CARDIFF BUSINESS SCHOOL
WORKING PAPER SERIES

Cardiff Economics Working Papers

Guangjie Li

Consistent Estimation, Model Selection and Averaging of Dynamic Panel Data Models with Fixed Effect

E2009/5

Cardiff Business School
Cardiff University
Colum Drive
Cardiff CF10 3EU
United Kingdom
t: +44 (0)29 2087 4000
f: +44 (0)29 2087 4419
www.cardiff.ac.uk/carbs

ISSN 1749-6101
March 2009

Consistent Estimation, Model Selection and Averaging of Dynamic Panel Data Models with Fixed Effect*

Guangjie LI†
Cardiff Business School
Cardiff University

Abstract

In the context of an autoregressive panel data model with fixed effect, we examine the relationship between consistent parameter estimation and consistent model selection. Consistency in parameter estimation is achieved by using the tansformation of the fixed effect proposed by Lancaster (2002). We find that such transformation does not necessarily lead to consistent estimation of the autoregressive coefficient when the wrong set of exogenous regressors are included. To estimate our model consistently and to measure its goodness of fit, we argue for comparing different model specifications using the Bayes factor rather than the Bayesian information criterion based on the biased maximum likelihood estimates. When the model uncertainty is substantial, we recommend the use of Bayesian Model Averaging. Finally, we apply our method to study the relationship between financial development and economic growth. Our findings reveal that stock market development is positively related to economic growth, while the effect of bank development is not as significant as the classical literature suggests.

JEL Classification Code: C52, C11, C13, C15
Keywords: dynamic panel data model with fixed effect, incidental parameter problem, consistency in estimation, model selection, Bayesian Model Averaging, finance and growth

*The author wishes to thank Roberto Leon Gonzalez for his patient and insightful guidance and Gary Koop for his long-term encouragement and helpful advice. The author would also like to thank Ross Levine for the data set provided. An earlier draft of this paper was presented at the 62nd European Meeting of the Econometric Society and the 2nd Japanese-European Bayesian Econometrics and Statistics Meeting. Comments from the participants are gratefully acknowledged. The author is responsible for all the remaining errors in the paper.

†email address: ligj@cf.ac.uk

1 Introduction

For a panel data linear regression model with lags of the dependent variable as regressors and agent specific fixed effects, the maximum likelihood estimate (MLE) of the common parameter is inconsistent when the number of time periods is small and fixed regardless of the cross section sample size. Nerlove (1968) showed in Monte Carlo simulations that the MLE is severely downward biased. Nickell (1981) derived the analytical form of the bias for the first order autoregression (AR) model. This problem, known as the "incidental parameter problem", due to the fixed effect parameter (incidental parameter), whose dimension will increase with the cross section sample size has been reviewed by Lancaster (2000). The current econometric literature focuses mainly on deriving consistent estimator for the common parameter. See, for example, Arellano and Bond (1991), Blundell and Bond (1998), Gourieroux et al. (2006) and Hahn and Newey (2004). Little attention is given to model specification comparison in the presence of incidental parameter.

Cox and Reid (1987) found that when the nuisance parameter[1] is information orthogonal[2] to the common parameter, it is more preferable to construct a statistical test for the common parameter, especially for exponential family likelihood models, based on the conditional likelihood given the maximum likelihood estimaor for the nuisance parameter than on the profile likelihood. Following the line of information orthogonalization, Lancaster (2002) proposed a Bayesian procedure to obtain consistent inference on the common parameter. Compared to the classical methods, it is relatively straightforward to unify parameter estimation and model comparison under a Bayesian framework. In this paper, we argue that parameter estimation and model comparison should not be treated as two different issues, which is the predominant practice in the linear dynamic panel model literature. Our arguments are as follows. First, from an application point of view, researchers are often confronted with a large set of possible regressors in the panel model. In such situations, it is hard for indirect inference and moment methods to examine what model specification performs better than the others and whether some regressors can robustly explain the dependent variable. Second, as shown in this paper, likelihood based correction approach (including Bayesian) will not always lead to consistent estimation of the common parameter when the wrong set of exogenous regressors are included. We show that consistent estimation is the result of certain regularity conditions. Since model uncertainty can increase our estimation risk, we should consider comparing different model specifications. We find that consistency in estimation and consistency in model selection are interrelated. If we base our model selection decision on the Bayes factor, which is derived from Lancaster's reparameterization of the fixed effect, we tend to pick up the true model when the cross section sample size increases. However, the model selection performance of the Bayesian information criterion (BIC) based on the biased MLE

[1] Incidental parameter refers to the nuisance parameter which is of less interest to the researcher and whose dimension will increase with the sample size.

[2] See the appendix for the details.

is very poor both for small and big sample sizes. The BIC will asymptotically choose the wrong model for some situations[3]. Thirdly, for small sample size, when model uncertainty is substantial, we argue for the use of Bayesian model averaging (BMA) to reduce estimation risk[4]. Apart from the theoretical results, in the end of the paper we provide an example of finance and economic growth to show that our method is flexible enough to accommodate real world problems and handle issues like unbalanced panel.

The plan of the paper is as follows. Section 2 summarizes our model and the posterior results. Section 3 describes our motivation to compare different model specifications and shows when our posterior estimators will be consistent. Section 4 presents the conditions under which the Bayes factor and the BIC can lead to consistency in model selection followed by a short description of the BMA method. In section 6, we carry out simulation studies to check our Propositions. Section 7 then gives an example of application in finance and growth before Section 8 concludes.

2 The Model and the Posterior Results

Consider the model

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

Here we are investigating the case of first order autoregression linear panel, where ρ is a scalar and $x_{i,t}$ is a $k \times 1$ vector. Denote u_i as $[u_{i,1}, u_{i,2}, \dots, u_{i,T}]'$ and $X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,T})'$. We assume $u_i | f_i, X_i \sim N(0, \sigma^2 I_T)$ where I_T is an identity matrix with dimension T. Our assumption states that the error term is homoscedastic and our regressors, X_i, are strictly exogenous. It is well known in dynamic panel model literature, see Nickell (1981) and Lancaster (2000), that for a fixed T (the number of observations for each economic agent), the maximum likelihood estimators of ρ, β and σ^2 will not be consistent even if N (the number of economic agents) tends to infinity. This is due to the incidental parameter $f'_i s$, whose number will increase with the cross section sample size, N. Let us denote the common parameter $\theta = (\rho, \beta, \sigma^2)'$, whose dimension will not change with the sample size. To obtain consistent estimators for θ, Lancaster (2002) suggested an information orthogonal reparameterization of the fixed effect $f_i = f(\theta, g_i)$ such that the new fixed effect (g_i) is information orthogonal to the rest of the parameters (θ)[5]. However, this idea cannot lead to any valid reparameterization. By drawing analogy from two simpler cases, Lancaster instead found the following way to reparameterize the fixed effect:

$$f_i = g_i \exp\left[-b\left(\rho\right)\right] - \frac{1}{T}\iota' X_i \beta, \tag{2}$$

[3]For example, consider two models with the same exogenous regressors: one has the lag term of the dependent variable as a regressor and one does not. The BIC will asymptotically choose the model with the lag when the true model should be the one without the lag.

[4]Here it refers to the risk of using the estimates from a misspecified model.

[5]See the appendix for the details.

where $b(\rho)$ is defined as

$$b(\rho) = \frac{1}{T}\sum_{t=1}^{T-1}\frac{T-t}{t}\rho^t, \tag{3}$$

Let us transform our model accordingly as

$$y_i = g_i \exp\left[-b(\rho)\right]\iota + y_{i\_}\rho + HX_i\beta + u_i, \tag{4}$$

where $y_i = [y_{i,1}, y_{i,2}, \ldots, y_{i,T}]'$, $y_{i\_} = [y_{i,0}, y_{i,1}, \ldots, y_{i,T-1}]'$ and H is the demean matrix of dimension $T \times T$ equal to $I_T - \frac{\iota\iota'}{T}$ with ι as a vector of ones. Note that $y_{i,0}$ is viewed as known and our posterior results will be conditional on it.

The structure of the prior distribution for θ and $g = (g_1, g_2, ..., g_N)'$ is

$$\begin{aligned} p(g,\theta) &= p(g,\rho,\sigma^2,\beta) = p(g_1)...p(g_N)p(\rho)p(\sigma^2)p(\beta|\sigma^2) \\ &\propto \frac{1}{\sigma^2} I(-1<\rho<1)p(\beta|\sigma^2), \end{aligned} \tag{5}$$

which means we adopt independent improper priors for parameters other than β and ρ. The prior of ρ follows a uniform distribution between -1 and 1, which is the stationary region.

In regard to the conditional prior of β given σ^2, we want to have a proper distribution so that Bayes factors can lead to the selection of the true model as the cross section sample size increases. We can see this point more clearly later in Section 4. The prior we use takes the following g-prior form, proposed by Zellner (1986):

$$\beta|\sigma^2 \sim N\left(0, \sigma^2(\eta\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i)^{-1}\right), \tag{6}$$

where $\tilde{X}_i = HX_i$. The strength of the prior depends on the value of η. The smaller the value is, the less informative is our prior. We will give more details about the choice of η later. With the parameter priors given in (5) and (6), we can derive the posterior distributions of the parameters shown in Proposition 2.1.

Proposition 2.1. *The posterior distributions for the parameters in our model will take the following form:*

$$g_i|Y, y_{i,0}, \sigma^2, \rho \sim N\left(e^{b(\rho)}\frac{\iota'(y_i - y_{i\_}\rho)}{T}, \frac{\sigma^2}{T}\exp[2b(\rho)]\right), \tag{7}$$

$$\beta|\sigma^2,\rho,Y,Y_0 \sim N\left(\frac{1}{\eta+1}\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i, \sigma^2\left((\eta+1)\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\right), \tag{8}$$

$$\sigma^2|\rho,Y,Y_0 \sim IW(N(T-1), A), \tag{9}$$

$$\rho|Y, Y_0 \propto I(-1 < \rho < 1) \exp[Nb(\rho)] |A|^{-\frac{N(T-1)}{2}}, \tag{10}$$

where $\tilde{w}_i = H(y_i - y_{i\_}\rho)$, $A = \sum_{i=1}^{N} \tilde{w}_i{}'\tilde{w}_i - \frac{1}{\eta+1} \sum_{i=1}^{N} \tilde{w}_i'\tilde{X}_i \left(\sum_{i=1}^{N} \tilde{X}_i'\tilde{X}_i \right)^{-1} \sum_{i=1}^{N} \tilde{X}_i'\tilde{w}_i$, *and* Y_0 *is the collection of the initial observations from each economic agent while* Y *is the vector of observations excluding the initial observations.*

We can see that the posterior distributions of the parameters have a hierarchical structure. The conditional posterior distributions of all the parameters other than ρ are commonly known standard distributions, while at the bottom of the hierarchy the posterior distribution of ρ is not standard. To make draws of all the parameters from the posterior distributions, we first need to draw from this nonstandard posterior distribution of ρ. One way to do it is as follows. We first split the interval $(-1, 1)$ into small partitions $-1, \rho_1, \rho_2, \ldots, 1$ and then use some deterministic numerical method (such as Gaussian quadrature) to calculate the value of the cumulative distribution function at each partition point, i.e. $F(-1), F(\rho_1), F(\rho_2), \ldots, F(1)$. Next we draw a random variable u from uniform distribution $U[0, 1]$ and deliver $F^{-1}(u)$ as a draw of ρ from the nonstandard distribution. $F^{-1}(u)$ is obtained from piecewise cubic Hermite interpolation, see for example Süli and Mayers (2003).

3 Motivation to Compare Different Model Specifications

Lancaster (2002) showed that without model misspecification if we adopt the fixed effect reparameterization and the prior $p(g, \theta) \propto \frac{1}{\sigma^2}$, the mode of the marginal posterior for θ will be consistent. The difference adopted here is the g-prior we use for $p\left(\beta|\sigma^2\right)$ in (6). As long as we specify η as a function of the cross section sample size N such that $\lim_{N\to\infty} \eta(N) = 0$, our posterior results will be identical to Lancaster's for big cross section sample size. However, we cannot expect our model will always be correctly specified, i.e. the true regressors used to generate the data are always included in the regression. Here in proposition 3.1 we show the conditions under which we can obtain consistent posterior estimates for σ^2 and ρ even if we include the wrong set of exogenous regressors.

Proposition 3.1. *The posterior estimates from (9) to (10) are consistent if we have either*

$$\frac{-(T-1)h_2(\beta, \underline{\rho})}{h_3(\beta)} = h(\underline{\rho}) \tag{11}$$

or

$$h_2(\beta, \underline{\rho}) = h_3(\beta) = 0, \tag{12}$$

where

$$h(\rho) = \sum_{t=1}^{T-1} \frac{T-t}{T} \rho^{t-1} = \frac{d\, b(\rho)}{d\, \rho}. \tag{13}$$

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

$$h_3(\beta) = \underset{N\to\infty}{plim}\frac{1}{N}\left[\sum_{i=1}^{N} \beta' X_i' H X_i\beta - \frac{1}{\eta+1}\sum_{i=1}^{N} \beta' X_i' H \underline{X_i}\left(\sum_{i=1}^{N}\underline{X_i}' H \underline{X_i}\right)^{-1}\sum_{i=1}^{N}\underline{X_i}' H X_i\beta\right]. \tag{14}$$

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

Note that $0 < h(\rho) < \frac{T-1}{2}$ and it is monotonically increasing for $\rho \in (-1,1)$. For $h_2(\beta,\underline{\rho}) = h_3(\beta) = 0$ to be satisfied, it is enough that the true regressors X are a subset of $\underline{X}$. For $\frac{-(T-1)h_2(\beta,\underline{\rho})}{h_3(\beta)} = h(\underline{\rho})$ to hold, one example could be that no serial correlation and collinearity exist among the true regressors and the included regressors have zero correlation with the true regressors.[6] Proposition 3.1 tells us that if neither (11) nor (12) is satisfied, our posterior estimates of σ^2 and ρ will not be consistent even if we have a large cross section sample size when the number of observations for each economic agent is small in the panel. This is one of the major reasons why we need to compare different model specifications. Due to Bartlett's paradox[7], if we want to compare different models, we need to have a proper prior[8] for parameters not common to all the models. That is why we adopt the prior for β in (6).

In empirical applications, such as that of the growth theory, we will often have many possible regressors suggested by different theories to be included in the regression in (1). In a case like this, the number of potential exogenous regressors will be large. In addition to the concern over inconsistent estimation, we may want to know which combination of these regressors can best explain our data. The predominant GMM method in the literature to estimate the fixed effect model provides little information in this respect. Classical diagnostic tool such as R-square is not well defined. In a Bayesian framework such as ours, we can evaluate how good the model fits the data by looking at the posterior model probability. In our context, different models are defined by different combinations of the regressors and by whether or not we have a lag term of the dependent variable in the regression. So the total number of models is 2^{K+1}, where K stands for the number of all the potential exogenours regressors. The

[6] The proof is trivial and available upon request from the author, though the author admits that such case sounds impractical in reality.

[7] See for example Poirier (1995). To summarize it briefly, the problem here is that under an improper prior (the integral of which is not finite), the most restricted model will have the highest posterior model probability no matter whether it is true or not.

[8] Our prior is informative and proper in the sense that we have introduced the parameter η and $\eta \neq 0$.

posterior model probability of model i is calculated as

$$\begin{aligned} p\left(M_i|Y,Y_0\right) &= \frac{p\left(M_i\right)p\left(Y|Y_0,M_i\right)}{p\left(Y|Y_0\right)} \\ &= \frac{p\left(M_i\right)p\left(Y|Y_0,M_i\right)}{\sum_{j=1}^{2^{K+1}} p\left(M_j\right)p\left(Y|Y_0,M_j\right)}. \end{aligned} \tag{15}$$

where $p\left(M_i\right)$ is the prior model probability. Here we just assume all the models are equally possible a priori such that the posterior model probability only depends on the marginal likelihood, $p\left(Y|Y_0,M_i\right), j = 1,2,\ldots,2^{K+1}$. We can see in (15) that to evaluate the posterior probability of a single model we have to calculate the marginal likelihood of all the models. However, from the derivation of Proposition 2.1, we can only know the product of the marginal likelihood and the posterior of ρ:

$$\begin{aligned} p(\rho|Y,Y_0)p(Y|Y_0) =& \frac{1}{2}I(-1<\rho<1)\left(\frac{\eta}{\eta+1}\right)^{\frac{k}{2}}|A|^{-\frac{N(T-1)}{2}} \\ & \Gamma\left[\frac{N(T-1)}{2}\right]T^{-\frac{N}{2}}(\pi)^{-\frac{N(T-1)}{2}}\exp\left(Nb(\rho)\right) \end{aligned} \tag{16}$$

To calculate the marginal likelihood, we can use the same numerical techniques as we calculate the posterior cumulative distribution function of ρ. By integrating ρ out of the product, we can obtain $p(Y|Y_0,M_i)$. If the total number of models is not large, say less than 2^{20}, it is possible to use any mainstream PC of today to calculate the marginal likelihood of all the models and then use (15) to find the posterior model probability for each of them. For large set of models beyond the computation power of today, we can use the method of Markov Chain Monte Carlo Model Composition (MC^3) developed by Madigan and York (1995).

4 Consistency in Model Selection

In this section, we show that in our setting, how the posterior model probability can lead us to locate the true model when the cross section sample size tends to infinity and certain regularity conditions are met. That is, if Y is indeed generated by some combination of the potential regressors in the linear model, the posterior model probability of this combination, which is obtained by integrating out ρ in (16), will tend to 1 when N tends to infinity. In the end of this section, we will also analyze whether the Bayesian information criterion (BIC) based on the biased MLE can lead to consistency in model selection.

In the simpler case in which the true value of ρ is known to be zero (i.e. static panel data models), the consistency in model selection easily follows from the analysis by Fernandez et al. (2001). In our context, all we need to ensure consistency is to set η as a function of N such that $\lim_{N\to\infty}\eta(N) = 0$. One possible

choice could be $\eta = O(\frac{1}{N})$. As for the BIC, it is consistent in model selection for the static panel.

Let us now consider the case when our candidate model contains a lag term of the dependent variable. We can either compare it against a model without the lag term and with different regressors or a model with the lag term and with different regressors. The Bayes factor, which is defined as the ratio between the marginal likelihoods of the two models, looks like the following respectively.

$$\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)} = \left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}} \frac{\frac{1}{2}\int_{-1}^{1}\exp\left[Nb(\rho)\right]\left[\sum_{i=1}^{N}w_i'Hw_i - \frac{1}{\eta+1}\sum_{i=1}^{N}w_i'HX_{i1}\left(\sum_{i=1}^{N}X_{i1}'HX_{i1}\right)^{-1}\sum_{i=1}^{N}X_{i1}'Hw_i\right]^{-\frac{N(T-1)}{2}} d\rho}{\left[\sum_{i=1}^{N}y_i'Hy_i - \frac{1}{\eta+1}\sum_{i=1}^{N}y_i'HX_{i0}\left(\sum_{i=1}^{N}X_{i0}'HX_{i0}\right)^{-1}\sum_{i=1}^{N}X_{i0}'Hy_i\right]^{-\frac{N(T-1)}{2}}} \tag{17}$$

$$\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)} = \left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}} \frac{\int_{-1}^{1}\exp\left[Nb(\rho)\right]\left[\sum_{i=1}^{N}w_i'Hw_i - \frac{1}{\eta+1}\sum_{i=1}^{N}w_i'HX_{i1}\left(\sum_{i=1}^{N}X_{i1}'HX_{i1}\right)^{-1}\sum_{i=1}^{N}X_{i1}'Hw_i\right]^{-\frac{N(T-1)}{2}} d\rho}{\int_{-1}^{1}\exp\left[Nb(\rho)\right]\left[\sum_{i=1}^{N}w_i'Hw_i - \frac{1}{\eta+1}\sum_{i=1}^{N}w_i'HX_{i0}\left(\sum_{i=1}^{N}X_{i0}'HX_{i0}\right)^{-1}\sum_{i=1}^{N}X_{i0}'Hw_i\right]^{-\frac{N(T-1)}{2}} d\rho} \tag{18}$$

where $w_i = y_i - y_{i\_}\rho$, k_1 and k_0 are the dimensions of X_{i1} and X_{i0}, which denote the regressors included under M_1 and M_0 respectively. To simplify (17) and (18), we need to simplify the integrals that appear in the numerator and the denominator. Let us first define the following quantities:

$$\begin{aligned}
a =& \sum_{i=1}^{N}y_{i_}'Hy_{i_} - \frac{1}{\eta+1}\sum_{i=1}^{N}(y_{i_}'H\underline{X}_i)\left(\sum_{i=1}^{N}\underline{X}_i'H\underline{X}_i\right)^{-1}\sum_{i=1}^{N}\left(\underline{X}_i'Hy_{i_}\right),\\
b =& \sum_{i=1}^{N}y_{i_}'Hy_i - \frac{1}{\eta+1}\sum_{i=1}^{N}(y_{i_}'H\underline{X}_i)\left(\sum_{i=1}^{N}\underline{X}_i'H\underline{X}_i\right)^{-1}\sum_{i=1}^{N}\left(\underline{X}_i'Hy_i\right),\\
c =& \sum_{i=1}^{N}y_i'Hy_i - \frac{1}{\eta+1}\sum_{i=1}^{N}(y_i'H\underline{X}_i)\left(\sum_{i=1}^{N}\underline{X}_i'H\underline{X}_i\right)^{-1}\sum_{i=1}^{N}\left(\underline{X}_i'Hy_i\right).
\end{aligned} \tag{19}$$

Here we assume y_i and $\underline{X}_i$ have finite second moments so that the following probability limits exist.

$$
\begin{aligned}
&\underset{N\to\infty}{plim}\frac{1}{N}a = \underline{a}\\
&\underset{N\to\infty}{plim}\frac{1}{N}b = \underline{a}(\underline{\rho}+NB)\\
&\underset{N\to\infty}{plim}\frac{1}{N}c = \underline{\rho}^2\underline{a}+2\underline{a}\underline{\rho}NB+h_3(\beta)+(T-1)\sigma^2\\
NB &= \underset{N\to\infty}{plim}\frac{\left\{\begin{array}{c}\sum_{i=1}^{N}y_{i\_}'HX_i\beta-\frac{1}{\eta+1}\sum_{i=1}^{N}y_{i\_}'H\underline{X}_i\left(\sum_{i=1}^{N}\underline{X}_i'H\underline{X}_i\right)^{-1}\sum_{i=1}^{N}\underline{X}_i'HX_i\beta+\\ \sum_{i=1}^{N}y_{i\_}'Hu_i-\frac{1}{\eta+1}\sum_{i=1}^{N}y_{i\_}'H\underline{X}_i\left(\sum_{i=1}^{N}\underline{X}_i'H\underline{X}_i\right)^{-1}\sum_{i=1}^{N}\underline{X}_i'Hu_i\end{array}\right\}}{\sum_{i=1}^{N}y_{i\_}'Hy_{i\_}-\frac{1}{\eta+1}\sum_{i=1}^{N}(y_{i\_}'H\underline{X}_i)\left(\sum_{i=1}^{N}\underline{X}_i'H\underline{X}_i\right)^{-1}\sum_{i=1}^{N}(\underline{X}_i'Hy_{i\_})}\\
&= \frac{h_2(\beta,\underline{\rho})-\sigma^2h(\underline{\rho})}{\underline{a}}.
\end{aligned}
\tag{20}
$$

If the true model is either M_1 or M_0, we can show the conditions in Proposition 4.1 and 4.2 under which the Bayes factors in (17) and (18) can lead to the selection of the right model asymptotically.

Proposition 4.1. *When M_1 is the true model, i.e. $\underline{\rho}\neq 0$ and $X_{i1}'s$ are the true regressors to generate Y (which means X_{i0} is the same as $\underline{X}_i$ in (14)), as the cross section sample size increases, $\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}$ in (17) will tend to infinity if the following holds,*

$$
\begin{aligned}
&z(\underline{\rho}) = b(\underline{\rho})+\\
&\frac{T-1}{2}\ln\left[\frac{\underline{a}_{|M_0}\underline{\rho}^2-2\underline{\rho}\sigma^2h(\underline{\rho})+2\underline{\rho}h_{2|M_0}(\beta,\underline{\rho})+h_{3|M_0}(\beta)+(T-1)\sigma^2}{(T-1)\sigma^2}\right]>0.
\end{aligned}
\tag{21}
$$

When M_0 is the true model, i.e. $X_{i0}'s$ are the true regressors to generate Y and $\underline{\rho}=0$, as the cross section sample size increases, $\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}$ in (17) will tend to 0 if either of the following is satisfied:

1. *Under M_1, $\underset{N\to\infty}{plim}\, f(\rho)$ has a unique maximum ρ^* in (−1,1), where $f(\rho)$ is defined as*

$$f(\rho) = b(\rho)-\frac{T-1}{2}\ln\left(\rho^2-2\frac{b}{a}\rho+\frac{c}{a}\right) \tag{22}$$

and

$$b(\rho^*)+\frac{T-1}{2}\ln\frac{(T-1)\sigma^2}{d(\rho^*|M_1)}<0 \tag{23}$$

where

$$\begin{aligned} d(\rho|M_i) =& \underline{a}_{|M_i}\rho^2 - 2\underline{a}_{|M_i}(\underline{\rho} + NB_{|M_i})\rho + \\ & \underline{a}_{|M_i}\underline{\rho}^2 + 2\underline{a}_{M_i}\underline{\rho}NB_{|M_i} + (T-1)\sigma^2 + h_{3|M_i}(\beta). \end{aligned} \tag{24}$$

2. *Though M_1 is misspecified, it can still lead to the consistent estimation of ρ, i.e. either (11) or (12) holds.*

Proposition 4.2. *When M_1 is the true model, as the cross section sample size increases, $\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}$ in (18) will tend to infinity if any of the following holds:*

1. *Under M_0, $\underset{N\to\infty}{plim}\, f(\rho)$ has a unique maximum ρ^* in $(-1,1)$ and*

$$b(\underline{\rho}) - b(\rho^*) + \frac{T-1}{2}\ln\frac{d(\rho^*|M_0)}{(T-1)\sigma^2} > 0 \tag{25}$$

2. *Either (11) or (12) holds.*

In addition to the Bayes factor calculated based on our parameterization of the fixed effect, we may be interested in knowing whether or not the Bayesian information criterion based on the biased MLE will lead to consistency in model selection. The results are shown in Proposition 4.3.

Proposition 4.3. *For the comparison of the two models in (17), when M_1 is the true model, BIC is consistent in model selection if the following condition is met,*

$$h_{3|M_0}(\beta) + \underline{a}_{|M_0}\underline{\rho}^2 + 2\underline{\rho}h_{2|M_0}(\beta,\underline{\rho}) - 2\underline{\rho}\sigma^2 h(\underline{\rho}) + \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}_{|M_1}} > 0 \tag{26}$$

However, when $\underline{\rho} + NB_{|M_1} = 0$ and $X_{i1} = X_{i0}$, BIC is inconsistent. When M_0 is the true model, BIC is consistent if the following is satisfied

$$\frac{\left[h_{2|M_1}(\beta,0) - \sigma^2\frac{T-1}{T}\right]^2}{\underline{a}_{|M_1}} - h_{3|M_1}(\beta) < 0 \tag{27}$$

However, if we have $h_{3|M_1}(\beta) = 0$[9], BIC is inconsistent.

For the comparison of the two models in (18), when M_1 is the true model, BIC is consistent in model selection if the following holds

$$\underline{a}_{|M_1}\underline{a}_{|M_0}h_{3|M_0}(\beta) + \underline{a}_{|M_0}\sigma^4 h^2(\underline{\rho}) - \underline{a}_{|M_1}\left[h_{2|M_0}(\beta,\underline{\rho}) - \sigma^2 h(\underline{\rho})\right]^2 > 0 \tag{28}$$

Moreover, if X_{i0} nests the true set of regressors, i.e. $h_{2|M_0}(\beta,\underline{\rho}) = h_{3|M_0}(\beta) = 0$ and $\underline{a}_{|M_1} = \underline{a}_{|M_0}$, BIC will be consistent.

[9]For example, X_{i1} nests the true set of regressors or $\beta = 0$.

5 Motivations of Bayesian Model Averaging

Our method allows us to compare the goodness of fit of different model specifications. However, as Raftery and Zheng (2003) and Yuan and Yang (2005) point out, if there is substantial model uncertainty, model averaging is more preferable than model selection. In regard to our empirical application of finance and growth, the data set we have is relatively small (such as the one in Section 7, with cross section sample size equal to 40), which implies model uncertainty for estimation. When we want to study the relationship between economic growth and other variables from the panel data, it should be more appropriate to consider different model specifications than just drawing our conclusions based on a single model so that we can reduce the estimation risk in the presence of substantial model uncertainty. This point will be made more clear in the subsequent sections. At the moment, we will just briefly talk about the Bayesian model averaging (BMA) approach.

From different model specifications, we can have different estimates of θ.[10] Essentially, BMA consists in mixing the posterior distributions of θ from all different models according to their posterior model probabilities in (15). Inference about θ is drawn from its posterior distribution unconditional on the model space, which takes the following form.

$$p(\theta|Y, Y_0) = \sum_{i}^{2^{K+1}} p(\theta|Y, Y_0, M_i)p(M_i|Y, Y_0) \tag{29}$$

We then can use the posterior mean as the BMA point estimate for θ. To measure the importance of certain element in θ (say, θ_j), we can use the posterior inclusion probability defined as the following,

$$\sum_{i}^{2^{K+1}} I(\theta_j \in M_i)p(M_i|Y, Y_0). \tag{30}$$

We can see that it is a sum of the posterior model probabilities of the models which leave θ_j unrestricted.

6 Simulation Studies

In this section we will show the evidence for model selection consistency of our method based on simulated data sets. Here we try to make our simulation close to our application of the finance and growth example in the next section. We set $t = 4$ (the number of observations for each economic agent) and the number of possible regressors to 8. We draw independently the fixed effect f from U[-1,1]. For each iteration in the simulation, we do the following:

[10] Different models are defined by resticting different elements of θ, such as ρ or β to 0.

1. We first generate the potential regressors ($X_i's$) from the uniform distribution $U[-4,4]$. We then make these regressors correlated with each other and we also introduce serial correlation in our regressors.

2. We draw the model by selecting each regressor with the probability of 50%, (i.e. all possible models have the same probability of being selected). The element(s) of β are drawn from $U[-2,2]$. If our model includes the lag term of the dependent variable, we set $\underline{\rho} = 0.9$.[11]

3. We calculate the posterior model probabilities of all the models and compare the one with the highest model probability to the true model.

In Proposition 3.1 we show that we cannot have a consistent estimate of ρ when neither (11) nor (12) holds. We want to check whether we can still select the right model asymptotically using Lancaster's transformation of the fixed effect. That is why in step 1 we want to add collinearity and serial correlation to our regressors. To achieve this, we first make each two neighboring period observations correlated with each other as follows,

$$x_{t,s} = s_{t-1}x_{t-1,s} + \bar{s}_t x_{t,ns}, \tag{31}$$

where $x_{t,ns}$ has no serial correlation and is generated from the i.i.d. uniform distribution U[-4,4]. We set $s_{t-1} = \frac{s'_{t-1}}{\sqrt{s'^2_{t-1}+s'^2_t}}$ and $\bar{s}_t = \frac{s'_t}{\sqrt{s'^2_{t-1}+s'^2_t}}$. For s'_{t-1} and s'_t, we generate them from $i.i.d.U[-2.5, 2.5]$. In doing so, the correlation matrix for the serially correlated $[x_{1,s}, x_{2,s}, \ldots, x_{T,s}]'$ is

$$S = \begin{pmatrix} 1 & s_1 & \cdots & \prod_{i=1}^{T-1} s_i \\ s_1 & 1 & \cdots & \prod_{i=2}^{T-1} s_i \\ s_2 s_1 & s_2 & \cdots & \prod_{i=3}^{T-1} s_i \\ \cdots & \cdots & \cdots & \cdots \\ \prod_{i=1}^{T-1} s_i & \prod_{i=2}^{T-1} s_i & \cdots & 1 \end{pmatrix} \tag{32}$$

We can see that $\{x_t\}$ generated in such a way is not covariance stationary. Moreover, for small T[12], the distribution of $x's$ will change with t. However, if T is sufficiently large[13], the final few points of $x's$ at the end of the series will approximately follow, due to the central limit theorem, a normal distribution with the same mean (0) and the same variance (around 5.3) as the uniform distribution. We just use the final 4 observations from the series for our study.

[11] We have also set the lag coefficient to other value, such as 0.5. The results, which are available from the author upon request, do not change much.

[12] Here T denotes the sample size of the generated series.

[13] We choose T to be 100 for the results to be presented later. We have also used small value of T to generate the data, all the results are similar and neither (21) nor (25) is violated. These results are available upon request from the author.

Next we introduce correlation among the regressors by using a linear combination of those we just made serially correlated.

$$X_{j,c} = \sum_{i=1}^{K} q_{j,i} X_{i,nc} \qquad j = 1, 2, \ldots, K \tag{33}$$

where $X_{i,nc}$ denotes the regressor without collinearity and we set $q_{j,i} = \frac{q'_{j,i}}{\sqrt{\sum_{i=1}^{K} q'^2_{j,i}}}$ and $q'_{j,i} \sim i.i.d.U[-2.5, 2.5]$. Note that the L^2-norm of $[q_{j,1}, q_{j,2}, \ldots, q_{j,K}]'$ is equal to 1 so that we can preserve the same variance as that from the uniform distribution we use to generate x at the very beginning. Note that the correlation coefficient of any two elements of X_i is the same across different individuals and can be calculated as

$$corr(X_{t,k}, X_{t',k'}) = S(t, t') \sum_{i=1}^{K} q_{k,i} q_{k',i} \qquad t = 1, 2, \ldots, T \quad k = 1, 2, \ldots, K. \tag{34}$$

where $S(t, t')$ denote the (t, t') element in S and K is the potential number of regressors. Through such data generating mechanism we can explicitly calculate the values of $h_2(\beta, \underline{\rho})$ and $h_3(\beta)$, $\underline{a}$ and NB in (14) and (20) respectively. Hence we can check whether condition (21) and (25) are violated or not when there is an error in our model selection based on posterior model probability.

We run the experiment for 200 times. At first we set $\eta = \frac{1}{N}$ and $\sigma^2 = 1$. The results are presented in Table 1. The ER (error rate) column tells us how often the model with the highest posterior model probability ends up being different from the true model. When the cross section sample size is 40 (the same as our application later), the Bayes factor criterion fails to pick up the true model by 86 out of 200 simulations. However, we can see that the error rate tends to decrease with cross section sample size, which is a sign of model selection consistency. One thing to note is that we generate β from $U[-2, 2]$. When the values of some elements in β are very close to zero, it is virtually equivalent to the case when the true model does not include the corresponding regressors. In Table 1, the column "nest" denotes how often the top model is nested inside the true model (including the case when the top model is the true model). We can find that this number generally rises with cross section sample size. The column "nouni" checks among the errors from the Bayes factor criterion how many of them is related to the fact that either there is no solution or there are more than one solutions in (−1,1) for the equation $\underset{N\to\infty}{plim}\, f'(\rho) = 0$, where $f(\rho)$ is defined in (22). We show in the proof of Proposition 4.2 that when $\underset{N\to\infty}{plim}\, f(\rho)$ does not have a global maximum in the stationary region, we cannot use Laplace method to approximate the integral(s) in the Bayes factor. Hence the condition in (21), (23) and (25) do not hold. Under our simulating data generating mechanism, such situations do not exist. The columns of "no(21)", "no(23)" and "no(25)"

denote the error rates with the violation of (21), (23) and (25) respectively. We can see that the numbers of the columns are all zeros, which means all our errors are fixable with a large cross section sample size. The columns of "topprob" and "top10prob" are the average of the posterior model probabilities of the top model and the sum of the top ten models in the simulation. If these two numbers are far below 1, it is a sign of model uncertainty. As the cross section sample size increases, model uncertainty diminishes. If we raise the variance of the disturbance, model uncertainty will increase. The results are shown in Table 2 where we set the variance of the disturbance to 4. Comparing Table 2 to Table 1, we can see that the error rate is higher and the rest of the three columns are generally smaller for a particular cross section sample size. As for the model selection performance of BIC based on the biased maximum likelihood estimates, we list the results in Table 3 and Table 4. We can see that the BIC performance is much worse than our Bayes factor method. The error rates stay above 50% for different cross section sample sizes. Even for $N = 1000$, there is not much improvement. In addition to the error rate, the top model is not very often nested inside the true model as compared with the Bayes factor method. Again, it does not improve much with the cross section sample size. Moreover, the column headed with "no(27)" shows how many errors violate condition (27). Such errors are not fixable even if we have infinite cross section sample size according to Proposition 4.3. Note that around 50% of the true models do not have the lag term of the dependent variable under our simulation set-up. Also note that under our data generating scheme, we can be almost sure that $\underline{\rho} + NB = 0$ will not occur. Hence condition (26) will almost surely not be violated. It could be true to say that the error rate for the BIC would approach 50% in the limit since it is always possible for condition (27) to be violated while condition (26) and (28) hold. When the true model does not have a lag term of the dependent variable as the regressor, it is always possible to find a candidate model with both the lag term and exactly the same set of exogenous regressors as the true model such that condition (27) will be violated. When we compare them, we will choose the candidate model over the true model as the cross section sample size increases. Also we could expect that the percentage under $no(27)$ in Table 3 and 4 should rise with cross section sample size.

Table 1: Simulation results when $\sigma^2 = 1$

N	**ER**	**nest**	**topprob**	**top10prob**	**no(21)**	**no(23)**	**no(25)**	**nouni**
40	0.40	0.83	0.38	0.85	0.00	0.00	0.00	0.00
100	0.29	0.86	0.56	0.94	0.00	0.00	0.00	0.00
200	0.31	0.88	0.62	0.96	0.00	0.00	0.00	0.00
500	0.14	0.94	0.74	0.99	0.00	0.00	0.00	0.00
1000	0.10	0.97	0.81	0.99	0.00	0.00	0.00	0.00

Judging from the previous simulation results, we can find that if we simply select the model with the highest model probability to provide estimates of our interest, chances are high that the model selected is not the true model. Note that the top model probability for $N = 40$ is about 32% while it is about 78%

Table 2: Simulation results when $\sigma^2 = 4$

N	**ER**	**nest**	**topprob**	**top10prob**	**no(21)**	**no(23)**	**no(25)**	**nouni**
40	0.61	0.77	0.32	0.80	0.00	0.00	0.00	0.00
100	0.43	0.86	0.50	0.93	0.00	0.00	0.00	0.00
200	0.36	0.88	0.58	0.95	0.00	0.00	0.00	0.00
500	0.28	0.92	0.69	0.98	0.00	0.00	0.00	0.00
1000	0.16	0.96	0.78	0.99	0.00	0.00	0.00	0.00

Table 3: Simulation results for BIC when $\sigma^2 = 1$

N	**error rate**	**nest**	**no(26)**	**no(27)**	**no(28)**
40	0.78	0.34	0.00	0.46	0.00
100	0.69	0.42	0.00	0.60	0.00
200	0.69	0.41	0.00	0.51	0.00
500	0.54	0.51	0.00	0.65	0.00
1000	0.58	0.47	0.00	0.61	0.00

Table 4: Simulation results for BIC when $\sigma^2 = 4$

N	**error rate**	**nest**	**no(26)**	**no(27)**	**no(28)**
40	0.88	0.34	0.00	0.28	0.00
100	0.77	0.41	0.00	0.36	0.00
200	0.74	0.38	0.00	0.34	0.00
500	0.65	0.43	0.00	0.40	0.00
1000	0.55	0.425	0.00	0.65	0.00

for $N = 1000$ when we set $\sigma^2 = 4$. To account for such model uncertainty, we recommend averaging the estimates from every model. We argue that BMA can reduce our estimation risk when there is substantial model uncertainty. To illustrate this, next we carry out another simulation, in which we set the β's to fixed values along with ρ (we set it to 0.9 as in our previous simulation). Then we use the posterior means to estimate these values. Table 5 shows the root mean squared errors (RMSE) from different point estimators based on 200 iterations with the cross section sample size (N) as 40. The true values of ρ and β's are shown under the column "TRUE", where the first number is the value of ρ. The column "TOP" shows the RMSE resulting from the posterior mean estimator of the top model, which has the highest posterior model probability, while the column "BMA" uses the posterior mean in (29). To evaluate the significance of a regressor coefficient, we calculate the sum of the posterior model probabilities of all the models which include the corresponding regressor. If the inclusion probability for a regressor is too low, we may be better off by viewing the coefficient for this regressor as zero. In Table 5, we try to give some ideas on how to interpret such inclusion probabilities which we will use in our application later. The RMSE in the columns headed with a percentage number are derived based on certain inclusion probability criterion. For each simulated data set, if the inclusion probability for a regressor is lower than the percentage number on top of the column, we will simply use zero as its point estimate. In the last row of Table 5, we sum up the RMSE in each column. We can see that the overall performances of BMA and various inclusion probability criteria are all better than that of the top model criterion in terms of smaller total RMSE. Such performance seems to fare best when we set our inclusion probability to 50%. We can also see that higher inclusion probability criterion tends to give us smaller RMSE when the true value of the parameter is exactly zero and higher RMSE when the true value is not zero while the BMA seems to give us a safer option for almost all the parameter estimates.

Table 5: The RMSE of the point estimates when N=40 and $\sigma^2 = 1$

TRUE	TOP	BMA	10%	20%	30%	40%	50%	60%
0.9	0.034	0.034	0.034	0.034	0.034	0.034	0.034	0.034
0.1	0.112	0.090	0.090	0.092	0.096	0.102	0.105	0.108
0.3	0.139	0.132	0.132	0.134	0.137	0.142	0.148	0.164
0	0.065	0.054	0.054	0.053	0.050	0.042	0.038	0.031
0	0.069	0.057	0.057	0.054	0.050	0.044	0.039	0.037
1	0.127	0.133	0.133	0.133	0.133	0.133	0.133	0.145
0	0.054	0.068	0.068	0.067	0.065	0.036	0.032	0.029
0	0.076	0.075	0.075	0.074	0.047	0.044	0.030	0.026
2	0.134	0.122	0.122	0.122	0.122	0.122	0.122	0.122
SUM	0.810	0.765	0.765	0.765	0.734	0.700	0.683	0.697

To add more insight into how to use inclusion probability to determine the significance of a regressor coefficient, we presents the error rates of including the

wrong regressor due to different inclusion probability criteria[14] in Table 6. We can see that the overall error rates based on the 10% criterion is the highest. All the errors are from those parameters whose values are actually zeros. Again, the 50% criterion shows reasonably good performance, although the 60% criterion is slightly better. One thing to note is that the top model criterion has smaller overall error rate than nearly all inclusion probability criteria. Hence it seems to be a useful tool in terms of making the decision on whether to include a particular regressor or not.

Table 6: The error rates of whether to include a regressor when N=40 and $\sigma^2 = 1$

TRUE	TOP	10%	20%	30%	40%	50%	60%
0.9	0	0	0	0	0	0	0
0.1	0.835	0	0.36	0.62	0.78	0.86	0.915
0.3	0.22	0	0.03	0.075	0.14	0.215	0.295
0	0.05	1	0.425	0.195	0.09	0.065	0.02
0	0.095	1	0.475	0.245	0.155	0.095	0.035
1	0	0	0	0	0	0	0
0	0.065	1	0.495	0.205	0.09	0.06	0.015
0	0.045	1	0.51	0.195	0.085	0.035	0.015
2	0	0	0	0	0	0	0
SUM	1.31	4	2.295	1.535	1.34	1.33	1.295

Next we increase model uncertainty by increasing the variance of the disturbance to 4. Point estimate performances based on different criteria are shown in Table 7. When the model uncertainty is larger, the advantage of the averaging estimators becomes more obvious. Though their performances are quite alike, the 50% inclusion probability criterion still gives us the smallest overall RMSE. The error rates of whether to include a regressor are presented in Table 8. Now none of the inclusion probability criteria can have smaller overall error rates than that of the top model criterion. It seems that the criteria using inclusion probability above 40% (along with the top model criterion) have problems with the parameter whose value is 0.1 (close to zero).

When we set the variance of the disturbance to 1 and the cross section sample size to 1000 such that the model uncertainty is not substantial, the performance of different criteria will be similar, which is shown in Table 9 and Table 10. However, the averaging estimators still fare slightly better in point estimates and the top model criterion is reasonably good in deciding whether or not to include a variable.

In terms of the point estimation, BMA seems to be more preferable than simply using the estimates from the top model since it takes account of model uncertainty explicitly. Moreover, in Bayesian Econometrics we have many sen-

[14] If a regressor has no less than the given inclusion probability, we include it, which may not be one of the true regressors. The top model criterion means we only include the regressors in the top model.

Table 7: The RMSE when N=40 and $\sigma^2 = 4$

TRUE	TOP	BMA	10%	20%	30%	40%	50%	60%
0.9	0.066	0.065	0.065	0.065	0.065	0.065	0.065	0.065
0.1	0.156	0.115	0.115	0.115	0.116	0.116	0.113	0.105
0.3	0.299	0.211	0.211	0.215	0.225	0.237	0.249	0.257
0	0.147	0.098	0.098	0.097	0.092	0.085	0.079	0.064
0	0.184	0.123	0.123	0.123	0.119	0.115	0.109	0.106
1	0.312	0.240	0.240	0.241	0.241	0.251	0.263	0.295
0	0.193	0.118	0.118	0.117	0.114	0.110	0.101	0.097
0	0.236	0.146	0.146	0.145	0.142	0.137	0.129	0.124
2	0.273	0.222	0.222	0.222	0.222	0.222	0.222	0.246
SUM	1.866	1.338	1.338	1.339	1.336	1.339	1.331	1.359

Table 8: The error rates of whether to include a regressor when N=40 and $\sigma^2 = 4$

TRUE	TOP	10%	20%	30%	40%	50%	60%
0.9	0	0	0	0	0	0	0
0.1	0.915	0	0.445	0.76	0.855	0.95	0.975
0.3	0.5	0	0.155	0.31	0.455	0.55	0.615
0	0.045	1	0.465	0.16	0.085	0.03	0.015
0	0.06	1	0.455	0.185	0.095	0.05	0.035
1	0.035	0	0	0.005	0.025	0.035	0.04
0	0.05	1	0.52	0.185	0.095	0.045	0.025
0	0.045	1	0.42	0.175	0.075	0.04	0.03
2	0	0	0	0	0	0	0
SUM	1.65	4	2.46	1.78	1.685	1.7	1.735

Table 9: The RMSE of the point estimates when N=1000 and $\sigma^2 = 1$

TRUE	TOP	BMA	10%	20%	30%	40%	50%	60%
0.9	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007
0.1	0.036	0.031	0.032	0.032	0.033	0.035	0.035	0.038
0.3	0.024	0.023	0.023	0.023	0.023	0.023	0.023	0.023
0	0.018	0.015	0.014	0.014	0.013	0.013	0.012	0.012
0	0.011	0.008	0.008	0.008	0.007	0.007	0.007	0.006
1	0.029	0.026	0.026	0.026	0.026	0.026	0.026	0.026
0	0.014	0.012	0.012	0.011	0.011	0.011	0.010	0.010
0	0.015	0.010	0.010	0.010	0.009	0.009	0.008	0.006
2	0.027	0.026	0.026	0.026	0.026	0.026	0.026	0.026
SUM	0.181	0.158	0.158	0.156	0.155	0.156	0.154	0.153

Table 10: The error rates of whether to include a regressor when N=1000 and $\sigma^2 = 1$

TRUE	TOP	10%	20%	30%	40%	50%	60%
0.9	0	0	0	0	0	0	0
0.1	0.075	0.025	0.05	0.055	0.06	0.075	0.095
0.3	0	0	0	0	0	0	0
0	0.015	0.17	0.08	0.045	0.015	0	0
0	0.02	0.125	0.055	0.025	0.02	0.01	0
1	0	0	0	0	0	0	0
0	0.005	0.14	0.055	0.03	0.01	0	0
0	0	0.145	0.065	0.035	0.02	0	0
2	0	0	0	0	0	0	0
SUM	0.115	0.605	0.305	0.19	0.125	0.085	0.095

sible tools to help us understand our data. As will be shown in the application later, our inference is based on the posterior distribution of the parameter unconditional on the model space, which gives us information on what we are more sure of and of what we are less sure.

In our previous simulation studies, we adopt the g-prior and set its coefficient $\eta = \frac{1}{N}$, which should lead to consistency in model selection. Our previous simulation results seem to have confirmed this. In addition to setting $\eta = O(\frac{1}{N})$, Fernandez et al. (2001) also suggest setting $\eta = \frac{1}{K^2}$ for linear model of non-panel data, where K is the number of potential regressors. Their recommendation is

$$\eta = \begin{cases} \frac{1}{N} & \text{if } N > K^2 \\ \frac{1}{K^2} & \text{if } N \leq K^2 \end{cases}$$

In our context, K is the number of potential regressors plus 1 (the lag term). We can see that Fernandez et al. (2001) basically recommend a more non-informative prior. They argue that although the second prior is inconsistent[15], it may perform better than the first one for small samples. In contrast to Table 1 and Table 2, Table 11 and Table 12 present the results under the second prior. It suggests that when $N = 40$ (the cross section sample size in our application), the second prior seems to do much better for smaller disturbance variance in terms of whether the true model is nested inside the top model and it also has higher posterior top model probability. However, it fares more or less the same as the first prior for bigger disturbance variance. As the sample size increases, the improvement of the second prior does not seem to be as big as that under

[15]The inconsistency in model selection under the second prior here means the posterior model probability of the true model will not tend to 1 with increasing sample size. However, when the true model does not have a lag term as regressor, the Bayes factor under the second prior will still favour the true model, i.e. the true model still has higher model probability than the other models. For more details, see Fernandez et al. (2001). When the model has a lag term, as long as relevant conditions in Proposition 4.1 and 4.2 hold, consistency in model selection will follow. That is why we can still see improved performance under the second prior over larger sample size.

the first prior. For large sample size (such as $N = 1000$), the first prior is more preferable than the second.

Table 11: Simulation results when the variance of the disturbance is 1 and under the prior $\eta = \frac{1}{K^2}$

N	ER	nest	topprob	top10prob	no(21)	no(23)	no(25)	nouni
40	0.43	0.9	0.45	0.90	0	0	0	0
100	0.3	0.88	0.46	0.91	0	0	0	0
200	0.27	0.9	0.48	0.92	0	0	0	0
500	0.15	0.92	0.53	0.94	0	0	0	0
1000	0.17	0.9	0.49	0.92	0	0	0	0

Table 12: Simulation results when the variance of the disturbance is 4 and under the prior $\eta = \frac{1}{K^2}$

N	ER	nest	topprob	top10prob	no(21)	no(23)	no(25)	nouni
40	0.59	0.78	0.40	0.85	0	0	0	0
100	0.48	0.8	0.42	0.88	0	0	0	0
200	0.34	0.88	0.44	0.89	0	0	0	0
500	0.39	0.77	0.47	0.92	0	0	0	0
1000	0.2	0.89	0.47	0.91	0	0	0	0

Results on the point estimation performance under the second prior are presented in Table 13, Table 14 and Table 15. In comparison with Table 5 to Table 9, the performance under the second prior does not seem to differ much, though when the sample size is small and model uncertainty is large, for the column of the top model criterion, the second prior seems to do better than the first prior. But all the averaging estimators still tend to dominate the top model criterion for small samples. For large sample size, such dominance of averaging estimators seems to diminish and their performances are quite close to that from the top model. In Table 15, under the second prior, the BMA estimates even have higher RMSE than the top model criterion. Again, the first prior is more preferable than the second for large samples in terms of smaller RMSE from the averaging estimators.

7 An Application Example of Financial Development and Economic Growth

The model in our application is slightly different from (1) and it takes the following form.

$$\begin{gathered} y_{i,t} - y_{i,t-1} = f_i + y_{i,t-1}\rho + x'_{i,t}\beta + u_{it}, \\ i = 1 \dots N, \quad t = 1 \dots T. \end{gathered} \tag{35}$$

Table 13: The RMSE of the point estimates when N=40 and $\sigma^2 = 1$ under the prior $\eta = \frac{1}{K^2}$

TRUE	TOP	BMA	10%	20%	30%	40%	50%	60%
0.9	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033
0.1	0.118	0.092	0.092	0.094	0.098	0.101	0.102	0.103
0.3	0.170	0.143	0.143	0.144	0.149	0.155	0.167	0.173
0	0.097	0.061	0.061	0.060	0.058	0.051	0.048	0.038
0	0.097	0.060	0.060	0.059	0.050	0.047	0.031	0.019
1	0.116	0.113	0.113	0.113	0.113	0.113	0.113	0.113
0	0.058	0.068	0.068	0.068	0.061	0.038	0.031	0.020
0	0.058	0.051	0.051	0.048	0.043	0.041	0.038	0.038
2	0.144	0.132	0.132	0.132	0.132	0.132	0.132	0.132
SUM	0.890	0.754	0.754	0.752	0.739	0.711	0.695	0.671

Table 14: The RMSE of the point estimates when N=40 and $\sigma^2 = 4$ under the prior $\eta = \frac{1}{K^2}$

TRUE	TOP	BMA	10%	20%	30%	40%	50%	60%
0.9	0.074	0.075	0.075	0.075	0.075	0.075	0.075	0.075
0.1	0.146	0.118	0.118	0.119	0.120	0.115	0.107	0.104
0.3	0.280	0.228	0.228	0.233	0.242	0.251	0.261	0.272
0	0.102	0.109	0.109	0.108	0.101	0.095	0.078	0.073
0	0.238	0.118	0.118	0.117	0.112	0.108	0.090	0.066
1	0.316	0.258	0.258	0.263	0.267	0.274	0.301	0.313
0	0.233	0.148	0.148	0.146	0.143	0.140	0.124	0.120
0	0.117	0.111	0.111	0.109	0.104	0.075	0.069	0.064
2	0.271	0.241	0.241	0.241	0.241	0.241	0.241	0.271
SUM	1.776	1.407	1.407	1.411	1.405	1.374	1.347	1.358

Table 15: The RMSE of the point estimates when N=1000 and $\sigma^2 = 1$ under the prior $\eta = \frac{1}{K^2}$

TRUE	TOP	BMA	10%	20%	30%	40%	50%	60%
0.9	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009
0.1	0.029	0.028	0.028	0.028	0.028	0.029	0.031	0.034
0.3	0.026	0.025	0.025	0.025	0.025	0.025	0.025	0.025
0	0.012	0.013	0.013	0.012	0.012	0.011	0.010	0.010
0	0.008	0.020	0.020	0.019	0.009	0.007	0.005	0.005
1	0.026	0.028	0.028	0.028	0.028	0.028	0.028	0.028
0	0.011	0.018	0.018	0.018	0.013	0.007	0.006	0.004
0	0.022	0.023	0.023	0.022	0.018	0.017	0.016	0.016
2	0.036	0.039	0.039	0.039	0.039	0.039	0.039	0.039
SUM	0.178	0.201	0.201	0.200	0.181	0.172	0.169	0.171

Here $y_{i,t}$ is the log of GDP per capita, f_i is the country-specific fixed effect and $x_{i,t}$ is a vector of the explanatory variables as before. So on the left hand side of the equation is the economic growth per capita, which we are using the lag of the logged GDP per capita along with other variables to explain on the right hand side of the equation. The framework we developed in the previous sections is still applicable here given necessary adjustments. It can be shown that the Jacobian from Y conditional on Y_0 to $Y - Y_-$ is one, where Y_- is the collection of all the lag terms of the dependent variables for different individuals. To apply our method from the previous sections to the real data, we need to make the following small modifications.

The data we use are taken from Beck and Levine (2004) and are available from Levine's website. There are altogether 40 (N) countries and the panel covers the period from 1976 to 1998. Eight potential explanatory variables ($x_{i,t}$) have been proposed in the literature. Details of the variables can be found in Table 16. Here we just follow the practice of Beck and Levine (2004) on how the variables enter equation (35). Our focus is on the variables measuring the development of stock market and banking sector. We also include three dummy variables for each period as our potential explanatory variables.[16] Hence the total number of possible regressors is 11. Since we are studying the long run relationship between economic growth and other economic variables, we average the data over every five years. Due to missing data and the dynamic nature of our model, we can only use 143 observations in the panel. Since it is an unbalanced panel, i.e. not every country in the panel has the same number of observations (T), we have to replace some quantities that appear in the previous sections as the following,

$$\begin{aligned} T^{-\frac{N}{2}} &: \prod_{i=1}^{N} T_i^{-\frac{1}{2}}, \\ NT &: \sum_{i=1}^{N} T_i, \\ Nb(\rho) &: \sum_{i=1}^{N} b(\rho, T_i). \end{aligned}$$

There are 4,096 possible model specifications in total. Here in Table 17, we just present the top ten models with the highest posterior model probabilities. We can see that the top model is nested in most of the top ten models and it just accounts for around 6.4% posterior model probability while the model probability of the top 10 models in total is about 30%. The result is quite different from the simulation studies in the previous section when we have a true model to generate our data. We find that in simulation the top model alone (in many cases, the true one) usually accounts for above 30%. This confirms the fact that there is substantial model uncertainty in our data. To study the

[16] At most there are 5 observations for each country. Due to the dynamic nature of our model, we have to take away one observation. Therefore we have three dummy variables.

Table 16: Details of the explanatory variables

1. START: the per capita GDP at the starting year of the five years. It enters the equation of (35) in natural log.
2. PRIV: the measure of bank development, calculated from bank claims on the private sector by deposit money banks divided by GDP. It enters the equation in log.
3. PI: the inflation rate. It enters the equation as log(1+PI).
4. GOV: the ratio of government expenditure to GDP. It enters the equation in log.
5. TRADE: the shares of exports and imports to GDP. It enters the equation in log.
6. TOR: the measure of stock market development, which equals the value of traded shares on domestic exchanges divided by the total value of listed shares. It enters the equation in log.
7. BMP: the black market premium. It enters the equation as log(1+BMP).
8. SCHOOL: average years of schooling. It enters the equation as log(1+SCHOOL).

relationship of economic growth and different economic variables, BMA should be a more preferable approach.

Table 17: Posterior Model Probabilities of the Top Ten Models

Ranking	**Model**	**Posterior Model Probability**
1	0,1,6[a]	0.064
2	0,1,6,9	0.057
3	0,1,5,6,9	0.037
4	0,1,4,6	0.029
5	0,1,4,6,9	0.025
6	0,1,3,6,9	0.021
7	0,1,5,6	0.0183
8	0,1,4,5,6,9	0.0176
9	0,1,3,6	0.016
10	0,1,3,4,6,9	0.014

[a] See the description of the set of explanatory variables. 0 stands for the GDP of one period lag. Number 9 to 11 denote the dummy variables.

The BMA point estimates of the slope parameters from the posterior distribution in equation (8) are shown in Table (18), where we omit the results for the dummy variables. The estimates are based on 10,000 draws from the posterior model and parameter space. The column headed by "slope" presents the posterior mean of β in (35). The "NSE" column is the numerical standard error, which is a measure of accuracy of our simulated calculations. The true posterior means with around 95% confidence should lie in the range of [estimate$-1.96NSE$,estimate$+1.96NSE$] due to the central limit theorem. The inclusion probability is calculated as the sum of the model probabilities from the models that include the regressor. Finally, $prob < 0$ is the cumulative posterior probability of the parameter less than 0. It is based on the mixture of the models that include the regressor and can be viewed as how far away the posterior distribution is from 0. If our point estimate is negative (positive) and its posterior distribution is far away from 0, we would expect $prob < 0$ to be close to 1 (0). Not surprisingly, the regressors with the highest inclusion probability are the initial GDP and the lagged GDP, which are closely related to our dependent variable, the per capita GDP growth. The turnover of stock market also has high inclusion probability, about 78% and it is positively related with economic growth and its posterior mean is around 1.28. This confirms the finding by Beck and Levine (2004), whose GMM point estimates of stock market turnover are significant and they range from 0.95 to 1.7 under the inclusion of different sets of exogenous variables. However, it is a surprise to see that bank credit to private sector, which is a measure of bank development, has the lowest inclusion probability among all the regressors and its point estimate is quite close to 0. Moreover, the column of $prob < 0$ tells us that the posterior distribution of stock market turnover is far away from 0 while the posterior distribution of

bank credit has its center near 0. It seems that bank development is not that important for economic growth. This finding seems to contradict the results based on the GMM approach in Beck and Levine (2004).

Table 18: Estimates of the Slope Parameters

regressor	slope	NSE	inclusion probability	$prob < 0$
START	0.74	0.08	1	0
PRIV	0.055	0.04	0.14	0.38
PI	-1.19	0.07	0.27	0.89
GOV	-2.24	0.06	0.37	0.95
TRADE	1.66	0.05	0.35	0.05
TOR	1.28	0.007	0.78	0.0057
BMP	-0.002	0.014	0.16	0.49
SCHOOL	-0.1	0.14	0.16	0.55
LAG	-0.82	0.0009	0.99	1

To verify our results, in Table 19 and Table 20 we present the highest (marginal) posterior probability intervals (HPDI) of bank private credit and stock market turnover respectively. Such intervals are calculated by a kernel smoothing package (ksdensity.m) in MatLab®. The package uses a normal kernel function to fit to certain number of draws from the parameter's posterior distribution. For bank private credit, the number of draws is 1,414 and the one for stock market turnover is 7,794. Note that the results are based on the models which include the regressor. The HPDI results confirm what we found previously, i.e. the posterior distribution for stock market turnover is far different from zero while bank private credit is not. We may conclude that stock market development is more important to economic growth than bank development based on our dataset.

Table 19: The highest posterior density intervals for the private credit

PRIV	lower bound	upper bound
99%	-3.45	4.21
95%	-2.70	3.34
90%	-2.08	2.82
80%	-1.48	2.31

Table 20: The highest posterior density intervals for the stock market turnover

TOR	lower bound	upper bound
99%	0.118	3.104
95%	0.432	2.83
90%	0.64	2.66
80%	0.84	8.56

Next from Table 21 to Table 24, we present the results under the g-prior $\eta = \frac{1}{K^2}$, where K is the number of potential explanatory variables plus one (the lag term of the dependent variable). As is shown in our simulation, this prior may have better performance when the cross section sample size is as small as in our application. We can see that the second prior mainly reconfirms our previous results. First there is substantial model uncertainty as shown by the top model probability.[17] Second the stock market development is more significant and the bank private credit is more insignificant under the second prior than the first prior. One difference under the second prior is that trade seems more important. The top model now consists of stock market development and trade.

Table 21: Posterior Model Probabilities of the Top Ten Models under the prior $\eta = \frac{1}{K^2}$

Ranking	**Model**	**Posterior Model Probability**
1	0,1,6,9	0.109
2	0,1,5,6,9	0.0965
3	0,1,6	0.0671
4	0,1,4,5,6,9	0.0572
5	0,1,4,6,9	0.056
6	0,1,4,6	0.034
7	0,1,3,4,5,6,9	0.025
8	0,1,3,6,9	0.023
9	01,3,4,6,9	0.0216
10	0,1,3,5,6,9	0.0214

[a] See the description of the set of explanatory variables. Number 0 stands for the GDP of one period lag. Number 9 to 11 denote the dummy variables.

Table 22: Estimates of the Slope Parameters under the prior $\eta = \frac{1}{K^2}$

regressor	slope	NSE	inclusion probability	$prob < 0$
START	0.84	0.079	1	0
PRIV	0.047	0.043	0.093	0.34
PI	-0.79	0.069	0.20	0.91
GOV	-2.48	0.052	0.39	0.97
TRADE	2.05	0.041	0.40	0.018
TOR	1.73	0.006	0.93	0.00086
BMP	0.01	0.015	0.09	0.40
SCHOOL	0.001	0.15	0.1	0.37
LAG	-0.93	0.00084	0.99	1

[17] The sum of the posterior top ten model probabilities is 51%.

Table 23: The highest posterior density intervals for the private credit under the prior $\eta = \frac{1}{K^2}$

PRIV	lower bound	upper bound
99%	-2.82	3.59
95%	-2.06	3.08
90%	-1.68	2.70
80%	-1.07	2.26

Table 24: The highest posterior density intervals for the stock market turnover under the prior $\eta = \frac{1}{K^2}$

TOR	lower bound	upper bound
99%	0.49	3.21
95%	0.82	2.92
90%	1.02	2.77
80%	1.18	2.57

8 Conclusion

In this paper, we investigate the information orthogonal method proposed by Lancaster (2002) to obtain consistent estimation of common parameters under a model comparison context. We found that under the linear dynamic panel model, when the wrong set of exogenous regressors are included, it is not necessarily true that Lancaster's fixed effect transformation will lead to consistent estimation of the autoregressive coefficient. To take into account the effect of model misspecification on parameter estimation and to provide a measure of goodness of fit, we advocate to compare different model specifications. In the paper, we use Lancaster's transformation to estimate the model and to calculate the marginal likelihood. We have shown the conditions under which the Bayes factor can lead to consistency in model selection. When the conditions are not obviously met, we rely on Monte Carlo experiments and find that the Bayes factor obtained from the transformation can still lead to the selection of the true model asymptotically. We also compare the BIC model selection performance based on the biased estimates with our Bayes factor method. It is found that the BIC performs very poorly and that some errors will not disappear with the increase of cross section sample size. This shows the relationship between parameter estimation and model selection. It will be more likely for us to obtain consistency in model selection if we can have consistency in parameter estimation. When model uncertainty is substantial, we argue for the use of Bayesian model averaging. Through Monte Carlo experiments, we have found that BMA will tend to produce smaller RMSE than if we simply select the model with the highest posterior model probability. Using the method developed, we study the connection of stock market and bank development with economic growth. Consistent with the results from the classical approach, our finding suggests

that stock market development is positively correlated with economic growth. However, the effect of bank development on economic growth is not significant, which contradicts the classical results. In our framework, we have restricted our attention to stationary data and strictly exogenous explanatory variables. Future research to relax such restrictions could be promising.

References

Arellano, M. and S. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58, 277–97.

Beck, T. and R. Levine (2004): "Stock markets, banks, and growth: Panel evidence," *Journal of Banking & Finance*, 28, 423–442.

Blundell, R. and S. Bond (1998): "Initial conditions and moment restrictions in dynamic panel data model," *Journal of econometrics*, 115–143.

Cox, D. R. and N. Reid (1987): "Parameter Orthogonality and Approximate Conditional Inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, 49, 1–39.

Fernandez, C., E. Ley, and M. F. Steel (2001): "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100, 381–427.

Gourieroux, C., P. C. B. Phillips, and J. Yu (2006): "Indirect Inference for Dynamic Panel Models," Cowles Foundation Discussion Papers 1550, Cowles Foundation, Yale University.

Hahn, J. and W. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72, 1295–1319.

Kass, R. E., L. Tierney, and J. B. Kadane (1990): "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*, ed. by S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, North-Holland.

Lancaster, T. (2000): "The incidental parameter problem since 1948," *Journal of Econometrics*, 95, 391–413.

——— (2002): "Orthogonal Parameters and Panel Data," *Review of Economic Studies*, 69, 647–666.

Madigan, D. and J. York (1995): "Bayesian graphical models for discrete data," *International Statistical Review*, 63, 215–232.

Nerlove, M. (1968): "Experimental Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross-Sections," *The Economic Studies Quarterly*, 18, 42–74.

Nickell, S. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49, 1417–1426.

Poirier, D. J. (1995): *Intermediate Statistics and Econometrics : A Comparative Approach*, Cambridge, Massachusetts, USA: MIT Press.

Raftery, A. E. and Y. Zheng (2003): "Long-Run Performance of Bayesian Model Averaging," Technical report no. 433, Department of Statistics, University of Washington.

Süli, E. and D. Mayers (2003): *An Introduction to Numerical Analysis*, The Edinburgh Building, Cambridge, UK: Cambridge University Press.

Tierney, L. and J. Kadane (1986): "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Yuan, Z. and Y. Yang (2005): "Combining Linear Regression Models: When and How?" *Journal of the American Statistical Association*, 100, 1202–1214.

Zellner, A. (1986): "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*, ed. by P. K. Goel and A. Zellner, Amsterdam: North-Holland, 233–243.

A Appendix

A.1 The Informatation Orthogonal Method

Here we briefly mentioned the information orthogonal method developed by Lancaster (2002). In general, we can separate the parameters in the model into two categories, the incidental parameter, f_i, for $i = 1, 2, \ldots, N$, where N is the sample size, and the common parameter, θ, whose dimension will stay the same regardless of the sample size. When we say f_i is information orthogonal to θ, we mean the following,

$$\begin{aligned} & E\left(\frac{\partial^2 \ln p(y_i|\theta, f_i)}{\partial f_i \partial \theta}\right)\Big|_{f_i=f_{i,true},\theta=\theta_{true}} \\ & = \int \frac{\partial^2 \ln p(y_i|\theta, f_i)}{\partial f_i \partial \theta} p\left(y_i|\theta, f_i\right)\, dY\Big|_{i=f_{i,true},\theta=\theta_{true}} = 0 \end{aligned} \tag{36}$$

Lancaster (2002) showed that if (36) evaluated at the true values of f_i and θ is satisfied, the mode of the marginal posterior of θ ($p(\theta|y) \propto \int \prod_{i=1}^{N} p(y_i|\theta, f_i) p(f_i|\theta) d\, f$), which is obtained by integrating out f_i with repect to the flat prior, $p(f_i\theta) \propto 1$[18], is a consistent estimator. When the the original incidental parameter is not information orthogonal to the common parameter, Lancaster (2002) suggested we reparameterize $f_i = (g_i, \theta)$ such that the new incidental parameter g_i is information orthogonal to θ. To find the new parameterization is equivalent to finding the solution for the following differential equation,

$$\frac{\partial f_i}{\partial \theta} = -\left(E_Y(\frac{\partial^2 \ln p(y_i|\theta, f_i)}{\partial f_i^2})\right)^{-1} E_Y\left(\frac{\partial^2 \ln p(y_i|\theta, f_i)}{\partial f_i \partial \theta}\right) \tag{37}$$

However, when θ is a vector, say $\theta = (\theta_1, \theta_2)$, there is no guarantee that (37) has a solution since the compatability condition $\frac{\partial^2 f_i}{\partial \theta_1 \partial \theta_2} = \frac{\partial^2 f_i}{\partial \theta_2 \partial \theta_1}$ may not be satisfied. For the linear AR(1) panel model, Lancaster (2002) showed that an information orthogonal parameterization can not be found.

A.2 Proof of Proposition 2.1

Denote y_i as $[y_{i,1}, y_{i,2}, \ldots, y_{i,T}]'$ and $y_{i\_}$ as $[y_{i,0}, y_{i,1}, \ldots, y_{i,T-1}]'$. We can rewrite model (1) as

$$y_i = f_i\iota + y_{i\_}\rho + X_i\beta + u_i \tag{38}$$

$$y_{i\_} = f_i c_1 + y_{i,0} c_2 + CX_i\beta + Cu_i \tag{39}$$

[18]Here we assume the flat prior is permissible in the sense that $\int p(y_i|\theta, f_i) d\, f_i < \infty$.

where

$$c_1 = \begin{pmatrix} 0 \\ 1 \\ 1+\rho \\ \cdots \\ 1+\rho+\rho^2+\cdots+\rho^{T-2} \end{pmatrix}, c_2 = \begin{pmatrix} 1 \\ \rho \\ \rho^2 \\ \cdots \\ \rho^{T-1} \end{pmatrix}, C = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \rho & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \rho^{T-2} & \rho^{T-3} & \cdots 1 & 0 \end{pmatrix}$$

and ι is a $T \times 1$ vector of ones. Note that $h(\rho)$ defined in 13 is equal to $\frac{1}{T}\iota' c_1 = -trace(C'H)$.

Lancaster (2002) finds the appropriate reparameterization of the fixed effect parameter by drawing analogy from two simpler cases, i.e. when the model only contains either the lag term of the dependent variable or the exogenous regressors. Here we provide a slightly different derivation of the reparameterization. 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. The solution turns out to be the same as Lancaster's. The derivation strategy adopted here is also needed for the proof of Proposition 3.1. To obtain such a correction function, let us first reparameterize the fixed effect as

$$f_i = g_i \underline{r}(\rho) - \frac{1}{T}\iota' X_i \beta \tag{40}$$

where $\underline{r}(\rho)$ is a function of ρ, which we will find out later.
Now we can transform our model as

$$y_i = g_i \underline{r}(\rho)\iota + y_{i\_}\rho + HX_i\beta + u_i, \tag{41}$$

The following is the derivation of the posterior distribution and the marginal likelihood.

Let us define $w_i = y_i - y_{i\_}\rho$. The product of the likelihood of the transformed model and the prior for θ is

$$\begin{aligned} p(\theta)p(Y|\theta, Y_0) = p(\beta|\sigma^2)\frac{1}{2}I(-1 \le \rho \le 1)(2\pi)^{-\frac{TN}{2}}\sigma^{2(-\frac{NT+2}{2})} \\ \prod_{i=1}^{N} \exp\left\{-\frac{1}{2\sigma^2}[w_i - g_i\underline{r}(\rho)\iota - HX_i\beta]'[w_i - g_i\underline{r}(\rho)\iota - HX_i\beta]\right\}, \end{aligned} \tag{42}$$

where $Y = (y_1, y_2, ..., y_N)$ excludes the first observations of all economic agents, i.e. $Y_0 = (y_{1,0}, y_{2,0}, ...y_{N,0})$.

Next we derive the posterior distribution of g_i. Note that $\iota' H = 0$ so we can rewrite equation (42) as

$$\begin{aligned} p(\theta)p(Y|\theta, Y_0) &= p(\beta|\sigma^2)\frac{1}{2}I(-1 \le \rho \le 1)(2\pi)^{-\frac{TN}{2}}\sigma^{2(-\frac{NT+2}{2})} \\ &\prod_{i=1}^{N} \exp\Big\{-\frac{1}{2\sigma^2}[(w_i - HX_i\beta)'(w_i - HX_i\beta) \\ &+ Tg_i^2\underline{r}^2(\rho) - 2\iota' w_i g_i \underline{r}^2(\rho)]\Big\}. \end{aligned}$$

Then we complete the square for $g_i\underline{r}(\rho)$ by adding $-\frac{(\iota' w_i)^2}{T} + \frac{(\iota' w_i)^2}{T}$ inside the exponential. So it becomes

$$\begin{aligned} p(\theta)p\left(Y|\theta, Y_0\right) &= p(\beta|\sigma^2)\frac{1}{2}I(-1 \leq \rho \leq 1)(2\pi)^{-\frac{TN}{2}}\sigma^{2(-\frac{NT+2}{2})} \\ &\prod_{i=1}^{N}\exp\Big\{-\frac{1}{2\sigma^2}[(w_i - \frac{\iota\iota' w_i}{T} - HX_i\beta)'(w_i - \frac{\iota\iota' w_i}{T} - HX_i\beta) \\ &+T(g_i\underline{r}(\rho) - \frac{\iota' w_i}{T})^2]\Big\}, \end{aligned}$$

or equivalently

$$\begin{aligned} p(\theta)p\left(Y|\theta, Y_0\right) &= p(\beta|\sigma^2)\frac{1}{2}I(-1 \leq \rho \leq 1)(2\pi)^{-\frac{TN}{2}}\sigma^{2(-\frac{NT+2}{2})} \\ &\prod_{i=1}^{N}\exp\Big\{-\frac{1}{2\sigma^2}[(w_i - X_i\beta)'H(w_i - X_i\beta) \\ &+T(g_i\underline{r}(\rho) - \frac{\iota' w_i}{T})^2]\Big\} \end{aligned}$$

Note that $Hw_i = H(y_i - y_{i\_}\rho)$ and $\frac{\iota' w_i}{T} = \frac{\iota'(y_i - y_{i\_}\rho)}{T}$. So we can have

$$\begin{aligned} p(\theta)p\left(Y|\theta, Y_0\right) &= p(\beta|\sigma^2)\frac{1}{2}I(-1 \leq \rho \leq 1)(2\pi)^{-\frac{TN}{2}}\sigma^{2(-\frac{NT+2}{2})} \\ &\prod_{i=1}^{N}\exp\left\{-\frac{\underline{r}^2(\rho)}{2\frac{\sigma^2}{T}}[g_i - \frac{\iota'(y_i - y_{i\_}\rho)}{T\underline{r}(\rho)}]^2\right\} \\ &\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - y_{i\_}\rho - X_i\beta)'H(y_i - y_{i\_}\rho - X_i\beta)\right] \end{aligned} \tag{43}$$

Remember $p(\beta|\sigma^2)$ does not involve parameters other than σ^2. Moreover, since we ignore the distribution of Y_0 and assume the prior of θ is independent of it, from (43) it is clear that the posterior distribution of g_i conditional on $y_{i,0}$, σ^2 and ρ is i.i.d. normal as in (7).

Next we go on to derive the posterior distributions for β and σ^2. First we can integrate out g in equation (43) to obtain

$$\begin{aligned} p(\rho, \beta, \sigma^2, Y|Y_0) &= p(\rho, \beta, \sigma^2|Y, Y_0)p(Y|Y_0) \\ &= p(\beta|\sigma^2)\frac{1}{2}I(-1 < \rho \leq 1)T^{-\frac{N}{2}}(2\pi)^{-\frac{N(T-1)}{2}}\sigma^{2\left[-\frac{N(T-1)+2}{2}\right]} \\ &\underline{r}^{-N}(\rho)\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - y_{i\_}\rho - X_i\beta)'H(y_i - y_{i\_}\rho - X_i\beta)\right]. \end{aligned} \tag{44}$$

Let us define a new function $r(\rho) = \underline{r}^{-N}(\rho)$, $\tilde{w}_i = H(y_i - y_{i\_}\rho)$ and $\tilde{X}_i = HX_i$. Incorporating the prior of β in (6) we can rewrite equation (44) as

$$p(\rho,\beta,\sigma^2|Y,Y_0)p(Y|Y_0) = \frac{1}{2}I(-1\leq\rho\leq 1)T^{-\frac{N}{2}}(2\pi)^{-\frac{N(T-1)+k}{2}}\cdot$$
$$\sigma^{2\left[-\frac{N(T-1)+2+k}{2}\right]}r(\rho)\left|\eta\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right|^{\frac{1}{2}}\cdot$$
$$\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}\tilde{w}_i'\tilde{w}_i+\beta'\sum_{i=1}^{N}(\eta+1)\tilde{X}_i'\tilde{X}_i\beta-2\sum_{i=1}^{N}\tilde{w}_i'\tilde{X}_i\beta\right]\right\}$$

Then completing the square of β yields

$$p(\rho,\beta,\sigma^2|Y,Y_0)p(Y|Y_0) = \frac{1}{2}I(-1\leq\rho\leq 1)T^{-\frac{N}{2}}(2\pi)^{-\frac{N(T-1)+k}{2}}\cdot$$
$$\sigma^{2\left[-\frac{N(T-1)+2+k}{2}\right]}r(\rho)\left|\eta\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right|^{\frac{1}{2}}\cdot$$
$$\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}\tilde{w}_i'\tilde{w}_i-\frac{1}{\eta+1}\sum_{i=1}^{N}\tilde{w}_i'\tilde{X}_i\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i\right]\right\}\cdot$$
$$\exp\left\{-\frac{1}{2\sigma^2}\left[\beta-\frac{1}{\eta+1}\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i\right]'\cdot\right.$$
$$\left.\left(\sum_{i=1}^{N}(\eta+1)\tilde{X}_i'\tilde{X}_i\right)\left[\beta-\frac{1}{\eta+1}\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i\right]\right\}$$

We can see that the posterior kernel for β is normal as in (8) and hence we can integrate it out. The posterior distribution for ρ and σ^2 is

$$p(\rho,\sigma^2|Y,Y_0)p(Y|Y_0) = \frac{1}{2}I(-1\leq\rho\leq 1)\left(\frac{\eta}{\eta+1}\right)^{\frac{k}{2}}T^{-\frac{N}{2}}(2\pi)^{-\frac{N(T-1)}{2}}$$
$$\sigma^{2\left[-\frac{N(T-1)+2}{2}\right]}r(\rho)\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}\tilde{w}_i'\tilde{w}_i\right.\right. \tag{45}$$
$$\left.\left.-\frac{1}{\eta+1}\sum_{i=1}^{N}\tilde{w}_i'\tilde{X}_i\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i\right]\right\}$$

It is also clear from equation (45) that conditional on ρ, σ^2 follows an inverted gamma distribution with mean $\frac{A}{N(T-1)-2}$ and degrees of freedom $N(T-1)$ as in (9).

Now we can integrate out σ^2 to obtain the posterior distribution of ρ as in (10). Another way to write the posterior of ρ is as follows

$$p(\rho|Y,Y_0) \propto I(-1<\rho<1)r(\rho)t(\frac{b}{a},\frac{c}{av}-\frac{b^2}{a^2v},v) \tag{46}$$

where

$$
\begin{aligned}
a &= \sum_{i=1}^{N} y_{i\_}' H y_{i\_} - \frac{1}{\eta+1} \sum_{i=1}^{N} (y_{i\_}' H X_i) \left(\sum_{i=1}^{N} X_i' H X_i \right)^{-1} \sum_{i=1}^{N} (X_i' H y_{i\_}) \\
b &= \sum_{i=1}^{N} y_{i\_}' H y_i - \frac{1}{\eta+1} \sum_{i=1}^{N} (y_{i\_}' H X_i) \left(\sum_{i=1}^{N} X_i' H X_i \right)^{-1} \sum_{i=1}^{N} (X_i' H y_i) \\
c &= \sum_{i=1}^{N} y_i' H y_i - \frac{1}{\eta+1} \sum_{i=1}^{N} (y_i' H X_i) \left(\sum_{i=1}^{N} X_i' H X_i \right)^{-1} \sum_{i=1}^{N} (X_i' H y_i) .
\end{aligned} \tag{47}
$$

Equation (46) tells us that when ρ is in the stationary region, its kernel of the posterior distribution can be viewed as the product of $r(\rho)$ and the t distribution with the mean parameter $\frac{b}{a}$ and the variance parameter $\frac{c}{av} - \frac{b^2}{a^2 v}$, where $v = N(T-1)-1$ is the degrees of freedom. Note that $\frac{b}{a}$ is the within-group estimator, which we could obtain if we operate on the first difference data and adopt a non-informative prior for ρ by assuming our model is stationary ($|\rho| < 1$) and the regressors are exogenous. This estimator is inconsistent and the bias is a function of the true value of ρ. If our posterior estimate of ρ is consistent, $r(\rho)$ should act as the correction term to the bias. Let us denote NB as the bias and $\underline{\rho}$ as the true value of the parameter. We will have the following[19]

$$
\begin{aligned}
&\underset{N\to\infty}{plim} \frac{b}{a} = \underline{\rho} + NB \\
&\underset{N\to\infty}{plim} \frac{1}{N} a = \underline{a} \\
&NB = \underset{N\to\infty}{plim} \frac{\sum_{i=1}^{N} y_{i\_}' H u_i - \sum_{i=1}^{N} y_{i\_}' H X_i \left(\sum_{i=1}^{N} X_i' H X_i \right)^{-1} \sum_{i=1}^{N} X_i' H u_i}{\sum_{i=1}^{N} y_{i\_}' H y_{i\_} - \sum_{i=1}^{N} y_{i\_}' H X_i \left(\sum_{i=1}^{N} X_i' H X_i \right)^{-1} \sum_{i=1}^{N} X_i' H y_{i\_}} \\
&\qquad = -\frac{\sigma^2 h(\underline{\rho})}{\underline{a}}, \\
&\underset{N\to\infty}{plim} \frac{1}{N} \left[\sum_{i=1}^{N} y_{i\_}' H u_i - \sum_{i=1}^{N} y_{i\_}' H X_i \left(\sum_{i=1}^{N} X_i' H X_i \right)^{-1} \sum_{i=1}^{N} X_i' H u_i \right] = -\sigma^2 h(\underline{\rho}), \\
&\underset{N\to\infty}{plim} \frac{1}{N} \left[\sum_{i=1}^{N} u_i' H u_i - \sum_{i=1}^{N} u_i' H X_i \left(\sum_{i=1}^{N} X_i' H X_i \right)^{-1} \sum_{i=1}^{N} X_i' H u_i \right] = (T-1)\sigma^2,
\end{aligned} \tag{48}
$$

where the function $h(\cdot)$ is given in (13). So we can obtain

$$
\underset{N\to\infty}{plim} \frac{c}{a} = cta = \underline{\rho}^2 + 2\underline{\rho} NB + \frac{(T-1)\sigma^2}{\underline{a}} \tag{49}
$$

[19] Recall that we have specified η as a function of N in a way such that $\eta(N)$ is $o(\frac{1}{N})$.

.
Hence when the cross section sample size tends to infinity, the posterior kernel of ρ can be written as

$$p(\rho|Y,Y_0) \propto I(-1<\rho<1)r(\rho)t(\underline{\rho}+NB,\frac{1}{v}\left(cta-(\underline{\rho}+NB)^2\right),v) \tag{50}$$

Recall that $v=N(T-1)-1$. If our estimate from the above kernel is consistent, the posterior distribution of ρ should become a spike at the true value of ρ (the mode of the kernel). The mode of the kernel in (50) can be obtained from the following first order condition,

$$\frac{1}{N}\frac{d\,\ln p(\rho|Y,Y_0)}{d\,\rho}=0.$$

So we will have

$$\frac{1}{N}\frac{d\,r(\rho)}{d\,\rho}=(T-1)\frac{\rho-\underline{\rho}-NB}{cta-(\underline{\rho}+NB)^2+(\rho-\underline{\rho}-NB)^2}. \tag{51}$$

If our specification of $r(\rho)$ leads to consistent estimator, the true value $\underline{\rho}$ should be a solution for the above differential equation. By using (48), we can obtain

$$Nh(\underline{\rho})d\,\underline{\rho}=\frac{1}{r(\underline{\rho})}d\,r. \tag{52}$$

Finally by using (13), we will have

$$\begin{aligned} r(\rho)&=exp\left(Nb(\rho)\right)\\ \underline{r}(\rho)&=exp\left(-b(\rho)\right), \end{aligned} \tag{53}$$

where $b(\rho)$ is given in (3). By inserting (53) into (40), we will get the transformation in (2). By replacing $\underline{r}(\rho)$ and $r(\rho)$ in our derivation, we will have exactly the same results as those from (7) to (10).

A.3 Proof of Proposition 3.1

When the regressors under the candidate model are neither perfectly correlated nor perfectly uncorrelated with those under the true model, we can define $h_2(\beta,\underline{\rho})$ and $h_3(\beta)$ as in (14) whereX_i and $\underline{X}_i$ denote the regressors under the true and the candidate model respectively. We can also rewrite (47) as (19) and in the limit we will have (20). We can still have (50), but the differential equation in (51) has now become

$$\frac{-N(T-1)\left[h_2(\beta,\underline{\rho})-\sigma^2h(\underline{\rho})\right]}{h_3(\beta)+(T-1)\sigma^2}d\,\underline{\rho}=\frac{1}{r(\underline{\rho})}d\,r \tag{54}$$

If the solution in (53) is still valid, we can insert (52) into (54) to obtain

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

It is obvious that unless we have either $\frac{-(T-1)h_2(\beta,\underline{\rho})}{h_3(\beta)} = h(\underline{\rho})$ or $h_2(\beta,\underline{\rho}) = h_3(\beta) = 0$, (53) is not a solution for (54). In other words, the reparameterization of the fixed effect in (2) cannot lead us to consistent estimation of ρ.[20] Generally speaking, if the candidate model does not nest the true model, it is likely that the reparameterization that will enable us to estimate ρ consistently will involve the true values of the common parameters (β, σ^2 and ρ).

In summary, it is not always true that Lancaster's parameterization of the fixed effect will lead to consistent estimation of the model when the model is misspecified. It therefore justifies our motivation to compare different model specifications.

A.4 Proof of Proposition 4.1

To prove Proposition 4.1 and 4.2, 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 Tierney and Kadane (1986) and Kass et al. (1990). To apply the method, we can first multiply both the numerator and the denominator by $\left(\frac{1}{N}\right)^{-\frac{N(T-1)}{2}}$. The integral appearing in the Bayes factor can be simplified as

$$
\begin{aligned}
&\left(\frac{1}{N}\right)^{-\frac{N(T-1)}{2}} \int_{-1}^{1} \exp\left[Nb(\rho)\right] \\
&\left[\sum_{i=1}^{N} w_i' H w_i - \frac{1}{\eta+1}\sum_{i=1}^{N} w_i' H \underline{X}_i \left(\sum_{i=1}^{N} \underline{X}_i' H \underline{X}_i\right)^{-1} \sum_{i=1}^{N} \underline{X}_i' H w_i\right]^{-\frac{N(T-1)}{2}} d\rho \\
&= \left(\frac{1}{N}\right)^{-\frac{N(T-1)}{2}} \int_{-1}^{1} \exp\left[Nb(\rho)\right] \left(a\rho^2 - 2b\rho + c\right)^{-\frac{N(T-1)}{2}} \, d\rho \qquad (55) \\
&= \left(\frac{a}{N}\right)^{-\frac{N(T-1)}{2}} \int_{-1}^{1} \exp\left(N\left[b(\rho) - \frac{T-1}{2}\ln(\rho^2 - 2\frac{b}{a}\rho + \frac{c}{a})\right]\right) d\rho \\
&= \left(\frac{a}{N}\right)^{-\frac{N(T-1)}{2}} \int_{-1}^{1} \exp\left[Nf(\rho)\right] \, d\rho
\end{aligned}
$$

where $f(\rho)$ and its derivatives are defined as follows,

[20] The inconsistency of the estimator for σ^2 follows since σ^2 is not independent from ρ (asymptotically) as can be seen from (9).

$$
\begin{aligned}
f(\rho) &= b(\rho) - \frac{T-1}{2}\ln(\rho^2 - 2\frac{b}{a}\rho + \frac{c}{a}),\\
f'(\rho) &= h(\rho) - \frac{(T-1)(\rho - \frac{b}{a})}{\rho^2 - 2\frac{b}{a}\rho + \frac{c}{a}},\\
f''(\rho) &= h'(\rho) - \frac{(T-1)(\rho^2 - 2\frac{b}{a}\rho + \frac{c}{a}) - 2(T-1)(\rho - \frac{b}{a})^2}{(\rho^2 - 2\frac{b}{a}\rho + \frac{c}{a})^2},
\end{aligned}
\tag{56}
$$

where $h'(\rho) = \sum_{i=1}^{T-2} \frac{i(T-i-1)}{T}\rho^{i-1} = \frac{1}{(1-\rho)^2} - \frac{(T+2)\rho^{T+1} - 2\rho^T - T\rho^{T-1} - 2\rho + 2}{T(1-\rho)^4}$. Based on (20), if we take the probability limit of (56), we can arrive at (57) as follows,

$$
\begin{aligned}
&\underset{N\to\infty}{plim}\, f(\rho) = b(\rho) - \frac{T-1}{2}\ln\left[\rho^2 - 2(\underline{\rho} + NB)\rho + \underline{\rho}^2 + 2\underline{\rho}NB + \frac{(T-1)\sigma^2 + h_3(\beta)}{\underline{a}}\right],\\
&\underset{N\to\infty}{plim}\, f'(\rho) = h(\rho) - \frac{(T-1)(\rho - \underline{\rho} - NB)}{\rho^2 - 2\rho(\underline{\rho} + NB) + \underline{\rho}^2 + 2\underline{\rho}NB + \frac{(T-1)\sigma^2 + h_3(\beta)}{\underline{a}}},\\
&\underset{N\to\infty}{plim}\, f''(\rho) = h'(\rho) -\\
&\frac{(T-1)\left[\rho^2 - 2\rho(\underline{\rho} + NB) + \underline{\rho}^2 + 2\underline{\rho}NB + \frac{(T-1)\sigma^2 + h_3(\beta)}{\underline{a}} - 2(\rho - \underline{\rho} - NB)^2\right]}{\left[\rho^2 - 2\rho(\underline{\rho} + NB) + \underline{\rho}^2 + 2\underline{\rho}NB + \frac{(T-1)\sigma^2 + h_3(\beta)}{\underline{a}}\right]^2}.
\end{aligned}
\tag{57}
$$

Now we can use Laplace's method to approximate the integral. Suppose for the equation $\underset{N\to\infty}{plim}\, f'(\rho) = 0$, there exists only one solution ρ^* in (−1,1) and $\underset{N\to\infty}{plim}\, f''(\rho^*) < 0$. For large N, the integral in (55) can be approximated by

$$
\begin{aligned}
&\left(\frac{1}{N}\right)^{-\frac{N(T-1)}{2}} \int_{-1}^{1} \exp\left[Nb(\rho)\right]\\
&\left[\sum_{i=1}^{N} w_i' H w_i - \frac{1}{\eta+1}\sum_{i=1}^{N} w_i' H \underline{X}_i \left(\sum_{i=1}^{N} \underline{X}_i' H \underline{X}_i\right)^{-1} \sum_{i=1}^{N} \underline{X}_i' H w_i\right]^{-\frac{N(T-1)}{2}} d\rho\\
&\approx \underline{a}^{-\frac{N(T-1)}{2}} \sqrt{\frac{2\pi}{N\left|f''(\rho^*)\right|}} \exp\left[Nf(\rho^*)\right]\\
&= \sqrt{\frac{2\pi}{N\left|f''(\rho^*)\right|}} \exp\left[Nb(\rho^*) - \frac{N(T-1)}{2}\ln d(\rho^*)\right],
\end{aligned}
\tag{58}
$$

where $d(\rho)$ is defined in (24).

Moreover, if our choice of the set of regressors included can lead us to consistent estimation of ρ, i.e. either (11) or (12) is satisfied, by substituting the true value of ρ (i.e. $\underline{\rho}$) into (57) we can obtain

$$
\begin{aligned}
&\underset{N\to\infty}{plim}\, f(\underline{\rho}) = b(\underline{\rho}) - \frac{T-1}{2}\ln\frac{(T-1)\sigma^2 + h_3(\beta)}{\underline{a}},\\
&\underset{N\to\infty}{plim}\, f'(\underline{\rho}) = 0,\\
&\underset{N\to\infty}{plim}\, f''(\underline{\rho}) = h'(\underline{\rho}) - \frac{\underline{a}(T-1)}{(T-1)\sigma^2 + h_3(\beta)} + \frac{2h^2(\underline{\rho})}{T-1}.
\end{aligned}
\tag{59}
$$

For large value of N, the integral in (55) can now be approximated by

$$
\begin{aligned}
&\left(\frac{1}{N}\right)^{-\frac{N(T-1)}{2}} \int_{-1}^{1} \exp\left[Nb(\rho)\right]\\
&\left[\sum_{i=1}^{N} w_i' H w_i - \frac{1}{\eta+1}\sum_{i=1}^{N} w_i' H \underline{X}_i \left(\sum_{i=1}^{N} \underline{X}_i' H \underline{X}_i\right)^{-1} \sum_{i=1}^{N} \underline{X}_i' H w_i\right]^{-\frac{N(T-1)}{2}} d\rho\\
&\approx \sqrt{\frac{2\pi}{N\left|h'(\underline{\rho}) - \frac{\underline{a}(T-1)}{(T-1)\sigma^2+h_3(\beta)} + \frac{2h^2(\underline{\rho})}{T-1}\right|}}\cdot\\
&\quad \exp\left[Nb(\underline{\rho}) - \frac{N(T-1)}{2}\ln\left((T-1)\sigma^2 + h_3(\beta)\right)\right]
\end{aligned}
\tag{60}
$$

Considering (17), if $X_{i1}'s$ are the true regressors to generate Y (so $h_2(\beta,\underline{\rho}) = h_3(\beta) = 0$), in the probability limit (17) can be approximated by

$$
\begin{aligned}
&\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}\\
\approx&\frac{1}{2}\left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}} \left[\underline{a}_{|M_0}\underline{\rho}^2 - 2\underline{\rho}\sigma^2 h(\underline{\rho}) + 2\underline{\rho}h_{2|M_0}(\beta,\underline{\rho}) + h_{3|M_0}(\beta) + (T-1)\sigma^2\right]^{\frac{N(T-1)}{2}}\\
&\sqrt{\frac{2\pi}{N\left|h'(\underline{\rho}) - \frac{a}{\sigma^2} + \frac{2h^2(\underline{\rho})}{T-1}\right|}} \exp\left[Nb(\underline{\rho}) - \frac{N(T-1)}{2}\ln(T-1)\sigma^2\right]\\
=&\frac{1}{2}\left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}} \sqrt{\frac{2\pi}{N\left|h'(\underline{\rho}) - \frac{a}{\sigma^2} + \frac{2h^2(\underline{\rho})}{T-1}\right|}} \exp\Bigg\{Nb(\underline{\rho}) +\\
&\frac{N(T-1)}{2}\ln\left[\frac{\underline{a}_{|M_0}\underline{\rho}^2 - 2\underline{\rho}\sigma^2 h(\underline{\rho}) + 2\underline{\rho}h_{2|M_0}(\beta,\underline{\rho}) + h_{3|M_0}(\beta) + (T-1)\sigma^2}{(T-1)\sigma^2}\right]\Bigg\}.
\end{aligned}
\tag{61}
$$

So we can guarantee $\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)} = \infty$ ($\underline{\rho} \neq 0$) as long as (21) holds. It does not matter whether we choose η to be $O(\frac{1}{N})$ or $\frac{1}{K^2}$ as used in the simulation studies.

Now let us consider the case when the true model is M_0 in (17), i.e. the true value of ρ is 0 and X_{i0} are the right regressors. Given the assumptions in Proposition 4.1, the probability limit of the Bayes factor in (17) takes the following form,

$$\begin{aligned}
&\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}\\
\approx&\frac{1}{2}\left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}}\left[(T-1)\sigma^2\right]^{\frac{N(T-1)}{2}}\\
&\sqrt{\frac{2\pi}{N\left|f''(\rho^*|M_1)\right|}}\exp\left[Nb(\rho^*)-\frac{N(T-1)}{2}\ln d(\rho^*)\right]\\
=&\frac{1}{2}\left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}}\sqrt{\frac{2\pi}{N\left|f''(\rho^*|M_1)\right|}}\exp\left[Nb(\rho^*)+\frac{N(T-1)}{2}\ln\left[\frac{(T-1)\sigma^2}{d(\rho^*|M_1)}\right]\right].
\end{aligned}\tag{62}$$

If (23) holds, then the Bayes factor in (17) will tend to 0 for large sample size. If M_1 is misspecified but it can still give consistent estimates of ρ, i.e. $\rho^*=0$ (either (11) or (12) holds), we can simplify (62) as

$$\begin{aligned}
&\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}\\
=&\frac{1}{2}\left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}}\sqrt{\frac{2\pi}{N\left|f''(0|M_1)\right|}}\exp\left[\frac{N(T-1)}{2}\ln\left[\frac{(T-1)\sigma^2}{(T-1)\sigma^2+h_{3|M_1}(\beta)}\right]\right].
\end{aligned}\tag{63}$$

If $h_{3|M_1}(\beta)>0$, the Bayes factor in (63) will be 0 when N tends to infinity. If $h_{3|M_1}(\beta)=0$, we should have $k_1-k_0>0$. Once again, the choice of η between $O(\frac{1}{N})$ and $\frac{1}{K^2}$ are not important here.

A.5 Proof of Proposition 4.2

For (18), suppose the true model is M_1 and M_0 despite being misspecified can still lead to consistent estimation of ρ, (18) can be approximated as

$$\begin{aligned}
&\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}\\
\approx&\left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}}\sqrt{\left|1+\frac{\frac{a}{\sigma^2}-\frac{a(T-1)}{(T-1)\sigma^2+h_3(\beta)}}{h'(\underline{\rho})-\frac{a}{\sigma^2}+\frac{2h^2(\underline{\rho})}{T-1}}\right|}\left\{\frac{(T-1)\sigma^2+h_3(\beta)}{(T-1)\sigma^2}\right\}^{\frac{N(T-1)}{2}}.
\end{aligned}\tag{64}$$

Since $h_3(\beta)$ is a semi-positive definite quadratic form of β, it should be greater than or equal to 0. It is 0 when M_0 nests M_1 ($k_1<k_0$). It is not hard

to see that $\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}=\infty$ when M_1 is the true model and we set η to be $O(\frac{1}{N})$. Under the choice of $\eta=\frac{1}{K^2}$, the Bayes factor in the limit will not tend to infinity, but rathe a constant, which is still possible to be greater than 1 and favours the true model.

If we are comparing the true model to a model under which we cannot obtain consistent estimate of ρ using the transformation of the fixed effect, we cannot use (60) to approximate the marginal likelihood of the misspecified model. In fact we may not be able to use Laplace's method to approximate the integral since $\underset{N\to\infty}{plim}\, f(\rho)$ may not have a unique maximum point in (-1,1). However, if $\underset{N\to\infty}{plim}\, f(\rho)$ has a nice bell shape in the stationary region, we can prove that when using the reparameterization of the fixed effect, Bayes factor can lead to the selection of the true model asymptotically under certain circumstances. To see this, we continue to suppose M_1 is the true model in (18) and denote ρ^* as our estimate of ρ under M_0. The Bayes factor (18) can be approximated by

$$\begin{aligned}&\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}\\&\approx\left(\frac{\eta}{\eta+1}\right)^{\frac{k-k_0}{2}}\sqrt{\left|\frac{f''(\rho^*)}{f''(\underline{\rho})}\right|}\exp\left\{N\left[b(\underline{\rho})-b(\rho^*)+\frac{(T-1)}{2}\ln\left(\frac{d(\rho^*)}{d(\underline{\rho})}\right)\right]\right\}\end{aligned}\tag{65}$$

Note that $d(\underline{\rho})=(T-1)\sigma^2$. So if (25) is satisfied, the Bayes factor is consistent in selecting the true model, as claimed by Proposition 4.2. It is difficult to interpret under what circumstances our data can satisfy (25). Note that the equation $\underset{N\to\infty}{plim}\, f'(\rho)=0$ generally do not have analytical solution when our model is misspecified and it does not nest the true model. Therefore it is hard to check (25) and we have to rely on simulation studies to shed some light on this issue.

A.6 Proof of Proposition 4.3

The likelihood function takes the following form,

$$\begin{aligned}p\left(Y|\theta,Y_0\right)=&(2\pi)^{-\frac{TN}{2}}\sigma^{2(-\frac{NT}{2})}\\&\prod_{i=1}^{N}\exp\{-\frac{1}{2\sigma^2}[y_i-y_{i\_}\rho-\iota f_i-X_i\beta]'\left[y_i-y_{i\_}\rho-\iota f_i-X_i\beta\right]\}.\end{aligned}\tag{66}$$

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

$$\begin{aligned}
\sigma^2 &= \frac{1}{NT}\sum_{i=1}^{N}[y_i - y_{i_}\rho - \iota f_i - X_i\beta]'\,[y_i - y_{i_}\rho - \iota f_i - X_i\beta]\,, \\
f_i &= \frac{\iota'(y_i - y_{i_}\rho - X_i\beta)}{T}, \\
\beta &= \sum_{i=1}^{N}(X_i'HX_i)^{-1}\sum_{i=1}^{N}X_i'H(y_i - yi_{_}\rho), \\
\rho &= \frac{b}{a},
\end{aligned} \tag{67}$$

where a and b are defined in (19) with $\eta = 0$. Based on the MLE, we can find the Bayesian information criterion (BIC) as the following,

$$BIC = NT\left(\ln\frac{c - \frac{b^2}{a}}{NT} + \ln 2\pi + 1\right) + (1 + k + N)\ln(NT). \tag{68}$$

A BIC value close to zero calculated under a model indicates evidence in favor of the model. Using (20), we can find the probability limit of BIC as

$$\begin{aligned}
\underset{N\to\infty}{plim}\, BIC =& NT\left(\ln\frac{\underset{N\to\infty}{plim}\,\frac{1}{N}c - (\underset{N\to\infty}{plim}\,\frac{1}{N}b)^2(\underset{N\to\infty}{plim}\,\frac{1}{N}a)^{-1}}{T} + \ln(2\pi) + 1\right) \\
&+ (1 + k + N)\ln(NT) \\
=& NT\left(\ln\frac{(T-1)\sigma^2 + h_3(\beta) - \underline{a}NB^2}{T} + \ln(2\pi) + 1\right) \\
&+ (1 + k + N)\ln(NT) \\
=& NT\left\{\ln\frac{(T-1)\sigma^2 + h_3(\beta) - \frac{[h_2(\beta,\underline{\rho}) - \sigma^2 h(\underline{\rho})]^2}{\underline{a}}}{T} + \ln(2\pi) + 1\right\} \\
&+ (1 + k + N)\ln(NT).
\end{aligned} \tag{69}$$

For the true model, its BIC value at the probability limit is

$$\underset{N\to\infty}{plim}\, BIC = NT\left(\ln\frac{(T-1)\sigma^2 - \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}}}{T} + \ln(2\pi) + 1\right) + (1 + k + N)\ln(NT). \tag{70}$$

For the model without the lag term of the dependent variable, the BIC at the probability limit is calculated as

$$\begin{aligned}\underset{N\to\infty}{plim}\, BIC &= NT\left(\ln\frac{\underset{N\to\infty}{plim}\,\frac{1}{N}c}{T}+\ln(2\pi)+1\right)+(k+N)\ln(NT)\\ &= NT\left(\ln\frac{(T-1)\sigma^2+h_3(\beta)+\underline{a}\underline{\rho}^2+2\underline{\rho}h_2(\beta,\underline{\rho})-2\underline{\rho}\sigma^2h(\underline{\rho})}{T}+\ln(2\pi)+1\right)\\ &\quad+(k+N)\ln(NT).\end{aligned} \tag{71}$$

Let us now look at the case of (17). When X_{i1} are the true regressors to generate Y_i, the difference between the BIC under M_0 and M_1 is

$$\begin{aligned}BIC_{|M_0}-BIC_{|M_1} =& \\ & NT\ln\frac{(T-1)\sigma^2+h_{3|M_0}(\beta)+\underline{a}_{|M_0}\underline{\rho}^2+2\underline{\rho}h_{2|M_0}(\beta,\underline{\rho})-2\underline{\rho}\sigma^2h(\underline{\rho})}{(T-1)\sigma^2-\frac{\sigma^4h^2(\underline{\rho})}{\underline{a}_{|M_1}}}\\ & +(k_0-k_1-1)\ln(NT)\end{aligned} \tag{72}$$

Clearly if we have $BIC_{|M_0}-BIC_{|M_1}>0$ for large N, which means M_1 is the preferred model, inside the natural log on the right hand side of the equation, the numerator should be larger than the denominator. In other words, we should have (26) stated in Proposition 4.3. If $\underline{\rho}=0$, it is clear that (26) can be satisfied and model selection is consistent. However, if $X_{i1}=X_{i0}$, we can have $\underline{a}_{|M_0}=\underline{a}_{|M_1}=\underline{a}$, $k_1=k_0$ and $h_{2|M_0}(\beta,\underline{\rho})=h_{3|M_0}(\beta)=0$. Hence we can simplify (72) as

$$\begin{aligned}BIC_{|M_0}-BIC_{|M_1} =& NT\ln\frac{(T-1)\sigma^2-\frac{\sigma^4h^2(\underline{\rho})}{\underline{a}}+\underline{a}\underline{\rho}^2-2\underline{\rho}\sigma^2h(\underline{\rho})+\frac{\sigma^4h^2(\underline{\rho})}{\underline{a}}}{(T-1)\sigma^2-\frac{\sigma^4h^2(\underline{\rho})}{\underline{a}}}\\ &+(k_0-k_1-1)\ln(NT)\\ =& NT\ln\left[1+\frac{\left(\underline{a}\underline{\rho}-\underline{\rho}\sigma^2h(\underline{\rho})\right)^2}{(T-1)\underline{a}\sigma^2-\sigma^4h^2(\underline{\rho})}\right]-\ln(NT)\end{aligned} \tag{73}$$

If $\underline{a}\underline{\rho}-\underline{\rho}\sigma^2h(\underline{\rho})=0$, i.e. $\underline{\rho}+NB=0$, we will always have $BIC_{|M_0}-BIC_{|M_1}<0$, which means we will always 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 also arises when M_0 is the true model. Now the

difference between the two BICs is

$$\begin{aligned} BIC_{|M_0} - BIC_{|M_1} =& NT \ln \frac{(T-1)\sigma^2}{(T-1)\sigma^2 + h_{3|M_1}(\beta) - \frac{\left[h_{2|M_1}(\beta,0) - \sigma^2 \frac{T-1}{T}\right]^2}{\underline{a}_{|M_1}}} \\ &+ (k_0 - k_1 - 1)\ln(NT). \end{aligned} \tag{74}$$

If we want to have M_0 preferred by BIC, we should have $BIC_{|M_0} - BIC_{|M_1} < 0$, which means we should have (27) claimed in Proposition 4.3. However, if we have $h_{3|M_1}(\beta) = 0$, which implies $k_1 \geq k_0$, (27) cannot be satisfied since $\frac{[h_{2|M_1}(\beta,0) - \sigma^2 \frac{(T-1)}{T}]^2}{\underline{a}_{|M_1}} \geq 0$. Again, this implies inconsistency in model selection.

For the case of (18), suppose M_1 is the true model, the difference between the BICs calculated under M_0 and M_1 is

$$\begin{aligned} BIC_{|M_0} - BIC_{|M_1} =& NT \ln \frac{(T-1)\sigma^2 + h_{3|M_0}(\beta) - \frac{\left[h_{2|M_0}(\beta,\underline{\rho}) - \sigma^2 h(\underline{\rho})\right]^2}{\underline{a}_{|M_0}}}{(T-1)\sigma^2 - \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}_{|M_1}}} \\ &+ (k_0 - k_1)\ln(NT) \\ =& NT \ln \left\{ 1 + \frac{h_{3|M_0}(\beta) + \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}_{|M_1}} - \frac{\left[h_{2|M_0}(\beta,\underline{\rho}) - \sigma^2 h(\underline{\rho})\right]^2}{\underline{a}_{|M_0}}}{(T-1)\sigma^2 - \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}_{|M_1}}} \right\} \\ &+ (k_0 - k_1)\ln(NT). \end{aligned} \tag{75}$$

If M_1 is the true model, (28) stated in Proposition 4.3 should hold. If X_{i0} nests the true set of regressors, i.e. $h_{2|M_0}(\beta,\underline{\rho}) = h_{3|M_0}(\beta) = 0$ and $\underline{a}_{|M_1} = \underline{a}_{|M_0}$, (75) is reduced to

$$BIC_{|M_0} - BIC_{|M_1} = (k_0 - k_1)\ln(NT) \tag{76}$$

Since $k_0 > k_1$, the difference between the two BICs will be greater than 0. Therefore, the BIC is consistent in model selection in this case.