Article

Consistency in Estimation and Model Selection of Dynamic Panel Data Models with Fixed Effects

Guangjie Li

Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff CF10 3EU, UK; E-Mail: ligj@cf.ac.uk; Tel.: +44-2920-874833

Academic Editor: Kerry Patterson

Abstract: We examine the relationship between consistent parameter estimation and model selection for autoregressive panel data models with fixed effects. We find that the transformation of fixed effects proposed by Lancaster (2002) does not necessarily lead to consistent estimation of common parameters when some true exogenous regressors are excluded. We propose a data dependent way to specify the prior of the autoregressive coefficient and argue for comparing different model specifications before parameter estimation. Model selection properties of Bayes factors and Bayesian information criterion (BIC) are investigated. When model uncertainty is substantial, we recommend the use of Bayesian Model Averaging to obtain point estimators with lower root mean squared errors (RMSE). We also study the implications of different levels of inclusion probabilities by simulations.

1. Introduction

For a panel linear regression model with lags of the dependent variable as regressors (dynamic panel model) and agent specific fixed effects, the maximum likelihood estimators (MLE) of the common parameters, whose number does not change with sample size, are inconsistent when the number of time periods is small and fixed, see Nerlove [1] and Nickell [2]. This problem, known as the "incidental parameter problem", has been reviewed by Lancaster [3]. A plethora of studies have been undertaken

to obtain consistent estimators for the common parameters in dynamic panel models. Among them there are two main approaches: one is to use the generalized method of moments (GMM), see the overview in Hsiao [4]; the other is based on modified profile or integrated likelihood, see e.g., the recent works by Bester and Hansen [5], Hahn and Kuersteiner [6], Arellano and Bonhomme [7], Dhaene and Jochmans [8]. Researchers using these two approaches usually presume the moment conditions or the parametric models are correctly specified and the issue of model selection has relatively attracted less attention. Correct model specification is very important, without which consistent parameter estimation can not be achieved. In Andrews and Lu [9], the authors proposed model and moment selection criteria (MMSC) under GMM context based on J-test statistics to address the issue. However, for dynamic panel models, GMM will suffer from the weak instrument problem when the coefficient for the lagged dependent variable is close to 1. Hence MMSC is unlikely to work for such situation.

Lee and Phillips [10] used the bias reducing prior from Arellano and Bonhomme [7] to develop integrated likelihood information criterion to study lag order selection in dynamic panel models. The prior in Arellano and Bonhomme [7] is designed to obtain first-order (in the time dimension) unbiased estimators.[1] Lancaster [11] suggested a way to reparameterize the fixed effects to achieve consistent estimation (not just first-order) in the panel. While Lee and Phillips [10] only considered stationary data in their application, we show that it is possible for Lancaster's method to handle non-stationary data. Different from Lee and Phillips [10], our paper focuses on the selection of exogenous regressors rather than lag order selection. For the purpose of model comparison, proper priors must be used for parameters not common to all the models to avoid Bartlett's paradox when Bayes factors are used (see e.g., [12]). Dhaene and Jochmans [8] found that the modified profile likelihood with Lancaster's correction term can be infinite for infinite parameter support, which implies a prior to ensure the posterior distribution to be proper should be used in a Bayesian context. We develop a data dependent proper prior to combine with Lancaster's reparameterization to calculate Bayes factors and find that model selection based on Bayes factors is inconsistent only for very extreme situations, such as when the number of time periods is 2 or when the true value of the lag coefficient is less than -1. On the other hand, model selection based on Bayesian information criterion (BIC) with the parameters evaluated at the biased MLE can be inconsistent under more common circumstances. From an empirical point of view, researchers could often be confronted with a large number of possible regressors and hence many possible models. Model uncertainty leads to estimation risks especially for small samples since the estimates that we use from a misspecified model could be far away from the true parameter values and hence misleading. From our simulations, we find that Bayesian model averaging (BMA) can reduce such risk and produce point estimators with lower root mean squared errors (RMSE).

The plan of the paper is as follows. Section 2 summarizes the model and the posterior results with the estimation strategies discussed. Section 3 gives our motivations to compare different model specifications and shows the conditions under which our estimator will be consistent when the model is misspecified. Section 4 presents the conditions under which Bayes factors and BIC can be

---

[1] They treat the bias as a result from finite time periods (finite sample bias) and remove the first order bias in the Taylor expansion.

consistent in model selection. In Section 5, we carry out simulation studies to verify our claims before Section 6 concludes.

2. The Model and the Estimation

Here we investigate the first order autoregression linear panel model with a fixed effect, f_i,

$$y_{i,t} = f_i + y_{i,t-1}\rho + x'_{i,t}\beta + u_{i,t},$$
$$i = 1 \ldots N, \quad t = 1 \ldots T. \tag{1}$$

where ρ is a scalar and $x_{i,t}$ is a $k \times 1$ vector of explanatory variables. Denote u_i as $(u_{i,1}, u_{i,2}, \ldots, u_{i,T})'$, $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,T})'$ and y_i as $(y_{i,1}, y_{i,2}, \ldots, y_{i,T})'$. We can rewrite Equation (1) into the vector form below,

$$y_i = f_i \iota + y_{i\_} \rho + X_i \beta + u_i, \tag{2}$$

where ι is a $T \times 1$ vector of ones. By repeated substitution, we can obtain

$$y_{i\_} = f_i \zeta_1 + y_{i,0} \zeta_2 + C X_i \beta + C u_i, \tag{3}$$

where $y_{i\_} = (y_{i,0}, y_{i,1}, \ldots, y_{i,T-1})'$,

$$\zeta_1 = \begin{pmatrix} 0 \\ 1 \\ 1 + \rho \\ \cdots \\ 1 + \rho + \rho^2 + \cdots + \rho^{T-2} \end{pmatrix}, \zeta_2 = \begin{pmatrix} 1 \\ \rho \\ \rho^2 \\ \cdots \\ \rho^{T-1} \end{pmatrix}, C = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \rho & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \rho^{T-2} & \rho^{T-3} & \cdots 1 & 0 \end{pmatrix}$$

The following are the assumptions we use throughout the paper.

Assumption 1. $u_i | X_i, f_i, y_{i,0}, \sigma^2, \rho, \beta \sim i.i.d.(0, \sigma^2 I_T)$ where I_T is an identity matrix with dimension $T \geq 2$.

Assumption 2. (a) $\{(X_i, f_i, y_{i,0})\}$ is a cross-sectionally independent sequence;

(b) $E(|y_{i,0}^2|^{1+\delta}) < \infty$, $E(|f_i^2|^{1+\delta}) < \infty$ and $E(|x_{i,t,h}^2|^{1+\delta}) < \infty$, for some $\delta > 0$, all $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$ and $h = 1, 2, \ldots, k$, where $x_{i,t,h}$ denotes the hth element in $x_{i,t}$;

(c) k and T are finite;

(d) $E\left(\frac{\sum_{i=1}^N X'_i H X_i}{N}\right)$ is finite and uniformly positive definite (see [13], p. 22) where $H = I_T - \frac{\iota\iota'}{T}$;

(e) For any finite values of ρ, the following expression is uniformly positive, *i.e.*, given sufficiently large N,

$$\frac{1}{N} \sum_{i=1}^N \left\{ E\left[(y_{i\_} - Cu_i)' H (y_{i\_} - Cu_i)\right] - \right.$$
$$\left. E\left[(y_{i\_} - Cu_i)' H X_i\right] E\left(\sum_{i=1}^N X'_i H X_i\right)^{-1} E\left[\sum_{i=1}^N X'_i H (y_{i\_} - Cu_i)\right] \right\} > 0. \tag{4}$$

Assumption 1 implies that X_i, f_i and $y_{i,0}$ are strictly exogenous. In comparison to the i.i.d. regularity conditions in Lancaster [11], Assumption 2 (a)–(d) allow the distribution of X_i, f_i and $y_{i,0}$ to be heterogeneous for cross sectional units with slightly more rigorous conditions on their moments such that the asymptotic results in the paper can hold. Assumption 2 (e) is used to simplify the proofs of Proposition 4 and Lemma 10 in Appendix D. Its purpose is to prevent the (within-group) regression of $(f_i\zeta_1 + y_{i,0}\zeta_2 + CX_i\beta)$ on fixed effects and X_i from having perfect fit asymptotically, *i.e.*, R-squared tends to 1 as N increases and to ensure the true value of ρ to be the local mode of its marginal posterior (discussed later) asymptotically. When β $= 0$ (no exogenous regressors in the model), Assumption 2 (e) rules out $f_i = 0$. When $T \geq 3$, if Assumption 2 (e) is satisfied and $\beta \neq 0$, as shown in Appendix D (Equation (53) and its discussion), the following probability limit should also be strictly positive,

$$\underset{N\to\infty}{plim}\frac{1}{N}\left[\sum_{i=1}^{N}\beta'X_i'C'M_\zeta CX_i\beta - \sum_{i=1}^{N}\beta'X_i'C'M_\zeta X_i\left(\sum_{i=1}^{N}X_i'M_\zeta X_i\right)^{-1}\sum_{i=1}^{N}X_i'M_\zeta CX_i\beta\right] > 0, \quad (5)$$

where $M_\zeta = I_T - \boldsymbol{\zeta}(\boldsymbol{\zeta}'\boldsymbol{\zeta})^{-1}\boldsymbol{\zeta}'$ and $\boldsymbol{\zeta} = (\zeta_1, \zeta_2)$. In practice, one could calculate the expression after *plim* in Equation (5) to check Assumption 2 (e) with ρ and β replaced by their consistent estimates. If the value of the expression decreases towards 0 with N,[2] there would be concern for Assumption 2 (e). We would think such case should be very rare with real data.

The MLE of the common parameters, ρ, β and σ^2, are not consistent due to the presence of the incidental parameter f_i, whose number will increase with N. For fixed T, it is impossible for the MLE of f_i to be consistent. When the predetermined regressor $y_{i,t-1}$ is included, the MLE for ρ will be correlated with that of f_i and will also be inconsistent. To obtain the consistent estimators for the common parameters, Lancaster [11] suggested the following way to reparameterize the fixed effect:

$$f_i = g_i \exp\left[-\phi\left(\rho\right)\right] - \frac{1}{T}\iota'X_i\beta, \quad (6)$$

where g_i is the new fixed effect, ι is a vector of ones and $\phi\left(\rho\right)$ is defined as

$$\phi(\rho) = \frac{1}{T}\sum_{t=1}^{T-1}\frac{T-t}{t}\rho^t. \quad (7)$$

We use the following prior for ρ, β and σ^2 and $g = (g_1, g_2, ..., g_N)'$:

$$p(g, \sigma^2, \rho, \beta) = p(g_1)...p(g_N)p(\sigma^2)p(\rho)p(\beta|\sigma^2) \propto \frac{1}{\sigma^2}I(\rho_L \leq \rho \leq \rho_U)p(\beta|\sigma^2, \boldsymbol{X}). \quad (8)$$

In other words, a flat prior for g and Jeffreys' prior for σ^2 are used. ρ is uniformly distributed over $[\rho_L, \rho_U]$. The specifications of ρ_L and ρ_U will be discussed in Proposition 4 later. $p(\beta|\sigma^2)$ takes the form of g-prior in [14]:

$$\beta|\sigma^2, \boldsymbol{X} \sim N\left(0, \sigma^2(\eta\sum_{i=1}^{N}X_i'HX_i)^{-1}\right), \quad (9)$$

---

2   One could withhold some sample while calculating the expression to see how its value changes with N.

where $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$. The strength of the prior depends on the value of η. The smaller is η, the less informative is our prior. As discussed in Section 4, to ensure model selection consistency we can choose $0 < \eta(N) = O(N^\alpha)$ for $\alpha < 0$. For our simulation studies below, we choose $\eta = \frac{1}{NT}$. The posterior results of the model are summarized below.

Proposition 3. *The posterior distributions of the parameters in our model will take the following forms:*

$$g_i | y_i, y_{i,0}, \sigma^2, \rho \sim i.i.d.N\left(e^{\phi(\rho)}\frac{\iota'(y_i - y_{i\_}\rho)}{T}, \frac{\sigma^2}{T}\exp[2\phi(\rho)]\right), \tag{10}$$

$$\beta | \sigma^2, \rho, Y, Y_0, \boldsymbol{X} \sim N\left(\frac{\left(\sum\limits_{i=1}^{N}X_i'HX_i\right)^{-1}\sum\limits_{i=1}^{N}X_i'H(y_i - y_{i\_}\rho)}{\eta + 1}, \frac{\sigma^2\left(\sum\limits_{i=1}^{N}X_i'HX_i\right)^{-1}}{\eta + 1}\right), \tag{11}$$

$$\sigma^2 | \rho, Y, Y_0, \boldsymbol{X} \sim IG(N(T-1), a\rho^2 - 2b\rho + c), \tag{12}$$

$$p(\rho | Y, Y_0, \boldsymbol{X}) \propto I(\rho_L < \rho < \rho_U)\exp\left(N\psi(\rho)\right), \tag{13}$$

where $Y_0 = (y_{1,0}, y_{2,0}, \ldots, y_{N,0})$, $Y = (y_1, y_2, \ldots, y_N)$ *and*

$$\psi(\rho) = \phi(\rho) - \frac{T-1}{2}\ln\left(\frac{a}{N}\rho^2 - 2\frac{b}{N}\rho + \frac{c}{N}\right), \tag{14}$$

$$a = \sum_{i=1}^{N}y_{i\_}'Hy_{i\_} - \frac{1}{\eta+1}\sum_{i=1}^{N}y_{i\_}'HX_i\left(\sum_{i=1}^{N}X_i'HX_i\right)^{-1}\sum_{i=1}^{N}X_i'Hy_{i\_}, \tag{15}$$

$$b = \sum_{i=1}^{N}y_{i\_}'Hy_i - \frac{1}{\eta+1}\sum_{i=1}^{N}y_{i\_}'HX_i\left(\sum_{i=1}^{N}X_i'HX_i\right)^{-1}\sum_{i=1}^{N}X_i'Hy_i, \tag{16}$$

$$c = \sum_{i=1}^{N}y_i'Hy_i - \frac{1}{\eta+1}\sum_{i=1}^{N}y_i'HX_i\left(\sum_{i=1}^{N}X_i'HX_i\right)^{-1}\sum_{i=1}^{N}X_i'Hy_i. \tag{17}$$

$IG(\cdot)$ denotes inverted gamma distribution with degrees of freedom $N(T-1)$ and mean $\frac{a\rho^2 - 2b\rho + c}{N(T-1)-2}$.

Note that a and c in Equations (15) and (17) are close to the sum of squared residuals (SSR) obtained by respectively regressing $y_{i\_}$ and y_i on fixed effects and X_i.[3] $\phi(\rho)$ in Equation (14) is the term from Lancaster's reparameterization, which corrects the marginal posterior local mode of ρ to make it consistent. Dhaene and Jochmans [8] showed that the modified profile likelihood function with $\phi(\rho)$ can be infinite for $\rho \to \infty$. Analogical to their results, we show that the marginal posterior of ρ will be infinite and hence improper when $\rho \to \infty$ or when T is odd and $\rho \to -\infty$ in Appendix D. Lancaster [11] noted such behaviour of ρ's marginal posterior in simulations, but did not discuss much on how to specify the boundary points. In Proposition 4, we provide a data dependent way to specify ρ_L and ρ_U, which is necessary for model comparison to avoid Bartlett's paradox. First note that the probability limit of $\psi(\rho)$ is

$$\underline{\psi}(\rho) = \underset{N\to\infty}{plim}\psi(\rho) = \phi(\rho) - \frac{T-1}{2}\ln d(\rho), \tag{18}$$

---

3   They are SSR when $\eta = 0$.

where

$$d(\rho) = \underset{N \to \infty}{plim} \left(\frac{a}{N}\rho^2 - 2\frac{b}{N}\rho + \frac{c}{N} \right). \tag{19}$$

Proposition 4. *If X_i are the true set of exogenous regressors used to generate y_i, under Assumption 1 and 2, asymptotically the marginal posterior of ρ in Equation (13) will have more than one stationary points satisfying $\underline{\psi}'(\rho) = 0$: 3 stationary points when T is odd and 2 when T is even, regardless of the true value of ρ. The local posterior mode, which is a consistent point estimator, is the stationary point nearest to the MLE satisfying $\underline{\psi}''(\rho) < 0$ asymptotically, which the other stationary point(s) do not satisfy. ρ_U can be specified as the stationary point on the right of the posterior mode. When T is odd, ρ_L can be specified as the stationary point on the left of the posterior mode; when T is even, ρ_L could be chosen as a function of N such that $\rho_L(N) < 0$ is sufficiently small.*[4]

Choosing the boundary points as in Proposition 4 can ensure the marginal posterior of ρ to be proper and its support to contain the true value of ρ asymptotically. ρ_L and ρ_U are different from the boundary points in the constrained maximization in Dhaene and Jochmans [8], who only considered parameter estimation. The interval of our boundary points is wider than theirs, since we want to preserve the bell-shaped part of the posterior density curve for model comparison. Another point to note is that when the true exogenous regressors are included, the local posterior mode will exist regardless of the true value of ρ (even if it is 1) due to Assumption 2.2 (e). Next we investigate the consequences when we can not include the true regressors.

3. Motivations and Methods to Compare Different Model Specifications

In empirical applications, researchers are often faced with many possible regressors suggested by different economic theories to be included into Equation (1). Different models are defined by the inclusion of different combinations of the exogenous regressors and by whether or not the lagged dependent variable is present. Proposition 5 below implies there is no guarantee that the posterior mode in Equation (13) is a consistent estimator if some true regressors are excluded from the model.

Proposition 5. *The posterior mode in Equation (13) is a consistent estimator for ρ if and only if we have either*

$$h_2(\beta, \underline{\rho}) = h_3(\beta) = 0, \tag{20}$$

or

$$\frac{-(T-1)h_2(\beta, \underline{\rho})}{h_3(\beta)} = h(\underline{\rho}) \tag{21}$$

where

$$h(\rho) = \sum_{t=1}^{T-1} \frac{T-t}{T}\rho^{t-1} = \frac{d\,\phi(\rho)}{d\,\rho} = \frac{1}{T}\iota'\zeta_1 = -trace(C'H), \tag{22}$$

---

[4] When T is even, $\underset{\rho \to \infty}{\lim}\underline{\psi}(\rho) = 0$. For model comparison, we need a proper prior for ρ and hence ρ_L must be finite for finite N. In practice, we can choose ρ_L to ensure that $[\rho_L, \rho_U]$ contains a unique posterior mode when the true set of regressors are included. In the subsequent simulations, we choose $\rho_L = -N$ when T is even.

$$h_2(\beta, \underline{\rho}) = \underset{N \to \infty}{plim} \frac{1}{N} \left[\sum_{i=1}^{N} y'_{i\_} H \underline{X}_i \beta - \sum_{i=1}^{N} y'_{i\_} H X_i \left(\sum_{i=1}^{N} X'_i H X_i \right)^{-1} \sum_{i=1}^{N} X'_i H \underline{X}_i \beta \right],$$

$$h_3(\beta) = \underset{N \to \infty}{plim} \frac{1}{N} \left[\sum_{i=1}^{N} \beta' \underline{X}'_i H \underline{X}_i \beta - \sum_{i=1}^{N} \beta' \underline{X}'_i H X_i \left(\sum_{i=1}^{N} X'_i H X_i \right)^{-1} \sum_{i=1}^{N} X'_i H \underline{X}_i \beta \right]. \quad (23)$$

Here $\underline{X}$ represents the regressors in the true model and X denotes the regressors we actually include in our candidate model, while $\underline{\rho}$ is the true value of ρ.

The values of $h_2(\beta, \underline{\rho})$ and $h_3(\beta)$ depend on how the true regressors and the included regressors are related, apart from the values of β and $\underline{\rho}$. For $h_2(\beta, \underline{\rho}) = h_3(\beta) = 0$ to be satisfied, it suffices that the true regressors $\underline{X}$ are a subset of X.[5] When some true regressors are excluded, the model will suffer omitted variable bias unless Equation (21) holds. Given Assumption 1 and 2, one example for Equation (21) to hold could be that all the true regressors are covariance stationary and they have no serial correlation; the included regressors have zero correlation with the true regressors; moreover, $\zeta'_1 H \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} E(f_i \underline{X}_i) \beta = \zeta'_2 H \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} E(y_{i,0} \underline{X}_i) \beta = 0$. For this restrictive case, it will be possible to estimate ρ consistently without any true regressors included.[6]

To avoid inconsistent estimation due to model misspecification, one could include all the potential regressors into the model. For finite sample, however, that could inflate the posterior variances for the coefficients of the true regressors if too many irrelevant regressors are included. The simulation studies in Section 5 reveal that while including all the regressors does not influence the estimation of ρ in comparison to other consistent approaches, it leads to substantially high RMSE when estimating β. Hence appropriate procedures for model selection are desirable. In a Bayesian framework, one can evaluate different model specifications, denoted by M_i below, by their posterior model probabilities, which can be calculated as

$$\begin{aligned} p(M_j | Y, Y_0, \boldsymbol{X}_j) &= \frac{p(M_j) p(Y | Y_0, \boldsymbol{X}_j, M_j)}{p(Y | Y_0)} \\ &= \frac{p(M_j) p(Y | Y_0, \boldsymbol{X}_j, M_j)}{\sum_{i=1}^{2^{K+1}} p(M_i) p(Y | Y_0, \boldsymbol{X}_i, M_i)}, \end{aligned} \quad (24)$$

where $\boldsymbol{X}_j$ denotes the regressors included under M_j and $p(Y | Y_0, \boldsymbol{X}_j, M_j)$ is the marginal likelihood, obtained by integrating out ρ in Equation (13) or (43) in Appendix C. K is the number of all potential exogenous regressors. The total number of models is therefore 2^{K+1}. $p(M_j)$ is the prior model probability of model j. For finite samples, Ley and Steel [15] showed that the choice of prior model probability will affect the posterior results to a large extent when the number of potential regressors is large compared to the sample size. In what follows, we focus on the asymptotic behaviour of posterior model probabilities and assume all the models are equally possible a priori. The posterior model probability $p(M_j | Y, Y_0, \boldsymbol{X}_j)$ will hence only depend on $p(Y | Y_0, \boldsymbol{X}_j, M_j)$.

---

[5]   Note that $h_3(\beta)$ is the probability limit of $\frac{1}{N}$ times the SSR obtained by regressing $\underline{X}_i \beta$ on fixed effects and X_i.

[6]   Due to the assumptions in the example, $\frac{-(T-1)h_2(\beta, \underline{\rho})}{h_3(\beta)} = \frac{-(T-1)trace\{C'H \sum_{i=1}^{N}[Var(\underline{X}_i \beta) + HE(\underline{X}_i \beta)E(\underline{X}_i \beta)']\}}{trace\{H \sum_{i=1}^{N}[Var(\underline{X}_i \beta) + HE(\underline{X}_i \beta)E(\underline{X}_i \beta)']\}}$. Since $\frac{\sum_{i=1}^{N} Var(\underline{X}_i \beta)}{N}$ is proportional to I_T and $HE(\underline{X}_i \beta)E(\underline{X}_i \beta)' = 0$. Hence $\frac{-(T-1)h_2(\beta, \underline{\rho})}{h_3(\beta)} = -trace(C'H) = h(\underline{\rho})$.

4. Consistency in Model Selection

In this section, we discuss the situations when the posterior model probability of the true model will tend to 1 as N tends to infinity. We will also analyze whether Bayesian information criterion (BIC) based on biased MLE is consistent in model selection.

For static panel models when the true value of ρ is zero and the lagged dependent variable is not included as a regressor, the analysis of Bayes factors is similar to that of Fernandez *et al.* [16]. In our context, we can ensure model selection consistency by setting η as a function of N with $0 < \eta(N) = O(N^\alpha)$ for $\alpha < 0$. As for BIC, it is consistent in model selection for static panel models.

Let us now consider the case when our candidate model (M_1) contains $y_{i\_}$ and X_{i1}. M_1 is compared to either M_0, which has X_{i0} but no $y_{i\_}$, or M_2, which has X_{i2} and $y_{i\_}$. X_{ij} denotes the exogenous regressors included under model M_j for $j = 0, 1, 2$, which satisfy Assumption 2.1 and 2.2. X_{i1} can be the same or different from X_{i0}, while X_{i2} is different from X_{i1}. The Bayes factors respectively are:

$$\frac{p(Y|Y_0, \boldsymbol{X}_1, M_1)}{p(Y|Y_0, \boldsymbol{X}_0, M_0)} = \left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}} \frac{\int_{\rho_{L_1}}^{\rho_{U_1}} \exp\left[N\psi(\rho|M_1)\right] d\rho}{(\rho_{U_1} - \rho_{L_1}) \left(\frac{c_{|M_0}}{N}\right)^{-\frac{N(T-1)}{2}}}, \tag{25}$$

$$\frac{p(Y|Y_0, \boldsymbol{X}_1, M_1)}{p(Y|Y_0, \boldsymbol{X}_2, M_2)} = \frac{\rho_{U_2} - \rho_{L_2}}{\rho_{U_1} - \rho_{L_1}} \left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_2}{2}} \frac{\int_{\rho_{L_1}}^{\rho_{U_1}} \exp\left[N\psi(\rho|M_1)\right] d\rho}{\int_{\rho_{L_2}}^{\rho_{U_2}} \exp\left[N\psi(\rho|M_2)\right] d\rho}, \tag{26}$$

where k_j denotes the number of columns in X_{ij}; $a_{|M_j}$, $b_{|M_j}$ and $c_{|M_j}$ in $\psi(\rho|M_j)$ defined in Equation (14) are calculated by replacing X_i with X_{ij} in Equations (15) to (17) for $j = 0, 1, 2$, which multiplied by $\frac{1}{N}$ have the following probability limits under Assumption 1 and 2 with $\eta(N) = O(N^\alpha)$ and $\alpha < 0$:

$$\plim_{N\to\infty} \frac{1}{N} a_{|M_j} = \underline{a}_{|M_j}, \tag{27}$$

$$\plim_{N\to\infty} \frac{1}{N} b_{|M_j} = \underline{b}_{|M_j} = \underline{a}_{|M_j}(\underline{\rho} + \gamma_{|M_j}), \tag{28}$$

$$\plim_{N\to\infty} \frac{1}{N} c_{|M_j} = \underline{c}_{|M_j} = \underline{\rho}^2 \underline{a}_{|M_j} + 2\underline{a}_{|M_j}\underline{\rho}\gamma_{|M_j} + h_3(\beta_{|M_j}|M_j) + (T-1)\sigma^2, \tag{29}$$

$$\gamma_{|M_j} = \frac{h_2(\beta_{|M_j}, \underline{\rho}|M_j) - \sigma^2 h(\underline{\rho})}{\underline{a}_{|M_j}}. \tag{30}$$

$\gamma_{|M_j}$ stands for the Nickell MLE bias of ρ under M_j. We can see the MLE bias results from two sources: the incidental parameter part ($\sigma^2 h(\underline{\rho})$) and the model misspecification part ($h_2(\beta_{|M_j}, \underline{\rho}|M_j)$). Proposition 4 shows that when the model is correctly specified, the local posterior mode is a consistent estimator for ρ. In the simulation studies in Section 5 we find that when some combination of the wrong exogenous regressors are included, the marginal posterior density of ρ can be either monotonically increasing or is U-shaped depending on the value of T and does not have a local maximum. When we find such a wrong model, we will assign 0 as its posterior model probability and will not estimate the model. In Proposition 6 and 7 below, we consider the cases when the local maximum exists for $\underline{\psi}(\rho|M_j)$

in Equation (18) and show the sufficient conditions under which the Bayes factors in Equations (25) and (26) can lead to the selection of the true model asymptotically. Denote $\rho^*_{|M_j}$ as the local maximum of $\underline{\psi}(\rho|M_j)$ in (ρ_{L_j}, ρ_{U_j}) with $\underline{\psi}''(\rho^*_{|M_j}|M_j) < 0$.

Proposition 6. *When M_1 is the true model, i.e., $\underline{\rho} \neq 0$ and X_{i1} is the set of true regressors to generate Y, as N increases, $\frac{p(Y|Y_0, \boldsymbol{X}_1, M_1)}{p(Y|Y_0, \boldsymbol{X}_0, M_0)}$ in Equation (25) will tend to infinity if the following holds,*

$$\phi(\underline{\rho}) + \frac{T-1}{2} \ln \frac{\underline{c}_{|M_0}}{(T-1)\sigma^2} > 0 \tag{31}$$

When M_0 is the true model, i.e., X_{i0} is the set of true regressors and $\underline{\rho} = 0$, as N increases, $\frac{p(Y|Y_0, \boldsymbol{X}_1, M_1)}{p(Y|Y_0, \boldsymbol{X}_0, M_0)}$ in Equation (25) will tend to 0 if either of the following is satisfied:

(a) we have

$$-\phi(\rho^*_{|M_1}) + \frac{T-1}{2} \ln \frac{d(\rho^*_{|M_1}|M_1)}{(T-1)\sigma^2} > 0; \tag{32}$$

If Equation (21) is true under M_1, Equation (32) will hold.

(b) Equation (20) is true under M_1. In this case, the left hand side of Equation (32) is equal to 0.

Proposition 7. *When M_2 is the true model and M_1 is the misspecified model, as N increases, $\frac{p(Y|Y_0, \boldsymbol{X}_1, M_1)}{p(Y|Y_0, \boldsymbol{X}_2, M_2)}$ in Equation (26) will tend to 0 if either of the following holds:*

(a) we have

$$\phi(\underline{\rho}) - \phi(\rho^*_{|M_1}) + \frac{T-1}{2} \ln \frac{d(\rho^*_{|M_1}|M_1)}{(T-1)\sigma^2} > 0, \tag{33}$$

If Equation (21) is true under M_1, Equation (33) will hold.

(b) Equation (20) is true under M_1. In this case, the left hand side of Equation (33) is equal to 0.

Since both Equations (20) and (21) imply that the local posterior mode in Equation (13) is a consistent estimator for ρ, from Proposition 6 and 7, we can see that if the posterior mode is consistent under the misspecified model, the misspecified model will not be chosen by the Bayes factor (model selection will be consistent). In Appendix D, we show that $h(\rho)$ is positive over $\mathbb{R}$ when T is an even number. This implies $\phi(\rho)$ is an increasing function over $\mathbb{R}$. Also note that $\phi(0) = 0$. Hence $\phi(\rho) < 0$ for $\rho < 0$ and it is possible for Equation (31) to be violated when T is even and $\underline{\rho}$ is a negative number. As shown in the last paragraph in Appendix E, though Equation (31) could be violated for the extreme case of $T = 2$ and $\underline{\rho} < 0$, fortunately, apart from this extreme case, violation of Equation (31) could only occur when $\underline{\rho} < -1$ for T being an even number greater than 2, which may not be relevant for most economic applications with $\underline{\rho} \in [-1, 1]$.

Note that Equation (32) is the special case of Equation (33) with $\underline{\rho} = 0$. When the posterior mode is not consistent under the misspecified model, it is difficult to state under what circumstances Equation (33) is or is not satisfied since ρ^* generally does not have closed form. By construction, $\rho^*_{|M_1}$ is a local minimum for the left of Equation (33). In our simulation studies in Section 5, we calculate Equation (32) or (33) under different settings when model selection errors based on Bayes factors occur. We cannot find a single occurrence of either Equation (32) or (33) being violated except the cases when Equation (20) is true, that is, when the candidate model nests the true model. It appears that the left hand

sides of Equation (33) could be interpreted as how close the candidate model is to nest the true model. Note that with real data, it is difficult to check Equation (20), but one can assess whether Equation (33) is violated by replacing $d(\rho|M_j)$ with $\frac{a_{|M_j}}{N}\rho^2 - 2\frac{b_{|M_j}}{N}\rho + \frac{c_{|M_j}}{N}$ and supplanting σ^2 and $\underline{\rho}$ by their consistent estimates e.g., those from the model including all the potential regressors.

Proposition 8 below shows when the BIC based on the biased MLE is consistent in model selection. BIC for the model with and without the lagged dependent variable is defined respectively below,

$$BIC_{lag} = NT\left[\ln\left(\frac{c}{NT} - \frac{b^2}{a \times NT}\right) + \ln 2\pi + 1\right] + (1 + k + N)\ln(NT), \qquad (34)$$

$$BIC_{\text{no lag}} = NT\left[\ln\frac{c}{NT} + \ln 2\pi + 1\right] + (k + N)\ln(NT), \qquad (35)$$

where a, b and c are defined respectively in Equations (15), (16) and (17) with $\eta = 0$, and k is the number of exogenous regressors included. The model with smaller BIC value will be preferred.

Proposition 8. *For the comparison of the two models in Equation (25), when M_0 is the true model, BIC will be consistent if the following is satisfied*

$$(T-1)\sigma^2 - \underline{c}_{|M_1} + \frac{\underline{b}^2_{|M_1}}{\underline{a}_{|M_1}} < 0. \qquad (36)$$

However, if Equation (20) is true under M_1, the left of Equation (36) will be greater than 0 and BIC will be inconsistent.

When M_1 is the true model, BIC will be consistent in model selection if the following condition is met:

$$\underline{c}_{|M_0} - \underline{c}_{|M_1} + \frac{\underline{b}^2_{|M_1}}{\underline{a}_{|M_1}} > 0. \qquad (37)$$

If X_{i0} is the same as X_{i1} and the probability limit of $\hat{\rho}_{MLE}$ is equal to 0, the left of Equation (37) will be 0 and BIC will be inconsistent.

For the comparison of the two models in Equation (26), when M_1 is the true model, BIC will be consistent in model selection if the following holds

$$\underline{c}_{|M_2} - \underline{c}_{|M_1} - \frac{\underline{b}^2_{|M_2}}{\underline{a}_{|M_2}} + \frac{\underline{b}^2_{|M_1}}{\underline{a}_{|M_1}} > 0. \qquad (38)$$

Conditions (36), (37) and (38) are just the sufficient but not necessary conditions for BIC to be consistent in model selection since BIC has a penalty term against over-parameterization (the last term in Equations (34) and (35)). Note that $\hat{\rho}_{MLE} = \frac{b}{a}$ and $\underset{N\to\infty}{plim}\hat{\rho}_{MLE} = \underline{\rho} + \gamma$ from Equations (27) and (28), where γ is the Nickell bias. The violation of Equations (36) and (37) is related to the hypothesis test, $H_0 : \rho = 0$. When $\underset{N\to\infty}{plim}\hat{\rho}_{MLE} = \gamma$ ($\underline{\rho} = 0$) and $X_{i1} = X_{i0}$, the probability of making type I errors based on classical test statistics, such as Wald or likelihood ratio (LR), will be 1 and BIC will choose M_1 asymptotically with the left hand side of Equation (36) being $\underline{a}_{|M_1}\gamma^2_{|M1} > 0$; when $\underset{N\to\infty}{plim}\hat{\rho}_{MLE} = 0$ and $X_{i1} = X_{i0}$, the probability of making type II errors will be 1 asymptotically and BIC will choose M_0 even if $\underline{\rho} \neq 0$ with the left hand side of Equation (37) being $\underline{a}_{|M_1}(\underline{\rho} + \gamma_{|M1})^2 = 0$. When incidental parameters are present, Cox and Reid [17] suggest using the likelihood conditional on the MLE of the

orthogonalized incidental parameters to construct LR statistics. In practice, if we find $\hat{\rho}_{MLE}$ is close to 0 or the estimated Nickell bias, we should be cautious to use BIC for model selection. For Equation (38), as shown in Appendix G, if M_1 is the true model and X_{i2} nests X_{i1}, the left hand side of Equation (38) will be less than or equal to 0 asymptotically depending on whether $\underline{a}_{|M_2}$ is less than or equal to $\underline{a}_{|M_1}$. Though Equation (38) is violated when $\underline{a}_{|M_2} = \underline{a}_{|M_1}$, BIC can still favour M_1 since there are more parameters under M_2. However, if $\underline{a}_{|M_2} < \underline{a}_{|M_1}$, which could happen when f_i is highly correlated with all the potential regressors, BIC will choose the wrong model M_2 asymptotically, as shown in Section 5.3. Given the SSR interpretation of a in Equation (15) (with $\eta = 0$), the practical implication of this result is that if BIC chooses the model with all the regressors included, which always has smaller $\frac{a}{N}$ for finite sample comparing to other models, we should be cautious with such choice in the application.

5. Simulation Studies

In this section we use Monte Carlo simulation to verify the claims in Proposition 6, 7 and 8 and investigate the impact of model uncertainty on point estimation. The number of simulations is 1000. We set $T = 4$, $\sigma^2 = 1$, $\eta = \frac{1}{NT}$, $\rho_L = -N$ when T is even and the number of possible regressors to 8. We select 4 regressors out of 8 (K) to generate the dependent variable. The coefficient values of the chosen regressors are 0.1 0.3, 1 and 2 respectively. We draw independently f_i and $y_{i,0}$ from $U[-4, 4]$. For each simulation, we calculate the posterior model probabilities and the BIC of all the models and evaluate the performances of the two criteria. In Proposition 5, we show that the posterior mode is not a consistent estimator of ρ when neither Equation (20) nor (21) holds, which is possible when the regressors have collinearity and serial correlation. We generate the potential regressors to be covariance stationary and make them serially correlated and correlated with each other. The details of the data generating process (DGP) can be found in Appendix A. There are three parameters controlling the properties of the regressors: $\sigma_X^2 = 5.33$ (the variance with the same value as those of f_i and $y_{i,0}$), $s = 0.5$ (the autocorrelation coefficient) and $\lambda = 1$ (between 0 and 1, the closer to 0, the higher the correlation among the regressors). The settings are the same for the subsequent simulation exercises unless otherwise stated. The results of robust checks with other values of σ_X^2, s and λ are shown in Appendix B.

5.1. When Model Selection is Consistent

The model selection performance results for different values of $\underline{\rho} > -1$ appear similar and are available upon request. If some true regressors are excluded, Equation (20) or (21) would be violated more often under $\underline{\rho} = -1$ than when $\underline{\rho}$ takes other positive values. The results presented in Table 1 are based on $\underline{\rho} = -1$. The "**ER**" column shows the error rates of Bayes factors[7] while the "**ERBIC**"column contains those of BIC. For $N = 40$, BIC performs better than Bayes factors. As the sample size increases, the error rates of the two criteria get closer and both decrease, which implies both are consistent in model selection. Note that the coefficient of one of the exogenous regressors in the true model is equal to 0.1,

---

[7]   That is how often the model with the highest posterior model probability is not the true model.

which is close to 0. Models selected based on Bayes factors often cannot pick up this regressor when $N = 40$. The column "**nest**" indicates how often the model chosen by Bayes factors only omits the regressor with coefficient 0.1 or is the same as the true model. In other words, the true model nests the chosen model. Comparing this column to "**nestbic**", we can see that the models chosen by Bayes factors are more often nested inside the true model with the less important regressor excluded than the models from BIC. Column "**ER11**" shows the proportions of errors committed when the true model and the model chosen by Bayes factors both include $y_{i\_}$ but have different exogenous regressors.[8] We can see that all the errors made by Bayes factors and BIC are due to the inclusion of the wrong set of exogenous regressors rather than omitting $y_{i\_}$. Hence there is no point to check whether or not Equation (31) or (37) is violated. When the errors of **ER11** or **ERBIC11** occurred, we checked whether Equation (33) or (38) was violated. For this and the following simulation exercises, we did not find any violations of Equation (33). For Equation (38), it is only violated with its left hand side being 0 when the chosen model (M_2) includes all the regressors of the true model (M_1). For this case, BIC is still consistent. In other words, the errors of **ER11** or **ERBIC11** are fixable with larger sample sizes for both selection criteria.

Table 1. Simulation results when both criteria are consistent in model selection.

N	ER	ERBIC	ER11	ERBIC11	Nest	Nestbic
40	0.834	0.762	1	1	0.799	0.704
100	0.543	0.510	1	1	0.829	0.777
200	0.300	0.299	1	1	0.862	0.830
500	0.122	0.110	1	1	0.902	0.907
1000	0.064	0.064	1	1	0.943	0.942

5.2. When Equation (31) is Violated for Bayes Factors

In Section 4, we mentioned that when T is even and $\underline{\rho}$ is a small negative number, it is possible for Equation (31) to be violated. Under the settings in Section A, the left hand side of Equation (31) often has a root of $\underline{\rho}$ between -7.4 and -7.2 when the true regressors are included.[9] If $\underline{\rho}$ is less than the root, Equation (31) will be violated. In our next exercise, we set $\underline{\rho} = -7.4$ and run the simulations again. The results are in Table 2. We can see that Bayes factors cannot select the true model for once out of the 1000 simulations for all sample sizes while the error rates of BIC gradually decrease with N. All the Bayes factors errors are made when the chosen model does not contain $y_{i\_}$ (see "**ER10**") and Equation (31) is violated. Similar problems with Bayes factors arise when $T = 2$ and $-1 < \underline{\rho} < 0$ as explained in Appendix E. Table 3 shows the simulation results for such situation when $\underline{\rho} = -0.9$, $\sigma^2 = 100$ and the true model does not have X_i while other settings are the same as before. Bayes factors again show no signs of model selection consistency almost always due to the violation of Equation (31). The "**noreg**"

---

[8] In the subsequent discussion, we use **ER10** to denote the proportion of errors made when the true model includes $y_{i\_}$ while the chosen model does not, and **ER01** for when the true model does not include $y_{i\_}$ while the chosen model does. Note that either **ER10** + **ER11** = 1 or **ER01** + **ER00** = 1. The notations for BIC are defined similarly.

[9] Since the correlation among different regressors is random in our simulation, $\underline{a}_{M_0}$ and the root are also random.

column shows how often in the errors made by Bayes factors, the chosen model only includes the fixed effects with no other regressors. As sample size increases, Bayes factors tend to make more such errors which BIC never commits.

Table 2. Simulation results when Equation (31) is violated with $\underline{\rho} = -7.4$.

N	ER	ERBIC	ER10	no(31)	ER10BIC	Nest	Nestbic
40	1	0.787	1	1	0	0	0.688
100	1	0.552	1	1	0	0	0.756
200	1	0.300	1	1	0	0	0.832
500	1	0.139	1	1	0	0	0.886
1000	1	0.064	1	1	0	0	0.941

Table 3. Simulation results when Equation (31) is violated with $\underline{\rho} = -0.9$, $T = 2$, $\sigma^2 = 100$ and no X_i in the true model.

N	ER	ERBIC	ER10	no(31)	ERBIC10	Noreg	Noregbic
40	0.910	0.696	0.936	0.998	0	0.621	0
100	0.873	0.559	0.956	1	0	0.751	0
200	0.854	0.440	0.977	1	0	0.833	0
500	0.858	0.360	0.986	1	0	0.893	0
1000	0.896	0.273	0.991	1	0	0.921	0

5.3. When Equations (36), (37) or (38) is Violated for BIC

Bayes factors perform poorly in model selection when Equation (31) is violated, which takes place under rather extreme situations. Next we show that BIC can perform poorly under more common circumstances, which are more possible for economic applications. As discussed in Proposition 8, if $\underline{\rho} = 0$, BIC could asymptotically choose the model with the true exogenous regressor(s) and $y_{i\_}$ over the true model. For the next simulation exercise, we change $\underline{\rho}$ to 0. The results are shown in Table 4. Bayes factors now have smaller error rates while BIC cannot identify the true model. As expected, BIC always chooses the models with $y_{i\_}$ (see "**ER01BIC**"), while the proportion of errors violating Equation (36), gets higher for bigger sample sizes. Column "**cnestbic**" shows how often the chosen model by BIC nests the true model when the errors of **ER01BIC** occur. The values in this column are just slightly smaller than those in "**no(36)**", which indicates a high proportion of the violation of Equation (36) happens when the chosen model nests the true model.

Another situation of poor BIC performance is when Equation (37) is close to violation. In Proposition 8, we mentioned that if $\plim_{N\to\infty} \hat{\rho}_{MLE} = 0$ under the true model, a candidate model with the same exogenous regressors as those of the true model will violate Equation (37). In our next experiment, we do not include any exogenous regressors into the true model and set $\underline{\rho} = 0.0756$ to make $\hat{\rho}_{MLE}$ close to 0. If the candidate model (M_0) only has fixed effects, the left hand side of Equation (37) is close to but slightly above 0. The simulation results are given in Table 5. We can see that BIC error rates gradually increase to near 1 with the sample size. Column "**noregbic**" indicates the proportion of BIC

errors committed when the chosen model only includes fixed effects. Note that the values in this column are the same as those in "**ER10BIC**", which also get closer to 1 with sample size. Clearly, the poor performance of BIC in this scenario should be related to Equation (37).

Table 4. Simulation results when Equation (36) is violated with $\underline{\rho} = 0$ and $T = 4$.

N	ER	ERBIC	ER01	ER01BIC	no(36)	Cnestbic	Nest	Nestbic
40	0.844	1	0	1	0.243	0.230	0.777	0
100	0.572	1	0	1	0.524	0.511	0.823	0
200	0.290	1	0	1	0.777	0.760	0.864	0
500	0.104	1	0	1	0.941	0.925	0.924	0
1000	0.042	1	0	1	0.996	0.985	0.965	0

Table 5. Simulation results when Equation (37) is violated with $\underline{\rho} = 0.0756$ and no exogenous regressors are included in the true model.

N	ER	ERBIC	ER10	no(37)	ER10BIC	Noreg	Noregbic
40	0.952	0.968	0.973	0	0.876	0.660	0.876
100	0.808	0.971	0.943	0	0.947	0.762	0.947
200	0.540	0.976	0.887	0	0.968	0.720	0.968
500	0.132	0.985	0.311	0	0.981	0.303	0.981
1000	0.056	0.991	0	0	0.990	0	0.990

Finally we show a case when Equation (38) is violated. Note that the left hand side of Equation (38) asymptotically depends on $\underline{a}_{|M_1}$ (calculated under the true model) and $\underline{a}_{|M_2}$ (calculated under the wrong candidate model). If $\underline{a}_{|M_1} > \underline{a}_{|M_2}$, BIC will be inconsistent, which could happen when M_2 nests M_1 and f_i is highly correlated with all the potential exogenous regressors. In our next exercise, we set $T = 3$, $\underline{\rho} = -1$ and generate $y_{i,0}$ and f_i^* from $U[-1, 1]$. When we generate $\tilde{x}_{i,t}$ in Equation (40), we set $s = -0.9$. f_i is generated as $f_i = f_i^* + 10\frac{1}{TK}\sum_{t=1}^{T}\sum_{h=1}^{K} x_{i,t,h}$. In the true model, no exogenous regressors are included, which implies any candidate model including $y_{i\_}$ nests or is the same as the true model. The results in Table 6 show that BIC is not consistent with increasing error rates as the sample size gets larger than 200 and all the errors are of type **ERBIC11**. For all the errors made by BIC, we have found that Equation (38) is violated with $\underline{a}_{|M_1} > \underline{a}_{|M_2}$. For a few cases, $\underline{a}_{|M_1}$ is very close to $\underline{a}_{|M_2}$. The column with the heading $\frac{a_{|M_2}}{a_{|M_1}} < 0.999$ in Table 6 indicates the percentage of the errors when $\underline{a}_{|M_2}$ is smaller than $\underline{a}_{|M_1}$ by more than 0.1% of its value. We can see that the majority of the errors happen when $\underline{a}_{|M_2}$ is smaller by more than a tiny fraction of $\underline{a}_{|M_1}$. The column headed with $E\left(\frac{a_{|M_2}}{a_{|M_1}}\right)$ shows the sample average of $\frac{a_{|M_2}}{a_{|M_1}}$ from all the errors, which gets smaller with the sample size. This implies BIC tends to choose the model with lower $\underline{a}$ in comparison to the true model with larger sample sizes. Note that the simulation results are sensitive to the parameter settings. If we change T to 4 while keeping other settings the same, among the BIC errors, $\underline{a}_{|M_2}$ will be virtually the same as $\underline{a}_{|M_1}$ and BIC will show decreasing error rates, which, though, are higher than those of Bayes factors for different sample sizes. In this case, we need to change $\underline{\rho}$ and s to make BIC inconsistent. The results are available upon request.

Table 6. Simulation results when Equation (38) is violated.

N	ER	ERBIC	ER11	ERBIC11	$\frac{a_{\|M_2}}{a_{\|M_1}} < 0.999$	$E\left(\frac{a_{\|M_2}}{a_{\|M_1}}\right)$
40	0.209	0.439	1	1	0.968	0.878
100	0.119	0.352	1	1	0.977	0.866
200	0.072	0.336	1	1	0.973	0.858
500	0.050	0.406	1	1	0.990	0.811
1000	0.039	0.600	1	1	0.993	0.782

To sum up, it is possible for Equation (36), (37) or (38) to be violated and BIC can be inconsistent in model selection under more common circumstances than Bayes factors.

5.4. Point Estimation

Judging from the previous simulation results, we can see that if we simply select the model with the highest posterior model probability to provide the estimates of our interest, the chances will be high that the model selected is not the true model especially when N is small regardless of which criterion we use. Next we investigate how model uncertainty impacts on point estimation. We set $\underline{\rho} = 1$ and the number of simulations equal to 2000. We then evaluate the performances of different consistent point estimators.[10]

Table 7 shows the root mean squared errors (RMSE) with the cross section sample size (N) equal to 40. The true values of ρ and β are shown under the column "**True**". There are 8 potential regressors, 4 of which are not included in the true model and hence have coefficients equal to 0. The column "**Top**" shows the RMSE resulting from the posterior mode of the model with the highest posterior model probability, the column "**All**" shows the results from the model which include all the potential regressors, while the values in the column "**BMA**" are from the posterior mode average of different models with the weights equal to the posterior model probabilities. To evaluate the significance of a regressor in the Bayesian context, we can calculate the sum of the posterior model probabilities of all the models which include the regressor. The RMSE in the columns headed with percentage numbers are calculated based on certain inclusion probability criterion. If the inclusion probability for a regressor is lower than the percentage number of the column, we will simply use zero as its point estimate. Otherwise, we will use the BMA estimate. From Table 7, we can see that the model including all the potential regressors has much higher RMSE for all the parameters except ρ than other methods. BMA has smaller RMSE for almost all the parameters than the top model criterion[11] and it tends to have lower RMSE than inclusion probability criteria for parameters different from 0 while larger RMSE for parameters equal to 0. Higher inclusion probability tends to give smaller RMSE when the true value of the parameter is 0 while higher RMSE for non-zero parameters. The last row of Table 7 shows the sum of RMSE in each column, which

---

[10] The absolute value of Nickell bias under $\underline{\rho} = 1$ is bigger than when $\underline{\rho} = -1$. None of the conditions for model selection consistency are found to be violated.

[11] We have also obtained the results of finite sample biases of different point estimators (available upon request), which show that the top model point estimators are generally less biased than other criteria and hence the higher top model RMSE should be due to larger estimator variances.

is a measure of the overall performances of different criteria. We can see that BMA and various inclusion probability criteria are all better than those of the top model and the all-inclusive model. The sum of RMSE is the smallest when we set the inclusion probability to 50%.

Table 7. root mean squared errors (RMSE) of point estimators when $N = 40$.

True		Top	BMA	All	30%	40%	50%	60%	70%	80%
ρ	**1**	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017
	0.1	0.110	0.090	4.954	0.096	0.099	0.101	0.102	0.103	0.103
	0.3	0.126	0.114	4.263	0.117	0.121	0.128	0.138	0.151	0.167
	0	0.072	0.066	3.567	0.062	0.060	0.058	0.056	0.049	0.046
	0	0.071	0.057	5.219	0.054	0.052	0.049	0.045	0.042	0.036
β	**0**	0.051	0.044	2.254	0.039	0.036	0.030	0.027	0.023	0.019
	1	0.119	0.104	6.033	0.105	0.105	0.106	0.111	0.114	0.121
	0	0.057	0.053	3.777	0.049	0.047	0.041	0.038	0.033	0.025
	2	0.118	0.108	6.573	0.108	0.108	0.113	0.113	0.120	0.128
Sum		0.739	0.652	36.656	0.647	0.643	0.643	0.646	0.653	0.661

To add more insights into inclusion probabilities, we present the error rates of in/excluding the wrong/right regressor based on different inclusion probability criteria in Table 8 and compare to those from the top model. Similar to the findings of RMSE, higher inclusion probabilities tend to give larger error rates for non-zero parameters while smaller error rates for the zero parameters. The last row shows the average error rates of different columns, of which the highest value appears when the 10% criterion is used and the majority of the errors are from the zero parameters. Note that for a particular regressor, the prior inclusion probability is 50% in our setting. If the posterior inclusion probability is no less than 50%, it implies the data confirm or strengthen the prior. The top model criterion has smaller average error rate than almost all the inclusion probability criteria except 40% and 50%.

Table 8. The error rates of excluding or including a regressor based on different criteria when $N = 40$.

True		Top	10%	20%	30%	40%	50%	60%	70%	80%
ρ	**1**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.1	0.826	0.181	0.526	0.682	0.789	0.856	0.894	0.927	0.955
	0.3	0.119	0.004	0.019	0.046	0.070	0.110	0.150	0.205	0.270
	0	0.053	0.647	0.209	0.105	0.060	0.042	0.026	0.016	0.011
	0	0.052	0.638	0.217	0.114	0.073	0.042	0.023	0.013	0.008
β	**0**	0.048	0.654	0.226	0.103	0.064	0.032	0.019	0.011	0.006
	1	0.003	0.000	0.001	0.001	0.001	0.002	0.003	0.004	0.006
	0	0.043	0.633	0.204	0.107	0.064	0.037	0.023	0.012	0.005
	2	0.001	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.002
Avg.		0.127	0.307	0.156	0.129	0.125	0.124	0.127	0.132	0.140

Table 9 presents the results of RMSE sums and average error rates under different sample sizes. BMA has smaller RMSE than the top model estimators for all sizes, while the top model average error rate is in general close to the minimum of various inclusion probability criteria. The minimums of RMSE sums are usually attained when the inclusion probability is above or equal to 50%, while the minimum average error rates appear at around 50%. Therefore, under our simulation settings, for point estimation, it seems sensible to use 50% inclusion probability to decide whether or not a regressor should be included.

Table 9. Sum of RMSE and averages of error rates.

	Sum of RMSE					Average Error Rates				
N	40	100	200	500	1000	40	100	200	500	1000
Top	0.739	0.535	0.299	0.189	0.107	0.127	0.083	0.046	0.015	0.005
BMA	0.652	0.459	0.288	0.171	0.101	N.A.	N.A.	N.A.	N.A.	N.A.
10%	0.653	0.460	0.287	0.171	0.100	0.307	0.178	0.116	0.062	0.029
20%	0.653	0.460	0.286	0.170	0.099	0.156	0.102	0.063	0.030	0.012
30%	0.647	0.461	0.284	0.169	0.098	0.129	0.085	0.049	0.021	0.008
40%	0.643	0.461	0.285	0.169	0.097	0.125	0.081	0.046	0.017	0.006
50%	0.643	0.429	0.284	0.168	0.096	0.124	0.082	0.048	0.015	0.005
60%	0.646	0.437	0.278	0.169	0.096	0.127	0.087	0.051	0.015	0.004
70%	0.653	0.447	0.279	0.170	0.098	0.132	0.092	0.056	0.017	0.005
80%	0.661	0.437	0.279	0.161	0.100	0.140	0.099	0.063	0.021	0.007
90%	0.685	0.439	0.305	0.171	0.102	0.153	0.111	0.077	0.028	0.008

6. Conclusions

In this paper, we investigated consistent parameter estimation and model selection for the linear dynamic panel model. We use the fixed effect reparameterization proposed by Lancaster [11] combined with our data dependent prior for estimation and calculate Bayes factors to compare different model specifications. We recommend model selection should precede parameter estimation, since Lancaster's fixed effect transformation may not necessarily lead to consistent estimation when some true exogenous regressors are excluded. We have given the conditions under which Bayes factors or BIC can lead to consistency in model selection and have shown that Bayes factors could be inconsistent in model selection when the number of time periods is 2, or when the true autoregressive coefficient is less than -1. Such situations could be rare for most economic applications. BIC based on the biased MLE can be inconsistent when the fixed effects are highly correlated with all the potential exogenous regressors or when the true autoregressive coefficient is 0 or its MLE is close to 0, which are more likely to happen in reality.

When model uncertainty is substantial, e.g. with small sample sizes, we argue for the use of Bayesian model averaging, which can produce point estimators with smaller RMSE than the model with the highest posterior model probability in our simulation exercises. Inclusion probability criteria can be helpful to reduce estimation risk and for deciding which regressor(s) should be chosen. We recommend using 50% (the prior inclusion probability) to decide the inclusion of a regressor, which usually produces the smallest RMSE and average error rates in our simulation exercises. It can be promising for future

research to extend Lancaster's reparameterization to account for higher order AR models and to consider lag order selection along with regressor selection.

Acknowledgments

Conflicts of Interest

The author declares no conflict of interest.

Appendix

A. The DGP of Exogenous Regressors

We first draw $\underset{K \times 1}{x_{i,t}^*}$ from $i.i.d. N(0, \sigma_X^2 I_K)$ and then generate $\tilde{x}_{i,t}$ as follows

$$\tilde{x}_{i,t} = s\, \tilde{x}_{i,t-1} + \sqrt{1-s^2} x_{i,t}^*, \tag{39}$$

with $\tilde{x}_{i,0} = x_{i,0}^*$. s is the first order autocorrelation. Denote $\tilde{X}_i = (\tilde{x}_{i,1}, \ldots, \tilde{x}_{i,T})'$, $\tilde{X}_{i,j}$ and $X_{i,h}$ to be the jth and hth column of $\tilde{X}_i$ and X_i respectively, where

$$X_{i,h} = \sum_{j=1}^{K} q_{h,j} \tilde{X}_{i,j} \qquad j = 1, 2, \ldots, K. \tag{40}$$

Define $\underset{K \times 1}{z} = (\frac{1}{\sqrt{K}}, \ldots, \frac{1}{\sqrt{K}})'$. $q_h = (q_{h,1}, \ldots, q_{h,K})'$ is drawn from angular central Gaussian distribution (ACG) with $q_h' q_h = 1$ and parameter $zz' + \lambda(I - zz')$ (see [18]). q_h can be viewed as an orientation (direction) in $\mathbb{R}^K$. If $\lambda = 1$, q_h will be uniformly distributed; if λ is closer to 0, q_h will be closer to the orientation of z, *i.e.*, the regressors generated thereby will have higher correlation.[12] Note that under our data generating design, any element in X_i will have mean 0 and variance σ_X^2. The correlation coefficient of any two elements in X_i is the same across i and can be calculated as

$$corr\left(X_{i,t,h}, X_{i,\tilde{t},\tilde{h}}\right) = s^{|t-\tilde{t}|} \sum_{j=1}^{K} q_{h,j} q_{\tilde{h},j} \qquad t = 1, 2, \ldots, T \quad h = 1, 2, \ldots, K. \tag{41}$$

---

[12] When $\lambda = 1(0.01)$, by simulation, we find that the 2.5%th, 50%th and 97.5%th quantile of $|q_h' z|$ respectively are around 0.012(0.12), 0.26(0.94) and 0.73(0.99).

B. Properties of the Exogenous Regressors in the Simulation

Here we will do some robustness checks of our simulation results under different settings. Apart from the conditions in Proposition 6 to 8, model selection performance of Bayes factors and BIC is also sensitive to the properties of X_i for different values of σ_X^2, s and λ. We first reduce σ_X^2 to 1.33 to obtain the results in Table 10. The error rates of the two criteria are all higher for different sample sizes than those in Table 1 while the nest rates are all lower. Similar model selection deterioration could also occur with inflated error variances (σ^2). Hence model selection performance is affected by the relative strength of the signal compared to the noise, which is determined by their variances.

Next we show that the levels of serial correlation and collinearity in the regressors also affect model selection performance. Recall from Section A that s is the first order autocorrelation and λ controls the level of collinearity. To have the regressors with no collinearity, we can set q_h to be the hth column of an identity matrix of dimension K. We set $\underline{\rho} = 1$ and $N = 200$. The error rates under different levels of serial correlation and collinearity are shown in Table 11. We can see that cross regressor correlation and positive serial correlation are harmful for model selection. If different regressors are orthogonal to each other, Bayes factors and BIC will have lower error rates than when collinearity is present, while the highest error rates appear when s is 0.9 under different levels of collinearity. One intriguing phenomenon is that negative serial correlation seems to enhance model selection performance for most cases in comparison to positive or no serial correlation.

Table 10. Simulation results when $\underline{\rho} = -1$ and $\sigma_X^2 = 1.33$.

N	ER	ERBIC	ER11	ERBIC11	Nest	Nestbic
40	0.962	0.958	1	1	0.501	0.497
100	0.886	0.872	1	1	0.736	0.716
200	0.767	0.747	1	1	0.827	0.812
500	0.500	0.470	1	1	0.869	0.856
1000	0.221	0.217	1	1	0.910	0.896

Table 11. Error rates under different levels of serial correlation and collinearity for $\underline{\rho} = 1$ and $N = 200$.

	Bayes Factors					BIC				
λs	**–0.9**	**–0.5**	**0**	**0.5**	**0.9**	**–0.9**	**–0.5**	**0**	**0.5**	**0.9**
orthogonal	0.045	0.039	0.061	0.085	0.569	0.098	0.081	0.089	0.127	0.543
$\lambda = 1$	0.135	0.154	0.174	0.276	0.779	0.148	0.178	0.186	0.278	0.762
$\lambda = 0.01$	0.638	0.661	0.696	0.803	0.972	0.600	0.627	0.678	0.794	0.966

C. Proof of Proposition 3 and Proposition 5

Here we use a different way of derivation from Lancaster [11]. In brief, we attempt to find a correction function attached to the marginal posterior density of ρ such that the mode of the marginal posterior is a consistent estimator for ρ. We first reparameterize the fixed effect as

$$f_i = g_i r(\rho) - \frac{1}{T} \iota' X_i \beta \tag{42}$$

where $r(\rho)$ is a function of ρ, which we will find out later. The derivation of the conditional posterior distribution $p(g_i, \beta, \sigma^2 | \rho, Y, Y_0)$ follows standard Bayesian techniques, see e.g., [19] Chapter 10. The details are available upon request. Here we just show the results after g_i, β and σ^2 are integrated out.

$$
\begin{aligned}
p(\rho|Y, Y_0)p(Y|Y_0) =& \Gamma \left[\frac{N(T-1)}{2} \right] T^{-\frac{N}{2}} (2\pi)^{-\frac{N(T-1)}{2}} \frac{I(\rho_L < \rho < \rho_U)}{\rho_U - \rho_L} \left(\frac{\eta}{\eta + 1} \right)^{\frac{k}{2}}, \\
& N^{-\frac{N(T-1)}{2}} \left(\frac{a}{N} \rho^2 - 2\frac{b}{N} \rho + \frac{c}{N} \right)^{-\frac{N(T-1)}{2}} r^{-N}(\rho).
\end{aligned}
\tag{43}
$$

Taking log and differentiating both sides with respect to ρ produces

$$\frac{d \ln p(\rho|Y, Y_0)}{d\rho} = -\frac{N(T-1)(a\rho - b)}{a\rho^2 - 2b\rho + c} - N\frac{d \ln r(\rho)}{d\rho}.$$

By setting the above equal to 0, we can obtain

$$\frac{d \ln r(\rho)}{d\rho} = -\frac{(T-1)(\frac{a}{N}\rho - \frac{b}{N})}{\frac{a}{N}\rho^2 - 2\frac{b}{N}\rho + \frac{c}{N}}.$$

Suppose for now we have included the true regressors in our model. Taking probability limit of the right hand side by using Equations (27), (28) and (29) and evaluating both sides at $\underline{\rho}$ gives

$$\frac{d \ln r(\underline{\rho})}{d\rho} = -h(\underline{\rho}). \tag{44}$$

Solving the above differential equation, we will have[13]

$$r(\rho) = exp\left[-\phi(\rho)\right], \tag{45}$$

where $\phi(\rho)$ is given in Equation (7). By replacing $r(\rho)$ with $exp\left[-\phi(\rho)\right]$ in Equation (43) and dropping the terms not involving ρ, we will have the result in Equation (13).

When some true regressors are excluded from the model, the differential Equation (44) now becomes

$$\frac{d \ln r(\underline{\rho})}{d\underline{\rho}} = \frac{(T-1)\left[h_2(\beta, \underline{\rho}) - \sigma^2 h(\underline{\rho})\right]}{h_3(\beta) + (T-1)\sigma^2} d\underline{\rho} \tag{46}$$

---

[13] Strictly speaking, the right hand side should be multiplied by an arbitrary constant not involving ρ.

If the solution in Equation (45) is still valid, we should have

$$\frac{-(T-1)h_2(\beta,\underline{\rho}) + (T-1)\sigma^2 h(\underline{\rho})}{h_3(\beta) + (T-1)\sigma^2} = h(\underline{\rho}).$$

So unless we have either $\frac{-(T-1)h_2(\beta,\underline{\rho})}{h_3(\beta)} = h(\underline{\rho})$ or $h_2(\beta,\underline{\rho}) = h_3(\beta) = 0$, Equation (45) will not be a solution for Equation (46). In other words, the reparameterization of the fixed effect in Equation (6) cannot lead to consistent estimation of ρ.

D. Proof of Proposition 4

To prove the claims in Proposition 4, we first need to prove Lemma 9 and Lemma 10.

Lemma 9. *For $T \geq 3$, when T is odd, the polynomial $h(\rho)$ is strictly increasing over $(-\infty,\infty)$ and has only one real root in $[-2,-1)$; when T is even, the polynomial $h(\rho)$ is greater than 0 with no real roots and is strictly decreasing for $\rho < -1$ and strictly increasing for $\rho > -1$ with -1 as the minimum point.*

Proof. When $T = 3$, we have $h(\rho) = \frac{1}{3}(\rho+2)$; when $T = 4$, we have $h(\rho) = \frac{1}{4}(\rho^2+2\rho+3)$. These two cases obviously satisfy the claims in Lemma 9. For $T > 4$, note that $h'(\rho) = \frac{d\,h(\rho)}{d\,\rho} = \sum_{t=1}^{T-2}\frac{t(T-t-1)}{T}\rho^{t-1} = \frac{(T-2)\rho^T - T\rho^{T-1} + T\rho - (T-2)}{T(\rho-1)^3}$ and the Sturm sequence of $Th'(\rho)(\rho-1)^3$ is

$$\left\{ Th'(\rho)(\rho-1) = \sum_{t=0}^{T-2}(2t-T+2)\rho^t, \right.$$
$$\left. \frac{1}{(\rho-1)^2}\frac{d\,h'(\rho)(\rho-1)^3}{d\,\rho} = \sum_{t=1}^{T-2}t\rho^{t-1}, \sum_{t=1}^{T-3}(T-t-2)\rho^{t-1}, -(T-2)^2 \right\}. \tag{47}$$

Table 12 shows the signs of the Sturm sequence for $\rho = \pm\infty$ and T being even. We can see that the difference between the number of sign changes when ρ changes from $-\infty$ to ∞ is 2. By Sturm's theorem, this implies there are two real roots for $Th'(\rho)(\rho-1)^3 = 0$ in $(-\infty,\infty)$. Clearly, $\rho = 1$ is one real root. In other words, $h'(\rho)$ has only one real root in $(-\infty,\infty)$, which is $\rho = -1$, and $h'(\rho) > 0$ for $\rho > -1$, $h'(\rho) < 0$ for $\rho < -1$. Therefore $h(\rho)$ is strictly decreasing for $\rho < -1$ and strictly increasing for $\rho > -1$ with $\rho = -1$ as the minimum point. Similarly, checking the difference between the number of sign changes in Table 13, we can find that $h'(\rho)$ has no real roots and $h'(\rho) > 0$. Hence $h(\rho)$ is strictly increasing over the real line when T is odd.

Table 12. Sturm sequence of $Th'(\rho)(\rho-1)^3$ when T is even and greater than 4.

ρ	$h'(\rho)$	$\sum_{t=0}^{T-2}(2t-T+2)\rho^t$	$\sum_{t=1}^{T-2}t\rho^{t-1}$	$\sum_{t=1}^{T-3}(T-t-2)\rho^{t-1}$	$-(T-2)^2$
$-\infty$	$-$	$+$	$-$	$+$	$-$
∞	$+$	$+$	$+$	$+$	$-$

Table 13. Sturm sequence of $Th'(\rho)(\rho-1)^3$ when T is odd and greater than 4.

ρ	$h'(\rho)$	$\sum_{t=0}^{T-2}(2t-T+2)\rho^t$	$\sum_{t=1}^{T-2}t\rho^{t-1}$	$\sum_{t=1}^{T-3}(T-t-2)\rho^{t-1}$	$-(T-2)^2$
$-\infty$	$+$	$-$	$+$	$-$	$-$
∞	$+$	$+$	$+$	$+$	$-$

We can write $h(\rho) = \frac{\rho^T - T\rho + T - 1}{T(\rho-1)^2}$ and the Sturm sequence of $Th(\rho)(\rho-1)^2$ is

$$\left\{ Th(\rho)(\rho-1) = \sum_{t=1}^{T-2}\rho^t - (T-1), \frac{1}{(\rho-1)}\frac{d\,h(\rho)(\rho-1)^2}{d\,\rho} = \sum_{t=0}^{T-2}\rho^t, T-1 \right\}. \tag{48}$$

From Table 14, we can see that $h(\rho)$ does not have real roots and hence it is greater than 0 when T is an even number. Table 15 shows $h(\rho)$ has only one real root when T is odd. Since $h(-2) = \frac{(-2)^T + 3T - 1}{9T} \le 0$, $h(-1) = \frac{T-1}{2T} > 0$ and $h(\rho)$ is strictly increasing when T is an odd number no less than 3, the real root of $h(\rho) = 0$ must lie in between -2 and -1.

Table 14. Sturm sequence of $Th(\rho)(\rho-1)^2$ when T is even and greater than or equal to 3.

ρ	$h(\rho)$	$\sum_{t=1}^{T-1}\rho^t - T + 1$	$\sum_{t=0}^{T-2}\rho^t$	$T-1$
$-\infty$	$+$	$-$	$+$	$+$
∞	$+$	$+$	$+$	$+$

Table 15. Sturm sequence of $Th(\rho)(\rho-1)^2$ when T is odd and greater than or equal to 3.

ρ	$h(\rho)$	$\sum_{t=1}^{T-1}\rho^t - T + 1$	$\sum_{t=0}^{T-2}\rho^t$	$T-1$
$-\infty$	$-$	$+$	$-$	$+$
∞	$+$	$+$	$+$	$+$

$\square$

Additionally, we need the following lemma to show that the true value of ρ is a local posterior mode asymptotically.

Lemma 10. *Under Assumption 1 and 2, we have* $\plim_{N\to\infty}\frac{a}{N} = \underline{a} > \sigma^2 trace(C'HC)$, *where ρ is evaluated at its true value ($\underline{\rho}$) in C. Also, we have* $trace(C'HC) \ge \frac{2h^2(\underline{\rho})}{T-1}$, *where the equal sign holds only for $T = 2$, and* $trace(C'HC) \ge \frac{2h^2(\underline{\rho})}{T-1} + h'(\underline{\rho})$, *where the equal sign holds for $T = 2$ or $\underline{\rho} = 1$. In other words, the following are true:*

$$\underline{a} > \frac{2\sigma^2 h^2(\underline{\rho})}{T-1}, \tag{49}$$

$$\underline{a} > \frac{2\sigma^2 h^2(\underline{\rho})}{T-1} + h'(\underline{\rho})\sigma^2. \tag{50}$$

Proof. Substituting Equation (3) into the right of Equation (15) gives

$$
a = \sum_{i=1}^{N} (f_i\zeta_1 + y_{i,0}\zeta_2 + CX_i\beta + Cu_i)' H (f_i\zeta_1 + y_{i,0}\zeta_2 + CX_i\beta + Cu_i)
$$

$$
-\frac{\sum_{i=1}^{N} [(f_i\zeta_1 + y_{i,0}\zeta_2 + CX_i\beta + Cu_i)' HX_i] \left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1} \sum_{i=1}^{N} [X_i'H(f_i\zeta_1 + y_{i,0}\zeta_2 + CX_i\beta + Cu_i)]}{\eta + 1}.
$$

(51)

Since we assume $E(u_i|X_i, f_i, y_{i,0}) = 0$ and set $\eta = O(N^\alpha)$ with $\alpha < 0$, $\frac{a}{N}$ is asymptotically equivalent to $\frac{\tilde{a}}{N}$, where $\tilde{a}$ is defined as

$$
\begin{aligned}
\tilde{a} =& \sum_{i=1}^{N} u_i'C'HCu_i + \sum_{i=1}^{N} (f_i\zeta_1 + y_{i,0}\zeta_2 + CX_i\beta)' H (f_i\zeta_1 + y_{i,0}\zeta_2 + CX_i\beta) \\
&- \sum_{i=1}^{N} [(f_i\zeta_1 + y_{i,0}\zeta_2 + CX_i\beta)' HX_i] \left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1} \sum_{i=1}^{N} [X_i'H(f_i\zeta_1 + y_{i,0}\zeta_2 + CX_i\beta)], \\
=& \sum_{i=1}^{N} u_i'C'HCu_i + \sum_{i=1}^{N} (y_- - Cu_i)' H (y_- - Cu_i) \\
&- \sum_{i=1}^{N} [(y_- - Cu_i)' HX_i] \left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1} \sum_{i=1}^{N} [X_i'H(y_- - Cu_i)].
\end{aligned}
$$

(52)

Note that $\frac{1}{N}\left(\tilde{a} - \sum_{i=1}^{N} u_i'C'HCu_i\right)$ is non-negative since it is equal to $\frac{1}{N}$ multiplied by the SSR obtained by regressing $f_i\zeta_1 + y_{i,0}\zeta_2 + CX_i\beta$ on fixed effects and X_i, *i.e.*,

$$
f_i\zeta_1 + y_{i,0}\zeta_2 + CX_i\beta = q_i\iota + X_i\vartheta + \varepsilon_i,
$$

(53)

where q_i denotes the fixed effect scalar and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT})'$ is the error term in the regression. Assumption 2 (e) rules out $plim_{N\to\infty}\frac{\sum_{i=1}^{N}\varepsilon_i'\varepsilon_i}{N} = 0$. Note that $(1-\rho)\zeta_1 + \zeta_2 = \iota$. When $T \geq 3$, pre-multiplying both sides of Equation (53) by M_ζ yields

$$
M_\zeta CX_i\beta = M_\zeta X_i\vartheta + M_\zeta\varepsilon_i.
$$

(54)

Therefore one can use Equation (5) to check Assumption 2 (e), which ensures the $\frac{SSR}{N}$ from Equation (53) to be strictly positive asymptotically and

$$
\underline{a} > plim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} u_i'C'HCu_i = \sigma^2 trace(C'HC).
$$

(55)

Hence Equation (49) is strict if $trace(C'HC) \geq \frac{2h^2(\rho)}{T-1}$. Similarly, to prove Equation (50), one needs to show $trace(C'HC) \geq \frac{2h^2(\rho)}{T-1} + h'(\rho)$. The proof of these two inequalities can be found in the proof of Lemma 3 in Dhaene and Jochmans [8], by noting $V_0^{LB} = \frac{trace(C'HC)}{T-1}$, $b_0 = -\frac{h(\rho)}{T-1}$ and $c_0 = -\frac{h'(\rho)}{T-1}$, where V_0^{LB}, b_0 and c_0 are the notations used in their paper. $\square$

From Equation (14), we can see that $\psi(\rho) = \phi(\rho) - l(\rho)$ and $l(\rho) = \frac{T-1}{2}\ln(\frac{a}{N}\rho^2 - 2\frac{b}{N}\rho + \frac{c}{N})$. The posterior mode is found by checking the intersection points of $h(\rho)$, *i.e.*, $\frac{d\phi(\rho)}{d\rho}$ and $l'(\rho)$, which is

$$l'(\rho) = \frac{(T-1)(\frac{a}{N}\rho - \frac{b}{N})}{\frac{a}{N}\rho^2 - 2\frac{b}{N}\rho + \frac{c}{N}}. \tag{56}$$

Assuming that the true exogenous regressors are included such that $h_2(\beta, \underline{\rho}) = h_3(\beta) = 0$ and using Equation (27) to (30), we can find that the probability limit of $l'(\rho)$, denoted as $\underline{l}'(\rho)$, is

$$\underline{l}'(\rho) = \underset{N\to\infty}{pliml}'(\rho) = \frac{(T-1)(\rho - \underline{\rho} - \gamma)}{(\rho - \underline{\rho} - \gamma)^2 + \frac{\sigma^2}{\underline{a}^2}\left[\underline{a}(T-1) - \sigma^2 h^2(\underline{\rho})\right]}. \tag{57}$$

From Equation (49) in Lemma 10, we know that $\frac{\sigma^2}{\underline{a}^2}\left[\underline{a}(T-1) - \sigma^2 h^2(\underline{\rho})\right] > \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}^2} \geq 0$. Hence the denominator in Equation (57) is positive and $\underline{l}'(\rho) \geq (<)0$ for $\rho \geq (<)\underline{\rho} + \gamma$. Moreover,

$$\underline{l}''(\rho) = \underset{N\to\infty}{pliml}''(\rho) = -(T-1)\frac{(\rho - \underline{\rho} - \gamma)^2 - \frac{\sigma^2}{\underline{a}^2}\left[\underline{a}(T-1) - \sigma^2 h^2(\underline{\rho})\right]}{\left((\rho - \underline{\rho} - \gamma)^2 + \frac{\sigma^2}{\underline{a}^2}\left[\underline{a}(T-1) - \sigma^2 h^2(\underline{\rho})\right]\right)^2}. \tag{58}$$

The denominator above is positive. The polynomial in the numerator has two roots: $\underline{\rho} + \gamma \pm \sqrt{\frac{\sigma^2}{\underline{a}^2}\left[\underline{a}(T-1) - \sigma^2 h^2(\underline{\rho})\right]}$. Using Equation (49) in Lemma 10 again, we have $\underline{\rho} + \gamma + \sqrt{\frac{\sigma^2}{\underline{a}^2}\left[\underline{a}(T-1) - \sigma^2 h^2(\underline{\rho})\right]} > \underline{\rho} - \frac{\sigma^2 h(\underline{\rho})}{\underline{a}} + \frac{\sigma^2}{\underline{a}}\left|h(\underline{\rho})\right|$ and $\underline{\rho} + \gamma - \sqrt{\frac{\sigma^2}{\underline{a}^2}\left[\underline{a}(T-1) - \sigma^2 h^2(\underline{\rho})\right]} < \underline{\rho} - \frac{\sigma^2 h(\underline{\rho})}{\underline{a}} - \frac{\sigma^2}{\underline{a}}\left|h(\underline{\rho})\right|$. Therefore the true value of ρ, *i.e.*, $\underline{\rho}$ lies in between the two roots of $\underline{l}''(\rho) = 0$, where $\underline{l}'(\rho)$ is increasing.[14] When ρ is larger (less) than the bigger (smaller) root, $\underline{l}'(\rho)$ will be decreasing. Since $\underline{l}'(\rho) \geq (<)0$ for $\rho \geq (<)\underline{\rho} + \gamma$, we can see that $\lim_{\rho\to\pm\infty}\underline{l}'(\rho) = 0$. Since $\underline{l}'(\underline{\rho}) = h(\underline{\rho})$, $h(\rho)$ therefore intersects $\underline{l}'(\rho)$ at $\underline{\rho}$. Define $\underline{\psi}(\rho) = \phi(\rho) - \underset{N\to\infty}{pliml}(\rho)$. Evaluating its second order derivative at the true value of ρ yields $\underline{\psi}''(\underline{\rho}) = h'(\underline{\rho}) - \frac{\left[(T-1)\underline{a} - 2\sigma^2 h^2(\underline{\rho})\right]}{(T-1)\sigma^2}$. Using Equation (50) in Lemma 10, we can find $\underline{\psi}''(\underline{\rho}) < 0$. In other words, $\underline{\rho}$ is a local maximum for $\underline{\psi}(\rho)$. When T is even, because $h(\rho) > 0$ is increasing after -1 and $\underline{l}'(\rho)$ is decreasing beyond the bigger root of $\underline{l}''(\rho) = 0$, $h(\rho)$ should intersect $\underline{l}'(\rho)$ again as at the point ρ_U in Figure 1, which is larger than $\underline{\rho}$ and the bigger root of $\underline{l}''(\rho) = 0$. We can see that $\underline{\psi}(\rho)$ has bell shape over $(-\infty, \rho_U]$ when T is even. However, for $\rho > \rho_U$, $\underline{\psi}(\rho)$ is increasing and hence $\lim_{\rho\to\infty}\underline{\psi}(\rho) = \infty$. To ensure the marginal posterior of ρ to be proper, we have to restrict the bounds of ρ in our estimation. Similarly, when T is odd, since $h(\rho)$ is strictly increasing, $h(\rho)$ should intersect $\underline{l}'(\rho)$ on the left hand side of $\underline{\rho}$. There should be three intersection points as shown in Figure 2. Since $\underline{\psi}(\rho)$ is decreasing for $\rho < \rho_L$ due to $\underline{\psi}'(\rho) < 0$, we have $\lim_{\rho\to-\infty}\underline{\psi}(\rho) = \infty$ when T is odd. Choosing ρ_L and ρ_U in the way described by Proposition 4 can ensure $\underline{\psi}(\rho)$ evaluated at the end points to be smaller than $\underline{\psi}(\underline{\rho})$ and the marginal posterior density of ρ to be proper.

---

[14] Note that $\underline{a}$ and σ^2 are positive. When $h(\underline{\rho}) \geq 0$, we can have $\underline{\rho} + \gamma + \sqrt{\frac{\sigma^2}{\underline{a}^2}\left[\underline{a}(T-1) - \sigma^2 h^2(\underline{\rho})\right]} > \underline{\rho}$ and $\underline{\rho} + \gamma - \sqrt{\frac{\sigma^2}{\underline{a}^2}\left[\underline{a}(T-1) - \sigma^2 h^2(\underline{\rho})\right]} < \underline{\rho} - \frac{2\sigma^2 h(\underline{\rho})}{\underline{a}} \leq \underline{\rho}$; when $h(\underline{\rho}) < 0$, we can have $\underline{\rho} + \gamma + \sqrt{\frac{\sigma^2}{\underline{a}^2}\left[\underline{a}(T-1) - \sigma^2 h^2(\underline{\rho})\right]} > \underline{\rho} - \frac{2\sigma^2 h(\underline{\rho})}{\underline{a}} > \underline{\rho}$ and $\underline{\rho} + \gamma - \sqrt{\frac{\sigma^2}{\underline{a}^2}\left[\underline{a}(T-1) - \sigma^2 h^2(\underline{\rho})\right]} < \underline{\rho}$.
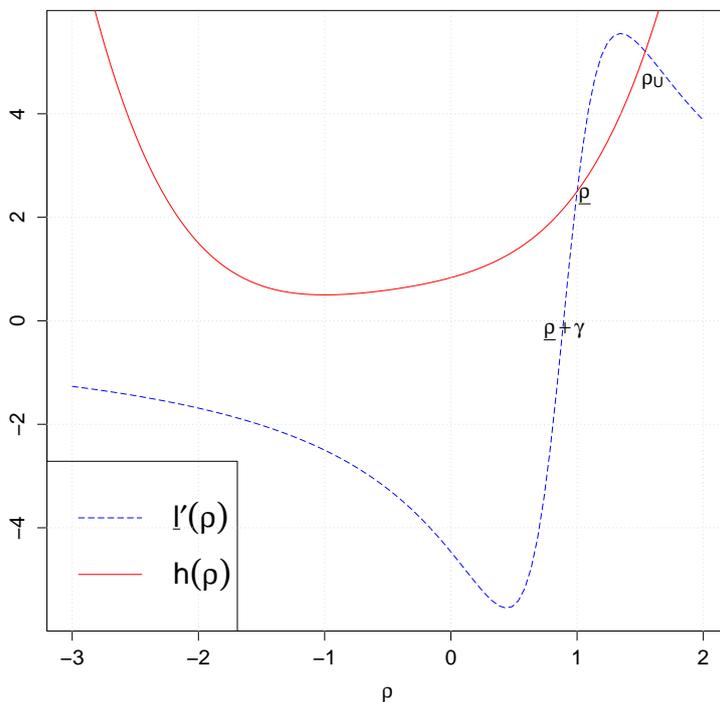
Figure 1. Intersection points of $h(\rho)$ and $\underline{l}'(\rho)$ when $T = 6$, $\underline{\rho} = 1$, $\sigma^2 = 1$, $\sigma_f^2 = 1$, $\sigma_{y_0}^2 = 0.1$, $E(f_i) = E(y_{i,0}) = E(f_i y_{i,0}) = 0$ and there are no exogenous regressors.
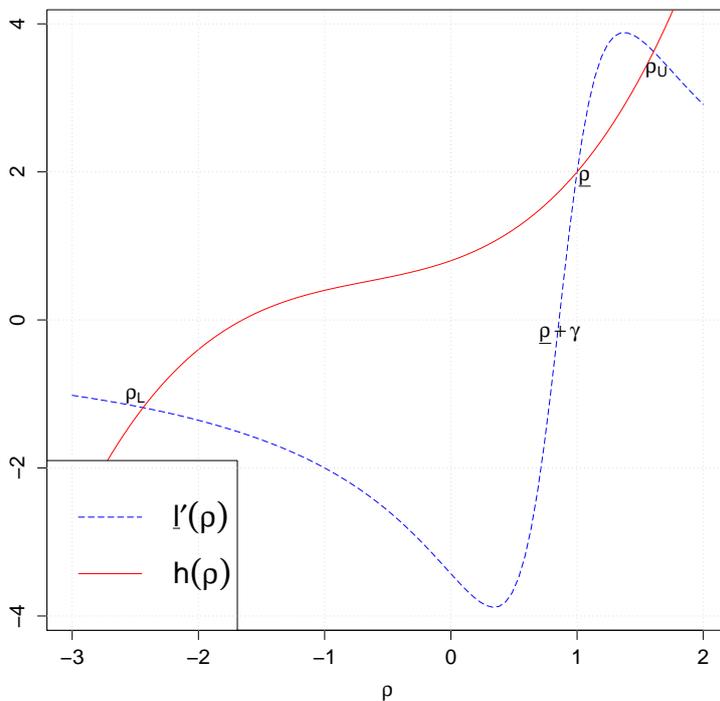
Figure 2. Intersection points of $h(\rho)$ and $\underline{l}'(\rho)$ when $T = 5$, $\underline{\rho} = 1$, $\sigma^2 = 1$, $\sigma_f^2 = 1$, $\sigma_{y_0}^2 = 0.1$, $E(f_i) = E(y_{i,0}) = E(f_i y_{i,0}) = 0$ and there are no exogenous regressors.

E. Proof of Proposition 6

To prove Proposition 6 and 7, essentially we need to simplify the integral(s) which appears in the Bayes factor. One way to do it is Laplace's method, the details of which can be found in [20,21]. Under the assumption that there exists only one solution ρ^* in (ρ_L, ρ_U) for $\psi'(\rho) = 0$ with $\psi''(\rho^*) < 0$, the integral appearing in the Bayes factor can be written as

$$\int_{\rho_L}^{\rho_U} \exp\left[N\psi(\rho)\right] = \sqrt{\frac{2\pi}{N\left|\psi''(\rho^*)\right|}} \exp\left[N\psi(\rho^*)\right]\left(1 + O(\frac{1}{N})\right) \tag{59}$$

Building on Equation (18), the first and second order derivatives of $\underline{\psi}(\rho)$ are

$$\underline{\psi}'(\rho) = h(\rho) - \frac{(T-1)(\rho - \underline{\rho} - \gamma)}{\left[\rho^2 - 2\rho(\underline{\rho} + \gamma) + \underline{\rho}^2 + 2\underline{\rho}\gamma + \frac{(T-1)\sigma^2 + h_3(\beta)}{a}\right]}, \tag{60}$$

$$\underline{\psi}''(\rho) = h'(\rho) - $$
$$\frac{(T-1)\left[\rho^2 - 2\rho(\underline{\rho} + \gamma) + \underline{\rho}^2 + 2\underline{\rho}\gamma + \frac{(T-1)\sigma^2 + h_3(\beta)}{a} - 2(\rho - \underline{\rho} - \gamma)^2\right]}{\left[\rho^2 - 2\rho(\underline{\rho} + \gamma) + \underline{\rho}^2 + 2\underline{\rho}\gamma + \frac{(T-1)\sigma^2 + h_3(\beta)}{a}\right]^2}. \tag{61}$$

If the chosen set of regressors can lead to consistent estimation of ρ, *i.e.*, either Equation (20) or (21) is satisfied, evaluating Equations (18), (60) and (61) at $\underline{\rho}$ will give

$$\underline{\psi}(\underline{\rho}) = \phi(\underline{\rho}) - \frac{T-1}{2}\ln\left[(T-1)\sigma^2 + h_3(\beta)\right],$$

$$\underline{\psi}'(\underline{\rho}) = 0,$$

$$\underline{\psi}''(\underline{\rho}) = h'(\underline{\rho}) - \frac{a(T-1)}{(T-1)\sigma^2 + h_3(\beta)} + \frac{2h^2(\underline{\rho})}{T-1}.$$

The Bayes factor in Equation (25) is

$$\frac{p(Y|Y_0, M_1)}{p(Y|Y_0, M_0)} = \left(\frac{\eta}{\eta+1}\right)^{\frac{k_1 - k_0}{2}} \sqrt{\frac{2\pi}{N\left|\psi''(\rho^*_{|M_1}|M_1)\right|}} \frac{\exp\left[N\psi(\rho^*_{|M_1}|M_1)\right]}{(\rho_{U_1} - \rho_{L_1})\left(\frac{c_{|M_0}}{N}\right)^{-\frac{N(T-1)}{2}}}\left(1 + O(\frac{1}{N})\right). \tag{62}$$

Asymptotically, replacing $\psi(\rho^*_{|M_1}|M_1)$, $\psi''(\rho^*_{|M_1}|M_1)$ and $\frac{c_{|M_0}}{N}$ by their probability limits and $\frac{\eta}{\eta+1}$ by $O(N^\alpha)$ with $\alpha < 0$ (our prior choice for η) should not affect the analysis of the Bayes factor. Define

$$\xi_{10} = \frac{O\left(N^{\frac{\alpha(k_1 - k_0)}{2}}\right)}{\rho_{U_1} - \rho_{L_1}}\sqrt{\frac{2\pi}{N\left|\underline{\psi}''(\rho^*_{|M_1}|M_1)\right|}}\exp\left[N\left(\underline{\psi}(\rho^*_{|M_1}|M_1) + \frac{T-1}{2}\ln\underline{c}_{|M_0}\right)\right], \tag{63}$$

which should have the same asymptotic behaviour as Equation (62). If X_{i1} is the true set of regressors to generate Y (so $h_2(\beta, \underline{\rho}|M_1) = h_3(\beta|M_1) = 0$ and $\rho^*_{|M_1} = \underline{\rho}$), ξ_{10} can be written as

$$\xi_{10} = \frac{O\left(N^{\frac{\alpha(k_1 - k_0)}{2}}\right)}{\rho_U - \rho_L}\sqrt{\frac{2\pi}{N\left|h'(\underline{\rho}) - \frac{a_{|M_1}}{\sigma^2} + \frac{2h^2(\underline{\rho})}{T-1}\right|}}\exp\left\{N\phi(\underline{\rho}) + \frac{N(T-1)}{2}\ln\left[\frac{\underline{c}_{|M_0}}{(T-1)\sigma^2}\right]\right\}. \tag{64}$$

So we can guarantee $\frac{p(Y|Y_0, M_1)}{p(Y|Y_0, M_0)}$ tends to infinity given $\underline{\rho} \neq 0$ as long as Equation (31) holds.

Now let us consider the case when the true model is M_0 in Equation (25), *i.e.*, $\underline{\rho}$ is 0 and X_{i0} is the set of true regressors. ξ_{10} takes the following form,

$$
\begin{aligned}
\xi_{10} =& \frac{[(T-1)\sigma^2]^{\frac{N(T-1)}{2}}}{\rho_{U_1} - \rho_{L_1}} O\left(N^{\frac{\alpha(k_1-k_0)}{2}}\right) \sqrt{\frac{2\pi}{N \left|\underline{\psi}''(\rho^*_{|M_1}|M_1)\right|}} \exp\left[N\psi(\rho^*_{|M_1}|M_1)\right] \\
=& O\left(N^{\frac{\alpha(k_1-k_0)}{2}}\right) \sqrt{\frac{2\pi}{N \left|\underline{\psi}''(\rho^*_{|M_1}|M_1)\right|}} \frac{\exp\left[N\phi(\rho^*_{|M_1}) + \frac{N(T-1)}{2}\ln\frac{(T-1)\sigma^2}{d(\rho^*_{|M_1}|M_1)}\right]}{\rho_{U_1} - \rho_{L_1}}.
\end{aligned}
\tag{65}
$$

If Equation (32) holds, then the Bayes factor in Equation (25) will tend to 0 for large sample size. If M_1 is misspecified, but ρ can be consistently estimated, *i.e.*, $\rho^*_{|M_1} = 0$, Equation (65) can be simplified as

$$
\xi_{10} = O\left(N^{\frac{\alpha(k_1-k_0)}{2}}\right) \sqrt{\frac{2\pi}{N \left|\underline{\psi}''(0|M_1)\right|}} \frac{\exp\left[\frac{N(T-1)}{2}\ln\left[\frac{(T-1)\sigma^2}{(T-1)\sigma^2 + h_3(\beta|M_1)}\right]\right]}{\rho_{U_1} - \rho_{L_1}}.
\tag{66}
$$

If Equation (21) holds, some true regressors should be excluded from M_1 and hence Equation (32) is true with $h_3(\beta|M_1) > 0$. If Equation (20) holds, we have $h_3(\beta|M_1) = 0$ and $k_1 \geq k_0$. The Bayes factor will therefore tend to 0 when N tends to infinity.

Finally we show when Equation (31) will be violated. If M_0 only includes the true exogenous regressors, we should have $h_2(\beta, \underline{\rho}|M_0) = h_3(\beta|M_0) = 0$ and $\underline{a}_{|M_0} > \sigma^2 trace(C'HC)$ as in Equation (55). The following should be true:

$$
\begin{aligned}
&\phi(\underline{\rho}) + \frac{T-1}{2}\ln\left[\frac{\underline{a}_{|M_0}\underline{\rho}^2 - 2\underline{\rho}\sigma^2 h(\underline{\rho})}{(T-1)\sigma^2} + 1\right] > \upsilon(\underline{\rho}) = \\
&\phi(\underline{\rho}) + \frac{T-1}{2}\ln\left[1 + \frac{\underline{\rho}^2 \sum_{j=0}^{T-2}(T-j-1)\underline{\rho}^{2j}}{T-1} - \frac{\underline{\rho}^2 \sum_{j=0}^{T-2}\left(\sum_{i=0}^{j}\underline{\rho}^i\right)^2}{T(T-1)} - \frac{2\underline{\rho}h(\underline{\rho})}{T-1}\right].
\end{aligned}
\tag{67}
$$

When $T = 2$, the right hand side of Equation (67), *i.e.*, $\upsilon(\underline{\rho})$, will be $\frac{\rho}{2} + \frac{1}{2}\ln(1 - \underline{\rho} + \frac{\rho^2}{2})$ and it is an increasing function with $\upsilon(0) = 0$. Hence $\upsilon(\underline{\rho})$ is less than 0 when $\underline{\rho} < 0$. The left hand side of Equation (67) can be negative for $\underline{\rho} < 0$ if σ^2 is much larger than $E(f_i^2)$ and $E(y_{i,0}^2)$.[15] When T is an odd number great than or equal to 3, $\upsilon(\rho)$ is positive for $\rho \in \mathbb{R}$. When T is even and greater than 2, $\upsilon(\rho)$ is positive for $\rho \in (-1, \infty)$ and has a root less than -1. If $\underline{\rho}$ is less than the root, $\upsilon(\underline{\rho})$ will be negative. By direct calculation, as T increases, we find that the root of $\upsilon(\rho)$ will get closer to -1 from the left. Hence, to sum up, Equation (31) will hold for $\underline{\rho} \in (-1, \infty)$ when T is any integer greater than or equal to 3 and $h_2(\beta, \underline{\rho}|M_0) = h_3(\beta|M_0) = 0$.

---

[15] When $T = 2$, $E(f_i y_{i,0}) = 0$ and the true model contains no exogenous regressors, $\phi(\underline{\rho}) + \frac{T-1}{2}\ln\left[\frac{\underline{a}_{|M_0}\underline{\rho}^2 - 2\underline{\rho}\sigma^2 h(\underline{\rho})}{(T-1)\sigma^2} + 1\right]$

is equal to $\frac{\rho}{2} + \frac{1}{2}\ln\left(1 - \underline{\rho} + \frac{[\frac{E(f_i^2)}{2} + E(y_{i,0}^2)(\frac{\rho^2}{2} - \underline{\rho} + \frac{1}{2}) + \frac{\sigma^2}{2}]\underline{\rho}^2}{\sigma^2}\right)$, which approaches $\upsilon(\underline{\rho})$ as $\frac{E(f_i^2)}{\sigma^2}$ and $\frac{E(y_{i,0}^2)}{\sigma^2}$ get smaller.

F. Proof of Proposition 7

By Laplace's method, we can write Equation (26) as

$$
\frac{p(Y|Y_0, X_1, M_1)}{p(Y|Y_0, X_2, M_2)} = \frac{\rho_{U_2} - \rho_{L_2}}{\rho_{U_1} - \rho_{L_1}} \left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_2}{2}} \sqrt{\frac{\psi''(\rho^*_{|M_2}|M_2)}{\psi''(\rho^*_{|M_1}|M_1)}}
$$
$$
\exp\left[N\left(\psi(\rho^*_{|M_1}|M_1) - \psi(\rho^*_{|M_2}|M_2)\right)\right]\left(1 + O(\frac{1}{N})\right). \tag{68}
$$

Suppose the true model is M_2, similar to the previous section, by dropping $\frac{\rho_{U_2}-\rho_{L_2}}{\rho_{U_1}-\rho_{L_1}}\sqrt{\left|\frac{\psi''(\rho^*_{|M_2}|M_2)}{\psi''(\underline{\rho}|M_1)}\right|} = O(1)$, we can obtain the corresponding ξ_{12},

$$
\xi_{12} = O\left(N^{\frac{\alpha(k_1-k_2)}{2}}\right)\exp\left\{N\left[\phi(\rho^*_{|M_1}) - \phi(\underline{\rho}) + \frac{T-1}{2}\ln\frac{d(\underline{\rho}|M_2)}{d(\rho^*_{|M_1}|M_1)}\right]\right\}. \tag{69}
$$

Note that $d(\underline{\rho}|M_2) = (T-1)\sigma^2$. So if Equation (33) is satisfied, the Bayes factor is consistent in model selection. If M_2 despite being misspecified can still lead to consistent estimation of ρ, ξ_{12} will become

$$
\xi_{12} = O\left(N^{\frac{\alpha(k_1-k_2)}{2}}\right)\left\{\frac{(T-1)\sigma^2}{(T-1)\sigma^2 + h_3(\beta|M_1)}\right\}^{\frac{N(T-1)}{2}}. \tag{70}
$$

If Equation (21) holds, we will have $h_3(\beta|M_1) > 0$ and hence Equation (33); if Equation (20) holds, M_1 will nest M_2 $(k_1 > k_2)$ with $h_3(\beta|M_2) = 0$. For both cases, $\frac{p(Y|Y_0, X_1, M_1)}{p(Y|Y_0, X_2, M_2)}$ will tend to 0.

G. Proof of Proposition 8

The likelihood function takes the following form,

$$
p(Y|\theta, Y_0) = (2\pi)^{-\frac{TN}{2}}\sigma^{2(-\frac{NT}{2})}
$$
$$
\prod_{i=1}^{N}\exp\{-\frac{1}{2\sigma^2}[y_i - y_{i\_}\rho - \iota f_i - X_i\beta]'[y_i - y_{i\_}\rho - \iota f_i - X_i\beta]\}. \tag{71}
$$

By taking log of the likelihood function and solving the first order condition, we can obtain the maximum likelihood estimators as the following,

$$
\widehat{\sigma^2} = \frac{1}{NT}\sum_{i=1}^{N}[y_i - y_{i\_}\hat{\rho} - \iota\hat{f}_i - X_i\hat{\beta}]'\left[y_i - y_{i\_}\hat{\rho} - \iota\hat{f}_i - X_i\hat{\beta}\right], \tag{72}
$$

$$
\hat{f}_i = \frac{\iota'(y_i - y_{i\_}\hat{\rho} - X_i\hat{\beta})}{T}, \tag{73}
$$

$$
\hat{\beta} = \left(\sum_{i=1}^{N}X_i'HX_i\right)^{-1}\sum_{i=1}^{N}X_i'H(y_i - y_{i\_}\hat{\rho}), \tag{74}
$$

$$
\hat{\rho} = \frac{b}{a}, \tag{75}
$$

Substituting the above into the log of (71) multiplied by -2 and adding the appropriate constants (number of parameters multiplied by the natural log of the sample size) yields the BIC Equations in (34)

and (35). A smaller BIC value indicates evidence in favour of the model. Let us now look at the case of Equation (25). When X_{i1} are the true regressors to generate Y_i, the BIC difference between M_0 and M_1 is

$$BIC_{|M_0} - BIC_{|M_1} = NT\left[\ln\frac{c_{|M_0}}{N} - \ln\left(\frac{c_{|M_1}}{N} - \frac{(b_{|M_1}/N)^2}{a_{|M_1}/N}\right)\right] + (k_0 - k_1 - 1)\ln(NT). \quad (76)$$

Asymptotically speaking, replacing $\frac{a}{N}$, $\frac{b}{N}$ and $\frac{c}{N}$ by $\underline{a}$, $\underline{b}$ and $\underline{c}$ defined in Equation (27) to (30) respectively should not affect the analysis. Define ω_{01} as

$$\begin{aligned}\omega_{01} =& NT\ln\frac{\underline{c}_{|M_0}}{\underline{c}_{|M_1} - \frac{\underline{b}^2_{|M_1}}{\underline{a}_{|M_1}}} + (k_0 - k_1 - 1)\ln(NT)\\ =& NT\ln\frac{(T-1)\sigma^2 + h_3(\beta|M_0) + \underline{a}_{|M_0}\underline{\rho}^2 + 2\underline{\rho}h_2(\beta,\underline{\rho}|M_0) - 2\underline{\rho}\sigma^2 h(\underline{\rho})}{(T-1)\sigma^2 + h_3(\beta|M_1) - \frac{[h_2(\beta,0|M_1) - \sigma^2 h(\underline{\rho})]^2}{\underline{a}_{|M_1}}}\\ &+ (k_0 - k_1 - 1)\ln(NT).\end{aligned} \quad (77)$$

If M_1 is the true model, we should have $\omega_{01} > 0$, $h_3(\beta|M_1) = h_2(\beta,0|M_1) = 0$ and inside the natural log of the first term, the numerator should be larger than the denominator. That is we should have Equation (37) stated in Proposition 8 or

$$h_3(\beta|M_0) + \underline{a}_{|M_0}\underline{\rho}^2 + 2\underline{\rho}h_2(\beta,\underline{\rho}|M_0) - 2\underline{\rho}\sigma^2 h(\underline{\rho}) + \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}_{|M_1}} > 0. \quad (78)$$

If X_{i0} is the same as X_{i1}, we will have $\underline{a}_{|M_0} = \underline{a}_{|M_1} = \underline{a}$, $k_1 = k_0$ and $h_2(\beta,\underline{\rho}|M_0) = h_3(\beta|M_0) = 0$. The left of Equation (78) will become $\underline{a}\left(\underline{\rho} - \frac{\sigma^2 h(\underline{\rho})}{\underline{a}}\right)^2$. If $\underline{\rho} - \frac{\sigma^2 h(\underline{\rho})}{\underline{a}} = 0$, *i.e.*, $\underline{\rho} + \gamma = \underset{N\to\infty}{plim}\hat{\rho}_{MLE} = 0$, we will have $BIC_{|M_0} - BIC_{|M_1} < 0$ asymptotically, which means we will prefer M_0 over M_1 even if $\underline{\rho} \neq 0$. In a situation like this, model selection is not consistent. The problem with BIC will also arise when M_0 is the true model with $\underline{\rho} = 0$. Now ω_{01} is

$$\omega_{01} = NT\ln\frac{(T-1)\sigma^2}{\underline{c}_{|M_1} - \frac{\underline{b}^2_{|M_1}}{\underline{a}_{|M_1}}} + (k_0 - k_1 - 1)\ln(NT), \quad (79)$$

To have $\omega_{01} < 0$, we should have Equation (36) in Proposition 8 or

$$\frac{\left[h_2(\beta,0|M_1) - \frac{(T-1)\sigma^2}{T}\right]^2}{\underline{a}_{|M_1}} - h_3(\beta|M_1) < 0. \quad (80)$$

If $h_2(\beta,0|M_1) = h_3(\beta|M_1) = 0$, we will have $\omega_{01} > 0$, which implies inconsistency in model selection. For the case of Equation (26), the corresponding ω_{21} is

$$\omega_{21} = NT\ln\frac{\underline{c}_{|M_2} - \frac{\underline{b}^2_{|M_2}}{\underline{a}_{|M_2}}}{\underline{c}_{|M_1} - \frac{\underline{b}^2_{|M_1}}{\underline{a}_{|M_1}}} + (k_2 - k_1)\ln(NT). \quad (81)$$

BIC will be consistent if we have Equation (38) stated in Proposition 8 or

$$\underline{a}_{|M_1}\underline{a}_{|M_2}h_3(\beta|M_2) + \underline{a}_{|M_2}\sigma^4 h^2(\underline{\rho}) - \underline{a}_{|M_1}\left[h_2(\beta,\underline{\rho}|M_2) - \sigma^2 h(\underline{\rho})\right]^2 > 0. \quad (82)$$

If X_{i2} nests X_{i1}, the left of Equation (82) can be simplified as $(\underline{a}_{|M_2} - \underline{a}_{|M_1})\sigma^4 h^2(\rho)$ with $\underline{a}_{|M_2} \leq \underline{a}_{|M_1}$. When $\underline{a}_{|M_2} = \underline{a}_{|M_1}$, ω_{21} will become $(k_2 - k_1)\ln(NT)$ with $k_2 > k_1$. BIC is therefore consistent. But if $\underline{a}_{|M_2} < \underline{a}_{|M_1}$, ω_{21} in Equation (81) will be dominated by the first term (negative) asymptotically and hence BIC is inconsistent.

References

1. Nerlove, M. Experimental Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross-Sections. *Econ. Stud. Quart.* **1968**, *18*, 42–74.
2. Nickell, S. Biases in Dynamic Models with Fixed Effects. *Econometrica* **1981**, *49*, 1417–1426.
3. Lancaster, T. The incidental parameter problem since 1948. *J. Econom.* **2000**, *95*, 391–413.
4. Hsiao, C. *Analysis of Panel Data*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2003.
5. Bester, C.; Hansen, C. A Penalty Function Approach to Bias Reduction in Nonlinear Panel Models with Fixed Effects. *J. Bus. Econ. Stat.* **2009**, *27*, 131–148.
6. Hahn, J.; Kuersteiner, G. Bias reduction for dynamic nonlinear panel models with fixed effects. *Econ. Theory* **2011**, *27*, 1152–1191.
7. Arellano, M.; Bonhomme, S. Robust Priors in Nonlinear Panel Data Models. *Econometrica* **2009**, *77*, 489–536.
8. Dhaene, G.; Jochmans, K. *Likelihood Inference in an Autoregression with Fixed Effects. Discussion Paper, Sciences Po*; Cambridge University Press: Cambridge, UK, 2013.
9. Andrews, D.; Lu, B. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Biometrika* **2001**, *101*, 123–164.
10. Lee, Y.; Phillips, P.C. Model Selection in the Presence of Incidental Parameters. *J. Econ.* **2014**, doi:10.1016/j.jeconom.2015.03.012.
11. Lancaster, T. Orthogonal Parameters and Panel Data. *Rev. Econ. Stud.* **2002**, *69*, 647–666.
12. Poirier, D.J. *Intermediate Statistics and Econometrics : A Comparative Approach*; MIT Press: Cambridge, MA, USA, 1995.
13. White, H. *Asymptotic Theory for Econometricians*; Prentice Hall: Upper Saddle River, NJ, USA, 2001.
14. Zellner, A. On Assessing Prior Distributions and Bayesian Regression Analysis with G-prior Distribution. In *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*; Goel, P.K., Zellner, A., Eds.; North-Holland: Amsterdam, The Netherlands, 1986; pp. 233–243.
15. Ley, E.; Steel, M.F. On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression. *J. Appl. Econom.* **2009**, *24*, 651–674.
16. Fernandez, C.; Ley, E.; Steel, M.F. Benchmark Priors for Bayesian Model Averaging. *J. Econom.* **2001**, *100*, 381–427.
17. Cox, D.R.; Reid, N. Parameter Orthogonality and Approximate Conditional Inference. *J. R. Stat. Soc. Ser. B* **1987**, *49*, 1–39.
18. Chikuse, Y. *Statistics on Special Manifolds*; Springer: Berlin, Germany; Heidelberg, Germany; 2003.

19. Koop, G.; Poirier, D.J.; Tobias, J.L. *Bayesian Econometric Methods*; Cambridge University Press: Cambridge , UK, 2007.

20. Tierney, L.; Kadane, J. Accurate Approximations for Posterior Moments and Marginal Densities. *J. Am. Stat. Assoc.* **1986**, *81*, 82–86.

21. Kass, R.E.; Tierney, L.; Kadane, J.B. The Validity of Posterior Expansions Based on Laplace's Method. In *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George a Barnard*; Geisser, S., Hodges, J.S., Press, S.J., Zellner, A., Eds.; North-Holland: Amsterdam, The Netherlands, 1990.