Lag Length Selection in Panel Autoregression*

Chirok Han
Korea University

Peter C. B. Phillips
Yale University, University of Auckland,
University of Southampton & Singapore Management University

Donggyu Sul
University of Texas at Dallas

July 2012, Revision July 2013

Abstract

Model selection by BIC is well known to be inconsistent in the presence of incidental parameters. This paper shows that, somewhat surprisingly, even without fixed effects in dynamic panels BIC is inconsistent and overestimates the true lag length with considerable probability. The reason for the inconsistency is explained and the probability of overestimation is found to be 50% asymptotically. Three alternative consistent lag selection methods are considered. Two of these modify BIC and the third involves sequential testing. Simulations evaluate the performance of these alternative lag selection methods in finite samples.

Keywords: BIC, Dynamic panel, Lag selection, X-differencing, Sequential testing
JEL Classification Number: C33.

*We thank two referees, an associate editor, and the editor for helpful comments. Chengcheng Jia provided excellent research assistance. Phillips acknowledges support from the NSF under Grant Nos SES 09-56687 and 12-58285. Han thanks support from the National Research Foundation of Korea Grant funded by the Korean Government (2012S1A5A8025004).

1 Dynamic Panel Lag Order Estimation

Specification of the appropriate lag order to capture response time and feedback is a delicate econometric issue in time series models. Some early work by Peter Schmidt (1971, 1973, 1974) and Schmidt and Sickles (1975) partly addressed this problem in the context of Almon distributed lag models and suggested various solutions. In dynamic panel models the problem is known to be even more complex in part because of the presence of fixed effects which mean that the dimension of the parameter space increases with the sample size.

Stone (1979) first demonstrated the inconsistency of the Schwarz (1978) information criterion (hereafter BIC) in a simple incidental parameter context. Since then some generalized criteria have been developed for this problem that have better properties and correspond more closely to Bayes factors (Berger et al, 2003; Chakrabarti and Ghosh, 2006; Lee, 2011). The inconsistency of BIC that was studied in Stone (1979) arises specifically because of the presence of incidental parameters. That outcome seems unsurprising at least in panel models given the well known bias effects of incidental parameters in dynamic panels.

Much more surprising, however, is the fact that BIC fails to produce a consistent lag order estimator in simple dynamic panels. The present paper shows, somewhat remarkably, that BIC is inconsistent for lag order estimation even in panel models with no fixed effects. Thus, the large sample good behavior of BIC is compromised in dynamic panel models even in the absence of an incidental parameter problem. The reason for the failure of BIC even in simple dynamic models with no fixed effects is that the BIC penalty is too small to compensate for the additional terms from cross section averaging ($O(n)$ such terms) that enter into the BIC model fit comparison when overfitting.[1] These additional terms arise from differences in the number of time series observations used in the calculation of the residual variance estimates $(\hat{\sigma}_k^2, \hat{\sigma}_{k_0}^2)$ in a panel model with k and k_0 lags. As we show, they satisfy a CLT and are of $O_p(\frac{1}{\sqrt{nT}})$ in relation to the BIC penalty of $O(\frac{\log nT}{nT})$. So they produce a strong tendency to overfit the panel autoregression as $n \to \infty$. The overfitting tendency is as high as 50% asymptotically.

To address the inconsistency of BIC, the paper develops some modified information criteria that are consistent in dynamic panels. These criteria involves simple modifications to BIC and are easy to implement in practice. They are compared in simulations to assess finite sample performance of the various criteria. Some comparisons are also made with standard sequential testing procedures

[1] These terms also arise in conventional time series applications of BIC, but produce no overfitting tendency because there are only a finite number of these terms. In a panel context this finite number is scaled by the number of cross section observations, thereby disturbing the asymptotic properties of BIC.

for lag order determination.

For brevity we consider the following simple panel AR(k) process

$$(1) \quad y_{it} = \sum\nolimits_{s=1}^{k} \rho_s y_{it-s} + \varepsilon_{it}, \text{ where } \varepsilon_{it} \sim iid\ N(0,\sigma^2),\ i = 1,..,n;\ t = 1,...,T$$

which will be sufficient to make the main points of the paper. Let k_0 be the true value of the lag order in (1). Define $X_{k,it} = (y_{it-1}, \ldots, y_{it-k})'$ and $\beta_k = (\rho_1, \ldots, \rho_k)'$. Conditioning on the initial observations $\{y_{i1}, \ldots, y_{ik}\}$, the Gaussian log-likelihood is

$$(2) \quad \ln L(\beta_k, \sigma^2) = -\frac{nT_k}{2} \ln 2\pi - \frac{nT_k}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{t=k+1}^{T} (y_{it} - \beta_k' X_{k,it})^2,$$

where $T_k = T - k$. In view of (2), the maximum likelihood estimator (MLE) of β_k is the same as pooled least squares (OLS), viz., $\hat{\beta}_k = (\sum_{i=1}^n \sum_{t=k+1}^T X_{k,it} X_{k,it}')^{-1} (\sum_{i=1}^n \sum_{t=k+1}^T X_{k,it} y_{it})$, with corresponding error variance estimator $\hat{\sigma}_k^2 = (nT_k)^{-1} \sum_{i=1}^n \sum_{t=k+1}^T \hat{\varepsilon}_{k,it}^2$, where $\hat{\varepsilon}_{k,it} = y_{it} - X_{k,it}' \hat{\beta}_k$.

Let k_0 be the true lag length in the model (1), i.e., $k_0 = \min\{k : \rho_k \neq 0,\ \rho_j = 0\ \forall j > k\}$. The order parameter k is frequently estimated using an information criterion (IC) according to the typical extremum rule $\hat{k} = \arg\min_{k \leq k_{\max}} IC_0(k)$ for some given $k_{\max} \geq k_0$ where IC commonly satisfies

$$(3) \quad IC_0(k) - IC_0(k_0) = \ln(\hat{\sigma}_k^2 / \hat{\sigma}_{k_0}^2) + (k - k_0) c_{nT},$$

and c_{nT} is some penalty function. The BIC penalty has the typical form

$$(4) \quad c_{nT} = \ln(nT)/nT,$$

which reflects the overall sample size nT in this panel data case.

To fix ideas and provide a rigorous development we make the following high level assumptions, which are easily shown to hold for stationary and asymptotically stationary panels.

Assumption A. *(i)* $(nT_k)^{-1} \sum_{i=1}^n \sum_{t=k+1}^T X_{k,it} X_{k,it}'$ *converges in probability to a positive definite matrix for all fixed* k*;*

(ii) $(nT_k)^{-1/2} \sum_{i=1}^n \sum_{t=k+1}^T X_{k,it} \varepsilon_{it} = O_p(1)$ *for all* k*;*

(iii) $\hat{\beta}_k - \beta_k = O_p(n^{-1/2} T_k^{-1/2})$ *for* $k \geq k_0$.

These conditions can be considerably relaxed at the cost of additional complexity. For instance, the zero intercept and normality in (1) are unnecessary and the iid error condition can be

replaced with independence over i and uniformly bounded heteroskedasticity and higher moments ($\sup_{i,t} \mathbb{E}(|\varepsilon_{it}|^{4+\delta}) = M < \infty$ for some $\delta > 0$). Under uniformly bounded fourth moments, the means of the second moments of ε_{it}^2 are well defined so that the main results of this paper go through.[2] While normality is not needed for the limit theory, it is conventionally employed to justify the form of the IC criterion (3) by means of the explicit likelihood (2). That formula can, of course, be easily generalized to allow for nonnormality by using an asymptotic development of the Bayes factor (e.g., Hartigan 1983; Phillips, 1996; Phillips and Ploberger, 1996) or by other mechanisms that may be more expressive functions of the whole data distribution (as noted by Ebrahimi et al, 1999, in their discussion of entropy measures in ranking distributions). Also, while Assumption A does not hold for nonstationary panels with a unit root in (1), we expect that all our main results continue to apply in that case under a suitably modified form of Assumption A with convergence rates adjusted for the directions of nonstationarity and stationarity – see Phillips (2007) and Cheng and Phillips (2009, 2012) for related time series model selection cases.

Lag order may also be selected by sequential (general to specific, hereafter GS) t-testing in which case $\hat{k}$ is determined as

$$\text{(5)} \quad \hat{k} = \max\left\{k : \left|t_{\hat{\rho}_k}\right| \geq d \text{ and } |t_{\hat{\rho}_j}| < d \text{ for all } j = k+1, ..., k_{\max}\right\}$$

where $t_{\hat{\rho}_k} = \hat{\rho}_k / se(\hat{\rho}_k)$ and d is the critical value used in the test sequence. This GS testing procedure will be used in simulations later in the paper for comparisons with BIC and its various consistent modifications.

2 Asymptotics of Information Criteria

The maximal log-likelihood in (2) leads to the usual formulation of the BIC criterion as $IC_0(k) = \ln \hat{\sigma}_k^2 + k \ln(nT)/(nT)$ or $IC_0^*(k) = \ln \hat{\sigma}_k^2 + k \ln(nT_k)/(nT_k)$, after adjusting for degrees of freedom. This traditional form of BIC prevents under-estimation as desired but typically overestimates k_0 with considerable probability, as we now discuss.

We start with two useful preliminary lemmas that lead to Theorem 1 below. These results hold as $nT \to \infty$, covering cases of fixed T and $T \to \infty$. Proofs of these lemmas and the subsequent

[2] It is sufficient for our main result that the following CLT

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{t=k_0+1}^{k} \left(\varepsilon_{it}^2 - \mathbb{E}\varepsilon_{it}^2\right) \Rightarrow N\left(0, \sum_{t=k_0+1}^{k} \lim \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(\varepsilon_{it}^2 - \mathbb{E}\varepsilon_{it}^2\right)^2\right)$$

holds for all fixed k which is so by virtue of independence over i and the existence of fourth moments of ε_{it}.

theorems are given in the Appendix.

Lemma 1. *For $k_0 \geq 1$ and $k < k_0$, $\operatorname{plim}_{nT\to\infty}(\hat{\sigma}_k^2 - \hat{\sigma}_{k_0}^2) > 0$.*

Lemma 2. *(i) For $k > k_0$ and T fixed as $n \to \infty$,*

$$\sqrt{n}T_k(\hat{\sigma}_k^2 - \hat{\sigma}_{k_0}^2) \Rightarrow N\left(0, 2\sigma^4 (k-k_0)\left(1 + \frac{k-k_0}{T_k}\right)\right).$$

(ii) For $k > k_0$, as $n, T \to \infty$, $\sqrt{n}T_k(\hat{\sigma}_k^2 - \hat{\sigma}_{k_0}^2) \Rightarrow N\left(0, 2\sigma^4 (k-k_0)\right)$.

The variance expressions in these limit distributions of $\sqrt{n}T_k(\hat{\sigma}_k^2 - \hat{\sigma}_{k_0}^2)$ hold when $\varepsilon_{it} \sim iid\ N(0, \sigma^2)$ and clearly need adjustment in heterogeneous and non-normal error cases.

Theorem 1 (Inconsistency of BIC). *Let $\hat{k} = \arg\min_{0\leq k\leq k_{\max}} IC_0(k)$. Then, as $n \to \infty$ and provided $\frac{\ln(nT)}{\sqrt{n}} \to 0$,*

(i) $P\{\hat{k} < k_0\} \to 0$,

(ii) $P\{\hat{k} > k_0\} \to 0.5$.

The heuristics of Theorem 1 are as follows. By virtue of the central limit theory of Lemma 2, $\sqrt{n}T_k \ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2) \sim \sqrt{n}T_k(\hat{\sigma}_k^2 - \hat{\sigma}_{k_0}^2)/\sigma^2$ which converges weakly to a centered Gaussian distribution, whereas $\sqrt{n}T\ln(nT)/(nT) \to 0$ as $n \to \infty$ for any T satisfying the condition $\frac{\ln(nT)}{\sqrt{n}} \to 0$. Thus for $k > k_0$, $\sqrt{n}T[IC_0(k) - IC_0(k_0)]$ converges to a centered normal distribution for which there will be an asymptotic 50% chance that $IC_0(k) < IC_0(k_0)$ as $n \to \infty$. In effect, the probability of overestimation can be as large as 50% as $n \to \infty$. The underlying reason for the overestimation is that, when $k > k_0$, the residual variance estimates $\hat{\sigma}_k^2$ and $\hat{\sigma}_{k_0}^2$ can contain many terms that are mutually independent. In particular, $\hat{\sigma}_{k_0}^2$ contains innovations that relate to $t = k_0+1, ..., k$ none of which enter the formula for $\hat{\sigma}_k^2$. In a panel model, there are a total of $n(k-k_0)$ of such terms (as compared with $k - k_0$ such terms in a simple time series model[3]), which is comparable in magnitude to nT unless $T \to \infty$. In consequence, $\ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2) \sim (\hat{\sigma}_k^2 - \hat{\sigma}_{k_0}^2)/\hat{\sigma}_{k_0}^2 = O_p\left(n^{-1/2}T\right)$ rather than $O_p(n^{-1}T)$. The result is that the order of the BIC penalty term $\ln(nT)/(nT)$ is dominated by $\ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2)$ as $n \to \infty$ for $k > k_0$ and the BIC penalty term does not prevent overestimation. Note that a degrees of freedom adjustment in the penalty does not change this outcome.

[3]Note that when $k_{\max} \to \infty$ the number of such terms potentially becomes large in a time series setting.

There are two obvious solutions to correct the criteria and avoid the problem of overestimation. First, the penalty can be adjusted so that it decreases slowly enough to dominate $\ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2)$ for $k > k_0$ as n increases. Since $\ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2)$ is of order $\sqrt{n}T$, we may correspondingly adjust the penalty to $\sqrt{n}T\ln(\sqrt{n}T)$. This adjustment is designed to deal with the difficulty explained in the preceding paragraph.

A second solution is to truncate the sample so that both $\hat{\sigma}_k^2$ and $\hat{\sigma}_{k_0}^2$ are computed using the same observations. That is, for all k we estimate β_k and σ^2 using $t = k_{\max}+1,\ldots,T$ (instead of using $t = k+1,\ldots,T$). Let these estimates be denoted by $(\tilde{\beta}_k, \tilde{\sigma}_k^2)$ and $(\tilde{\beta}_{k_0}, \tilde{\sigma}_{k_0}^2)$, i.e., $\tilde{\sigma}_k^2 = (nT_*)^{-1}\sum_{i=1}^n\sum_{t=k_{\max}+1}^T \tilde{\varepsilon}_{k,it}^2$, where $T_* = T - k_{\max}$, $\tilde{\varepsilon}_{k,it} = y_{it} - X_{k,it}'\tilde{\beta}_k$ and $\tilde{\beta}_k = (\sum_{i=1}^n\sum_{t=k_{\max}+1}^T X_{k,it}X_{k,it}')^{-1}(\sum_{i=1}^n\sum_{t=k_{\max}+1}^T X_{k,it}y_{it})$ for all k. While the original BIC criterion is inconsistent and overestimates k_0 frequently, these modified BIC criteria are designed to produce consistent lag order estimators, as we now demonstrate.

To fix ideas suppose IC_0 is the original panel BIC criterion and let IC_1 use $\sqrt{n}T\ln(\sqrt{n}T)$ as the penalty, and IC_2 truncate the data so that observations for $t = k_{\max}+1,\ldots,T$ are used in the regressions for all k. Define

$$\begin{aligned} IC_1(k) &= \ln\hat{\sigma}_k^2 + k\ln(\sqrt{n}T_k)/(\sqrt{n}T_k), \\ IC_2(k) &= \ln\tilde{\sigma}_k^2 + k\ln(nT_*)/(nT_*), \text{ where } T_* = T - k_{\max}. \end{aligned}$$

It is asymptotically unimportant, but we may also use the correct degrees of freedom for the computation of $\hat{\sigma}_k^2$ and $\tilde{\sigma}_k^2$ by using the standardizations $nT_k - k - 1$ and $nT_* - k - 1$, respectively, in these estimates. Let $\hat{k}_{(j)} = \arg\min_{0\le k\le k_{\max}} IC_j(k)$ for some given $k_{\max} \ge k_0$ and $j = 1, 2$. Both IC_1 and IC_2 are consistent.

Theorem 2 (Consistency of Modified BIC). *Under Assumption A,* $\mathbb{P}\{\hat{k}_{(j)} = k_0\} \to 1$ as $nT_* \to \infty$ for $j = 1, 2$.

Some remarks and discussion of this result now follow.

Remark 1 (Local to zero coefficients). It is well known that model selection criteria are blind to local alternatives (Phillips and Ploberger, 2003; Leeb and Pötscher, 2005). Hence, in stationary time series models with sample size T, BIC is unable to identify the correct lag order if ρ_{k_0} is in an $O(T^{-1/2})$ neighborhood of zero. For example, when $k_0 = 1$ and $\rho_1 = O(T^{-1/2})$ in the model above, we have $\hat{\sigma}_k^2 = \frac{1}{T}\sum_{t=1}^T \varepsilon_t^2 + O_p(T^{-1})$ for both $k = 0$ and $k = 1$ so that $\ln\left(\hat{\sigma}_0^2/\hat{\sigma}_1^2\right) = O_p\left(T^{-1}\right)$. This variance ratio fails to dominate the penalty $(\ln T)/T$ and so BIC

systematically under-estimates the lag order. For panel data information accumulates with n, and eventually the probability of under-estimation diminishes to zero for every T as $n \to \infty$. But when the autoregressive parameter is close to zero, the cross-sectional dimension n required to avoid under-estimation with reasonable probability can be impractically large, especially for IC_1 as the following remark discusses.

Remark 2 (Small-sample performance of IC1). For an AR(1), IC_1 can under-estimate the lag order with high probability compared to IC_0 or IC_2 when the autoregressive parameter (ρ) is close to zero. Because $\hat{\sigma}_1^2 = \sigma^2 + O_p(\frac{1}{\sqrt{nT}})$ and $\hat{\sigma}_0^2 = \frac{\sigma^2}{1-\rho^2} + O_p(\frac{1}{\sqrt{nT}})$, we have $IC_1(1) - IC_1(0) = \ln(1-\rho^2) + \frac{\ln\sqrt{nT}}{\sqrt{nT}} + O_p(\frac{1}{\sqrt{nT}})$. So, loosely speaking, n and T should be such that $\frac{\ln\sqrt{nT}}{\sqrt{nT}} < -\ln(1-\rho^2)$ in order to avoid under-estimation with non-trivial probability. For example, if $\rho = 0.1$ (so $-\ln(1-\rho^2) \simeq \rho^2 = 0.01$), then $\sqrt{nT}$ needs to be at least 644. For $T = 10$, this means that n should be at least as large as 4200. According to simulations, even for $n = 5000$ and $T = 10$, $\frac{\ln\sqrt{nT}}{\sqrt{nT}} \simeq 0.0093$ and the probability of under-estimation is still about 50%. (With $n = 10,000$ and $T = 10$, $\frac{\ln\sqrt{nT}}{\sqrt{nT}} \simeq 0.007$ and the probability of under-estimation by IC_1 falls to about 5%.) This is because $\frac{\ln\sqrt{nT}}{\sqrt{nT}}$ decreases very slowly as n increases while the variance ratio is distributed around a value close to unity when $\rho \simeq 0$. When the true parameter is $\rho = 0.05$, in order to expect performance of IC_1 similar to the case $n = 4200$, $T = 10$ and $\rho = 0.1$, we would need n to be larger than 100,000 (with $T = 10$)!

Remark 3 (Impact of over-estimation). Under-estimation is usually considered more problematic than over-estimation because under-estimation causes inconsistency. Theorem 1 indicates that IC_0 does not under-estimate lag length asymptotically. Thus, some practitioners may be comfortable using IC_0 in practice. On the other hand, we lose nk observations for an AR(k) specification and the efficiency loss due to unnecessarily large k can be substantial especially in short panels.

Remark 4 (The unit root case). Suppose that $y_{it} = y_{it-1} + \varepsilon_{it}$ and $n, T \to \infty$. Then $\hat{\beta}_k - \beta = O_p\left(n^{-1/2}T^{-1}\right)$ for $k = k_0 = 1$ and $\beta_k - \beta = O_p\left(n^{-1/2}T^{-1/2}\right)$ for $k > k_0 = 1$. Also, for both $k = 1, 2$ we find that (c.f., Phillips, 2008)

$$\hat{\sigma}_k^2 = \frac{1}{nT_k}\sum_{i=1}^{n}\sum_{t=k+1}^{T}\hat{\varepsilon}_{it}^2 = \frac{1}{nT_k}\sum_{i=1}^{n}\sum_{t=k+1}^{T}\varepsilon_{it}^2 + O_p\left(\frac{1}{nT}\right).$$

Hence, for $k = 2$, $\ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2) = O_p\left(n^{-1/2}T^{-1}\right)$ as in the proof of Lemma 2. Meanwhile the penalty function for IC_1, $\ln\left(n^{1/2}T\right) / \left(n^{1/2}T\right)$ goes to zero much slower than $\ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2)$. In other

words, $\Pr(\hat{k} > k_0) \to 0$ as $nT \to \infty$. For IC_2, we have $\ln(\tilde{\sigma}_k^2/\tilde{\sigma}_{k_0}^2) = O_p\left(n^{-1}T^{-1}\right)$ as in the proof of theorem 2. But the penalty function for IC_2 is $O\left(\ln\left(nT\right)/\left(nT\right)\right)$. Hence both IC_1 and IC_2 estimate k_0 consistently.

Remark 5 (Models with fixed effects). For panel dynamic models with fixed effects, it is well known that the within-group (WG) estimator is inconsistent and the bias is $O(T^{-1})$. In this case, we expect none of the above methods to work well unless T is large. The WG estimator has downward bias of order $O(T^{-1})$ so the zeros of ρ_j for $j > k_0$ are likely to be estimated by negative numbers of order $1/T$. Thus, for $k > k_0$, there can be $O(T^{-1})$ differences between $\ln\hat{\sigma}_k^2$ and $\ln\hat{\sigma}_{k_0}^2$, while the penalties decrease as $n \to \infty$. Thus, for large n, the penalty may be dominated by the differences in $\ln\hat{\sigma}_k^2$, in which case for any given T the considered information criteria will lead to over-estimation. For the panel AR(1) model, IC_2 asymptotically selects $k_{\max}$ as $n/T \to \infty$. The general-to-specific sequential testing procedure that we explain below behaves similarly. It seems of little interest to analyze the properties of lag selection methods that are based on inconsistent estimators, especially when there are alternative consistent procedures. We can instead use the consistent estimation method based on X-differencing recently proposed in Han, Phillips and Sul (2011, 2012). For other recent work on dynamic panels with fixed effects that utilize the results of the present paper to achieve consistent lag order selection, see Lee and Phillips (2013).

Lag Selection Using Sequential Testing

An obvious alternative approach that avoids the data loss involved in IC_2 is a general-to-specific (GS) sequential modeling procedure. This selection procedure may be implemented in the usual way. The sequence begins by estimating the largest model – the panel AR($k_{\max}$) model for some given $k_{\max}$ – and tests the significance of $\hat{\rho}_{k_{\max}}$. If the null hypothesis that $\rho_{k_{\max}} = 0$ is not rejected at the chosen level, then the panel AR($k_{\max} - 1$) model is fitted and the null hypothesis $\rho_{k_{\max}-1} = 0$ is tested. This sequential process of estimating and testing is continued until the null hypothesis is rejected, and $\hat{k}$ is defined as the largest k value such that the regressor y_{it-k} is significant, as specified in (5). All available time series observations are fully utilized in this process, giving the approach a finite sample advantage over IC_2.

The GS methodology applies conventional statistical tests. If the significance level of the tests is fixed, then the order estimator inevitably allows for a nonzero probability of overestimation. Furthermore, as is typical in sequential tests, this overestimation probability is bigger than the significance level when there are multiple steps in the order reductions from $k_{\max}$ because the

probability of false rejection accumulates as k step downs from $k_{\max}$ to $\hat{k}$.

These problems can be mitigated (and overcome at least asymptotically) by letting the level of the test be dependent on the sample size. More precisely, following Bauer, Pötscher and Hackl (1988), we can set the critical value d_{nT} in such a way that (i) $d_{nT} \to \infty$, and (ii) $r_{nT}^{-1} d_{nT} \to 0$ as $n, T \to \infty$, where r_{nT} is the convergence rate of the estimator. (Here, condition (i) prevents over-estimation and condition (ii) prevents underestimation.) The critical value in this case corresponds to the standard normal critical value for the significance level $\alpha_{nT} = 2[1 - \Phi(d_{nT})]$, where $\Phi(\cdot)$ is the standard normal cdf.

The following rule was found to work well in our simulations:

(6) $\quad \alpha_{nT} = \exp\{\ln(p)\sqrt{nT}/10\}, \quad p = 0.25.$

This choice of α_{nT} delivers a nominal size of 25% for $nT = 100$, so under-estimation is prevented at the cost of over-fitting for small samples. Because $\ln p < 0$, we have $\alpha_{nT} \to 0$ as $nT \to \infty$, and the associated critical value $d_{nT} = \Phi^{-1}(1 - \alpha_{nT}/2)$ satisfies Bauer et al.'s (1988) conditions stated above. Note that under a local alternative in which the long run autoregressive coefficient has the form $\rho = \sum_{j=1}^{p} \rho_j = c/\sqrt{T}$, the GS method identifies the true length asymptotically well as long as $n \to \infty$ irrespective of the size of T, which is corroborated in the simulation results that follow.

3 Simulations

We use two data generating processes to examine the finite sample performance of the suggested methods: a panel AR(1) and panel AR(3) specified as follows:

(7) $\quad y_{it} = \sum_{j=1}^{p} \rho_j y_{it-j} + u_{it}, \;\; u_{it} \sim iid\, N(0, 1)\,.$

We discard the first 100 observations to avoid the impact of the initial observation on estimation.

Table 1 reports the simulation results for an AR(1) coefficient of $\rho = 0.1$, which is intentionally small in order to give an exacting test of the procedures. The maximal lag order $k_{\max}$ is set to 2 for this experiment (results for larger values of $k_{\max}$ are reported in Table 3 and 4. We discuss the performance of the BIC criteria first. The first 9 columns show the under-, exact- and over-estimation frequencies of the BIC criteria IC_0, IC_1 and IC_2. Note that the conventional BIC criterion IC_0 estimates the true lag length consistently only when $T \to \infty$ with n fixed. The first four rows in Table 1 corroborate the good performance of IC_0 in this case for small fixed n. All lag selection methods estimate the true lag consistently as $T \to \infty$ but there are differences

in performance for moderate T. When $k_{\max} = 2$ the GS method is marginally superior but the performance of all the other estimators is also good. When T is small and n is larger, the four order estimators show major differences. Notably, IC_0 seriously overestimates the true lag as $n \to \infty$, in some cases by over 40%, corroborating Theorem 1. The finite sample performance of IC_1 is somewhat disappointing even though IC_1 is consistent. In particular, when T is small, IC_1 underestimates the lag length with significant probability as n increases. Only when T is large enough (for example $T = 30$), does the performance of IC_1 substantially improve with very large n, as suggested in Remark 2 to Theorem 1. In contrast, IC_2 performs very well as an order estimator. When either n or T increases, the finite sample performance of IC_2 noticeably improves and by a significant margin.[4]

The last 9 columns in Table 1 show the performance of various versions of the GS method. To highlight the differences, we show the consistent data dependent rule (6) as well as GS order selection applied with fixed critical values at the 5% and 25% levels. Obviously with 5% and 25% significance levels, the over-estimation probability converges to 0.05 and 0.25, respectively. Later we will consider the impact of varying $k_{\max}$ on GS methods with fixed significance levels. Compared to the inconsistency of GS methods based on fixed significance levels, the data dependent rule (6) exhibits its consistent behavior as either n or T increases. In fact, except for a couple of cases, the performance of the data determined GS selector dominates the BIC methods.

Table 2 shows results for the local to zero case where the AR(1) coefficient is set to $1/\sqrt{T}$. As we discussed in Remark 1, all methods fail to identify the true lag length in this case with univariate time series because information criteria are blind to local departures. As Table 2 shows, this behavior is manifest for small n ($n = 5$), where the under-estimation probability approaches one for all methods, especially IC_1. However as n increases, performance improves and for large enough, all of the consistent methods estimate the true lag length with high probability. This simulation evidence corroborates Theorem 1 and the discussion in Remarks 1 and 2.

Table 3 demonstrates the impact of $k_{\max}$ on the performance of both BIC and the GS methods. Somewhat surprisingly, the finite sample performance of the GS data dependent rule is little affected by the larger maximum lag length. However, the performance of IC_2 is more seriously influenced, especially with small n and T. This outcome is explained by the fact that IC_2 suffers a loss of an additional $4n$ observations when $k_{\max} = 6$ comparing to when $k_{\max} = 2$. Nonetheless,

[4] A referee suggested the Hannan-Quinn (1979, HQ hereafter) penalty function $\ln(\ln\sqrt{n}T_k) / (\sqrt{n}T_k)$ instead of IC_1. The HQ penalty is much weaker than IC_1. We examined their respective finite sample performance and found that the HQ criteria performs better than IC_1 only when n is large. Moreover, as discussed shortly, IC_2 outperforms IC_1 and the HQ criteria. Hence, the finite sample performance of the HQ criterion is not reported here.

both under-estimation and over-estimation rates go to zero quickly as n or T increases. On the other hand, the GS selector with fixed significance levels is heavily dependent on the choice of $k_{\max}$ and, as $k_{\max}$ increases, the probability of over-estimation increases.

Table 4 considers the AR(3) model with $\rho_1 = \rho_2 = \rho_3 = 0.1$ and $k_{\max} = 6$. Apparently, the finite sample performances worsen as the true lag length increases. This holds for all methods and comparisons among the methods is not clear cut for small n and T. However as n or T increases, both IC_2 and the GS data dependent selector work well.

Table 5 shows the impact of a unit root on the performance of both BIC and GS methods for models with fixed effects. In the experiment here we consider only the consistent X-differencing estimator. For when $\rho < 1$, we found that the finite sample performance of IC_2 and the GS selector using X-differencing is similar to that of pooled OLS estimator without fixed effects. Hence we do not report results for the stationary case. And we report results only for the data dependent GS selector in view of its better performance. Interestingly the over-estimation probabilities of IC_2 are much higher than those of GS. However as T increases, the over-estimation probabilities of IC_2 gradually decrease to zero.

Table 6 reports the impact of lag selection on panel estimation bias and variance of the X-differenced estimator when $\rho = 1$. As Table 5 reveals, the under-estimation probability of all methods goes to zero quickly as T increases. Correspondingly, the evidence in Table 6 confirms that the bias of the X-differencing estimator also tends to zero as T increases. However for small T, the biases arising from estimation based on IC_1 and IC_2 model selection are larger in absolute value than those based on IC_0 selection. Overall the data based GS selection leads to estimation with the minimum bias. As noted in Remark 2, over-estimation affects variance. Since the over-estimation probability under GS selection is smallest, coefficient estimation variance based on GS is correspondingly smallest. The main differences arise for small T. For moderate values of T there is little difference in either estimation bias or variance among the procedures.

While these simulations cover a range of interesting alternative models and procedures, the results in this section apply only to the considered data generating processes and further studies are warranted for a more thorough comparison.

4 Concluding Remarks

Practical empirical work with dynamic panel models relies on the choice of lag order in the dynamics. Test outcomes, consistency, and estimation efficiency are all likely to be dependent on

correct lag length selection. While it is well known that the presence of incidental parameters like fixed effects and incidental trends disturb model selection procedures and can lead to inconsistencies in order estimation, the present paper shows that these difficulties also arise in the absence of such effects. In particular, application of the conventional BIC selection criterion in dynamic panels with no intercepts yields inconsistent lag order selection and typically leads to considerable overestimation of lag order. This result may be surprising to many, given that received wisdom has primarily focused on the obstacles posed by fixed effects and other incidental parameters in dynamic panel estimation. The reason for the failure of BIC even in simple dynamic models with no fixed effects is that the BIC penalty is too small to compensate for the additional terms from cross section averaging that enter into the model fit comparison $\ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2)$ when $k > k_0$ in the BIC criterion, producing a strong tendency to overfit the panel autoregression as $n \to \infty$.

To address the deficiency of BIC, three alternative lag selection methods are suggested here, each of which is consistent. The first two methods modify BIC by increasing the penalty and by adjusting the sample fit comparisons so that they are homogeneous in the sample observations used by means of sample truncation. The final method involves GS sequential testing and our suggested procedure involves a data-determined critical value that ensures consistent order selection. Simulation findings indicate that modified BIC using sample truncation and data-determined GS lag order selection both perform well in finite samples for a range of different sample sizes (n, T), including cases with small T, and models with a unit root.

References

Bauer, P., Pötscher, B. M., Hackl, P. (1988). Model selection by multiple test procedures. *Statistics* 19:39–44.

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

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

Cheng, X., Phillips, P. C. B. (2009). Semiparametric cointegrating rank selection. *The Econometrics Journal* 12:S83–S104.

Cheng, X., Phillips, P. C. B. (2012). Cointegrating rank selection in models with time-varying variance. *Journal of Econometrics* 169:155–165.

Ebrahimi, N., E. Maasoumi, Soofi, E. S. (1999). Ordering univariate distributions by entropy and variance. *Journal of Econometrics*, 90, 317-336.

Han, C., Phillips, P. C. B., Sul, D. (2011). Uniform Asymptotic Normality in Stationary and Unit Root Autoregression. *Econometric Theory* 27:1117–1151.

Han, C., Phillips, P. C. B., Sul, D. (2012). X-Differencing and Dynamic Panel Model Estimation. forthcoming in *Econometric Theory*.

Hannan, E. J., and B. G. Quinn (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society*, B, 41, 190-195.

Hartigan, J. A. (1983). *Bayes Theory*. New York: Springer–Verlag.

Lee, Y. (2012). Model selection in the presence of incidental parameters. Unpublished working paper, University of Michigan.

Lee, Y., Phillips, P. C. B. (2013). Model Selection in the Presence of Incidental Parameters. Unpublished working paper, University of Michigan.

Leeb, H., Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* 21:21–59.

Phillips, P. C. B. (1996). "Econometric Model Determination", *Econometrica*, , Vol. 64, No. 4, July 1996, pp. 763-812.

Phillips, P. C. B. (2008). Unit root model selection, *Journal of the Japan Statistical Society* 38:65–74.

Phillips P. C. B. and W. Ploberger (1996). "An Asymptotic Theory of Bayesian Inference for Time Series", *Econometrica*, 64, 381-413.

Ploberger, W. Phillips, P. C. B. (2003). Empirical limits for time series econometric models. *Econometrica* 71:627–673.

Pötscher, B. M. (1983). Order estimation in ARMA-models by Lagrangian Multiplier tests. *Annals of Statistics* 11:872–885.

Schmidt, P. (1971). Estimation of a distributed lag model with second-order autoregressive disturbances: A Monte Carlo experiment. *International Economic Review* 12:372–380.

_____ (1973). On the difference between conditional and unconditional asymptotic distributions of estimates in distributed lag models with integer-valued parameters. *Econometrica* 41:165–169

_____ (1974). A modification of the Almon distributed lag. *Journal of the American Statistical Association* 69:679–681.

Schmidt, P., Sickles, R. (1975). On the efficiency of the Almon lag technique. *International Economic Review*, 16. 792–795.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6, 461-464.

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

A Appendix

We use the notation $T_k = T - k$ for all $k \geq 0$ and $T_* = T - k_{\max}$. Recall that $X_{k,it} = (y_{it-1}, \ldots, y_{it-k})'$, $\hat{\varepsilon}_{k,it} = y_{it} - X_{k,it}'\hat{\beta}_k$ and $\hat{\sigma}_k^2 = \frac{1}{nT_k}\sum_{i=1}^n \sum_{t=k+1}^T \hat{\varepsilon}_{k,it}^2$.

A.1 Inconsistency of IC_0

Proof of Lemma 1. For $k < k_0$ define $\hat{\beta}_k^+ = (\hat{\beta}_k', 0_{k_0-k}')'$ so $X_{k,it}'\hat{\beta}_k = X_{it}'\hat{\beta}_k^+$ for all $t > k$, where $X_{it} = X_{k_0,it}$ for notational brevity. (Note that the identity holds even though some elements of X_{it} are unobservable for $t \leq k_0$.) As $\hat{\varepsilon}_{k,it} = \varepsilon_{it} - (X_{k,it}'\hat{\beta}_k - X_{it}'\beta) = \varepsilon_{it} - X_{it}'(\hat{\beta}_k^+ - \beta)$, we have

$$\hat{\sigma}_k^2 = \frac{1}{nT_k}\sum_{i=1}^n \sum_{t=k+1}^T \varepsilon_{it}^2 - \frac{2}{nT_k}\sum_{i=1}^n \sum_{t=k+1}^T \varepsilon_{it} X_{it}'(\hat{\beta}_k^+ - \beta) + (\hat{\beta}_k^+ - \beta)'\hat{Q}_k(\hat{\beta}_k^+ - \beta). \tag{8}$$

The first term converges in probability to σ^2 as $nT_k \to \infty$, the second term is $O_p(1/\sqrt{nT_k})$ by Assumption A because $\hat{\beta}_k^+ - \beta$ is stochastically bounded, and the third term is asymptotically strictly positive because $\operatorname{plim}\hat{\beta}_k^+ \neq \beta$ (since $\rho_{k_0} \neq 0$ by assumption) and $\hat{Q}_k$ is asymptotically nonsingular. The stated result then holds as $nT \to \infty$, and in particular as $n \to \infty$ for both fixed T and as $T \to \infty$. ■

Proof of Lemma 2. (i): For all $k \geq k_0$ we have $\hat{\varepsilon}_{k,it} = \varepsilon_{it} - X_{it}'(\hat{\beta}_k - \beta_k)$ so that

$$\hat{\sigma}_k^2 = \frac{1}{nT_k}\sum_{i=1}^n \sum_{t=k+1}^T \varepsilon_{it}^2 + (\hat{\beta}_k - \beta_k)'Q_k(\hat{\beta}_k - \beta_k) - \frac{2}{nT_k}\sum_{i=1}^n \sum_{t=k+1}^T \varepsilon_{it} X_{k,it}'(\hat{\beta}_k - \beta_k), \tag{9}$$

where $Q_k = (nT_k)^{-1}\sum_{i=1}^n \sum_{t=k+1}^T X_{k,it}X_{k,it}'$. When $\hat{\beta}_k - \beta_k = O_p(n^{-1/2}T_k^{-1/2})$, the second and third terms are $O_p(n^{-1}T_k^{-1})$, and thus

$$\begin{aligned}
\hat{\sigma}_{k_0}^2 - \hat{\sigma}_k^2 &= \frac{1}{nT_{k_0}}\sum_{i=1}^n \sum_{t=k_0+1}^T \varepsilon_{it}^2 - \frac{1}{nT_k}\sum_{i=1}^n \sum_{t=k+1}^T \varepsilon_{it}^2 + O_p(n^{-1}T_*^{-1}) \\
&= \frac{1}{n}\sum_{i=1}^n \xi_{iT} + O_p(n^{-1}T_*^{-1}), \qquad (10)
\end{aligned}$$

where $T_* = T - k_{\max}$ as before and

$$\begin{aligned}
\xi_{iT} &= \frac{1}{T_{k_0}}\sum_{t=k_0+1}^T \varepsilon_{it}^2 - \frac{1}{T_k}\sum_{t=k+1}^T \varepsilon_{it}^2 \\
&= \frac{1}{T_{k_0}}\sum_{t=k_0+1}^k \varepsilon_{it}^2 + \frac{T_k - T_{k_0}}{T_{k_0}T_k}\sum_{t=k+1}^T \varepsilon_{it}^2 \\
&= \frac{1}{T_{k_0}}\sum_{t=k_0+1}^k \left(\varepsilon_{it}^2 - \sigma^2\right) - \frac{k - k_0}{T_{k_0}T_k}\sum_{t=k+1}^T \left(\varepsilon_{it}^2 - \sigma^2\right) \qquad (11)
\end{aligned}$$

The mean of ξ_{iT} is zero and the variance of $n^{-1}\sum_{i=1}^{n}\xi_{iT}$ is

$$\begin{aligned}\frac{1}{n}\mathbb{E}\xi_{iT}^2 &= \frac{(k-k_0)}{nT_{k_0}^2}\text{var}\left(\varepsilon_{it}^2\right)+\left(\frac{k-k_0}{T_{k_0}T_k}\right)^2\frac{T_k}{n}\text{var}\left(\varepsilon_{it}^2\right)\\ &= \frac{(k-k_0)+T_k^{-1}\left(k-k_0\right)^2}{nT_{k_0}^2}\text{var}\left(\varepsilon_{it}^2\right)\\ &= \frac{(k-k_0)}{nT_{k_0}^2}\text{var}\left(\varepsilon_{it}^2\right)\left(1+\frac{k-k_0}{T_k}\right)\end{aligned}$$

which shows that $\sqrt{n}T(\hat{\sigma}_{k_0}^2-\hat{\sigma}_k^2)=O_p\left(1\right)$. The result holds as $n\to\infty$ for both fixed T and as $T\to\infty$. Next, using (10), (11), and standard central limit arguments as $n\to\infty$ with T fixed

$$\begin{aligned}\sqrt{n}T_{k_0}(\hat{\sigma}_{k_0}^2-\hat{\sigma}_k^2) &= \frac{T_{k_0}}{\sqrt{n}}\sum_{i=1}^{n}\xi_{iT}+O_p(n^{-1/2})\\ &\Rightarrow\ \zeta_{k-k_0,T}=_d N\left(0,2\sigma^4\left(k-k_0\right)\left\{1+\frac{k-k_0}{T_k}\right\}\right),\end{aligned}\tag{12}$$

giving (i) as $n\to\infty$. When $n\to\infty$ and $T\to\infty$ we have

$$\begin{aligned}\sqrt{n}T(\hat{\sigma}_{k_0}^2-\hat{\sigma}_k^2) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{t=k_0+1}^{k}\left(\varepsilon_{it}^2-\sigma^2\right)+o_p\left(1\right)\\ &\Rightarrow\ \zeta_{k-k_0}=_d N\left(0,2\sigma^4\left(k-k_0\right)\right),\end{aligned}$$

giving (ii). ■

Proof of Theorem 1. (i): This follows by Lemma 1 and the fact that $\ln(1+x)>0$ for all $x>0$. Thus,

$$\begin{aligned}IC_0(k)-IC_0(k_0) &= \ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2)+(k-k_0)\left(\ln nT\right)/(nT)\\ &= \ln\left(1+\frac{\hat{\sigma}_k^2-\hat{\sigma}_{k_0}^2}{\hat{\sigma}_{k_0}^2}\right)+O\left(\frac{\ln nT}{nT}\right)\end{aligned}$$

so that $\mathbb{P}(\hat{k}<k_0)\to 0$ as $n\to\infty$ for both fixed T and as $T\to\infty$.

(ii): For $k>k_0$, we have $\hat{\sigma}_k^2-\hat{\sigma}_{k_0}^2=o_p\left(1\right)$ and

$$\begin{aligned}\sqrt{n}T[IC_0(k)-IC_0(k_0)] &= \sqrt{n}T\ln\left\{1+\frac{\hat{\sigma}_k^2-\hat{\sigma}_{k_0}^2}{\hat{\sigma}_{k_0}^2}\right\}+(k-k_0)\frac{\ln\left(nT\right)}{\sqrt{n}}\\ &= \hat{A}_{nT}\left\{1+o_p(1)\right\}+(k-k_0)\frac{\ln\left(nT\right)}{\sqrt{n}},\end{aligned}$$

where $\hat{A}_{nT} = \sqrt{n}T(\hat{\sigma}_k^2 - \hat{\sigma}_{k_0}^2)/\sigma^2 = O_p(1)$ by virtue of Lemma 2. Further, by (12) and Lemma 2 we deduce that

$$\begin{aligned}\hat{A}_{nT} &= -\frac{T}{\sqrt{n}\sigma^2}\sum_{i=1}^{n}\xi_{iT} + O_p(n^{-1/2}) \\ &\Rightarrow \begin{cases} N\left(0, 2(k-k_0)\left\{1+\frac{k-k_0}{T_k}\right\}\right) := \zeta_{k-k_0,T} & \text{for } T \text{ fixed} \\ N(0, 2(k-k_0)) := \zeta_{k-k_0} & \text{when } T \to \infty. \end{cases}\end{aligned}$$

Thus, provided $\ln(nT)/\sqrt{n} \to 0$ we have

$$\sqrt{n}T[IC_0(k) - IC_0(k_0)] \Rightarrow \zeta^* = \zeta_{k-k_0,T}1_{T \text{ fixed}} + \zeta_{k-k_0}1_{T\to\infty},$$

and then

$$\mathbb{P}\{IC_0(k) < IC_0(k_0)\} \to \mathbb{P}\{\zeta^* < 0\} = 0.5.$$

This implies that $\lim P\{\hat{k} > k_0\} > 0$ and proves the stated result for both fixed T and $T \to \infty$ provided $\frac{\ln(nT)}{\sqrt{n}} \to 0$. On the other hand, if $(\ln nT)/\sqrt{n} \to \infty$ or equivalently $e^{\sqrt{n}}/T \to 0$, then $\mathbb{P}\{IC_0(k) > IC_0(k_0)\} \to 1$, implying that $\mathbb{P}\{\hat{k} > k_0\} \to 0$. Thus, BIC is consistent only if T tends to infinity extremely rapidly relative to n. ■

A.2 Consistency of IC_1 and IC_2

Proof of Theorem 2. Lemma 1 continues to apply for $j = 1$. With minor adjustments to the proof of Lemma 1, we find that for $j = 2$, $k_0 \geq 1$, and $k < k_0$ we have $\operatorname{plim}_{nT_*\to\infty}(\tilde{\sigma}_k^2 - \tilde{\sigma}_{k_0}^2) > 0$. It therefore suffices to show that $\mathbb{P}\{IC_j(k) > IC_j(k_0)\} \to 1$ for $j = 1, 2$ when $k > k_0$. For $j = 1$, and $k > k_0$, we find by virtue of the proof of Lemma 2 that

$$\sqrt{n}T[IC_1(k) - IC_1(k_0)] = \hat{A}_{nT}\{1 + o_p(1)\} + (k - k_0)\ln(nT) \to \infty,$$

as $nT \to \infty$ so that $\mathbb{P}\{IC_1(k) > IC_1(k_0)\} \to 1$ for $k > k_0$. In a similar fashion we have

$$nT_*[IC_2(k) - IC_2(k_0)] = \tilde{A}_{nT}\{1 + o_p(1)\} + (k - k_0)\ln(nT_*),$$

where $\tilde{A}_{nT} = nT_*(\tilde{\sigma}_k^2 - \tilde{\sigma}_{k_0}^2)/\sigma_{k_0}^2$, for $k > k_0$. Now, proceeding as in the proof of Lemma 2, we find that

$$nT_*(\tilde{\sigma}_k^2 - \tilde{\sigma}_{k_0}^2) = nT_*\left\{\frac{1}{nT_*}\sum_{i=1}^{n}\sum_{t=k_{\max}+1}^{T}\varepsilon_{it}^2 - \frac{1}{nT_*}\sum_{i=1}^{n}\sum_{t=k_{\max}+1}^{T}\varepsilon_{it}^2\right\} + O_p(1) = O_p(1).$$

Hence

$$nT_*[IC_2(k) - IC_2(k_0)] = O_p(1) + (k - k_0)\ln(nT_*) \to \infty,$$

from which it follows that $\mathbb{P}\{IC_2(k) > IC_2(k_0)\} \to 1$, giving the required result. ■

Table 1: Finite Sample Performance of BIC, Modified BIC, and GS with no Fixed Effects under Fixed Alternative:
(AR(1), $\rho = 0.1$, kmax = 2)

		IC0			IC1			IC2			GS			GS with 5%			GS with 25%		
N	T	k<1	k=1	k>1	k<1	k=1	k>1	k<1	k=1	k>1	k<1	k=1	k>1	k<1	k=1	k>1	k<1	k=1	k>1
5	50	72.5	22.5	5.1	93.6	6.2	0.3	80.8	18.3	1.0	47.3	40.3	12.4	63.5	30.4	6.2	26.0	48.0	26.0
5	100	58.6	36.4	5.1	86.8	13.1	0.2	63.6	35.8	0.6	41.1	53.6	5.4	39.3	55.1	5.7	12.4	61.4	26.3
5	200	34.0	60.8	5.3	71.3	28.6	0.2	32.7	66.7	0.7	27.5	71.5	1.1	12.6	82.5	5.0	1.7	73.3	25.0
5	1000	0.0	96.5	3.5	0.3	99.7	0.1	0.0	99.9	0.2	0.1	99.9	0.0	0.0	95.9	4.1	0.0	77.1	23.0
50	5	60.6	26.7	12.8	98.8	1.2	0.0	93.3	6.7	0.1	52.7	36.4	11.0	69.9	25.0	5.1	29.5	44.2	26.4
50	10	50.4	32.6	17.0	98.5	1.5	0.0	73.3	26.5	0.3	43.6	51.9	4.6	41.5	53.6	5.0	12.8	61.2	26.1
50	20	37.7	43.8	18.6	97.6	2.4	0.0	36.8	62.9	0.4	28.7	70.3	1.1	13.2	81.3	5.5	2.6	71.6	25.9
50	30	23.3	55.7	21.1	96.5	3.6	0.0	16.6	83.1	0.4	17.2	82.5	0.4	3.5	91.6	5.0	0.2	73.4	26.4
100	5	52.4	31.4	16.3	99.8	0.2	0.0	87.9	12.0	0.2	48.8	46.6	4.7	47.5	47.4	5.2	14.6	61.0	24.5
100	10	39.4	38.6	22.1	99.4	0.6	0.0	46.6	53.0	0.5	31.6	67.2	1.3	14.6	80.7	4.8	3.1	72.7	24.3
100	20	19.6	54.7	25.7	97.6	2.4	0.0	8.2	91.3	0.6	10.6	89.0	0.5	0.8	93.6	5.7	0.1	76.6	23.3
100	30	8.8	63.1	28.1	94.5	5.6	0.0	0.8	98.8	0.4	3.3	96.7	0.1	0.0	95.2	4.9	0.0	75.5	24.5
200	5	43.8	35.7	20.5	100.0	0.1	0.0	70.6	29.3	0.1	36.5	62.5	1.0	18.2	76.9	5.0	3.5	72.8	23.7
200	10	25.7	45.6	28.7	99.4	0.6	0.0	13.7	85.8	0.6	11.4	88.3	0.3	0.9	93.7	5.5	0.1	76.2	23.8
200	20	6.7	60.0	33.4	96.6	3.4	0.0	0.0	99.6	0.4	0.7	99.3	0.0	0.0	94.9	5.2	0.0	75.2	24.9
200	30	1.5	66.4	32.2	87.8	12.2	0.0	0.0	99.7	0.4	0.1	99.9	0.0	0.0	95.5	4.5	0.0	75.3	24.7
1000	5	17.5	50.6	32.0	100.0	0.0	0.0	1.9	97.9	0.3	1.3	98.7	0.1	0.0	93.8	6.2	0.0	74.8	25.3
1000	10	2.0	57.6	40.5	98.8	1.2	0.0	0.0	99.9	0.1	0.0	100.0	0.0	0.0	94.5	5.6	0.0	75.7	24.4
1000	20	0.1	58.8	41.2	67.2	32.9	0.0	0.0	99.9	0.1	0.0	100.0	0.0	0.0	95.3	4.8	0.0	74.4	25.6
1000	30	0.0	60.2	39.9	9.8	90.2	0.0	0.0	99.9	0.1	0.0	100.0	0.0	0.0	95.4	4.6	0.0	75.8	24.3

Table 2: Finite Sample Performance of BIC, Modified BIC, and GS with no Fixed Effects under Local to Zero:
(AR(1), $\rho = 1/\sqrt{T}$, kmax = 2)

		IC0			IC1			IC2			GS			GS with 5%			GS with 25%		
N	*T*	*k*<1	*k*=1	*k*>1	*k*<1	*k*=1	*k*>1	*k*<1	*k*=1	*k*>1	*k*<1	*k*=1	*k*>1	*k*<1	*k*=1	*k*>1	*k*<1	*k*=1	*k*>1
5	50	54.0	39.2	6.8	82.9	16.7	0.5	59.6	39.2	1.3	25.5	62.2	12.4	39.9	54.0	6.2	11.9	62.0	26.2
5	100	58.6	36.4	5.1	86.8	13.1	0.2	63.6	35.8	0.6	41.1	53.6	5.4	39.3	55.1	5.7	12.4	61.4	26.3
5	200	65.0	31.3	3.8	91.5	8.6	0.0	69.0	30.6	0.5	63.4	35.7	1.0	40.2	55.0	4.9	11.5	63.7	24.9
5	1000	71.7	27.0	1.3	95.8	4.3	0.0	75.6	24.3	0.1	95.8	4.2	0.0	38.4	57.2	4.5	12.0	65.4	22.7
50	5	0.3	79.1	20.6	15.0	84.9	0.2	0.2	99.2	0.7	0.0	88.6	11.5	0.0	94.8	5.2	0.0	74.6	25.4
50	10	0.2	75.4	24.5	16.5	83.4	0.2	0.0	99.2	0.9	0.0	95.6	4.5	0.0	95.0	5.0	0.0	74.3	25.7
50	20	0.3	75.9	23.9	23.0	77.0	0.1	0.0	99.5	0.6	0.0	99.0	1.1	0.0	94.5	5.5	0.0	74.0	26.0
50	30	0.2	74.9	24.9	26.7	73.4	0.0	0.0	99.6	0.5	0.0	99.6	0.4	0.0	95.1	5.0	0.0	73.3	26.7
100	5	0.0	75.2	24.8	1.6	98.5	0.0	0.0	99.4	0.6	0.0	95.4	4.6	0.0	95.2	4.9	0.0	76.8	23.2
100	10	0.0	70.8	29.3	2.0	98.1	0.0	0.0	99.2	0.8	0.0	98.5	1.5	0.0	95.6	4.4	0.0	76.0	24.1
100	20	0.0	70.8	29.2	3.4	96.7	0.0	0.0	99.4	0.7	0.0	99.6	0.4	0.0	94.3	5.8	0.0	77.2	22.9
100	30	0.0	70.7	29.3	3.9	96.1	0.0	0.0	99.4	0.6	0.0	99.9	0.1	0.0	95.2	4.8	0.0	75.6	24.5
200	5	0.0	73.5	26.6	0.0	100.0	0.0	0.0	99.7	0.3	0.0	98.8	1.2	0.0	95.1	5.0	0.0	75.3	24.7
200	10	0.0	66.3	33.7	0.0	100.0	0.0	0.0	99.6	0.4	0.0	99.8	0.3	0.0	95.0	5.0	0.0	75.1	25.0
200	20	0.0	65.8	34.3	0.0	100.0	0.0	0.0	99.5	0.5	0.0	100.0	0.0	0.0	94.9	5.1	0.0	74.8	25.2
200	30	0.0	67.5	32.5	0.1	100.0	0.0	0.0	99.7	0.4	0.0	100.0	0.0	0.0	95.8	4.3	0.0	75.0	25.0
1000	5	0.0	64.8	35.3	0.0	100.0	0.0	0.0	99.8	0.3	0.0	100.0	0.1	0.0	94.2	5.8	0.0	73.1	27.0
1000	10	0.0	59.4	40.7	0.0	100.0	0.0	0.0	99.9	0.2	0.0	100.0	0.0	0.0	94.3	5.8	0.0	75.8	24.2
1000	20	0.0	58.9	41.2	0.0	100.0	0.0	0.0	99.9	0.1	0.0	100.0	0.0	0.0	95.0	5.0	0.0	75.3	24.7
1000	30	0.0	60.0	40.0	0.0	100.0	0.0	0.0	99.9	0.1	0.0	100.0	0.0	0.0	95.8	4.3	0.0	76.1	24.0

Table 3: Role of Kmax Value on Lag Selection with no Fixed Effects under Fixed Alternative:
(AR(1), $\rho = 0.1$, kmax = 6)

		IC0			IC1			IC2			GS			GS with 5%			GS with 25%		
N	*T*	*k*<1	*k*=1	*k*>1	*k*<1	*k*=1	*k*>1	*k*<1	*k*=1	*k*>1	*k*<1	*k*=1	*k*>1	*k*<1	*k*=1	*k*>1	*k*<1	*k*=1	*k*>1
5	50	71.5	21.7	6.8	93.6	6.2	0.3	81.9	17.0	1.2	29.8	24.9	45.4	51.9	25.0	23.2	8.2	16.5	75.4
5	100	58.3	36.0	5.8	86.8	13.1	0.2	65.4	33.8	0.9	33.9	44.8	21.3	31.7	45.0	23.4	3.7	19.7	76.7
5	200	33.7	60.5	5.9	71.3	28.6	0.2	33.7	65.5	0.8	26.2	68.2	5.7	10.6	67.6	21.9	0.8	22.4	76.8
5	1000	0.0	95.7	4.3	0.3	99.7	0.1	0.0	99.8	0.3	0.1	99.8	0.2	0.0	77.8	22.3	0.0	21.6	78.4
50	10	45.4	27.9	26.8	98.5	1.5	0.0	88.4	11.4	0.3	35.3	42.9	21.9	32.3	43.6	24.2	3.6	18.1	78.3
50	20	35.0	37.5	27.6	97.6	2.4	0.0	51.0	48.7	0.4	27.2	66.2	6.7	10.4	65.7	24.0	0.8	22.8	76.4
50	30	21.3	48.4	30.4	96.5	3.6	0.0	22.9	76.7	0.5	16.9	81.3	1.9	2.9	74.4	22.8	0.0	22.5	77.5
100	10	34.4	32.3	33.3	99.4	0.6	0.0	79.8	19.7	0.6	30.0	63.7	6.4	11.8	65.1	23.2	0.8	22.4	76.9
100	20	17.4	45.3	37.3	97.6	2.4	0.0	18.2	81.2	0.7	10.6	88.1	1.4	0.8	77.3	22.0	0.1	24.2	75.8
100	30	7.8	51.5	40.8	94.5	5.6	0.0	2.1	97.3	0.6	3.2	96.7	0.1	0.0	77.7	22.4	0.0	24.7	75.3
200	10	20.6	35.3	44.2	99.4	0.6	0.0	55.3	44.7	0.1	11.4	87.6	1.1	0.7	76.5	22.8	0.0	24.4	75.7
200	20	4.9	44.8	50.3	96.6	3.4	0.0	0.8	98.6	0.7	0.7	99.3	0.0	0.0	77.4	22.7	0.0	23.7	76.3
200	30	1.1	49.4	49.6	87.8	12.2	0.0	0.0	99.6	0.4	0.1	99.9	0.0	0.0	77.9	22.2	0.0	24.6	75.5
1000	10	1.2	35.6	63.3	98.8	1.2	0.0	0.1	99.8	0.1	0.0	100.0	0.0	0.0	76.8	23.2	0.0	25.0	75.1
1000	20	0.0	38.1	62.0	67.2	32.9	0.0	0.0	99.9	0.2	0.0	100.0	0.0	0.0	76.5	23.5	0.0	22.6	77.4
1000	30	0.0	39.6	60.5	9.8	90.2	0.0	0.0	99.9	0.1	0.0	100.0	0.0	0.0	77.1	23.0	0.0	24.4	75.6

Table 4: Role of AR order on Lag Selection with no Fixed Effects under Fixed Alternative:
(AR(3), $\rho_1 = \rho_2 = \rho_3 = 0.1$, kmax = 6)

		IC0			IC1			IC2			GS			GS with 5%			GS with 25%		
N	*T*	*k*<3	*k*=3	*k*>3	*k*<3	*k*=3	*k*>3	*k*<3	*k*=3	*k*>3	*k*<3	*k*=3	*k*>3	*k*<3	*k*=3	*k*>3	*k*<3	*k*=3	*k*>3
5	50	88.6	8.7	2.8	99.3	0.7	0.1	93.7	6.0	0.4	38.8	31.0	30.3	59.3	27.1	13.7	16.4	25.9	57.8
5	100	72.3	23.9	3.9	97.6	2.4	0.1	78.6	21.0	0.5	38.2	48.7	13.1	36.0	49.5	14.5	6.0	35.6	58.4
5	200	39.7	55.3	5.0	84.0	16.0	0.1	36.4	63.0	0.7	25.7	70.6	3.7	10.6	75.9	13.6	0.8	41.0	58.3
5	1000	0.0	96.0	4.1	0.3	99.7	0.1	0.0	99.7	0.4	0.2	99.8	0.1	0.0	85.6	14.5	0.0	39.1	60.9
50	10	71.6	15.4	13.1	100.0	0.0	0.0	97.1	2.8	0.2	47.9	37.9	14.3	46.0	38.3	15.8	10.7	30.9	58.5
50	20	48.0	30.3	21.7	100.0	0.0	0.0	61.4	38.0	0.6	32.6	63.1	4.4	14.6	70.3	15.2	1.9	39.3	58.9
50	30	30.2	45.0	24.9	99.9	0.1	0.0	24.7	75.0	0.4	19.3	79.8	0.9	3.6	82.5	14.0	0.3	41.0	58.7
100	10	56.5	25.2	18.4	100.0	0.0	0.0	90.7	9.3	0.1	44.3	52.2	3.6	20.7	64.6	14.7	2.8	39.5	57.8
100	20	26.3	41.9	31.9	100.0	0.0	0.0	18.3	81.2	0.5	14.9	84.6	0.6	1.5	85.1	13.5	0.1	42.6	57.3
100	30	9.2	53.6	37.3	99.7	0.3	0.0	1.9	97.9	0.3	4.1	95.9	0.1	0.0	86.8	13.2	0.0	41.8	58.2
200	10	38.1	32.5	29.4	100.0	0.0	0.0	61.9	37.8	0.3	25.4	74.2	0.5	3.2	82.9	14.0	0.3	43.6	56.2
200	20	8.8	47.5	43.8	99.9	0.1	0.0	0.6	99.1	0.3	1.8	98.3	0.0	0.0	86.6	13.5	0.0	42.5	57.5
200	30	1.9	53.4	44.8	97.9	2.2	0.0	0.0	99.8	0.2	0.0	100.0	0.0	0.0	86.7	13.4	0.0	43.9	56.2
1000	10	5.7	43.6	50.8	100.0	0.0	0.0	0.2	99.7	0.1	0.1	100.0	0.0	0.0	85.6	14.5	0.0	44.0	56.1
1000	20	0.0	44.6	55.4	91.3	8.7	0.0	0.0	99.8	0.3	0.0	100.0	0.0	0.0	85.1	14.9	0.0	41.7	58.4
1000	30	0.0	43.7	56.3	24.4	75.7	0.0	0.0	99.9	0.2	0.0	100.0	0.0	0.0	85.5	14.5	0.0	41.9	58.2

Table 5: Lag Selection with X-differencing under fixed effects
AR(1), Unit Root Case, kmax = 2, N= 200

	IC0			IC1			IC2			GS		
	k<1	*k=1*	*k>1*	*k<1*	*k=1*	*k>1*	*k<1*	*k=1*	*k>1*	*k<1*	*k=1*	*k>1*
5	1.4	0.0	98.7	53.2	0.0	46.8	0.0	57.0	43.1	0.0	98.3	1.8
6	0.6	1.2	98.3	37.2	15.5	47.4	0.0	61.6	38.4	0.0	99.3	0.8
7	0.0	5.6	94.4	0.1	67.9	32.0	0.0	63.5	36.6	0.0	99.6	0.4
8	0.0	10.7	89.4	0.0	82.3	17.7	0.0	62.1	38.0	0.0	99.6	0.4
10	0.0	21.2	78.9	0.0	94.1	5.9	0.0	64.7	35.3	0.0	99.8	0.3
20	0.0	42.5	57.6	0.0	99.9	0.2	0.0	72.7	27.4	0.0	100.0	0.0
30	0.0	52.3	47.7	0.0	100.0	0.0	0.0	78.6	21.4	0.0	100.0	0.0
50	0.0	59.3	40.8	0.0	100.0	0.0	0.0	84.2	15.8	0.0	100.0	0.0

Table 6: Impact of Lag Selection on Biases and Variances
with X-differencing under fixed effects
AR(1), Unit Root Case, kmax = 2, N= 200.

	Bias				Variance			
	IC0	IC1	IC2	GS	IC0	IC1	IC2	GS
5	-0.0124	-0.5362	-0.0531	-0.0003	3.2973	25.3016	0.8830	0.6279
6	-0.0067	-0.3800	-0.0275	-0.0003	1.2215	23.0854	0.3586	0.2684
7	-0.0022	-0.0098	-0.0168	0.0003	0.3007	0.2953	0.1996	0.1604
8	-0.0054	-0.0061	-0.0138	-0.0008	0.1772	0.1333	0.1379	0.1160
10	-0.0048	-0.0024	-0.0087	-0.0011	0.0855	0.0684	0.0763	0.0668
20	-0.0019	-0.0007	-0.0026	-0.0007	0.0146	0.0134	0.0143	0.0134
30	-0.0009	-0.0004	-0.0012	-0.0004	0.0059	0.0057	0.0059	0.0057
50	-0.0004	-0.0002	-0.0005	-0.0002	0.0020	0.0019	0.0019	0.0019