CARDIFF BUSINESS SCHOOL
WORKING PAPER SERIES

Cardiff Economics
Working Papers

Guangjie Li and Roberto Leon-Gonzalez

A Correction Function Approach to Solve the Incidental Parameter Problem

E2009/6

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

A Correction Function Approach to Solve the Incidental Parameter Problem *

Guangjie Li[†]
Cardiff Business School
Cardiff University

Roberto Leon-Gonzalez[‡§]
National Graduate Institute of Policy Studies

Abstract

Following Lancaster (2002), we propose a strategy to solve the incidental parameter problem. The method is demonstrated under a simple panel Poisson count model. We also extend the strategy to accomodate cases when information orthogonality is unavailable, such as the linear AR(p) panel model. For the AR(p) model, there exists a correction function to fix the incidental parameter problem when the model is stationary with strictly exogenous regressors. MCMC algorithms are developed for parameter estimation and model comparison. The results based on the simulated data sets suggest that our method could achieve consistency in both parameter estimation and model selection.

JEL Classification Code: C52, C11, C12, C13, C15
Keywords: dynamic panel data model with fixed effect, incidental parameter problem, consistency in estimation, model selection, Bayesian model averaging, Markov chain Monte Carlo (MCMC)

*The authors wish to thank Gary Koop for his long-term encouragement and helpful comments. Early draft of this paper has been presented in the 2008 Far Eastern and South Asian Meeting of the Econometric Society and the 63rd European Meeting of the Econometric Society. The authors are responsible for all the remaining errors in the paper.
[†]email address: ligj@cf.ac.uk
[‡]email address: rlg@grips.ac.jp
[§]Fellow of the Rimini Centre of Economic Analysis

1 Introduction

In microeconomic and other applications, we often see models with some parameters whose number will increase with the sample size and other parameters whose number will remain the same. We call those parameters whose number will change with the sample size incidental parameters. They capture the heterogeneity of economic agents. Those parameters whose size remains the same are called common parameters. It is well known in the literature that the maximum likelihood estimates (MLE) of the common parameters are not consistent due to the presence of the incidental parameters. Such problems are documented as incidental parameter problems, see e.g. Nerlove (1968), Nickell (1981) and Lancaster (2000). The failure of the likelihood method has driven researchers to look for valid instruments and orthogonality conditions to estimate the common parameters through generalized method of moments (GMM), see e.g. Arellano and Bond (1991) and Blundell and Bond (1998). However, when the instruments are weak predictors of the endogenous variables, the GMM estimators may have poor finite sample properties and are not free from bias. Such problems have been pointed out by Alonso-Borrego and Arellano (1999) and Stock et al. (2002). A more recent paper by Bun and Windmeijer (2007) showed that both the GMM estimators proposed by Arellano and Bond (1991) and Blundell and Bond (1998) are not free from weak instrument problems for the linear AR(1) panel model when the data are persistent. Moreover, the GMM statistics could have non-normal distributions, even for large sample size. The conventional IV or GMM inferences are hence misleading. Another problem with GMM is that it is hard for researchers to decide whether some set of the moment conditions are more superior than the others when both can pass the overidentification test. In this regard, the GMM framework provides little information on model comparison and selection.

While GMM seems to be the dominant method in most economic applications, there are some researchers who stick to the likelihood based methods to find solutions. The most common practice may be to treat the incidental parameters as random variables from certain distribution and to transform the estimation problem to estimating the common parameters along with the parameters in the distribution of the incidental parameters. It is known as the random effect model in the classical literature, see e.g. Wooldridge (2005). However, the viability of such method depends heavily on the correct specification of the incidental parameter distribution. Hsiao et al. (2002) got around the incidental parameter problem in MLE by assuming certain conditions on the data generating processes of the exogenous regressors. Hahn and Newey (2004) and Arellano and Hahn (2006) developed the bias reduction approach. This approach tries to first estimate the first order bias of the MLE and then remove the estimated bias from the estimator. Another important stream of the likelihood approach is the conditional likelihood method, or the modified profile likelihood developed by Cox and Reid (1987), who found that when the incidental parameters and the common parameters are information orthogonal, an approximation is available for the conditional likelihood given the maximum likelihood estimator

of the incidental parameter. This method attempts to fix the bias of the profile likelihood by introducing information orthogonality. Lancaster (2002) further developed this idea under the Bayesian framework and found the priors which lead to consistent estimation for a few models. However, information orthogonality is not available for all models, such as the linear autoregressive (AR) panel model with fixed effect and exogenous regressors. Arellano and Bonhome (2006) tried to find the first order bias reduction prior and their results showed that such prior will generally involve the dependent variable(s).

In this paper, we propose a strategy to derive the same prior found in Lancaster (2002). Our strategy is related to finding the Jacobian from the old incidental parameters, which are not information orthogonal to the common parameters, to the new information orthogonal incidental parameters and hence the correction function required for consistent estimation. We also extend our strategy to find the bias reducing prior for linear AR panel data model of order more than one. Our results show that the correction function happens to have closed form for this model and it involves only the common parameters in concern. The specific form of the correction function will change with the number of observations for each economic agent and the number of lags in the AR model. With the correction function, the posterior distribution of the common parameters is generally not a standard one. Therefore to estimate the model, we propose a Metropolis-Hastings algorithm. The results from the simulated datasets show strong signs of estimation consistency of our method. A very important issue related to the likelihood based bias correction method raised in Li (2009) is that consistent parameter estimation is related to consistent model selection. For the linear panel AR model, when we include the wrong set of exogenous regressors, we may not be able to obtain consistent estimate for the autoregresive coefficient. Therefore, parameter estimation and model selection should be carried out simultaneously. To compare different model specifications, we use the Bayes factor calculated through the method proposed by Chib and Jeliazkov (2001) and a reversible jump algorithm. The results from the simulated datasets suggest that the Bayes factor criterion could achieve consistency for model selection.

The setup of the paper is as follows. Section 2 gives a Bayesian perspective on the incidental parameter problem and our strategy to find the correction function to solve the problem. Section 3 demonstrates how our strategy is applied to the linear panel AR model of order more than one to derive the correction function. Section 3.2 and Section 3.3 discuss the algorithms to carry out point estimation and model comparison, while Section 3.4 and Section 3.5 give the respective examples using simulated datasets before Section 4 concludes.

2 A Possible Way to Solve the Incidental Parameter Problem

Let us put the parameters to be estimated into two categories: the common parameter, denoted by θ, whose dimension is the same regardless of the sample size, and the incidental parameter, f, whose dimension will increase with the sample size. The Bayesian way to estimate θ is to integrate f out of the likelihood function $p(Y|\theta, f)$ with respect to the prior $p(f|\theta)$ and then the estimation results are drawn from the marginal posterior distribution of θ,

$$\begin{aligned} p(\theta|Y) &\propto \int_F p(\theta, f)p(Y|\theta, f)\, df \\ &\propto \int_F p(\theta)p(f|\theta)p(Y|\theta, f)\, df. \end{aligned} \tag{1}$$

Here we use Y to stand for the collection of the dependent variable(s) and $p(f|\theta)$ is a permissible prior function[1] with support F. The problem with the Bayesian method is that there is no guarantee for us to obtain consistent estimates of θ for arbitrary specification of the prior function, $p(f|\theta)$[2]. That is, the posterior function $p(\theta|Y)$ will become a spike at a point different from the true value of θ (denoted by θ_{true}) as the sample size, N, increases[3]. Denote ν as the probability measure, of which $p(\theta|Y)$ is the density. Further assume that θ has the support Θ. If Ω represents any subset of Θ, we have the following,

$$\nu(\Omega) = \int_{\theta\in\Omega} p(\theta|Y)\, d\theta. \tag{2}$$

The incidental parameter problem now can be interpreted as

$$\underset{N\to\infty}{plim}\, \nu(\Omega) = I(\theta_b \in \Omega) \tag{3}$$

where $I(\cdot)$ is the indicator function and $\theta_b \neq \theta true$. The Bayesian method could be viewed as related to the random effect model in the classical literature, in which $p(f|\zeta, \theta)$[4] is assumed to be the correct distribution for f. In a situation like this, we have a new parameter ζ, whose dimension will not change with the sample size. We then need to estimate it along with θ after we integrate f out of the likelihood with respect to $p(f|\zeta, \theta)$. The difference between $p(f|\zeta, \theta)$ and $p(f|\theta)$ in (1) does not just lie in the introduction of a new parameter. For the random effect model to work well, the assumed $p(f|\zeta, \theta)$ has to be a proper

[1] A permissible prior function means that it should satisfy $p(Y|\theta) = \int_F p(f|\theta)p(Y|\theta, f)\, df < \infty$ for fixed sample size. Note that all proper priors are permissible while improper priors may or may not be permissible. For more details, see Bernardo (2005).

[2] It is shown by Hahn (2004) that the Jeffrey's prior is generally not bias reducing.

[3] We assume that the prior function $p(\theta)$ is non-dogmatic throughout. That is, the integrated likelihood function $p(Y|\theta)$ will asymptotically be dominant in the posterior function.

[4] The conditional density function can also possibly depend on the exogenous regressors.

density[5] and a good approximation of the underlying distribution for the incidental parameter. However, for most situations, it is unlikely for researchers to have such "prior" knowledge about the form of the true incidental parameter distribution. On the other hand, the prior used in a Bayesian framework does not have to be a proper probability measure. There is a large literature on the use of objective priors or so-called reference priors, which only depend on the assumed model and the available data (see Bernardo, 2005). Liseo (2006) found that such priors are able to solve or alleviate the incidental parameter problem for a few specific examples. However, the reference prior is not inherently designed to solve the incidental parameter problem. For some situations, there is not a clear guideline on the choice of bias-reducing prior.

To see why a prior, $p_r(f|\theta)$, can remove the bias, we can compare it to a bias prior, $p_b(f|\theta)$[6] which has the incidental parameter problem described in (3). Here we implicitly assume both priors are permissible. Then the marginal posterior density functions of θ implied by the two priors through the Bayes Theorem can be linked by a function, $p_r(\theta|y) \propto r(\theta)p_b(\theta|y)$[7], where

$$r(\theta) = \frac{\int_F p_r(f|\theta)p(Y|\theta,f)d\,f}{\int_F p_b(f|\theta)p(Y|f,\theta)d\,f}. \tag{4}$$

It is not hard to see that $r(\theta)$ serves as a correction function and is a non-negative and integrable (with respect to ν) function, which can induce another probability measure ν^r,

$$\nu^r(\Omega) = \int\limits_{\theta\in\Omega} k \cdot r(\theta)p_b(\theta|Y)\,d\theta = \int\limits_{\theta\in\Omega} k \cdot r(\theta)\,d\nu. \tag{5}$$

where k is a normalizing constant not depending on θ, such that

$$\underset{N\to\infty}{plim}\,\nu^r(\Omega) = I(\theta_{true} \in \Omega). \tag{6}$$

The problem now is to find the permissible and bias reducing prior, $p_r(f|\theta)$. Here we follow the information orthogonal argument used by Lancaster (2002) to find such prior. If f is information orthogonal to θ, i.e.

$$E_Y\left(\frac{\partial^2 \ln p(Y|\theta,f)}{\partial f\partial\theta}\right) = \int \frac{\partial^2 \ln p(Y|\theta,f)}{\partial f\partial\theta} p\left(Y|\theta,f\right)\,dY = 0 \tag{7}$$

we can just use a flat prior $p(f|\theta) \propto 1$[8] to integrate out the incidental parameter and the resulting marginal posterior mode of θ is a consistent estimator (given that $p(\theta)$ is non-dogmatic). This result holds since the Bayesian integrated likelihood obtained from a flat prior is asymptotically equivalent to the modified

[5] It means $\int_F p(f|\zeta,\theta)d\,f = 1$.

[6] For many cases, it is convenient to choose $p(f|\theta) \propto 1$ as a reference given that it is permissible, though this flat prior could be bias free in some case.

[7] We use the same marginal prior of θ under the two different conditional priors.

[8] We must assume here that the flat prior is a permissible prior.

profile likelihood in Cox and Reid (1987), see also Sweeting (1995). The modified profile likelihood was derived by Cox and Reid (1987) as an approximation to the conditional likelihood given the maximum likelihood estimator of the incidental parameter (as a function of the common parameter) when the incidental parameter is information orthogonal to the common parameter. We can understand this approach from the fact that consistent estimator of the common parameter can be obtained from maximizing the conditional likelihood given the sufficient statistic for the incidental parameter, see Lancaster (2000). If the orginal parameterization does not lead to information orthogonality, Lancaster (2002) suggested that we can reparameterize f as $f(g,\theta)$ such that the new incidental parameter g (with the same dimension as f) is information orthogonal to θ and the integrated likelihood $\int_G p\left(Y|f(g,\theta),\theta\right) d\, g$ can yield consistent estimation of θ. Lancaster (2002) showed that to find the information orthogonal reparameterization amounts to solving the following differential equation

$$\frac{\partial f}{\partial \theta} = -\left(E_Y(\frac{\partial^2 \ln p(Y|\theta,f)}{\partial f \partial f'})\right)^{-1} E_Y\left(\frac{\partial^2 \ln p(Y|\theta,f)}{\partial f \partial \theta}\right) \tag{8}$$

The new incidental parameter g can be recovered as the constant term in the solution. Under the flat prior $p(g|\theta) \propto 1$, the integrated likelihood can lead to consistent estimation of θ. In terms of the original parameterization, the integrated likelihood can be represented as $\int_F |det(\frac{\partial g}{\partial f'})| p\left(Y|f,\theta\right) d\, f$ for $p(g|\theta)|det(\frac{\partial g}{\partial f'})| = p(f^{-1}(g,\theta)|\theta)|det(\frac{\partial g}{\partial f'})| = p(f|\theta) \propto |det(\frac{\partial g}{\partial f'})|$. Hence to find the bias reducing prior is equivalent to finding the Jacobian from the old incidental parameter to the new incidental parameter. If we can assume different individuals ($y_i's$) are conditionally independent, since the bias reducing prior is proportional to the absolute value of the determinant of the Jacobian matrix, without loss of generality, we can assume $\frac{\partial g}{\partial f'}$ is diagonal, which means f_i is only related to g_i in addition to θ, such that $|det(\frac{\partial g}{\partial f'})| = \prod_{i=1}^{N} |\frac{\partial g_i}{\partial f_i}|$. We can now rewrite (8) as

$$\frac{\partial f_i}{\partial \theta} = \chi(f_i,\theta) \tag{9}$$

where $\chi(f_i,\theta)$ is defined as

$$\chi(f_i,\theta) = -\left(E_y(\frac{\partial^2 \ln p(y_i|\theta,f_i)}{\partial f_i^2})\right)^{-1} E_y\left(\frac{\partial^2 \ln p(y_i|\theta,f_i)}{\partial f_i \partial \theta}\right). \tag{10}$$

Since f_i is defined implicitly as a one-one function of g_i, we can differentiate both sides of (9) with respect to g_i to obtain

$$\frac{\partial^2 f_i}{\partial \theta \partial g_i} = \frac{\partial \chi(f_i,\theta)}{\partial f_i}\frac{\partial f_i}{\partial g_i},$$

which is equivalent to

$$-\frac{\partial \ln |\frac{\partial g_i}{\partial f_i}|}{\partial \theta} = \frac{\partial \ln |\frac{\partial f_i}{\partial g_i}|}{\partial \theta} = \frac{\partial^2 f_i}{\partial \theta \partial g_i}\left(\frac{\partial f_i}{\partial g_i}\right)^{-1} = \frac{\partial \chi(f_i,\theta)}{\partial f_i}. \tag{11}$$

Let us denote $\psi(f_i,\theta)=\frac{\partial\chi(f_i,\theta)}{\partial f_i}$ and $\lambda(f_i,\theta)=\ln|\frac{\partial g_i}{\partial f_i}|$. It is possible to find out $\lambda(f_i,\theta)$ and hence $|\frac{\partial g_i}{\partial f_i}|$ from (11) to solve the incidental parameter problem.

Example 1. *Let us consider a simple panel Poisson count model:* $y_{i,t}\sim i.i.d.Poisson\left(f_i\exp(x_{i,t}\theta)\right)$ *with* $t=1,2,\ldots,T$ *and* $i=1,2,\ldots,N$ *where* θ *is a scalar and* $f_i\exp(x_{i,t}\theta)$ *is the mean parameter in the Poisson distribution. Denote* $y_i=(y_{i,1},y_{i,2},\ldots,y_{i,T})'$*, the likelihood contribution of individual* i *is given by*

$$l_i(f_i,\theta)=p(y_i|f_i,\theta)\propto e^{-f_i\sum_t\exp(x_{it\theta})}f_i^{\sum_t y_{it}}e^{\theta\sum_t y_{it}x_{it}} \tag{12}$$

Note that we can choose the parameterization $f_i=g_i\left(\sum_t\exp(x_{it}\theta)\right)^{-1}$ *such that the individual likelihood can be decomposed into two functions of only* g_i *and* θ *respectively, i.e.* $l_i(f_i(g,\theta),\theta)=l_{i1}(g_i)l_{i2}(\theta)$*,*

$$l_i(f_i(g,\theta),\theta)\propto e^{-g_i}g_i^{\sum_t y_{it}}\times\frac{e^{\theta\sum_t y_{it}x_{it}}}{\left(\sum_t\exp(x_{it}\theta)\right)^{\sum_t y_{it}}} \tag{13}$$

which means g_i *and* θ *are orthogonal to each other and the MLE of* θ *is consistent. Due to the parameterization invariance property, the maximum likelihood estimator of* θ *is consistent under even the original parameterization. On the other hand, the flat prior* $p(f_i|\theta)\propto 1$ *can not lead to consistent estimation since the Bayesian integrated likelihood is*

$$p(y_i|\theta)\propto\frac{e^{\theta\sum_t y_{it}x_{it}}}{\left(\sum_t\exp(x_{it}\theta)\right)^{1+\sum_t y_{it}}}, \tag{14}$$

which is different from $l_{i2}(\theta)$ *in (13) and hence the posterior mode of* $p(\theta|y)$ *under the prior* $p(\theta)\propto 1$ *is not a consistent estimator. A natural choice of the correction function is* $r(\theta)=\sum_t\exp(x_{it}\theta)$*, by which (14) is multiplied to give the same form as* $l_{i2}(\theta)$*. We can also derive this correction function and the bias reducing prior from the Jacobian argument outlined before. First note that*

$$\begin{aligned}
E_y\left(\frac{\partial^2 l_i(f_i,\theta)}{\partial f_i\partial\theta}\right)&=-\sum_t x_{it}\exp(x_{it}\theta)\neq 0\\
E_y\left(\frac{\partial^2 l_i(f_i,\theta)}{\partial f_i^2}\right)&=E_y\left(-\frac{\sum_t y_{it}}{f_i^2}\right)=-\frac{\sum_t\exp(x_{it}\theta)}{f_i}\\
\chi(f_i,\theta)&=-\frac{f_i\sum_t x_{it}exp(x_{it}\theta)}{\sum_t exp(x_{it}\theta)}
\end{aligned} \tag{15}$$

We can see that f_i *is not information orthogonal to* θ *in the model. That is why the flat prior is not bias reducing in this case. Next we can see that* $\psi(f_i,\theta)=\frac{\partial\chi(f_i,\theta)}{\partial f_i}=-\frac{\sum_t x_{it}exp(x_{it}\theta)}{\sum_t exp(x_{it}\theta)}$*. Finally use (11) to find out that* $\lambda(f_i,\theta)=\ln(\sum_t exp(x_{it}\theta))$ *and hence the bias reducing prior* $p(f_i|\theta)\propto|\frac{\partial g_i}{\partial f_i}|=\sum_t exp(x_{it}\theta)$*, which is exactly the same as the correction function we found earlier.*

When the dimension of θ is more than one, say, $\theta = (\theta_1, \theta_2)$, there is no guarantee that we can find $\lambda(f_i, \theta)$ from the differential equation (11) since the compatibility condition $\frac{\partial^2 \psi(f_i,\theta)}{\partial \theta_1 \partial \theta_2} = \frac{\partial^2 \psi(f_i,\theta)}{\partial \theta_2 \partial \theta_1}$ may not be satisfied. That is why the information orthogonal reparameterization in general does not exist as pointed out by Lancaster (2002). For the linear dynamic panel AR(1) model, Lancaster found that information orthogonality is not necessary for consistent estimation of the common parameter. Note that if θ is a scalar, we can always find $\lambda(f_i, \theta)$ from (11). The idea proposed here is to break θ into blocks such that for the jth block we have the differential equation $\frac{\partial \lambda_j(f_i,\theta)}{\partial \theta_j} = \psi_j(f_i, \theta)$ which can be solved to obtain $\lambda_j(f_i, \theta)$. We then assemble all the solutions to yield the bias reducing prior as

$$p(f_i|\theta) \propto \exp\left[\lambda_1(f_i, \theta) + \lambda_2(f_i, \theta) + \dots\right]. \tag{16}$$

We will show in the next section that such strategy can produce the prior and the correction function needed to give consistent estimation for the linear dynamic AR(p) panel model.

3 The Linear AR(p) Panel Model with Fixed Effect

3.1 The Bias Reducing Prior and the Posterior Results

Suppose our model has p lags and can be written as

$$y_i = \iota f_i + Y_{i\_}\rho + X_i\beta + u_i \tag{17}$$

where y_i is $[y_{i,1}, y_{i,2}, \dots, y_{i,T}]'$, f_i is the fixed effect scalar, ι is a vector of ones, $Y_{i\_}$ is a $T \times p$ matrix, in which a typical row (the $j+1$th row) looks like $[y_{i,j}, y_{i,j-1}, \dots, y_{i,j-p+1}]$ (j=0,1,...,T-1), ρ is $[\rho_1, \rho_2, \dots, \rho_p]'$, X_i is a strictly exogenous regressor matrix of dimension $T \times K$ and u_i is a $T \times 1$ disturbance, for which we assume $u_i \sim i.i.d.N(0, \sigma^2 I_T)$.

In our model, it is obvious that f_i is the incidental parameter, or the fixed effect, which captures the heterogeneity of economic agents, while $\theta = (\rho', \beta', \sigma^2)'$ are the common parameters, which we want to have consistent estimates for. The dimension of θ is $p + K + 1$. Lancaster (2002) showed that there does not exist any information orthogonal reparameterization for this model. However, we can see that θ has naturally three blocks, ρ, β and σ^2. For each block, we may be able to solve the differential equation (11) to obtain $\lambda_\rho(f_i, \theta)$, $\lambda_\beta(f_i, \theta)$ and $\lambda_{\sigma^2}(f_i, \theta)$. Using the strategy mentioned in the previous section, the bias reducing prior could have the form:

$$p(f_i|\theta) \propto \exp\left[\lambda_\rho(f_i, \theta) + \lambda_\beta(f_i, \theta) + \lambda_{\sigma^2}(f_i, \theta)\right]. \tag{18}$$

We will show later that this is indeed the case for the model[9].

Note that the log likelihood contribution of individual i conditionnal on the initial p observations (denoted by $y_{i,-p}$) is the following,

$$l_i = \ln p(y_i|f_i, \theta, y_{i,-p}) \propto -\frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left(y_i - \iota f_i - Y_{i\_}\rho - X_i\beta\right)'\left(y_i - \iota f_i - Y_{i\_}\rho - X_i\beta\right). \tag{19}$$

To implement our strategy, we first need to calculate the following quantities,

$$E_y\left(\frac{\partial^2 l_i}{\partial f_i^2}\right) = -\frac{T}{\sigma^2}, \tag{20}$$

$$E_y\left(\frac{\partial^2 l_i}{\partial f_i \partial \beta}\right) = -\frac{T}{\sigma^2}X_i'\iota, \tag{21}$$

$$E_y\left(\frac{\partial^2 l_i}{\partial f_i \partial \sigma^2}\right) = -E_y\left[\frac{\left(y_i - \iota f_i - Y_i\rho - X_i\beta\right)'\iota}{(\sigma^2)^2}\right] = 0, \tag{22}$$

$$\begin{aligned} E_y\left(\frac{\partial^2 l_i}{\partial f_i \partial \rho}\right) &= -\frac{T}{\sigma^2}E_y(Y_{i\_}'\iota) \\ &= -\frac{T}{\sigma^2}\left[Th(\rho)f_i + \omega_1(X_i\beta, \rho) + \omega_2(y_{i,-p}, \rho)\right], \end{aligned} \tag{23}$$

where $h(\cdot)$, $\omega_1(\cdot)$ and $\omega_2(\cdot)$ are all $p \times 1$ vector functions[10]. $\omega_1(\cdot)$ and $\omega_2(\cdot)$ are functions which do not involve f_i. From (22), we can see that f_i is information orthogonal to σ^2. The right hand side of (21) does not involve f_i. Hence we can have $\lambda_\beta(f_i, \theta) = 0_{K\times 1}$ and $\lambda_{\sigma^2}(f_i, \theta) = 0_{1\times 1}$, which implies that we can just use a flat prior $p(f_i|\beta, \sigma^2) \propto 1$ to obtain consistent estimation of β and σ^2 when the model does not have the lag term, i.e. $\rho = 0$.[11] With the lag term, to find $\lambda_\rho(f_i, \theta)$, we need to solve the following differential equation system,

$$\frac{\partial \lambda_\rho(f_i, \theta)}{\partial \rho} = h(\rho). \tag{24}$$

We show in the appendix that (24) has a solution, $\lambda_\rho(f_i, \theta) = \tau(\rho)$, which is a function of ρ only. The functional form of $\tau(\rho)$ depends on T and p. Table 1 shows some forms of $\tau(\rho)$ under different values of T and p. For specific values of T and p, we refer the readers to the appendix of this paper and a Maplet program written by the author (available on request) for the exact form of $\tau(\rho)$. Since our posterior results are conditional on the initial p observations, the actual number of time periods for an economic agent is $T + p$. Under our setup, estimation is only possible if $T \geq 2$. When T takes a particular value, the form for $\tau(\rho)$ will not change for $p \geq T - 1$. Finally the bias reducing prior,

[9] In the appendix, we show that the true values of the common parameters constitute a local stationary point asymptotically for the integrated likelihood under the solution obtained in this way.

[10] See appendix for the detailed forms of the functions.

[11] It is well known that the within group estimator of β under static panel model is consistent. Under the Bayesian framework, the integrated likelihood will give the correct degrees of freedom for the estimator of σ^2.

Table 1: The functional form of $\tau(\rho)$ under different values of T and p

p \ T	2	3	4
1	$\frac{1}{T}\sum_{t=1}^{T-1}\frac{T-t}{t}\rho_1^t$		
2	$\frac{1}{2}\rho_1$	$\frac{1}{3}\sum_{t=1}^{2}\frac{3-t}{t}\rho_1^t+\frac{1}{3}\rho_2$	$\frac{1}{4}\sum_{t=1}^{3}\frac{4-t}{t}\rho_1^t+\frac{1}{4}\rho_1\rho_2+\frac{1}{2}\rho_2$
3	$\frac{1}{2}\rho_1$	$\frac{1}{3}\sum_{t=1}^{2}\frac{3-t}{t}\rho_1^t+\frac{1}{3}\rho_2$	$\frac{1}{4}\sum_{t=1}^{3}\frac{4-t}{t}\rho_1^t+\frac{1}{4}\rho_1\rho_2+\frac{1}{2}\rho_2+\frac{1}{4}\rho_3$
4	$\frac{1}{2}\rho_1$	$\frac{1}{3}\sum_{t=1}^{2}\frac{3-t}{t}\rho_1^t+\frac{1}{3}\rho_2$	$\frac{1}{4}\sum_{t=1}^{3}\frac{4-t}{t}\rho_1^t+\frac{1}{4}\rho_1\rho_2+\frac{1}{2}\rho_2+\frac{1}{4}\rho_3$

$p(f_i|\theta)$ under our strategy in (18) is

$$p(f_i|\theta) = p(f_i|\rho) \propto exp(\tau(\rho)). \tag{25}$$

Note that this prior involves ρ only. The correction function defined in (4) is therefore

$$r(\theta) = r(\rho) = exp[N\tau(\rho)]. \tag{26}$$

For the linear panel AR(p) model, it happens that the conditional prior of f given θ does not involve f in both the numerator and the denominator on the left hand side of (4). That is why the correction function in (26) has closed form. It is possible that the bias reducing prior defined in (16) can involve f in other cases[12] and the correction function does not have closed form.

Next we need to specify the prior, $p(\theta)$ for our Bayesian analysis. The structure of the prior distribution of (f, θ) looks like the following,

$$\begin{aligned} p(f,\theta) &= p(f,\rho,\beta,\sigma^2) = p(f_1|\rho)\dots p(f_N|\rho)p(\rho)p(\sigma^2)p(\beta|\sigma^2) \\ &\propto r(\rho)\frac{1}{\sigma^2}I(\rho\in S)\frac{1}{m(S)}p(\beta|\sigma^2) \end{aligned} \tag{27}$$

where the set S denotes the stationary region of ρ, $I(\cdot)$ is the indicator function and $m(S)$ is the measure of the volume of S[13]. The general form of $m(S)$ can be found in Piccolo (1982). Here we adopt the uniform prior restricted to the stationary region for ρ. We use the g-prior for the conditional prior of β on σ^2, which is asymptotically non-informative if we set $\eta = \eta(N)$ such that $\lim_{N\to\infty}\eta(N) = 0$[14],

$$\beta|\sigma^2 \sim N\left(0, \sigma^2(\eta\sum_{i=1}^{N}X_i'HX_i)^{-1}\right), \tag{28}$$

where the demean matrix H is equal to $I_T - \frac{\iota\iota'}{T}$.

[12] The binary logistic model is such an example.
[13] For example, if $p = 1$, then $\rho \in (-1, 1)$ and hence $m(S)$=2.
[14] Note also that β and σ^2 are asymptotically independent in our prior.

Proposition 3.1. *Conditional on the initial p observations of the dependent variable, using the bias reducing prior (25) and the priors described in (27) and (28), we can obtain the following posterior distributions,*

$$f_i|Y, y_{i,0}, \sigma^2, \rho, \beta \sim N\left(\frac{\iota'(y_i - Y_{i\_}\rho - X_i\beta)}{T}, \frac{\sigma^2}{T}\right), \tag{29}$$

$$\beta|Y, Y_0, \sigma^2, \rho \sim N\left(\frac{1}{\eta+1}\left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1}\sum_{i=1}^{N} X_i'H(y_i - Y_{i\_}\rho), \sigma^2\left((\eta+1)\sum_{i=1}^{N} X_i'HX_i\right)^{-1}\right), \tag{30}$$

$$\sigma^2|\rho, Y, Y_0 \sim IG(N(T-1), \Delta), \tag{31}$$

where

$$\begin{aligned}\Delta =& \sum_{i=1}^{N}(y_i - Y_{i\_}\rho)'H(y_i - Y_{i\_}\rho) - \\ & \frac{1}{\eta+1}\sum_{i=1}^{N}(y_i - Y_{i\_}\rho)'HX_i\left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1}\sum_{i=1}^{N} X_i'H(y_i - Y_{i\_}\rho).\end{aligned} \tag{32}$$

Moreover, after we integrate out f, β and σ^2, we can have

$$\rho|Y, Y_0 \propto I(\rho \in S)r(\rho)t(A^{-1}b, \frac{1}{N(T-1)-p}(c - b'A^{-1}b)A^{-1}, N(T-1) - p) \tag{33}$$

where

$$\begin{aligned}\underset{p\times p}{A} &= \sum_{i=1}^{N} Y_{i\_}'HY_{i\_} - \frac{1}{\eta+1}\sum_{i=1}^{N}(Y_{i\_}'HX_i)\left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1}\sum_{i=1}^{N}(X_i'HY_{i\_}) \\ \underset{p\times 1}{b} &= \sum_{i=1}^{N} Y_{i\_}'Hy_i - \frac{1}{\eta+1}\sum_{i=1}^{N}(Y_{i\_}'HX_i)\left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1}\sum_{i=1}^{N}(X_i'Hy_i) \\ \underset{1\times 1}{c} &= \sum_{i=1}^{N} y_i'Hy_i - \frac{1}{\eta+1}\sum_{i=1}^{N}(y_i'HX_i)\left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1}\sum_{i=1}^{N}(X_i'Hy_i).\end{aligned} \tag{34}$$

Equation (33) tells us that the kernel of the posterior distribution of ρ can be viewed as the product of $r(\rho)$ and the multivariate t distribution with $N(T-1)-p$ degrees of freedom, mean parameter $A^{-1}b$ and covariance matrix $\frac{1}{N(T-1)-p}(c-b'A^{-1}b)A^{-1}$, which we could have obtained using the flat prior $p(f|\theta) \propto 1$. Note that $A^{-1}b$ is the within group estimator in the classical literature, which is inconsistent. The function $r(\rho)$ serves as the correction function to fix such inconsistency.

3.2 Estimation Algorithm

Our estimation is based on the draws of the parameters from their posterior distributions. From (29), (30) and (31) we can see that the posterior distributions of g, β and σ^2 all depend on ρ. Once we have posterior draws of ρ, we can have draws of other parameters. We can see that the posterior distribution of ρ in (33) is not standard and we can not directly draw from it. Before we get into the details of the posterior estimation, let us recap the prior of ρ in (27). The prior of ρ is a uniform distribution in the stationary region. Barndorff-Nielsen and Schou (1973) found that there is a one-to-one differentiable mapping between the partial autocorrelations (PAC) and the slope coefficients (ρ) for the stationary AR model. Let us denote the PAC as $\pi_{p\times 1} = (\pi_1, \ldots, \pi_p)'$ and introduce the quantities $\kappa^{(k)} = (\kappa_1^{(k)}, \ldots, \kappa_k^{(k)})'$, $k = 1, \ldots, p$. Then the mapping from PAC to ρ can be recovered from

$$\kappa_i^{(k)} = \kappa_i^{(k-1)} - \pi_k \kappa_{k-i}^{(k-1)}, \quad i = 1, \ldots, k-1, \tag{35}$$

with $\kappa_k^{(k)} = \pi_k$ and $\rho = \kappa^{(p)}$. The Jacobian of the transformation is

$$J(\pi) = \prod_{k=2}^{p} (1-\pi_k)^{[\frac{k}{2}]} (1+\pi_k)^{[\frac{k-1}{2}]} \tag{36}$$

On the other hand, the mapping from ρ to π can be obtained by

$$\kappa_i^{(k-1)} = \frac{\kappa_i^{(k)} + \kappa_k^{(k)} \kappa_{k-i}^{(k)}}{1 - \left(\kappa_k^{(k)}\right)^2} \tag{37}$$

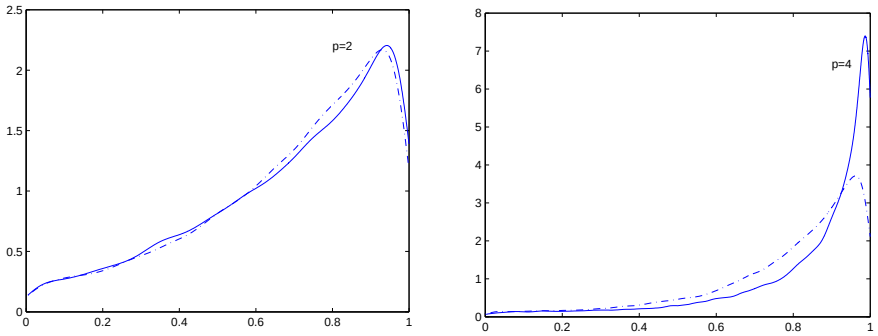
As Jones (1987) showed, if ρ follows a uniform distribution in the stationary region, PAC will be related to a beta distribution as follows,

$$\frac{\pi_k + 1}{2} \quad \sim \quad i.i.d. Beta\left([\frac{1}{2}(k+1)], [\frac{1}{2}k] + 1\right) \tag{38}$$

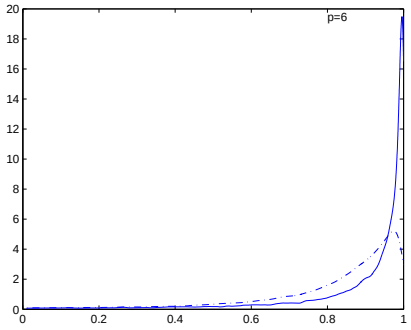
where $[x]$ denotes the integer part of x. Moreover, for the AR model to be stationary, the absolute values of all its partial autocorrelations must be less than 1. A more formal proof can be found in Ramsey (1974). It is also possible to adopt a uniform prior for the PAC instead, see Philippe (2006). However, through simulations we find that these two priors are very different. The second prior has a higher tendency to choose the models bordering the unit root circle as the lag order increases. Results are shown in Figure 1[15], We can see that as the number of lags increases, the moduli of the characteristic roots[16] from the AR model under the second prior tends more to be close to 1. Here we do

[15] Here and in the subsequent sections, we use a nonparametric package (ksdensity.m) from MatLab® to make such plots based on the simulated draws from the corresponding distributions.

[16] The roots are obtained from the characteristic equation: $x^p - \rho_1 x^{p-1} - \cdots - \rho_p = 0$



(a) p=2 (b) p=4



(c) p=6

Figure 1: The kernel plots of the characteristic roots moduli. The dashed lines represent the case when we use uniform prior for ρ and the solid lines denote the case when we use uniform prior for PAC.

not want to assume a prori that our model is close to the unit circle. Hence we choose the uniform prior for ρ in the stationary region.

Now we can turn to the details of how to take draws of ρ from (33), which can be rewritten as,

$$\begin{aligned} p(\rho|Y, Y_0) &\propto I(\rho \in S) \exp\left\{ N \left[\tau(\rho) - \frac{T-1}{2} \ln(\rho' A \rho - 2\rho' b + c) \right] \right\} \\ &\propto I(\rho \in S) \exp\left[N \vartheta(\rho) \right] \end{aligned} \tag{39}$$

where

$$\vartheta(\rho) = \tau(\rho) - \frac{T-1}{2} \ln(\rho' A \rho - 2\rho' b + c). \tag{40}$$

Since the mode of the posterior distribution is a consistent estimator, we can expect $\vartheta(\rho)$ has a unique global maximum in the stationary region when N tends to infinite. Under certain regularity conditions, the posterior distribution will converge to a normal distribution as the sample size N increases, see Bernardo and Smith (section 5.3 1994). It is sensible to use the following truncated normal distribution to approximate the posterior:

$$\rho|Y, Y_0 \quad \overset{a}{\sim} \quad I(\rho \in S) N\left(\hat{\rho}, \frac{1}{N} \left[-\vartheta''(\hat{\rho}) \right]^{-1} \right). \tag{41}$$

where the mean of the normal distribution, i.e. $\hat{\rho}$, is the maximum of $\vartheta(\rho)$ in the stationary region, which can be estimated by Newton's method, and $\vartheta''(\hat{\rho})$ denotes the Hessian matrix evaluated at $\hat{\rho}$. Algorithm 3.2 in the following is a Metropolis-Hastings (MH) algorithm, which makes draws from (39) using (41) as the proposal distribution. We refer the reader to Chib and Greenberg (1995) for the details on the convergence of MCMC estimates. Note that the truncated normal distribution is a good approximation to the true posterior only in large sample. To take account of such scale errors, in practice when we propose a draw from (41), we could replace N in the denominator of the variance by $v \cdot N$. The value of v is at our discretion. The variance in the proposal distribution is scaled in this way such that we can sample from a wide range of the parameter space.

Algorithm 3.2. *Starting from the current value of $\rho_0 \in S$, we repeat the following steps.*

1. *We propose a draw ρ_c from (41).*
2. *We accept ρ_c as a draw from the posterior distribution (39) with the probability*

$$\alpha(\rho_0, \rho_c) = \min\left(1, \frac{\exp\left[N\vartheta(\rho_c) \right] q(\rho_0)}{\exp\left[N\vartheta(\rho_0) \right] q(\rho_c)} \right) \tag{42}$$

   where $q(\cdot)$ is the density function of the truncated normal distribution (41).
3. *If we accept ρ_c as our new draw, we replace ρ_0 with ρ_c; otherwise we keep it the same. Then we go back to step 1.*

After we obtain enough draws from the posterior distribution, we can also use the mean of the draws as our point estimator and construct the highest posterior density interval to make inference.

The above algorithm should work for most circumstances. However, there are still some issues remaining. One potential problem is that when p is large but N is small, the Newton's method may not be efficient in finding the maximum point of the posterior distribution. For such situation, we may try many initial values but they may converge to different points through the Newton's method. A possible way to tackle the problem is to have a pilot run of Algorithm 3.2 after we obtain a crude estimate of the maximum point from the Newton's method. Then we could improve the estimation by using the Newton's method again on a selection of the posterior draws, such as those with high posterior density. We can repeat such processes until we find the satisfactory global maximum point.

Another potential problem has been noticed by Lancaster (2002). When N is small for the case of one lag, the posterior density function of ρ may not have a bell shape. Figure 2 shows such a case. We can see that the maximum is not close to the true value (0.6) but on the unit circle instead. More importantly, the second order derivative of the density function at the maximum is positive, which means the truncated distribution in (41) has a negative (definite) variance. Although such situation does not always arise, it is not hard to imagine that

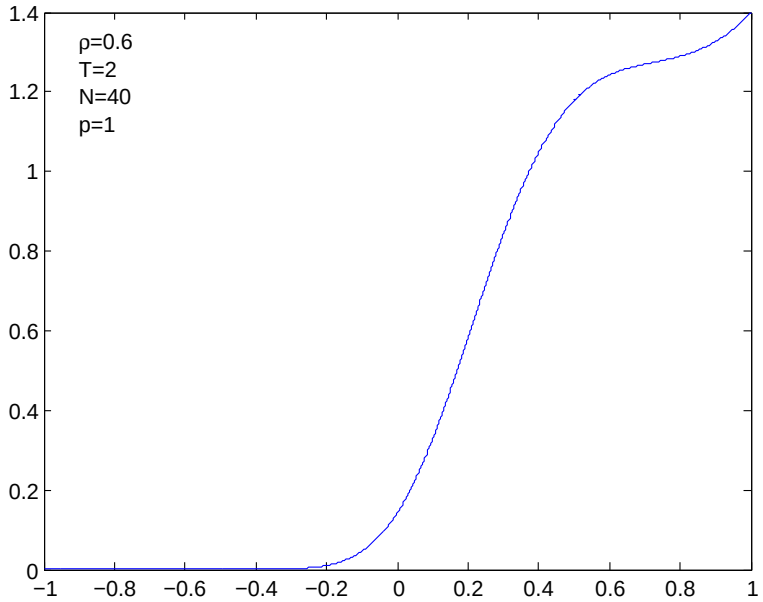


Figure 2: The plot of a non-bell shape posterior density function of ρ

when p gets larger and N is small, it could happen more often. Therefore it should be sensible for us to take precaution against such case in our algorithm. One way is to replace the negative definite variance matrix in (41) by a positive

definite variance matrix, such as $\frac{1}{N(T-1)-p}(c-b'A^{-1}b)A^{-1}$ in (33). Again, we can multiply the variance matrix by $\frac{1}{v}$ to control the acceptance rate such that our algorithm can explore a wide range of the parameter space.

3.3 Comparison of Different Model Specifications

Li (2009) noticed that when our model is misspecified, such as the case when we include the wrong set of exogenous regressors, the solution for (24) may not enable us to obtain consistent estimate of ρ under the AR(1) panel model. Therefore Li (2009) suggested comparing different model specifications using the Bayes factor and showed certain regularity conditions under which the Bayes factor is consistent in model selection. Drawing the analogy, we also recommend comparing different model specifications here. We propose two algorithms to achieve this.

Different model specifications are defined by different lag orders (p) and the inclusion of different sets of regressors in (17). They are compared based on their posterior model probabilities. We use a K by 1 vector ix, whose elements are either 0 or 1, to denote the exclusion or the inclusion of a particular exogenous regressor. If we denote the maximum AR order by P,[17] the total number of models will be $(P+1)2^K$. Suppose for our dataset, there are T_{true} observations for each economic agent. Since our estimation is conditional on the first p observations, the dimension of y_i (T) in (17) and the maximum AR order (P) must satisfy $P+T=T_{true}$. When we compare different model specifications, T does not change for different models. The posterior model probability of model i is defined as

$$\begin{aligned} p\left(M_i|Y,Y_0\right) &= \frac{p\left(M_i\right)p\left(Y|Y_0,M_i\right)}{p\left(Y|Y_0\right)} \\ &= \frac{p\left(M_i\right)p\left(Y|Y_0,M_i\right)}{\sum_{j=1}^{(P+1)2^K} p\left(M_j\right)p\left(Y|Y_0,M_j\right)}. \end{aligned} \tag{43}$$

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, i.e.

$$\begin{aligned} & p\left(Y|Y_0,M_i\right) \\ &= \int p(g,\theta|Y_0,M_i)p(Y|g,\theta,Y_0,M_i)dg\,d\theta \\ &= \int_{\rho\in S} p(\rho|Y_0,M_i)p(Y|\rho,Y_0,M_i)d\rho \end{aligned} \tag{44}$$

[17] In the case of $p=0$, we define $\tau(\rho)=0$, $A=0$ and $b=0$. When ix is a vector of zeros, we have $A=\sum_{i=1}^{N} Y_{i\_}'HY_{i\_}$, $b=\sum_{i=1}^{N} Y_{i\_}'Hy_i$ and $c=\sum_{i=1}^{N} y_i'Hy_i$.

Therefore the comparison of two different models depends on the Bayes factor, $\frac{p(Y|Y_0,M_i)}{p(Y|Y_0,M_j)}$.

If the number of models under consideration is not large, we can calculate the marginal likelihood for all of them. The method due to Chib and Jeliazkov (2001) can help us in this regard. Recall that the marginal likelihood for model i can be also calculated as,

$$p(M_i|Y,Y_0) = \frac{p(\rho^*|M_i)p(Y|\rho^*,Y_0,M_i)}{p(\rho^*|M_i,Y,Y_0)} \tag{45}$$

For ρ^* we can choose arbitrary value in the stationary region, but for estimation efficiency, the estimated mode of ρ from (41) is preferred. According to Chib and Jeliazkov (2001), $p(\rho|M_i,Y,Y_0)$ can be estimated by

$$\hat{p}(\rho|M_i,Y,Y_0) = \frac{K^{-1}\sum_{k=1}^{K}\alpha(\rho^{(k)},\rho^*)q(\rho^{(k)},\rho^*)}{J^{-1}\sum_{j=1}^{J}\alpha(\rho^*,\rho^{(j)})} \tag{46}$$

As in Algorithm 3.2 before, $\alpha(\rho^*,\rho^{(j)})$ and $q(\rho^*,\rho^{(j)})$ respectively stand for the acceptance probability and the proposal density function moving from ρ^* to $\rho^{(j)}$ in the Markov chain.[18] In addition to that, $\{\rho^{(k)}\}$ are the sample draws from the posterior distribution and $\{\rho^{(j)}\}$ are the draws from $q(\rho^*,\rho^{(j)})$ (the proposal density). Carlin and Luis (2000) recommend the Chib's method for calculating the marginal likelihood since it is safe and relatively easy to implement. For our algorithm, we find that the estimates of the marginal likelihood are quite stable once we set up the proposal density appropriately. However, the Chib's method can evaluate only one model each time we use it. When the number of models under consideration is huge, it is computationally prohibitive to evaluate all the models. Next we propose the reversible jump algorithm (Algorithm 3.3) which samples the parameter space and the model space at the same time.

Algorithm 3.3. *Starting from the current status* $(p^{(0)}, ix^{(0)}, \rho^{(0)})$*, we repeat the following steps.*

1. *From* $p^{(0)}$ *and* $ix^{(0)}$*, we propose* $p^{(c)}$ *and* $ix^{(c)}$*. The details of the proposal will be discussed later.*

2. *Depending on the values of* $p^{(c)}$ *and* $ix^{(c)}$*, we propose* $\rho^{(c)}$ *and calculate the acceptance probability according to the following:*

   - *If* $p^{(c)} > p^{(0)}$*, we first use (37) to transform* $\rho^{(0)}$ *into* $\pi^{(0)}$ *and then draw a* $(p^{(c)} - p^{(0)}) \times 1$ *vector* u*, whose elements follow i.i.d.*$U(-1,1)$*. Finally* $\rho^{(c)}$ *is obtained by transforming* $(\pi^{(0)}, u)'$ *through (35). The acceptance probability is calculated as*

[18] In our context, $q(\rho^*,\rho^{(j)}) = q(\rho^{(j)})$.

$$\min\Bigg(1, \left(\frac{\eta}{\eta+1}\right)^{(k^{(c)}-k^{(0)})} \cdot \frac{m\left(S^{(0)}\right)\exp\left[N\vartheta(\rho^{(c)}|ix^{(c)})\right]q(c,0)}{m\left(S^{(c)}\right)\exp\left[N\vartheta(\rho^{(0)}|ix^{(0)})\right]2^{p^{(0)}-p^{(c)}}q(0,c)}\left|\frac{\partial\rho^{(c)}}{\partial(\rho^{(0)\prime},u')}\right|\Bigg), \tag{47}$$

where $\vartheta(\cdot)$ is defined in (40). $q(x,y)$ denotes the probability of jumping to model y given that the chain is now at model x and $\left|\frac{\partial\rho^{(c)}}{\partial(\rho^{(0)\prime},u')}\right|$ is the Jacobian from $(\rho^{(0)\prime},u')$ to $\rho^{(c)}$. We can calculate the Jacobian as

$$\left|\frac{\partial\rho^{(c)}}{\partial(\rho^{(0)\prime},u')}\right| = \prod_{i=1}^{p^{(c)}-p^{(0)}}(1+u_i)^{[\frac{p^{(0)}+i-1}{2}]}(1-u_i)^{[\frac{p^{(0)}+i}{2}]}, \tag{48}$$

where $[x]$ denotes the integer part of x. (See the appendix for the proof.)

- *If $p^{(0)} > p^{(c)}$, we first transform $\rho^{(0)}$ to $\pi^{(0)}$ and $\rho^{(c)}$ is obtained from transforming $(\pi_1^{(0)},\ldots,\pi_{p^{(c)}}^{(0)})$. The acceptance probability is calculated as*

$$\min\Bigg(1, \left(\frac{\eta}{\eta+1}\right)^{(k^{(c)}-k^{(0)})}\frac{m\left(S^{(0)}\right)\exp\left[N\vartheta(\rho^{(c)}|ix^{(c)})\right]2^{p^{(c)}-p^{(0)}}q(c,0)}{m\left(S^{(c)}\right)\exp\left[N\vartheta(\rho^{(0)}|ix^{(0)})\right]q(0,c)}\cdot \left|\frac{\partial\rho^{(0)}}{\partial(\rho^{(c)\prime},\pi_{p^{(c)}+1}^{(0)},\ldots,\pi_{p^{(0)}}^{(0)})}\right|^{-1}\Bigg). \tag{49}$$

where the Jacobian takes the following form

$$\left|\frac{\partial\rho^{(0)}}{\partial(\rho^{(c)\prime},\pi_{p^{(c)}+1}^{(0)},\ldots,\pi_{p^{(0)}}^{(0)})}\right| = \prod_{i=p^{(c)}+1}^{p^{(0)}}(1+\pi_i^{(0)})^{[\frac{i-1}{2}]}(1-\pi_i^{(0)})^{[\frac{i}{2}]}. \tag{50}$$

- *If the values of $p^{(0)}$ and $p^{(c)}$ are the same, then we deliver $\rho^{(c)}=\rho^{(0)}$ and the acceptance probability is calculated from*

$$\min\left(1, \frac{\exp\left[N\vartheta(\rho^{(0)}|ix^{(c)})\right]}{\exp\left[N\vartheta(\rho^{(0)}|ix^{(0)})\right]}\right). \tag{51}$$

3. *If we accept $\rho^{(c)}$ as our new draw, we also replace $p^{(0)}$ and $ix^{(0)}$ with $p^{(c)}$ and $ix^{(c)}$. If we reject the proposed model and the parameter value, we use Algorithm 3.2 to update $\rho^{(0)}$ under the old model. Then we go back to step 1.*

The reversible jump algorithm, first proposed by Green (1995), can be seen as an extension of the MH algorithm when the dimension of the parameter space under consideration varies in the Markov chain. The rationale behind the updating scheme of ρ in step 2 is that when we increase (reduce) the dimension of ρ, we at the same time increase (reduce) the dimension of the PAC (π) in the model. The way of updating in step 2 means when we increase the dimension of ρ, we deliver $(\pi, u)'$ as our new PAC; for the dimension reduction, we deliver $(\pi_1, \ldots, \pi_{p^{(c)}})$ as our new PAC.

Now we go back to discuss how we propose to change the parameter dimension, i.e., how we propose $p^{(c)}$ and $ix^{(c)}$ in step 1 of Algorithm 3.3 above. The bottom line here is that we want our algorithm to move quickly enough to sample the model space (especially when it is large) and to overcome the problem of multi-modes. Similar practices can be seen in Ehlers and Brooks (2002). We propose $p^{(c)}$ and $ix^{(c)}$ independently. To propose $p^{(c)}$, we use the discretized Laplacian distribution so that the density for $p^{(c)}$ conditional on $p^{(0)}$ ($q(p^{(0)}, p^{(c)})$) is given by

$$q(p^{(0)}, p^{(c)}) \propto \exp\left(-\varsigma|p^{(c)} - p^{(0)}|\right), \quad p^{(c)}, p^{(0)} \in [1, \ldots, P], \tag{52}$$

where $p^{(0)}$ stands for the current value of p and $\varsigma \geq 0$ denotes a scale parameter. For $\varsigma = 0$, the proposal is a uniform distribution not depending on the current status of the chain, while for bigger values of ς, the models further away from $p^{(0)}$ are less likely to be proposed.

As for ix, we wish that it should change more often since the potential number of regressors is generally large. We may like every proposed model to be different from the old model. A simple way to achieve this is to first use a truncated binomial distribution[19] to generate the number of elements in ix to be changed. Then we draw the elements uniformly without replacement. For the selected elements, we change them to 1 (0) if they are originally 0 (1). Let us denote the number of elements to be changed by k and it has the probability function $q(k)$,

$$q(k) = \binom{K}{k} \gamma^k (1-\gamma)^{K-k} \left(1 - (1-\gamma)^K\right)^{-1} \tag{53}$$

where $\gamma \in (0, 1)$ is the scale parameter. Taking $\gamma = \frac{1}{2}$, we have the uniform distribution for all the potential models under consideration. For small values of γ, we prefer small changes while for big values of γ, we prefer big changes.

Through the study of the simulated dataset later, we find that the results obtained through our reversible jump algorithm are quite similar to the results

[19] We do not include 0 in the support for the proposal.

from the Chib's method, although the reversible jump may sometimes have difficulty in separating two models with close posterior probabilities.

3.4 Demonstration Examples for Estimation

In this section, we use simulated data to demonstrate the performance of our methods developed above. We want to show our methods can still work for a rather difficult case.

First we use the techniques in Section 3.2 to estimate a model with three lags and no exogenous regressors. Suppose there are T_{true} observations for each economic agent in our panel. Recall that P (the maximum lag) and T (the observations we use for estimation) must satisfy $T + P = T_{true}$. The lowest value for T is 2 according to Table 1. In the simulated dataset, we first set $T_{true} = 5$ and set $\sigma^2 = 1$, $\rho_1 = -1.1718$, $\rho_2 = 0.17399$ and $\rho_3 = 0.49181$ (Table 2). Such setting implies that the true value of ρ is near the unit circle in the stationary region. The largest modulus of the characteristic root is 0.9196, which is fairly close to 1. We estimate our model with different Ns (cross section sample sizes). The results are shown in Table 3. As we can see, for N=50 and 100, both the posterior mode and mean are very different from the true values, though the posterior mean seems to be closer than the mode. Note that the largest moduli of the characteristic roots obtained based on the posterior modes for these two cases are 0.9998 and 0.9999, which are virtually equal to 1. This should remind us of Figure 2 when the maximum point of the density function is obtained on the unit circle and the density function does not have a bell shape. In fact, evaluated at the posterior mode under N=50 and 100, the Hessian matrix of $\vartheta(\rho)$ is positive definite, which means the variance matrix of the proposal density in (41), i.e. $\frac{1}{N}\left[-\vartheta''(\hat{\rho})\right]^{-1}$, is negative definite and has to be replaced by a positive definite matrix. When N is increased to 200 and 1000, such problems disappear. The largest moduli are 0.8807 and 0.9282 respectively, which means the posterior modes for these cases are inside the stationary region. Moreover, the Hessian matrix of $\vartheta(\rho)$ is now negative definite. As for $N = 200$, the estimated mode and mean are already much closer to the true value of ρ than those for N=50 and 100. For ρ_1 and ρ_3 under $N = 1000$, our estimates look quite near to the true values. However, there is still some difference for ρ_2. We may say that when T is 2 and the true value is near the unit circle, consistency results may require huge N to achieve. When we have bigger values of T, our estimators could be dramatically improved, as will be shown later. We also put down the maximum likelihood estimates here under the header "MLE" for comparison. The MLE are much further away from the true values for all cases and none of the elements are close even for $N = 1000$.

Though point estimates could be important, sometimes we may be more interested in knowing the uncertainty surrounding our estimators. Figure 3 shows the posterior marginal density plots for ρ_1, ρ_2 and ρ_3 under different cross secion sample sizes. We can see that for $N = 50$ and 100, the marginal densities are quite skewed and show signs of non-normality. When $N = 200$, the marginal density already looks rather symmetrical. It looks more like normal distribution

Table 2: The true value of ρ in the simulation and the moduli of the characteristic roots

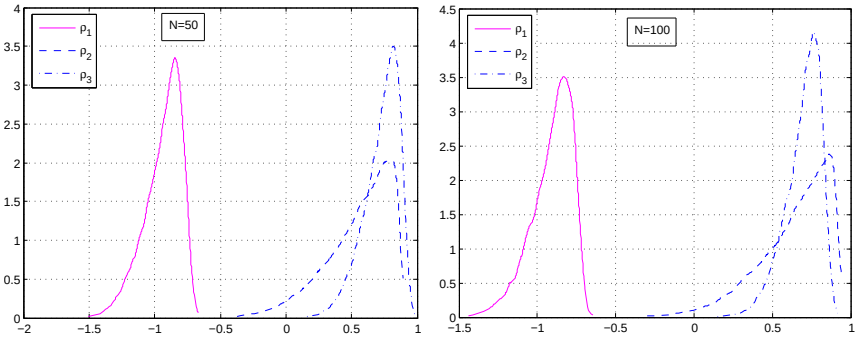
ρ_{true}	root moduli
-1.1718	0.9196
0.1740	0.9196
0.4918	0.5816

Table 3: Point Estimation Results for $T = 2$

$N = 50$				$N = 100$			
mode	root moduli	mean	MLE	mode	root moduli	mean	MLE
-0.7657	0.9998	-0.94	-2.157	-0.758	0.9999	-0.91	-2.145
0.8687	0.9469	0.55	-1.594	0.9203	0.9152	0.636	-1.548
0.8965	0.9469	0.73	-0.38	0.8376	0.9152	0.695	-0.325
$N = 200$				$N = 1000$			
mode	root moduli	mean	MLE	mode	root moduli	mean	MLE
-1.28	0.8807	-1.24	-1.942	-1.27	0.9282	-1.26	-1.943
-0.07	0.8807	-0.02	-1.217	0.03	0.9282	0.04	-1.191
0.32	0.4182	0.35	-0.227	0.44	0.5050	0.44	-0.176

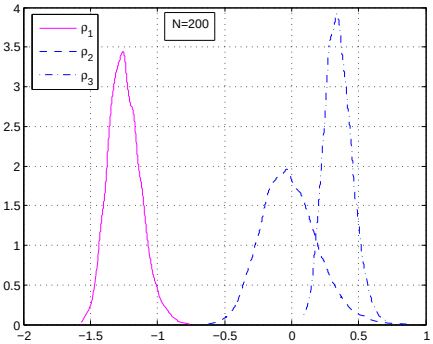
under $N = 1000$. Table 4 shows the highest posterior density intervals (HPDI) of the marginal distributions and the confidence intervals based on the MLE. Under $N = 50$, though the posterior means and the modes are very different from the true values of ρ, there is a high degree of uncertainty surrounding our estimators. As we can see, the posterior distributions have very long tails. The true values of ρ are within the 99% and 95% HPDI, and they are near the border of the 90% HPDI. When N equals 100, the situation is similar, though our point estimates are better than those under $N = 50$. As the sample size increases, the posterior distributions get more symmetrical. When $N = 200$, we start to see that not only can we get better point estimates, we can also have better interval coverage. The HPDIs become narrower with the true values inside as the cross section sample size increases. However, as for the MLE confidence intervals, the true values are far away from the intervals for any cross section sample size, which implies such intervals based on biased estimates could be very misleading.

Now we increase T and repeat the experiment above. As far as the MLE is concern, again, the estimates are poor even for $T = 10$ and $N = 1000$, which have still quite a distance from our true values. Although the confidence intervals are closer to the true values, none of them can have the true values inside for different cross section sample sizes. As for our correction function method, under $T = 4$, even for $N = 50$, the mode of the posterior distribution for ρ is no longer on the unit circle as before and the marginal distributions are all quite symmetrical. Though the posterior mode and the mean are still fairly different from the true values, compared to the case of $T = 2$, they already

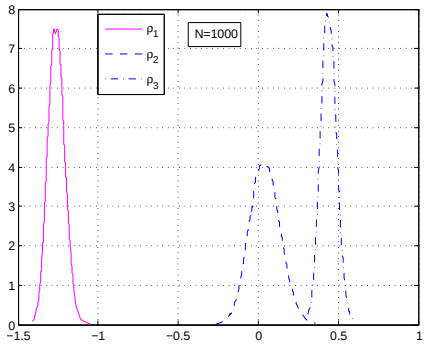


(a) N=50

(b) N=100



(c) N=200



(d) N=1000

Figure 3: The marginal density plots of the posterior draws of ρ for $T = 2$

Table 4: HPDI and Confidence Intervals for $T = 2$

$N = 50$	ρ_1 HPDI		ρ_1 MLE		ρ_2 HPDI		ρ_2 MLE		ρ_3 HPDI		ρ_3 MLE	
99%	-1.363	-0.679	-2.415	-1.9	-0.201	0.890	-2.07	-1.12	0.328	0.966	-0.64	-0.12
95%	-1.251	-0.711	-2.35	-1.96	0.046	0.890	-1.95	-1.24	0.443	0.942	-0.58	-0.18
90%	-1.173	-0.725	-2.32	-1.99	0.179	0.890	-1.89	-1.29	0.510	0.926	-0.55	-0.21
80%	-1.083	-0.740	-2.285	-2.029	0.358	0.875	-1.83	-1.36	0.588	0.909	-0.501	-0.25
N=100	ρ_1 HPDI		ρ_1 MLE		ρ_2 HPDI		ρ_2 MLE		ρ_3 HPDI		ρ_3 MLE	
99%	-1.293	-0.691	-2.33	-1.96	-0.070	0.938	-1.88	-1.21	0.348	0.896	-0.5	-0.149
95%	-1.179	-0.703	-2.28	-2.01	0.187	0.938	-1.8	-1.29	0.464	0.878	-0.458	-0.19
90%	-1.118	-0.710	-2.26	-2.03	0.324	0.938	-1.76	-1.33	0.528	0.862	-0.437	-0.212
80%	-1.037	-0.729	-2.23	-2.06	0.456	0.930	-1.71	-1.38	0.589	0.846	-0.412	-0.237
N=200	ρ_1 HPDI		ρ_1 MLE		ρ_2 HPDI		ρ_2 MLE		ρ_3 HPDI		ρ_3 MLE	
99%	-1.512	-0.917	-2.06	-1.82	-0.487	0.548	-1.44	-0.997	0.113	0.628	-0.346	-0.11
95%	-1.464	-1.010	-2.03	-1.85	-0.412	0.406	-1.38	-1.05	0.156	0.559	-0.32	-0.14
90%	-1.443	-1.055	-2.02	-1.86	-0.362	0.321	-1.36	-1.08	0.177	0.514	-0.303	-0.15
80%	-1.403	-1.100	-2.002	-1.88	-0.292	0.222	-1.33	-1.11	0.210	0.466	-0.29	-0.17
N=1000	ρ_1 HPDI		ρ_1 MLE		ρ_2 HPDI		ρ_2 MLE		ρ_3 HPDI		ρ_3 MLE	
99%	-1.386	-1.135	-1.998	-1.889	-0.189	0.302	-1.29	-1.09	0.323	0.565	-0.23	-0.12
95%	-1.356	-1.160	-1.985	-1.9	-0.140	0.233	-1.27	-1.11	0.347	0.535	-0.22	-0.13
90%	-1.344	-1.178	-1.98	-1.908	-0.111	0.193	-1.26	-1.13	0.359	0.517	-0.21	-0.14
80%	-1.328	-1.197	-1.97	-1.92	-0.084	0.161	-1.24	-1.14	0.375	0.502	-0.2	-0.148

get much closer[20]. The interesting thing to note is that although we have better point estimates under $T = 4$, the coverage of the posterior marginal distributions does not seem to be as good as for $T = 2$. The true values of ρ are quite often outside even the 99% intervals.[21] The situation only starts to improve for $N = 200$. When N gets to 1000, the true values of ρ are fairly well within (or bordering) the HPDIs, which are the signs of estimation consistency. For $T = 10$, all results appear to be much nicer. Both the posterior mode and the mean are already quite near the true values even for $N = 50$. As for the posterior marginal distribution coverage, the true values are quite near the center of the marginal distributions. This strongly confirms the viability of our correction function method under the linear short panel context.

Table 5: Point Estimation Results for $T = 4$

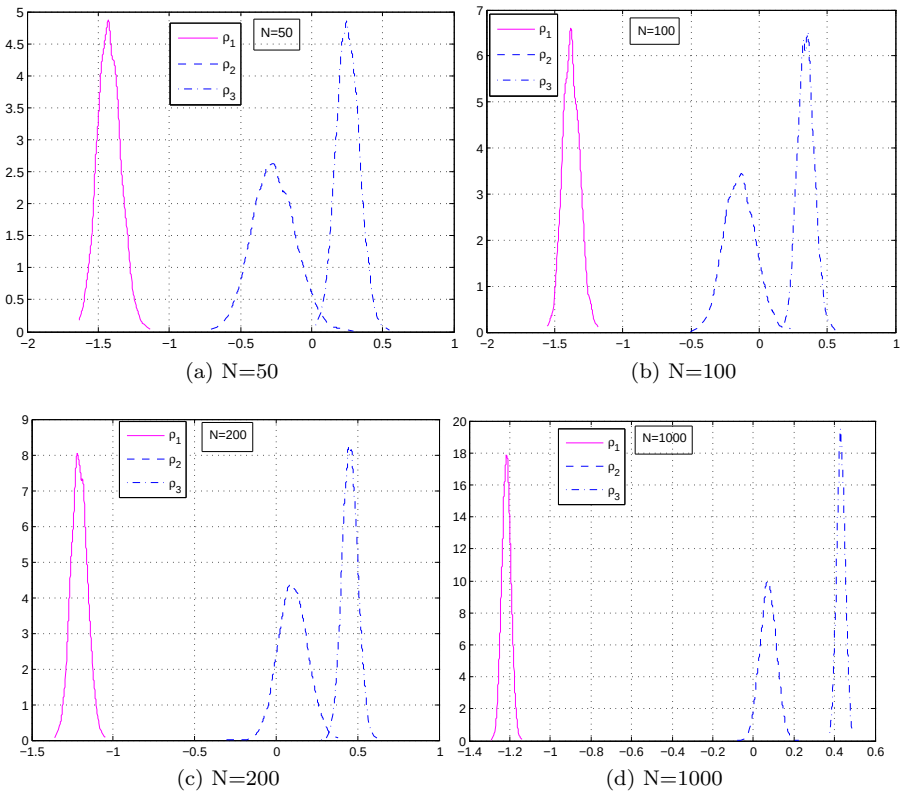
$N = 50$				$N = 100$			
mode	root moduli	mean	MLE	mode	root moduli	mean	MLE
-1.4303	0.9009	-1.4226	-1.7435	-1.3814	0.9253	-1.377	-1.712
-0.27864	0.9009	-0.2641	-0.8622	-0.14912	0.9253	-0.14	-0.7667
0.24896	0.3068	0.2563	-0.0491	0.34034	0.3975	0.3452	0.0239
$N = 200$				$N = 1000$			
mode	root moduli	mean	MLE	mode	root moduli	mean	MLE
-1.2123	0.9166	-1.2077	-1.6178	-1.2173	0.9111	-1.2164	-1.628
0.095423	0.9166	0.1039	-0.6581	0.072174	0.9111	0.0739	-0.665
0.44979	0.5354	0.4542	0.0578	0.43123	0.5195	0.4321	0.007

Table 6: HPDI and Confidence Intervals for $T = 4$

$N = 50$	ρ_1 HPDI		ρ_1 MLE		ρ_2 HPDI		ρ_2 MLE		ρ_3 HPDI		ρ_3 MLE	
99%	-1.629	-1.213	-1.92	-1.57	-0.657	0.140	-1.18	-0.55	0.066	0.459	-0.23	0.13
95%	-1.585	-1.259	-1.88	-1.61	-0.558	0.040	-1.1	-0.62	0.101	0.415	-0.18	0.085
90%	-1.556	-1.286	-1.86	-1.63	-0.512	-0.014	-1.06	-0.66	0.120	0.394	-0.16	0.064
80%	-1.530	-1.323	-1.83	-1.66	-0.463	-0.073	-1.02	-0.7	0.148	0.358	-0.137	0.039
$N = 100$	ρ_1 HPDI		ρ_1 MLE		ρ_2 HPDI		ρ_2 MLE		ρ_3 HPDI		ρ_3 MLE	
99%	-1.528	-1.213	-1.84	-1.59	-0.442	0.164	-0.99	-0.54	0.187	0.528	-0.1	0.15
95%	-1.498	-1.244	-1.81	-1.62	-0.358	0.093	-0.94	-0.596	0.226	0.472	-0.07	0.12
90%	-1.482	-1.273	-1.79	-1.63	-0.328	0.047	-0.91	-0.62	0.243	0.444	-0.056	0.1
80%	-1.456	-1.298	-1.77	-1.65	-0.289	-0.004	-0.88	-0.66	0.262	0.423	-0.038	0.09
$N = 200$	ρ_1 HPDI		ρ_1 MLE		ρ_2 HPDI		ρ_2 MLE		ρ_3 HPDI		ρ_3 MLE	
99%	-1.325	-1.089	-1.71	-1.53	-0.129	0.350	-0.82	-0.5	0.335	0.579	-0.034	0.15
95%	-1.301	-1.114	-1.69	-1.55	-0.076	0.287	-0.78	-0.54	0.366	0.548	-0.01	0.13
90%	-1.282	-1.133	-1.67	-1.56	-0.048	0.260	-0.76	-0.55	0.377	0.530	-0.001	0.12
80%	-1.270	-1.149	-1.66	-1.57	-0.018	0.219	-0.74	-0.58	0.395	0.511	0.012	0.1
$N = 1000$	ρ_1 HPDI		ρ_1 MLE		ρ_2 HPDI		ρ_2 MLE		ρ_3 HPDI		ρ_3 MLE	
99%	-1.275	-1.157	-1.67	-1.588	-0.029	0.180	-0.74	-0.59	0.381	0.483	0.029	0.11
95%	-1.260	-1.172	-1.66	-1.597	-0.004	0.152	-0.72	-0.61	0.393	0.471	0.039	0.1
90%	-1.254	-1.180	-1.65	-1.6	0.006	0.140	-0.71	-0.618	0.398	0.466	0.044	0.096
80%	-1.245	-1.189	-1.648	-1.61	0.021	0.126	-0.7	0.63	0.407	0.458	0.05	0.09

[20]The L^2 distance between the mode and the true value for $T = 2$ and $N = 50$ is 0.9, while for $T = 4$ and $N = 50$, it is 0.575.

[21]Note that these results are based on two particular datasets. If we want to investigate the HPDI coverage performance in more details, further simulation research needs to be carried out.



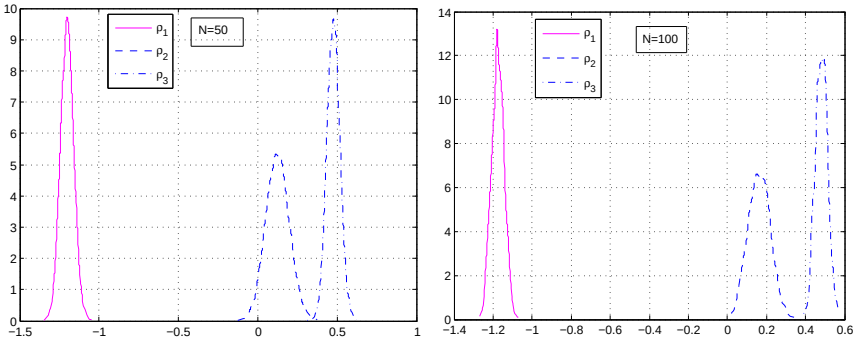
(a) N=50 (b) N=100

(c) N=200 (d) N=1000

Figure 4: The marginal density plots of the posterior draws of ρ for $T = 4$

Table 7: Point Estimation Results for $T = 10$

$N = 50$				$N = 100$			
mode	root moduli	mean	MLE	mode	root moduli	mean	MLE
-1.2032	0.9253	-1.2012	-1.362	-1.1758	0.91668	-1.1753	-1.354
0.12109	0.9253	0.1249	-0.177	0.16192	0.91668	0.1629	-0.176
0.47575	0.5556	0.4775	0.321	0.48155	0.57306	0.482	0.304
$N = 200$				$N = 1000$			
mode	root moduli	mean	MLE	mode	root moduli	mean	MLE
-1.1615	0.9086	-1.1607	-1.333	-1.1624	0.9173	-1.1624	-1.328
0.17642	0.9086	0.1777	-0.145	0.19369	0.9173	0.1938	-0.118
0.47595	0.5765	0.4765	0.31	0.49693	0.5905	0.497	0.335



(a) N=50 (b) N=100

(c) N=200 (d) N=1000

Figure 5: The marginal density plots of the posterior draws of ρ for $T = 10$

Table 8: HPDI and Confidence Intervals for $T = 10$

$N = 50$	ρ_1 HPDI		ρ_1 MLE		ρ_2 HPDI		ρ_2 MLE		ρ_3 HPDI		ρ_3 MLE	
99%	-1.311	-1.089	-1.47	-1.25	-0.066	0.310	-0.36	0.01	0.371	0.585	0.213	0.43
95%	-1.285	-1.115	-1.44	-1.28	-0.020	0.276	-0.32	-0.034	0.396	0.559	0.239	0.4
90%	-1.270	-1.133	-1.43	-1.29	0.002	0.246	-0.296	-0.057	0.409	0.545	0.25	0.39
80%	-1.256	-1.147	-1.42	-1.31	0.028	0.221	-0.27	-0.083	0.423	0.531	0.27	0.37
$N = 100$	ρ_1 HPDI		ρ_1 MLE		ρ_2 HPDI		ρ_2 MLE		ρ_3 HPDI		ρ_3 MLE	
99%	-1.256	-1.093	-1.433	-1.27	0.022	0.302	-0.31	-0.038	0.406	0.559	0.22	0.39
95%	-1.239	-1.114	-1.414	-1.29	0.051	0.268	-0.28	-0.071	0.424	0.542	0.24	0.37
90%	-1.230	-1.121	-1.4	-1.3	0.067	0.255	-0.26	-0.087	0.428	0.530	0.25	0.36
80%	-1.220	-1.133	-1.39	-1.31	0.086	0.235	-0.24	-0.11	0.440	0.521	0.26	0.34
$N = 200$	ρ_1 HPDI		ρ_1 MLE		ρ_2 HPDI		ρ_2 MLE		ρ_3 HPDI		ρ_3 MLE	
99%	-1.219	-1.105	-1.39	-1.28	0.067	0.282	-0.24	-0.05	0.424	0.532	0.26	0.37
95%	-1.204	-1.119	-1.38	-1.29	0.098	0.261	-0.22	-0.07	0.435	0.519	0.27	0.35
90%	-1.197	-1.125	-1.37	-1.3	0.112	0.244	-0.21	-0.08	0.442	0.511	0.28	0.345
80%	-1.189	-1.133	-1.36	-1.31	0.125	0.229	-0.19	-0.1	0.450	0.502	0.28	0.34
$N = 1000$	ρ_1 HPDI		ρ_1 MLE		ρ_2 HPDI		ρ_2 MLE		ρ_3 HPDI		ρ_3 MLE	
99%	-1.185	-1.138	-1.35	-1.3	0.153	0.237	-0.16	-0.077	0.475	0.521	0.31	0.36
95%	-1.180	-1.145	-1.346	-1.31	0.161	0.227	-0.15	-0.087	0.480	0.514	0.316	0.353
90%	-1.178	-1.148	-1.343	-1.313	0.166	0.221	-0.145	-0.092	0.483	0.511	0.319	0.35
80%	-1.174	-1.151	-1.34	-1.316	0.172	0.214	-0.139	-0.098	0.485	0.508	0.32	0.347

3.5 Demonstration Examples for Model Comparison

In this section, we show how well the algorithms developed in Section 3.3 work in some examples. As in the previous section, we also set the true values of ρ as $(-1.1718, 0.17399, 0.49181)'$, which indicates the model is fairly near the unit circle. For T, it is set to 4. We then include some exogenous regressors out of a group of potential regressors in our model. Similar to Li (2009)[22], we generate serially and cross-sectionally correlated exogenous regressors such that when we include the wrong set of regressors, the correction function is generally not a valid solution for the incidental parameter problem. We set the number of potential regressors to 6 and the maximum possible AR order to 3. Therefore the total number of models considered will be $(3+1)2^6 = 256$. For such scale of model space, both the Chib's method and the reversible jump are applicable for calculating the posterior model probabilities, though care should be taken in fine-tuning some parameter settings for the reversible jump method.

Table 9 shows the posterior model probabilities of the top models. The results from the Chib's method and the reversible jump are quite close. Most of the model rankings are the same, though some discrepancies exist for the posterior model probabilities. Such discrepancies may become more conspicuous for the models with low model probabilities, which, however, can be seen as unimportant for our analysis. As the cross section sample size becomes larger, the posterior model probability will concentrate more on the top models. Although for $N = 50$, the model with the highest posterior model probability is not the true model (the one with 3 lags and regressor 1,3,4 and 6). For bigger sample sizes, the top posterior model probability criterion successfully picks up the true model. This is the evidence supporting that our correction function method may not only lead to consistency in estimation, but also consistency in model selection.

In addition to calculating the posterior model probabilities, we use Bayesian model averaging (BMA) to estimate the coefficients for the exogenous regressors unconditional on any particular model (see Fernandez et al., 2001). We use the inclusion probability to measure the significance of each exogenous regressor[23]. Since we assume that all models are a priori equally probable, it is virtually equivalent as saying that the prior probability to include a particular regressor is 50%. If the posterior inclusion probability is above 50%, it could be interpreted as a sign that our data support or reinforce our prior and the exogenous regressor is significant. Since the posterior model probabilities based on the Chib's method and the reversible jump are quite close, we can use either of them for BMA. Table 10 shows the BMA results based on the reversible jump method, where the column under β shows the true values of the coefficients for the regressors included. The true model has regressor 1, 3, 4 and 6 included. The column under "inclp" is the inclusion probability obtained from the reversible jump method, while the column "inclpC" is calculated based on

[22] See Appendix for the details of the data generating process.

[23] It is the sum of posterior model probabilities of all the models with the exogenous regressor included.

Table 9: The top models for $T = 4$ (true model indicated by "R")

$N = 50$	Chib's Method		Reversible Jump		
Ranking	Model	Post Prob	Ranking	Model	Post Prob
1	1,4,6,$p = 3$	0.2433	1	1,4,6,p=3	0.23707
2	4,5,6,$p = 3$	0.19001	2	4,5,6,p=3	0.18852
3	1,2,4,6,$p = 3$	0.15484	3	1,2,4,6,p=3	0.14336
4	4,6,$p = 3$	0.076556	4	4,6,p=3	0.07719
5	3,4,5,6,$p = 3$	0.066607	5	3,4,5,6,p=3	0.07282
6(R)	1,3,4,6,$p = 3$	0.052044	6(R)	1,3,4,6,p=3	0.05234
7	2,4,5,6,$p = 3$	0.046289	7	2,4,5,6,p=3	0.0408
8	1,4,5,6,$p = 3$	0.036396	8	1,4,5,6,p=3	0.03418
9	1,2,3,4,6,p=3	0.030323	9	1,2,3,4,6,p=3	0.03318
10	1,2,4,5,6,p=3	0.025478	10	1,2,4,5,6,p=3	0.02379
$N = 200$	Chib's Method		Reversible Jump		
Ranking	Model	Post Prob	Ranking	Model	Post Prob
1(R)	1,3,4,6,p=3	0.52507	1(R)	1,3,4,6,p=3	0.54598
2	3,4,5,6,p=3	0.16266	2	3,4,5,6,p=3	0.15571
3	3,4,6,p=3	0.15322	3	3,4,6,p=3	0.14305
4	1,3,4,5,6,p=3	0.049876	4	1,3,4,5,6,p=3	0.0482
5	1,2,3,4,6,p=3	0.038934	5	1,2,3,4,6,p=3	0.03768
6	2,3,4,5,6,p=3	0.033994	6	2,3,4,5,6,p=3	0.03438
7	2,3,4,6,p=3	0.032771	7	2,3,4,6,p=3	0.0312
8	1,2,3,4,5,6,p=3	0.003463	8	1,2,3,4,5,6,p=3	0.0038
9	1,2,4,6,p=3	3.46E-06	9	1,2,3,6,p=3	0
10	4,5,6,p=3	1.10E-06	10	1,2,6,p=3	0
$N = 1000$	Chib's Method		Reversible Jump		
Ranking	Model	Post Prob	Ranking	Model	Post Prob
1(R)	1,3,4,6,p=3	0.8647	1(R)	1,3,4,6,p=3	0.88282
2	1,2,3,4,6,p=3	0.058693	2	1,2,3,4,6,p=3	0.04998
3	1,2,3,4,5,6,p=3	0.046229	3	1,2,3,4,5,6,p=3	0.04087
4	1,3,4,5,6,p=3	0.030118	4	1,3,4,5,6,p=3	0.02621
5	2,3,4,5,6,p=3	0.000183	5	2,3,4,5,6,p=3	7.00E-05
6	3,4,5,6,p=3	8.26E-05	6	3,4,5,6,p=3	5.00E-05
7	2,3,4,6,p=3	2.84E-10	7	3,4,6,p=1	0
8	1,2,4,5,6,p=3	4.01E-12	8	4,6,p=1	0
9	3,4,6,p=3	5.45E-17	9	2,3,4,6,p=1	0
10	1,2,4,6,p=3	9.90E-44	10	2,3,4,p=1	0

the Chib's method. Both the Chib's method and the reversible jump give us similar estimates. The coefficients of the true regressors, except regressor 3, all have inclusion probabilities higher than 50% under $N = 50$. When the cross section sample size increases, the true regressors will have higher inclusion probability; while for wrong regressors, the inclusion probabilities tend to decrease. For $N = 1000$, the BMA estimates are nearly equal to the true values of β. We can conclude that our method can not only achieve consistent estimates for ρ, but is also consistent for β.

Table 10: The BMA estimates for the exogenous regressors

N=50					
β	mean	nse	std	inclp	inclpC
0.1	0.095	0.000	0.129	0.567	0.586
0	-0.033	0.000	0.066	0.309	0.323
0.2	0.040	0.000	0.118	0.224	0.195
0.8	0.688	0.001	0.211	0.972	0.973
0	-0.034	0.000	0.112	0.422	0.427
1.6	1.504	0.000	0.129	1	1
N=200					
β	mean	nse	std	inclp	inclpC
0.1	0.063	0.000	0.057	0.636	0.617
0	-0.002	0.000	0.011	0.107	0.109
0.2	0.176	0.000	0.032	1	1
0.8	0.813	0.000	0.047	1	1
0	0.015	0.000	0.031	0.242	0.250
1.6	1.646	0.000	0.030	1	1
N=1000					
β	mean	nse	std	inclp	inclpC
0.1	0.105	0.000	0.038	1.000	1.000
0	-0.008	0.000	0.033	0.091	0.105
0.2	0.200	0.000	0.016	1	1
0.8	0.802	0.000	0.017	1	1
0	-0.004	0.000	0.022	0.067	0.077
1.6	1.598	0.000	0.022	1	1

Next we enlarge our model space by setting the potential regressors to 16 and choose 8 to include in the data generating process. Now there are 262144 models altogether. If we use the Chib's method to calculate the model probability for each model, it will take a mainstream PC $7-9$ days to run uninterruptedly to finish, which is rather impractical. The reversible jump is the only alternative, which only takes 1089 seconds for 20,000 draws. The point estimation results are shown in Table 11, which are quite good. All the true regressors have inclusion probabilities higher than 50% under $N = 50$ while the highest inclusion probabilities for the wrong regressors are below 40%. In terms of point estimates,

it appears to be better than the previous example with 6 potential regressors. Table 12 confirms the high level of model uncertainty when we enlarge the model space. The top twenty models only account for around 64% of posterior model probability compared to 92% taken up by the top ten models in the previous case under $N = 50$. However, the good thing for the true model with more exogenous regressors is that the true model has much higher model probability than any other potential models. This can again be viewed as signs of consistency in model selection.

Table 11: The BMA estimates for the exogenous regressors with a large model space

N=50				
β	**mean**	**nse**	**std**	**inclp**
0.1	0.0902	0.0016	0.0713	0.7710
0.2	0.1035	0.0024	0.1063	0.5960
0	0.0709	0.0025	0.1108	0.3845
0	-0.0009	0.0005	0.0240	0.0825
0	0.0016	0.0009	0.0410	0.1835
0.3	0.2637	0.0019	0.0837	0.9735
0.8	0.8538	0.0014	0.0618	1.0000
0.9	0.9308	0.0015	0.0681	1.0000
0	-0.0212	0.0011	0.0477	0.2830
1	1.0464	0.0015	0.0671	1.0000
0	0.0457	0.0020	0.0880	0.3335
0	-0.0013	0.0006	0.0266	0.1815
1.5	1.3995	0.0022	0.0967	1.0000
1.6	1.5485	0.0016	0.0719	1.0000
0	0.0264	0.0015	0.0670	0.2185
0	-0.0086	0.0010	0.0436	0.1445

4 Conclusion

In this paper, we propose a strategy to solve the incidental parameter problem. It involves finding the Jacobian from the incidental parameters, which are not information orthongonal to the common parameters, to the information orthogonal incidental parameters. The strategy is implemented under the original parameterization. No reparameterization of the incidental parameters is required. The strategy is demonstrated under a simple Poisson count model. We also extend our strategy to the case when information orthogonalization of the incidental parameters is not possible, such as the linear AR(p) panel model with fixed effect. We show that there exists a correction function to solve the incidental parameter problem for the model. It could be a function of the common parameters under concern and it does not necessarily depend on the dependent

Table 12: The top models for $T = 4$ with a large model space (true model indicated by "R")

$N = 50$	Reversible Jump	
Ranking	Model	Posterior Prob
1(R)	1,2,6,7,8,10,13,14,p=3	0.182
2	1,2,3,6,7,8,10,13,14,p=3	0.04
3	1,2,6,7,8,10,11,13,14,p=3	0.035
4	1,3,6,7,8,9,10,13,14,p=3	0.032
5	1,2,6,7,8,9,10,13,14,p=3	0.0305
6	1,6,7,8,10,11,13,14,p=3	0.029
7	1,2,3,6,7,8,10,12,13,14,p=3	0.0285
8	1,2,5,6,7,8,10,12,13,14,p=3	0.0285
9	1,2,5,6,7,8,9,10,13,14,p=3	0.0265
10	1,2,4,6,7,8,10,11,13,14,p=3	0.0225
11	1,3,6,7,8,9,10,12,13,14,p=3	0.0215
12	3,4,6,7,8,,9,10,13,14,15,p=3	0.021
13	1,6,7,8,10,11,13,14,15,p=3	0.021
14	1,2,4,6,7,8,10,13,14,16,p=3	0.02
15	1,6,7,8,10,11,13,14,16,p=3	0.0185
16	3,6,7,8,10,13,14,15,p=3	0.018
17	1,3,6,7,8,10,11,12,13,14,15,p=3	0.017
18	1,2,6,7,8,10,12,13,14,p=3	0.0165
19	2,3,5,6,7,8,10,13,14,15,p=3	0.016
20	1,2,6,7,8,10,11,13,14,16,p=3	0.016

variable when our model is correctly specified. We have also developed algorithms for estimation and to calculate the Bayes factors. Our results suggest that our method could achieve consistency in both parameter estimation and model selection.

Whether our approach will provide more solutions for other models with incidental parameter problem is still under research. Some assumptions for the panel AR model may be restrictive for application, such as the stationarity of the model, the strictly exogenous assumption for the regressors and the homoscedasticity. Future research to relax such assumptions and to investigate the correction function approach under a wider context may be productive.

References

Alonso-Borrego, C. and M. Arellano (1999): "Symmetrically Normalized Instrumental-Variable Estimation Using Panel Data," *Journal of Business & Economic Statistics*, 17, 36–49.

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.

Arellano, M. and S. Bonhome (2006): "Robust Priors in Nonlinear Panel Data Models," Working papers, CEMFI.

Arellano, M. and J. Hahn (2006): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments," in *Advances in Economics and Econometrics, Ninth World Congress*, ed. by R. Blundell, W. Newey, and T. Persson, Cambridge University Press.

Barndorff-Nielsen, O. E. and G. Schou (1973): "On the Parameterization of Autoregressive Models by Partial Autocorrelations," *Journal of Multivariate Analysis*, 408–419.

Bernardo, J. M. (2005): "Reference analysis," *Handbook of Statistics*, 25, 17–90.

Bernardo, J. M. and A. F. Smith (1994): *Bayesian Theory*, John Wiley & Sons Ltd.

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

Bun, M. J. and F. Windmeijer (2007): "The Weak Instrument Problem of the System GMM Estimator in Dynamic Panel Data Models," ESEM 2007 Conference paper.

Carlin, B. P. and T. A. Luis (2000): *Bayes and empirical Bayes methods for data analysis*, CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431, USA: Chapman & Hall/CRC.

Chib, S. and E. Greenberg (1995): "Understanding the Metropolis-Hastings algorithm," *American Statistician*, 49, 329–335.

Chib, S. and I. Jeliazkov (2001): "Marginal Likelihood From the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96, 270–281.

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.

Ehlers, R. S. and S. P. Brooks (2002): "Efficient Construction of Reversible Jump MCMC Proposals for Autoregressive Time Series Models," Tech. rep., University of Cambridge.

Fernandez, C., E. Ley, and M. F. J. Steel (2001): "Model uncertainty in cross-country growth regressions," *Journal of Applied Econometrics*, 16, 563–576.

Green, P. J. (1995): "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Hahn, J. (2004): "Does Jeffrey's prior alleviate the incidental parameter problem?" *Economics Letters*, 82, 135–138.

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

Hsiao, C., M. Hashem Pesaran, and A. Kamil Tahmiscioglu (2002): "Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods," *Journal of Econometrics*, 109, 107–150.

Jones, M. C. (1987): "Randomly Choosing Parameters from the Stationarity and Invertibility Region of Autoregressive-Moving Average Models," *Applied Statistics*, 36, 134–138.

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.

Li, G. (2009): "Consistent Estimation, Model Selection and Averaging of Dynamic Panel Data Models with Fixed Effect," Working paper, Cardiff Business School.

Liseo, B. (2006): "The elimination of nuisance parameters," *Handbook of Statistics*, 25.

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.

Philippe, A. (2006): "Bayesian Analysis of Autoregressive Moving Average Processes with Unnkown Orders," *Computational Statistics & Data Analysis*, 1904–1923.

Piccolo, D. (1982): "The Size of the Stationarity and Invertibility Region of an Autoregressive-Moving Average Process," *Journal of Time Series Analysis*, 3, 245–247.

Ramsey, F. L. (1974): "Characterization of the Partial Autocorrelation Function," *The Annals of Statistics*, 2, 1296–1301.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20, 518–29.

Sweeting, T. J. (1995): "A Bayesian approach to approximate conditional inference," *Biometrika*, 82, 25–36.

Wooldridge, J. (2005): "Simple Solutions to the Initial Conditions Prolem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics*, 20, 39–54.

A Appendix

A.1 Solution for (24)

By repetitive substitution, we can rewrite the model in (17) as the following,

$$[\mathbf{y}'_{i,-p}, y_{i,1}, y_{i,2}, \ldots, y_{i,T-1}]' = f_i c_1 + I_{T-1+p} \otimes y'_{i,-p} c_2 + CX_i\beta + Cu_i$$

$$\underset{p\times 1}{\mathbf{y}_{i,-p}} = \begin{pmatrix} y_{i,-p+1} \\ y_{i,-p+2} \\ \cdots \\ y_{i,-1} \\ y_{i,0} \end{pmatrix}, \underset{p\times p}{P} = \begin{pmatrix} \rho_1 & 1 & 0 & \ldots & 0 \\ \rho_2 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \\ \rho_{p-1} & 0 & 0 & \ldots & 1 \\ \rho_p & 0 & 0 & \ldots & 0 \end{pmatrix},$$

$$\underset{(T-1+p)\times 1}{c_1} = \begin{pmatrix} 0_{p\times 1} \\ 1 \\ P_{(1,1)} + 1 \\ P^2_{(1,1)} + P_{(1,1)} + 1 \\ \cdots \\ P^{T-2}_{(1,1)} + P^{T-3}_{(1,1)} + \cdots + P_{(1,1)} + 1 \end{pmatrix}, \underset{[p^2+(T-1)p]\times 1}{c_2} = \begin{pmatrix} vec(I_p) \\ P_{(:,1)} \\ P^2_{(:,1)} \\ \cdots \\ P^{T-1}_{(:,1)} \end{pmatrix},$$

$$\underset{(T-1+p)\times T}{C} = \begin{pmatrix} 0_{p\times 1} & 0_{p\times 1} & \ldots & 0_{p\times 1} & 0_{p\times 1} \\ 1 & 0 & \ldots & 0 & 0 \\ P_{(1,1)} & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ P^{T-2}_{(1,1)} & P^{T-3}_{(1,1)} & \ldots & 1 & 0 \end{pmatrix}. \tag{54}$$

where $P^n_{(1,1)}$ and $P^n_{(:,1)}$ denote the (1,1) element and the first column of the matrix P^n. To find $\underset{p\times 1}{E_y}(Y'_{i_-}\iota)$, we just need to make use of (54). For the convenience of subsequent exposition, we define $h : R^p \mapsto R^p$, $\omega_1 : R^{p+T} \mapsto R^p$ and $\omega_2 : R^{p+p} \mapsto R^p$ as

$$\underset{p\times 1}{h}(\underset{p\times 1}{\rho}) = \frac{1}{T}\begin{pmatrix} \iota' c_{1(p:T+p-1)} \\ \iota' c_{1(p-1:T+p-2)} \\ \cdots \\ \iota' c_{1(1:T)} \end{pmatrix} = -\begin{pmatrix} trace\left(HC_{(p:T+p-1,:)}\right) \\ trace\left(HC_{(p-1:T+p-2,:)}\right) \\ \cdots \\ trace\left(HC_{(1:T,:)}\right) \end{pmatrix}$$

$$\underset{p\times 1}{\omega_1}(\underset{T\times 1}{X_i\beta}, \underset{p\times 1}{\rho}) = \begin{pmatrix} \iota'(CX_i\beta)_{(p:T+p-1)} \\ \iota'(CX_i\beta)_{(p-1:T+p-2)} \\ \cdots \\ \iota'(CX_i\beta)_{(1:T)} \end{pmatrix} \tag{55}$$

$$\underset{p\times 1}{\omega_2}(\underset{p\times 1}{y_{i,-p}}, \underset{p\times 1}{\rho}) = \begin{pmatrix} \iota'(I_{T-1+p} \otimes y'_{i,-p}c_2)_{1(p:T+p-1)} \\ \iota'(I_{T-1+p} \otimes y'_{i,-p}c_2)_{1(p-1:T+p-2)} \\ \cdots \\ \iota'(I_{T-1+p} \otimes y'_{i,-p}c_2)_{1(1:T)} \end{pmatrix}$$

where $a_{1(1:T)}$ and $A_{(1:T,:)}$ denote the 1 to T elements and the 1 to T rows of a and A respectively. Note that since $E_y(Cu_i)$ is equal to zero, we can obtain $E_y(Y'_{i_-}\iota) = [Th(\rho)f_i + \omega_1(X_i\beta, \rho) + \omega_2(y_{i,-p}, \rho)]$ and hence (23).

Since the right hand side of (24) only involves ρ, we could assume $\underset{1\times 1}{\lambda_\rho}(f_i, \theta) =$

$\tau(\rho) + constant$, where the constant term could be any arbitrary function of f_i, β and σ^2. For simplicity, we choose the constant term to be 0.[24] The equation $\frac{\partial \tau(\rho)}{\partial \rho} = h(\rho)$ implies the following,

$$d\,\tau(\rho) = \sum_{k=1}^{p} h_k(\rho)\, d\rho_k. \tag{56}$$

To prove that $\tau(\rho)$ exists, we just need to prove the differential of $\tau(\rho)$ is exact. Before the proof, we need to establish Lemma A.1.

Lemma A.1.

$$\frac{\partial P_{(1,1)}^{i+j}}{\partial \rho_i} = \frac{\partial P_{(1,1)}^{i'+j}}{\partial \rho_{i'}} \tag{57}$$

where $i, i' = 1, 2, \ldots, p$ and j is zero or a positive integer. Without loss of generality, we can assume $i \leq i'$.[25]

Proof. First note that[26]

$$P_{(1,1)}^{n} = \sum_{k=1}^{p} \rho_k P_{(1,1)}^{n-k}. \tag{58}$$

The above equation implies $\frac{\partial P_{(1,1)}^{n}}{\partial \rho_i} = 0$ and $\frac{\partial P_{(1,1)}^{n}}{\partial \rho_i} = 1$ for $n < i$ and $n = i$ respectively. Then we can prove (57) by mathematical induction, which involves the following three steps:

1. We assume that for any integer less than j equation (57) holds. The left and right hand side of (57) can be rewritten as

$$\frac{\partial P_{(1,1)}^{i+j}}{\partial \rho_i} = \rho_1 \frac{\partial P_{(1,1)}^{i+j-1}}{\partial \rho_i} + \cdots + \frac{\partial \left(\rho_i P_{(1,1)}^{i+j-i}\right)}{\partial \rho_i} + \cdots + \rho_{i'} \frac{\partial P_{(1,1)}^{i+j-i'}}{\partial \rho_i} + \cdots + \rho_p \frac{\partial P_{(1,1)}^{i+j-p}}{\partial \rho_i} \tag{59}$$

$$\frac{\partial P_{(1,1)}^{i'+j}}{\partial \rho_{i'}} = \rho_1 \frac{\partial P_{(1,1)}^{i'+j-1}}{\partial \rho_{i'}} + \cdots + \rho_i \frac{\partial P_{(1,1)}^{i'+j-i}}{\partial \rho_{i'}} + \cdots + \frac{\partial \left(\rho_{i'} P_{(1,1)}^{i'+j-i'}\right)}{\partial \rho_{i'}} + \cdots + \rho_p \frac{\partial P_{(1,1)}^{i'+j-p}}{\partial \rho_{i'}} \tag{60}$$

   Due to our assumption[27], the following must hold

$$\rho_n \frac{\partial P_{(1,1)}^{i+j-n}}{\partial \rho_i} = \rho_n \frac{\partial P_{(1,1)}^{i'+j-n}}{\partial \rho_{i'}}, \tag{61}$$

[24] This choice indeed can produce the solution to achieve consistent estimation for this particular model. The authors are not entirely sure if $\frac{\partial \chi(f_i,\theta)}{\partial f_i}$, where $\chi(f_i, \theta)$ is defined in (10), involves all the common parameters and the incidental parameter, what strategy is required for consistent estimation. It should depend on the specific problems.

[25] It is obvious that if $i = i'$, equation (57) holds. Therefore in the following, we just need to prove the case when $i < i'$.

[26] We define $P_{(1,1)}^{n-k} = 1$ if $n - k = 0$ and $P_{(1,1)}^{n-k} = 0$ if $n - k < 0$.

[27] Note that $j - n < j$.

where $n \in \{1, 2, \ldots, p\} \setminus \{i, i'\}$. Now to prove (59) and (60) are equal to each other is reduced to proving

$$\frac{\partial \left(\rho_i P_{(1,1)}^{i+j-i}\right)}{\partial \rho_i} + \rho_{i'} \frac{\partial P_{(1,1)}^{i+j-i'}}{\partial \rho_i} = \rho_i \frac{\partial P_{(1,1)}^{i'+j-i}}{\partial \rho_{i'}} + \frac{\partial \left(\rho_{i'} P_{(1,1)}^{i'+j-i'}\right)}{\partial \rho_{i'}}, \tag{62}$$

which is equivalent to

$$P_{(1,1)}^{j} + \rho_i \frac{\partial P_{(1,1)}^{i+j-i}}{\partial \rho_i} + \rho_{i'} \frac{\partial P_{(1,1)}^{i+j-i'}}{\partial \rho_i} = P_{(1,1)}^{j} + \rho_i \frac{\partial P_{(1,1)}^{i'+j-i}}{\partial \rho_{i'}} + \rho_{i'} \frac{\partial P_{(1,1)}^{i'+j-i'}}{\partial \rho_{i'}}. \tag{63}$$

It is not hard to see that (63) is true due to our assumption. Finally we know that if (57) holds for any integer less than j, then it also holds for j.

2. The smallest possible number for j is 0, which indicates both sides of (57) are equal to 1. So (57) holds.

3. From the above two points, we know that Lemma A.1 is true.

□

Now we are ready to prove that there exists a solution for the partial differential equation system (56).

Proof. It can be seen from (56) that if the system has a solution, the differential of $\tau(\rho)$ must be exact, which implies the following must be satisfied,

$$\frac{\partial h_i(\rho)}{\rho_{i'}} = \frac{\partial h_{i'}(\rho)}{\rho_i} \tag{64}$$

Note that $h_i(\rho)$ and $h_{i'}(\rho)$ can take the following forms

$$h_i(\rho) = \frac{T-i}{T} + \frac{T-i-1}{T} P_{(1,1)} + \cdots + \frac{T-i-i'}{T} P_{(1,1)}^{i'} + \cdots + \frac{1}{T} P_{(1,1)}^{T-i-1}$$
$$h_{i'}(\rho) = \frac{T-i'}{T} + \frac{T-i'-1}{T} P_{(1,1)} + \cdots + \frac{T-i'-i}{T} P_{(1,1)}^{i} + \cdots + \frac{1}{T} P_{(1,1)}^{T-i'-1}.$$

To prove (64), we need to have

$$\frac{T-i-i'}{T} \frac{\partial P_{(1,1)}^{i'}}{\partial \rho_{i'}} + \cdots + \frac{1}{T} \frac{P_{(1,1)}^{i'+T-i-i'-1}}{\partial \rho_{i'}} = \frac{T-i'-i}{T} \frac{\partial P_{(1,1)}^{i}}{\partial \rho_i} + \cdots + \frac{1}{T} \frac{P_{(1,1)}^{i+T-i-i'-1}}{\partial \rho_i} \tag{65}$$

By Lemma A.1, we know that (65) is true. Hence (64) is true and $d\,\tau(\rho)$ is exact. So we can conculde that $\tau(\rho)$ exists and (56) has a solution. □

Next we go on to solve (56). A solution for $\tau(\rho)$ can take the following form,

$$\tau(\rho) = R_1(\rho) + \phi_1(\rho_{2:p}) \tag{66}$$

where $R_1(\rho) = \int h_1(\rho) d\rho_1$ and $\phi_1(\rho_{2:p})$ is a function involving all the elements in ρ except ρ_1. To derive $\phi_1(\rho_{2:p})$, we can use the following relationship

$$\frac{\partial \tau(\rho)}{\partial \rho_2} = h_2(\rho) = \frac{\partial R_1(\rho)}{\partial \rho_2} + \frac{\partial \phi_1(\rho_{2:p})}{\partial \rho_2}. \tag{67}$$

Hence

$$\phi_1(\rho_{2:p}) = \int \left(h_2(\rho) - \frac{\partial R_1(\rho)}{\partial \rho_2} \right) d\rho_2 + \phi_2(\rho_{3:p}). \tag{68}$$

where $\phi_3(\rho_{3:p})$ is a function of all the element of ρ except ρ_1 and ρ_2. We could denote $R_2(\rho_{2:p}) = \int \left(h_2(\rho) - \frac{\partial R_1(\rho)}{\partial \rho_2} \right) d\rho_2$. If we continue the above procedure p times, we could find out the general solution for $\tau(\rho)$ is

$$\tau(\rho) = \sum_{i=1}^{p} R_i(\rho_{i:p}) + k \tag{69}$$

where k is an arbitrary constant not depending on ρ and

$$R_i(\rho_{i:p}) = \int \left(h_i(\rho) - \sum_{j=1}^{i-1} \frac{\partial R_j(\rho_{j:p})}{\partial \rho_i} \right) d\rho_i \qquad for \quad i = 2, \ldots, p \tag{70}$$

with $R_1(\rho) = \int h_1(\rho) d\rho_1$. If we look at (69) more carefully, we can see that the general solution of $\tau(\rho)$ is obtained by summing up all the distinct terms in each element of $\int \underset{p\times 1}{h(\rho)}\, d\underset{p\times 1}{\rho}$ and an arbitrary constant (which we set to 0).

A.2 An Asymptotic Local Stationary Point of the Integrated Likelihood

In this subsection, we will prove that the true value, θ is a local stationary point asymptotically for the integrated likelihood function, $p(Y|\theta)$ obtained by integrating out f under the prior $p(f|\theta) = \prod_{i=1}^{N} p(f_i|\rho) \propto r(\rho) = exp[N\tau(\rho)]$. The natural log of the integrated likelihood function takes the following form (see the next subsection for derivation details),

$$\begin{aligned} \ln p(Y|r, b, s^2) &\propto Q_N(r, b, s^2) \\ &= -\frac{1}{2s^2} \sum_i (y_i - Y_{i\_} r - X_i b)' H (y_i - Y_{i\_} r - X_i b) - \frac{N(T-1)}{2} \ln s^2 + N\tau(r). \end{aligned} \tag{71}$$

where r, b and s^2 are the specific values that θ takes. Subsituting (17) into (71) yields

$$\begin{aligned}\ln p(Y|r,b,s^2) &\propto Q_N(r,b,s^2)\\ &= -\frac{1}{2s^2}\Big\{(\rho-r)'\sum_i Y_{i\_}'HY_{i\_}(\rho-r)+(\beta-b)'\sum_i X_i'HX_i(\beta-b)\\ &+\sum_i u_iH'u_i+2(\rho-r)'\sum_i Y_{i\_}'Hu_i+2(\rho-r)'\sum_i Y_{i\_}'HX_i(\beta-b)\\ &+2\sum_i u_i'HX_i(\beta-b)\Big\}-\frac{N(T-1)}{2}\ln s^2+N\tau(r).\end{aligned} \tag{72}$$

Next we assume the following probability limits exist:

$$\begin{aligned}&\underset{N\to\infty}{plim}\frac{1}{N}\sum_i^N Y_{i\_}'HY_{i\_}=\underset{p\times p}{Y\_Y\_}\\ &\underset{N\to\infty}{plim}\frac{1}{N}\sum_i^N Y_{i\_}'Hu_i=\sigma^2\begin{pmatrix}trace\left(HC_{(p:T+p-1,:)}\right)\\ trace\left(HC_{(p-1:T+p-2,:)}\right)\\ \cdots\\ trace\left(HC_{(1:T,:)}\right)\end{pmatrix}=-\sigma^2h(\rho)\\ &\underset{N\to\infty}{plim}\frac{1}{N}\sum_i^N Y_{i\_}'HX_i=\underset{p\times K}{Y\_X}\\ &\underset{N\to\infty}{plim}\frac{1}{N}\sum_i^N X_i'HX_i=\underset{K\times K}{XX}\end{aligned} \tag{73}$$

Hence the probability limit of $\frac{1}{N}Q_N(r,b,s^2)$ exists as the following[28],

$$\begin{aligned}\underset{N\to\infty}{plim}\frac{1}{N}Q_N(r,b,s^2)&=Q(r,b,s^2)\\ &=-\frac{1}{2s^2}\Big\{(\rho-r)'\left(Y\_Y\_\right)(\rho-r)+(\beta-b)'\left(XX\right)(\beta-b)+(T-1)\sigma^2\\ &-2\sigma^2(\rho-r)'h(\rho)+2(\rho-r)'\left(Y\_X\right)(\beta-b)\Big\}-\frac{T-1}{2}\ln s^2+\tau(r).\end{aligned} \tag{74}$$

[28] We also use the facts that $\underset{N\to\infty}{plim}\frac{1}{N}\sum_i^N u_iH'u_i=(T-1)\sigma^2$ and X_i and f_i are strictly exogenous.

Now we can differentiate $Q(r, b, s^2)$ to check the first order condition:

$$\begin{aligned}
\frac{\partial Q}{\partial r} &= \frac{1}{s^2}\left[(\rho - r)'\left(Y_Y_\right) - \sigma^2 h(\rho) + \left(Y_X\right)(\beta - b)\right] + h(r) \\
\frac{\partial Q}{\partial b} &= \frac{1}{s^2}\left[XX(\beta - b) + (\rho - r)'(Y_X)\right] \\
\frac{\partial Q}{\partial s^2} &= \frac{1}{2s^2}\Bigg\{(\rho - r)'\left(Y_Y_\right)(\rho - r) + (\beta - b)'\left(XX\right)(\beta - b) + (T-1)\sigma^2 \\
&\quad - 2\sigma^2(\rho - r)'h(\rho) + 2(\rho - r)'\left(Y_X\right)(\beta - b)\Bigg\} - \frac{(T-1)}{2s^2}.
\end{aligned} \tag{75}$$

We can see that $(r = \rho, b = \beta, s^2 = \sigma^2)$ can obviously solve the above three equations and hence is a local stationary point for the integrated likelihood asymptotically.

A.3 Proof of Proposition 3.1

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

$$\begin{aligned}
p(\theta)p(Y|\theta, Y_0) = \frac{1}{m(S)} I(\rho \in S) p(\beta|\sigma^2)(2\pi)^{-\frac{TN}{2}} \sigma^{2(-\frac{NT+2}{2})} r(\rho) \\
\prod_{i=1}^{N} \exp\left\{-\frac{1}{2\sigma^2}\left[w_i - f_i\iota - X_i\beta\right]'\left[w_i - f_i\iota - X_i\beta\right]\right\},
\end{aligned} \tag{76}$$

where $Y = (y_1, y_2, \ldots, y_N)'$ excludes the first observations of all economic agents, of which $Y_0 = (y_{1,0}, y_{2,0}, \ldots, y_{N,0})'$ is the collection.

Now we derive the posterior distribution of f_i. We can rewrite equation (76) as

$$\begin{aligned}
p(\theta)p\left(Y|\theta, Y_0\right) &= p(\beta|\sigma^2)\frac{1}{m(S)} I(\rho \in S)(2\pi)^{-\frac{TN}{2}} \sigma^{2(-\frac{NT+2}{2})} r(\rho) \\
&\quad \prod_{i=1}^{N} \exp\Bigg\{-\frac{1}{2\sigma^2}[(w_i - X_i\beta)'(w_i - X_i\beta) \\
&\quad + T f_i^2 - 2\iota'(w_i - X_i\beta) - f_i\Bigg\}.
\end{aligned}$$

We then complete the square for f_i by adding $-\frac{(\iota' w_i - X_i\beta)^2}{T} + \frac{(\iota' w_i - X_i\beta)^2}{T}$ inside the exponential. So it becomes

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

or equivalently

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

where $H=I_T-\frac{\iota\iota'}{T}$ is the demean matrix. Substituting $w_i=y_i-Y_{i\_}$ back into our equation, we can have

$$\begin{aligned} p(\theta)p\left(Y|\theta,Y_0\right) &= p(\beta|\sigma^2)\frac{1}{m(S)}I(\rho\in S)(2\pi)^{-\frac{TN}{2}}\sigma^{2(-\frac{NT+2}{2})}r(\rho) \\ &\prod_{i=1}^{N}\exp\left\{-\frac{1}{2\frac{\sigma^2}{T}}[f_i-\frac{\iota'(y_i-Y_{i\_}\rho-X_i\beta)}{T}]^2\right\} \\ &\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-Y_{i\_}\rho-X_i\beta)'H(y_i-Y_{i\_}\rho-X_i\beta)\right] \end{aligned} \tag{77}$$

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 (77) 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 (29).

Next we go on to derive the posterior distributions for β and σ^2. First we can integrate out g in equation (77) to obtain

$$\begin{aligned} p(\rho,\beta,\sigma^2,Y|Y_0) &= p(\rho,\beta,\sigma^2|Y,Y_0)p(Y|Y_0) \\ &= p(\beta|\sigma^2)\frac{1}{m(S)}I(\rho\in S)T^{-\frac{N}{2}}(2\pi)^{-\frac{N(T-1)}{2}}\sigma^{2\left[-\frac{N(T-1)+2}{2}\right]} \\ &r(\rho)\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-Y_{i\_}\rho-X_i\beta)'H(y_i-Y_{i\_}\rho-X_i\beta)\right]. \end{aligned} \tag{78}$$

If we define $\tilde{w}_i=H(y_i-y_{i\_}\rho)$ and $\tilde{X}_i=HX_i$, by incorporating the prior of β in (28) we can rewrite equation (78) as

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

Then completing the square of β yields

$$
\begin{aligned}
&p(\rho,\beta,\sigma^2|Y,Y_0)p(Y|Y_0)\\
=&\frac{1}{m(S)}I(\rho\in S)T^{-\frac{N}{2}}(2\pi)^{-\frac{N(T-1)+k}{2}}\sigma^{2\left[-\frac{N(T-1)+2+k}{2}\right]}r(\rho)\left|\eta\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right|^{\frac{1}{2}}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}\tilde{w}_i{}'\tilde{w}_i-\frac{1}{\eta+1}\sum_{i=1}^{N}\tilde{w}_i'\tilde{X}_i\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i\right]\right\}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[\beta-\frac{1}{\eta+1}\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i\right]'\right.\\
&\left.\left(\sum_{i=1}^{N}(\eta+1)\tilde{X}_i'\tilde{X}_i\right)\left[\beta-\frac{1}{\eta+1}\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i\right]\right\}
\end{aligned}
\tag{79}
$$

We can see that the conditional posterior of β follows a normal distribution as in (30). Now we can integrate out β in (79) to obtain the posterior distribution for ρ and σ^2 as the following,

$$
\begin{aligned}
p(\rho,\sigma^2|Y,Y_0)p(Y|Y_0)=&\frac{1}{m(S)}I(\rho\in S)\left(\frac{\eta}{\eta+1}\right)^{\frac{k}{2}}T^{-\frac{N}{2}}(2\pi)^{-\frac{N(T-1)}{2}}\\
&\sigma^{2\left[-\frac{N(T-1)+2}{2}\right]}r(\rho)\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}\tilde{w}_i{}'\tilde{w}_i\right.\right.\\
&\left.\left.-\frac{1}{\eta+1}\sum_{i=1}^{N}\tilde{w}_i'\tilde{X}_i\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i\right]\right\}.
\end{aligned}
\tag{80}
$$

It is clear from equation (80) that conditional on ρ, σ^2 follows an inverted gamma distribution with mean $\frac{\Delta}{N(T-1)-2}$ and degrees of freedom $N(T-1)$ as in (31),

Now we can integrate out σ^2 to obtain the posterior distribution of ρ as in (82).

$$
\begin{aligned}
p(\rho|Y,Y_0)p(Y|Y_0)=&\frac{1}{m(S)}I(\rho\in S)\left(\frac{\eta}{\eta+1}\right)^{\frac{k}{2}}(\Delta)^{-\frac{N(T-1)}{2}}\\
&\Gamma\left[\frac{N(T-1)}{2}\right]T^{-\frac{N}{2}}(\pi)^{-\frac{N(T-1)}{2}}r(\rho)
\end{aligned}
\tag{81}
$$

$$
p(\rho|Y,Y_0)\propto I(\rho\in S)r(\rho)\,(\Delta)^{-\frac{N(T-1)}{2}}, \tag{82}
$$

Another way to interpret the posterior of ρ is given in (33) under Proposition 3.1.

A.4 Proof of Equation (48)

Proof. Note that there is a differentiable mapping from $(\pi^{(0)\prime}, u')'$ to $\rho^{(c)}$, whose Jacobian is given in (36), and also from $\rho^{(0)}$ to $\pi^{(0)}$. Hence we can obtain

$$\begin{aligned}\left|\frac{\partial \rho^{(c)}}{\partial(\rho^{(0)\prime}, u')}\right| &= \left|\frac{\partial \rho^{(c)}}{\partial(\pi^{(0)\prime}, u')}\frac{\partial(\pi^{(0)\prime}, u')'}{\partial(\rho^{(0)\prime}, u')}\right| \\ &= \prod_{i=1}^{p^{(c)}-p^{(0)}} (1+u_i)^{[\frac{p^{(0)}+i-1}{2}]}(1-u_i)^{[\frac{p^{(0)}+i}{2}]}\left|\frac{\partial \rho^{(0)}}{\partial \pi^{(0)\prime}}\right|\left|\frac{\partial \pi^{(0)}}{\partial \rho^{(0)\prime}}\right| \\ &= \prod_{i=1}^{p^{(c)}-p^{(0)}} (1+u_i)^{[\frac{p^{(0)}+i-1}{2}]}(1-u_i)^{[\frac{p^{(0)}+i}{2}]}\end{aligned}$$

□

A.5 Data Generating Process for the Exogenous Regressors in Section 3.5

We go through the following steps to generate the exogenous regressors used in Section 3.5:

1. We generate the potential regressors $(X_i's)$ from the uniform distribution $U[-4, 4]$.

2. We make the regressors serially correlated with each other. We achieve this by first making each two neighboring period observations correlated with each other as follows,

$$x_{t,s} = s_{t-1}x_{t-1,s} + \bar{s}_t x_{t,ns}, \tag{83}$$

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

$$S = \begin{pmatrix} 1 & s_1 & \cdots & \prod_{i=1}^{T-1} s_i \\ s_1 & 1 & \cdots & \prod_{i=2}^{T-1} s_i \\ s_2 s_1 & s_2 & \cdots & \prod_{i=3}^{T-1} s_i \\ \cdots & \cdots & \cdots & \cdots \\ \prod_{i=1}^{T-1} s_i & \prod_{i=2}^{T-1} s_i & \cdots & 1 \end{pmatrix} \tag{84}$$

We can see that $\{x_t\}$ generated in such a way is not covariance stationary. Moreover, for small T[29], the distribution of $x's$ will change with t. However, if T is sufficiently large, 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[30]. We just use the final few observations from the series for our study.

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

$$X_{j,c} = \sum_{i=1}^{K} q_{j,i} X_{i,nc} \qquad j = 1, 2, \ldots, K \tag{85}$$

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

$$corr(X_{t,k}, X_{t',k'}) = S(t, t') \sum_{i=1}^{K} q_{k,i} q_{k',i} \qquad t = 1, 2, \ldots, T \quad k = 1, 2, \ldots, K. \tag{86}$$

where $S(t, t')$ denote the (t, t') element in S and K is the potential number of regressors.

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

[30] We choose T to be 100 for the results to be presented in the section so that $x's$ approximately converge to a normal distribution.