17	6.26	6.60	3.93	5.51	6.14	4.07	6.19
100	12	10.0	10.0	10.0	10.0	10.0	10.0	1.39	1.01	1.46
100	25	9.11	7.56	8.54	8.89	6.46	8.07	4.38	1.80	4.47
100	50	7.73	5.56	6.94	7.22	4.37	6.05	5.54	4.04	5.54

Table A.4: Average of lag order selections over 1000 iterations
$\big(p = 4;\ \overline{k} = 10\big)$

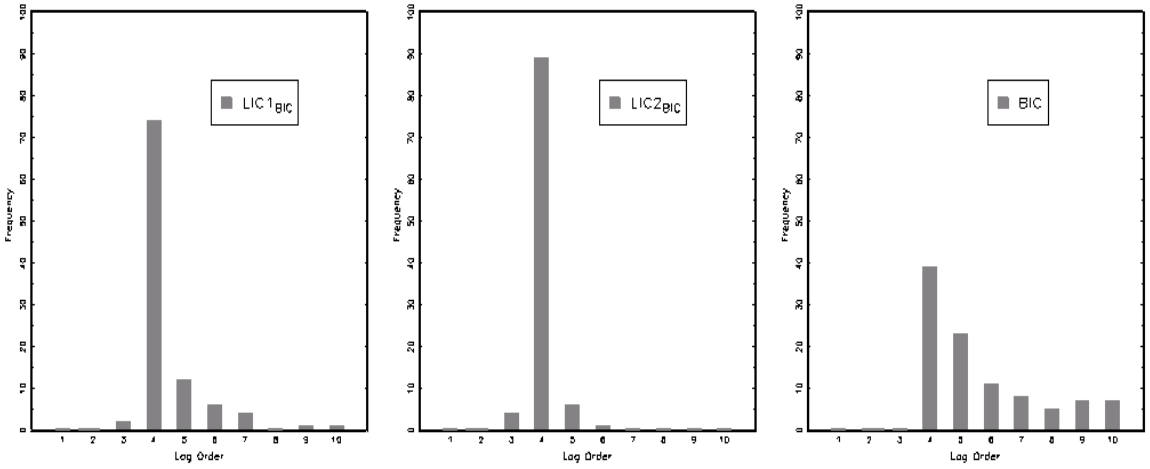

Figure A.4: Order selection frequencies over 1000 iterations
$(p = 4;\ (n, T) = (100, 50))$

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: Academiai Kiado.

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

Alvarez, J. and M. Arellano (2003). The time series and cross-section asymptotics of dynamic panel data estimators, *Econometrica*, 71, 1121-1159.

Arellano, M., and J. Hahn (2005). Understanding bias in nonlinear panel models: Some recent developments, unpublished manuscript.

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.

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

Bester, C.A. and C. Hansen (2007). A Penalty Function Approach to Bias Reduction in Nonlinear Panel Models with Fixed Effects, *Journal of Business and Economic Statistics*, forthcoming.

Choi, B. (1992). *ARMA Model Identification*, New York: Springer-Verlag.

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.

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 (2004). Bias reduction for dynamic nonlinear panel models with fixed effects, unpublished manuscript.

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

Hannan, E.J. and M. Deistler (1988). *The Statistical Theory of Linear Systems*, New York: John Wiley.

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

Hurvich, C.M. and, C.-L. Tsai (1989). Regression and time series model selection in small samples, *Boimetrika*, 76, 297-307.

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

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

Lee, Y. (2008). Nonparametric estimation of dynamic panel models with fixed effects, unpublished manuscript, University of Michigan.

Lee, Y. (2009). Bias in dynamic panel models under time series misspecification, unpublished manuscript, University of Michigan.

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.

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

Nickell, S. (1981). Biases in dynamic models with fixed effects, *Econometrica*, 49, 1417-1425.

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

Robinson, P.M. (1988). Root-N consistent semiparametric regression, *Econometrica*, 56, 931-954.

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

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, *The Annals of Statistics*, 20, 1768-1802.

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

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