Essays on Economic and Econometric Applications of Bayesian Estimation and Model Comparison

Guangjie LI

Thesis submitted for the degree of
Doctor of Philosophy
Department of Economics
University of Leicester

April 2008

To my family

Abstract

This thesis consists of three chapters on economic and econometric applications of Bayesian parameter estimation and model comparison. The first two chapters study the incidental parameter problem mainly under a linear autoregressive (AR) panel data model with fixed effect. The first chapter investigates the problem from a model comparison perspective. The major finding in the first chapter is that consistency in parameter estimation and model selection are interrelated. The reparameterization of the fixed effect parameter proposed by Lancaster (2002) may not provide a valid solution to the incidental parameter problem if the wrong set of exogenous regressors are included. To estimate the model consistently and to measure its goodness of fit, the Bayes factor is found to be more preferable for model comparson than the Bayesian information criterion based on the biased maximum likelihood estimates. When the model uncertainty is substantial, Bayesian model averaging is recommended. The method is applied to study the relationship between financial development and economic growth. The second chapter proposes a correction function approach to solve the incidental parameter problem. It is discovered that the correction function exists for the linear AR panel model of order p when the model is stationary with strictly exogenous regressors. MCMC algorithms are developed for parameter estimation and to calculate the Bayes factor for model comparison. The last chapter studies how stock return's predictability and model uncertainty affect a rational buy-and-hold investor's decision to allocate her wealth for different lengths of investment horizons in the UK market. The FTSE All-Share Index is treated as the risky asset, and the UK Treasury bill as the riskless asset in forming the investor's portfolio. Bayesian methods are employed to identify the most powerful predictors by accounting for model uncertainty. It is found that though stock return predictability is weak, it can still affect the investor's optimal portfolio decisions over different investment horizons.

Keywords: model comparison, model selection, consistency in estimation, incidental parameter problem, Bayesian model averaging (BMA), Markov chain Monte Carlo (MCMC), dynamic panel data model with fixed effect, finance and growth, seemingly unrelated regression (SUR) model, stock return predictability, portfolio choice

Preface

Compared to frequentist econometric methods, a marked difference of Bayesian econometrics is that it can provide a framework to unify parameter estimation and model comparison. Once an appropriate prior is set up, a Bayesian researcher can estimate the parameters in a model and compare different model specifications in a straightforward way. In comparison, frequentist econometricians tend to treat estimation and model comparison as two separate topics. Such difference can lead to quite different answers to some questions for these two approaches. Another feature of Bayesian econometrics is its ability to take account of model uncertainty. This issue is relevant especially when no single model specification is significantly better compared with others. Such situation may arise, e.g. when we have small data sample in our empirical study and there is too much noise in our data. Bayesian econometrics can explicitly handle this problem by using a technique called Bayesian Model Averaging (BMA). The thesis tries to demonstrate these Bayesian ideas by providing three chapters of econometric and economic applications.

The first and the second chapter study the incidental parameter problem. The maximum likelihood estimator (MLE) for the common parameters is not consistent while the generalized method of moments (GMM) is not very informative about model specification. An important finding in the first chapter is that consistent parameter estimation and consistent model selection are interrelated. With consistency in parameter estimation, it is likely that we can have consistency in model selection. On the other hand, model selection criterion based on inconsistent parameter estimates

tends to be misleading and performs poorly in application. Therefore we should consider these two issues simultaneously for our applications. The major contribution of the second chapter is to propose a new method from a Bayesian perspective to tackle the incidental parameter problem. A solution for linear autoregressive (AR) panel data model of order p with fixed effect is discovered.

The theoretical results from the first two chapters are based on large sample asymptotic theories. However, in real life applications, large data sample is rarely present and we still have to extract robust and useful patterns from our data. Forming our inference based on a single model may lead us to finding some accidental data patterns. The last chapter gives an example on this issue, in which the robust predictors for excess stock return need to be identified. By accounting for model uncertainty explicitly in our analysis, we will be able to reduce such risk. Bayesian methods form an elegant framework to allow us to consider model uncertainty by averaging estimation results from different possible models.

With the advancement of modern computing technology, Bayesian methods are becoming more influential in the field of econometrics than ever before. Programming is hence more and more important for Bayesian econometrics. Though it was initially difficult to learn such technique, I would consider my own learning experience worthwhile and rewarding. The programs used in the thesis are written in MatLab and Maple. In fact some useful ideas are discovered during the time of coding in these two languages.

The thesis is finished under the supervision of Gary Koop, Roberto Leon Gonzalez and Gianni De Fraja over the years of my PhD study at the University of Leicester. I wish to express my gratitude to Gary Koop for opening up the door of Bayesian econometrics so that it is possible for me to explore this wonderful world, to Roberto Leon Gonzalez for showing me research directions and giving me patient guidance and to Gianni De Fraja who kindly took over the role of supervisor in my final year. I would also like to extend my thanks to Sebastiano Manzan and Rodney Strachan for their advice on my third chapter, and to Ross Levine for sharing his dataset, which I use in my first chapter. Financial help from the Economics Department is grate-

fully acknowledged. The first chapter has benefited from the discussion with participants at the 62nd European Meeting of the Econometric Society and the 2nd Japanese-European Bayesian Econometrics and Statistics Meeting, while the third chapter has been improved based on the comments and suggestions received at the 2006 Far East Meeting of the Econometric Society and the 2nd PhD Conference at the University of Leicester. The errors that inevitably remain are solely my responsibility.

Leicester, England *Guangjie LI (Jack)*
April 2008 GL41@le.ac.uk

Contents

Chapter 1

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

1.1 Introduction

For a panel data linear regression model with lags of the dependent variable as regressors and agent specific fixed effects, the maximum likelihood estimate (MLE) of the common parameter is inconsistent when the number of time periods is small and fixed regardless of the cross section sample size. Nerlove (1968) showed in Monte Carlo simulations that the MLE is severely downward biased. Nickell (1981) derived the analytical form of the bias for the first order autoregression (AR) model. This problem, known as the "incidental parameter problem", due to the fixed effect parameter (incidental parameter), whose dimension will increase with the cross section sample size has been reviewed by Lancaster (2000). The current econometric

literature focuses mainly on deriving consistent estimator for the common parameter. See, for example, Arellano and Bond (1991), Blundell and Bond (1998), Gourieroux et al. (2006) and Hahn and Newey (2004). Little attention is given to model specification comparison in the presence of incidental parameter.

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

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

[2] See the appendix for the details.

from Lancaster's reparameterization of the fixed effect, we tend to pick up the true model when the cross section sample size increases. However, the model selection performance of the Bayesian information criterion (BIC) based on the biased MLE is very poor both for small and big sample sizes. The BIC will asymptotically choose the wrong model for some situations[3]. Thirdly, for small sample size, when model uncertainty is substantial, we argue for the use of Bayesian model averaging (BMA) to reduce estimation risk[4]. Apart from the theoretical results, in the end of the chapter we provide an example of finance and economic growth to show that our method is flexible enough to accommodate real world problems and handle issues like unbalanced panel.

The plan of the chapter is as follows. Section 1.2 summarizes our model and the posterior results. Section 1.3 describes our motivation to compare different model specifications and shows when our posterior estimators will be consistent. Section 1.4 presents the conditions under which the Bayes factor and the BIC can lead to consistency in model selection followed by a short description of the BMA method. In section 1.6, we carry out simulation studies to check our Propositions. Section 1.7 then gives an example of application in finance and growth before Section 1.8 concludes.

1.2 The Model and the Posterior Results

Consider the model

$$\begin{aligned} y_{i,t} &= f_i + y_{i,t-1}\rho + x'_{i,t}\beta + u_{i,t}, \\ i &= 1 \dots N, \quad t = 1 \dots T. \end{aligned} \tag{1.1}$$

Here we are investigating the case of first order autoregression linear panel, where ρ is a scalar and $x_{i,t}$ is a $k \times 1$ vector. Denote u_i as $[u_{i,1}, u_{i,2}, \dots, u_{i,T}]'$ and $X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,T})'$. We assume $u_i | f_i, X_i \sim N(0, \sigma^2 I_T)$ where I_T

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

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

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

$$f_i = g_i \exp\left[-b\left(\rho\right)\right] - \frac{1}{T}\iota' X_i \beta, \tag{1.2}$$

where $b\left(\rho\right)$ is defined as

$$b(\rho) = \frac{1}{T}\sum_{t=1}^{T-1}\frac{T-t}{t}\rho^t, \tag{1.3}$$

Let us transform our model accordingly as

$$y_i = g_i \exp\left[-b\left(\rho\right)\right]\iota + y_{i\_}\rho + HX_i\beta + u_i, \tag{1.4}$$

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

[5] See the appendix for the details.

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

$$\begin{aligned} p(g, \theta) &= p(g, \rho, \sigma^2, \beta) = p(g_1)...p(g_N)p(\rho)p(\sigma^2)p(\beta|\sigma^2) \\ &\propto \frac{1}{\sigma^2} I(-1 < \rho < 1)p(\beta|\sigma^2), \end{aligned} \tag{1.5}$$

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

In regard to the conditional prior of β given σ^2, we want to have a proper distribution so that Bayes factors can lead to the selection of the true model as the cross section sample size increases. We can see this point more clearly later in Section 1.4. The prior we use takes the following g-prior form, proposed by Zellner (1986):

$$\beta|\sigma^2 \sim N\left(0, \sigma^2(\eta \sum_{i=1}^{N} \tilde{X}_i'\tilde{X}_i)^{-1}\right), \tag{1.6}$$

where $\tilde{X}_i = HX_i$. The strength of the prior depends on the value of η. The smaller the value is, the less informative is our prior. We will give more details about the choice of η later. With the parameter priors given in (1.5) and (1.6), we can derive the posterior distributions of the parameters shown in Proposition 1.1.

Proposition 1.1. *The posterior distributions for the parameters in our model will take the following form:*

$$g_i|Y, y_{i,0}, \sigma^2, \rho \sim N\left(e^{b(\rho)} \frac{\iota'(y_i - y_{i\_}\rho)}{T}, \frac{\sigma^2}{T} \exp[2b(\rho)]\right), \tag{1.7}$$

$$\beta|\sigma^2, \rho, Y, Y_0 \sim N\left(\frac{1}{\eta+1}\left(\sum_{i=1}^{N} \tilde{X}_i'\tilde{X}_i\right)^{-1} \sum_{i=1}^{N} \tilde{X}_i'\tilde{w}_i, \sigma^2\left((\eta+1)\sum_{i=1}^{N} \tilde{X}_i'\tilde{X}_i\right)^{-1}\right), \tag{1.8}$$

$$\sigma^2|\rho, Y, Y_0 \sim IW(N(T-1), A), \tag{1.9}$$

$$\rho|Y, Y_0 \propto I(-1 < \rho < 1) \exp[Nb(\rho)] |A|^{-\frac{N(T-1)}{2}}, \tag{1.10}$$

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

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

1.3 Motivation to Compare Different Model Specifications

Lancaster (2002) showed that without model misspecification if we adopt the fixed effect reparameterization and the prior $p(g, \theta) \propto \frac{1}{\sigma^2}$, the mode of the marginal posterior for θ will be consistent. The difference adopted here is the g-prior we use for $p\left(\beta|\sigma^2\right)$ in (1.6). As long as we specify η as a function of the cross section sample size N such that $\lim_{N\to\infty} \eta(N) = 0$, our posterior results will be identical to Lancaster's for big cross section sample size. However, we cannot expect our model will always be correctly specified, i.e. the true regressors used to generate the data are always included in the

regression. Here in proposition 1.2 we show the conditions under which we can obtain consistent posterior estimates for σ^2 and ρ even if we include the wrong set of exogenous regressors.

Proposition 1.2. *The posterior estimates from (1.9) to (1.10) are consistent if we have either*

$$\frac{-(T-1)h_2(\beta,\underline{\rho})}{h_3(\beta)} = h(\underline{\rho}) \tag{1.11}$$

or

$$h_2(\beta,\underline{\rho}) = h_3(\beta) = 0, \tag{1.12}$$

where

$$h(\rho) = \sum_{t=1}^{T-1} \frac{T-t}{T}\rho^{t-1} = \frac{d\, b(\rho)}{d\,\rho}. \tag{1.13}$$

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

$$h_3(\beta) = \underset{N\to\infty}{plim}\frac{1}{N}\left[\sum_{i=1}^{N} \beta' X_i' H X_i\beta - \frac{1}{\eta+1}\sum_{i=1}^{N} \beta' X_i' H\underline{X_i}\left(\sum_{i=1}^{N}\underline{X_i}' H\underline{X_i}\right)^{-1}\sum_{i=1}^{N}\underline{X_i}' H X_i\beta\right]. \tag{1.14}$$

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

Note that $0 < h(\rho) < \frac{T-1}{2}$ and it is monotonically increasing for $\rho \in (-1,1)$. For $h_2(\beta,\underline{\rho}) = h_3(\beta) = 0$ to be satisfied, it is enough that the true regressors X are a subset of $\underline{X}$. For $\frac{-(T-1)h_2(\beta,\underline{\rho})}{h_3(\beta)} = h(\underline{\rho})$ to hold, one example could be that no serial correlation and collinearity exist among the true regressors and the included regressors have zero correlation with the true regressors.[6] Proposition 1.2 tells us that if neither (1.11) nor (1.12) is satisfied, our posterior estimates of σ^2 and ρ will not be consistent even if we

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

have a large cross section sample size when the number of observations for each economic agent is small in the panel. This is one of the major reasons why we need to compare different model specifications. Due to Bartlett's paradox[7], if we want to compare different models, we need to have a proper prior[8] for parameters not common to all the models. That is why we adopt the prior for β in (1.6).

In empirical applications, such as that of the growth theory, we will often have many possible regressors suggested by different theories to be included in the regression in (1.1). In a case like this, the number of potential exogenous regressors will be large. In addition to the concern over inconsistent estimation, we may want to know which combination of these regressors can best explain our data. The predominant GMM method in the literature to estimate the fixed effect model provides little information in this respect. Classical diagnostic tool such as R-square is not well defined. In a Bayesian framework such as ours, we can evaluate how good the model fits the data by looking at the posterior model probability. In our context, different models are defined by different combinations of the regressors and by whether or not we have a lag term of the dependent variable in the regression. So the total number of models is 2^{K+1}, where K stands for the number of all the potential exogenous regressors. The posterior model probability of model i is calculated as

$$\begin{aligned} p\left(M_i|Y,Y_0\right) &= \frac{p\left(M_i\right)p\left(Y|Y_0,M_i\right)}{p\left(Y|Y_0\right)} \\ &= \frac{p\left(M_i\right)p\left(Y|Y_0,M_i\right)}{\sum_{j=1}^{2^{K+1}} p\left(M_j\right)p\left(Y|Y_0,M_j\right)}. \end{aligned} \tag{1.15}$$

where $p\left(M_i\right)$ is the prior model probability. Here we just assume all the models are equally possible a priori such that the posterior model probability only depends on the marginal likelihood, $p\left(Y|Y_0,M_i\right), j=1,2,\ldots,2^{K+1}$.

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

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

We can see in (1.15) that to evaluate the posterior probability of a single model we have to calculate the marginal likelihood of all the models. However, from the derivation of Proposition 1.1, we can only know the product of the marginal likelihood and the posterior of ρ:

$$\begin{aligned} p(\rho|Y,Y_0)p(Y|Y_0) =& \frac{1}{2}I(-1<\rho<1)\left(\frac{\eta}{\eta+1}\right)^{\frac{k}{2}}|A|^{-\frac{N(T-1)}{2}} \\ & \Gamma\left[\frac{N(T-1)}{2}\right]T^{-\frac{N}{2}}(\pi)^{-\frac{N(T-1)}{2}}\exp\left(Nb(\rho)\right) \end{aligned} \tag{1.16}$$

To calculate the marginal likelihood, we can use the same numerical techniques as we calculate the posterior cumulative distribution function of ρ. By integrating ρ out of the product, we can obtain $p(Y|Y_0, M_i)$. If the total number of models is not large, say less than 2^{20}, it is possible to use any mainstream PC of today to calculate the marginal likelihood of all the models and then use (1.15) to find the posterior model probability for each of them. For large set of models beyond the computation power of today, we can use the method of Markov Chain Monte Carlo Model Composition (MC^3) developed by Madigan and York (1995).

1.4 Consistency in Model Selection

In this section, we show that in our setting, how the posterior model probability can lead us to locate the true model when the cross section sample size tends to infinity and certain regularity conditions are met. That is, if Y is indeed generated by some combination of the potential regressors in the linear model, the posterior model probability of this combination, which is obtained by integrating out ρ in (1.16), will tend to 1 when N tends to infinity. In the end of this section, we will also analyze whether the Bayesian information criterion (BIC) based on the biased MLE can lead to consistency in model selection.

In the simpler case in which the true value of ρ is known to be zero (i.e. static panel data models), the consistency in model selection easily follows from the analysis by Fernandez et al. (2001a). In our context, all we need to

ensure consistency is to set η as a function of N such that $\lim_{N\to\infty} \eta(N) = 0$. One possible choice could be $\eta = O(\frac{1}{N})$. As for the BIC, it is consistent in model selection for the static panel.

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

$$\frac{p(Y|Y_0, M_1)}{p(Y|Y_0, M_0)} = \left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}} \frac{\frac{1}{2}\int_{-1}^{1} \exp\left[Nb(\rho)\right] \left[\sum_{i=1}^{N} w_i' H w_i - \frac{1}{\eta+1}\sum_{i=1}^{N} w_i' H X_{i1} \left(\sum_{i=1}^{N} X_{i1}' H X_{i1}\right)^{-1} \sum_{i=1}^{N} X_{i1}' H w_i\right]^{-\frac{N(T-1)}{2}} d\rho}{\left[\sum_{i=1}^{N} y_i' H y_i - \frac{1}{\eta+1}\sum_{i=1}^{N} y_i' H X_{i0} \left(\sum_{i=1}^{N} X_{i0}' H X_{i0}\right)^{-1} \sum_{i=1}^{N} X_{i0}' H y_i\right]^{-\frac{N(T-1)}{2}}} \tag{1.17}$$

$$\frac{p(Y|Y_0, M_1)}{p(Y|Y_0, M_0)} = \left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}} \frac{\int_{-1}^{1} \exp\left[Nb(\rho)\right] \left[\sum_{i=1}^{N} w_i' H w_i - \frac{1}{\eta+1}\sum_{i=1}^{N} w_i' H X_{i1} \left(\sum_{i=1}^{N} X_{i1}' H X_{i1}\right)^{-1} \sum_{i=1}^{N} X_{i1}' H w_i\right]^{-\frac{N(T-1)}{2}} d\rho}{\int_{-1}^{1} \exp\left[Nb(\rho)\right] \left[\sum_{i=1}^{N} w_i' H w_i - \frac{1}{\eta+1}\sum_{i=1}^{N} w_i' H X_{i0} \left(\sum_{i=1}^{N} X_{i0}' H X_{i0}\right)^{-1} \sum_{i=1}^{N} X_{i0}' H w_i\right]^{-\frac{N(T-1)}{2}} d\rho} \tag{1.18}$$

where $w_i = y_i - y_{i\_}\rho$, k_1 and k_0 are the dimensions of X_{i1} and X_{i0}, which denote the regressors included under M_1 and M_0 respectively. To simplify (1.17) and (1.18), we need to simplify the integrals that appear in the nu-

merator and the denominator. Let us first define the following quantities:

$$
\begin{aligned}
a =& \sum_{i=1}^{N} y_{i\_}' H y_{i\_} - \frac{1}{\eta+1} \sum_{i=1}^{N} \left(y_{i\_}' H \underline{X}_i\right) \left(\sum_{i=1}^{N} \underline{X}_i' H \underline{X}_i\right)^{-1} \sum_{i=1}^{N} \left(\underline{X}_i' H y_{i\_}\right), \\
b =& \sum_{i=1}^{N} y_{i\_}' H y_i - \frac{1}{\eta+1} \sum_{i=1}^{N} \left(y_{i\_}' H \underline{X}_i\right) \left(\sum_{i=1}^{N} \underline{X}_i' H \underline{X}_i\right)^{-1} \sum_{i=1}^{N} \left(\underline{X}_i' H y_i\right), \\
c =& \sum_{i=1}^{N} y_i' H y_i - \frac{1}{\eta+1} \sum_{i=1}^{N} \left(y_i' H \underline{X}_i\right) \left(\sum_{i=1}^{N} \underline{X}_i' H \underline{X}_i\right)^{-1} \sum_{i=1}^{N} \left(\underline{X}_i' H y_i\right).
\end{aligned}
\tag{1.19}
$$

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

$$
\begin{aligned}
&\underset{N\to\infty}{plim} \frac{1}{N} a = \underline{a} \\
&\underset{N\to\infty}{plim} \frac{1}{N} b = \underline{a}(\underline{\rho} + NB) \\
&\underset{N\to\infty}{plim} \frac{1}{N} c = \underline{\rho}^2 \underline{a} + 2\underline{a}\underline{\rho} NB + h_3(\beta) + (T-1)\sigma^2 \\
&NB = \underset{N\to\infty}{plim} \frac{\left\{ \begin{array}{l} \sum_{i=1}^{N} y_{i\_}' H X_i \beta - \frac{1}{\eta+1} \sum_{i=1}^{N} y_{i\_}' H \underline{X}_i \left(\sum_{i=1}^{N} \underline{X}_i{}' H \underline{X}_i\right)^{-1} \sum_{i=1}^{N} \underline{X}_i' H X_i \beta + \\ \sum_{i=1}^{N} y_{i\_}' H u_i - \frac{1}{\eta+1} \sum_{i=1}^{N} y_{i\_}' H \underline{X}_i \left(\sum_{i=1}^{N} \underline{X}_i{}' H \underline{X}_i\right)^{-1} \sum_{i=1}^{N} \underline{X}_i{}' H u_i \end{array} \right\}}{\sum_{i=1}^{N} y_{i\_}' H y_{i\_} - \frac{1}{\eta+1} \sum_{i=1}^{N} \left(y_{i\_}' H \underline{X}_i\right) \left(\sum_{i=1}^{N} \underline{X}_i' H \underline{X}_i\right)^{-1} \sum_{i=1}^{N} \left(\underline{X}_i' H y_{i\_}\right)} \\
&\qquad = \frac{h_2(\beta, \underline{\rho}) - \sigma^2 h(\underline{\rho})}{\underline{a}}.
\end{aligned}
\tag{1.20}
$$

If the true model is either M_1 or M_0, we can show the conditions in Proposition 1.3 and 1.4 under which the Bayes factors in (1.17) and (1.18) can lead to the selection of the right model asymptotically.

Proposition 1.3. *When M_1 is the true model, i.e. $\underline{\rho} \neq 0$ and $X'_{i1}s$ are the true regressors to generate Y (which means X_{i0} is the same as $\underline{X_i}$ in (1.14)), as the cross section sample size increases, $\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}$ in (1.17) will tend to infinity if the following holds,*

$$z(\underline{\rho}) = b(\underline{\rho}) + \frac{T-1}{2} \ln \left[\frac{\underline{a}_{|M_0} \underline{\rho}^2 - 2\underline{\rho}\sigma^2 h(\underline{\rho}) + 2\underline{\rho} h_{2|M_0}(\beta, \underline{\rho}) + h_{3|M_0}(\beta) + (T-1)\sigma^2}{(T-1)\sigma^2} \right] > 0. \tag{1.21}$$

When M_0 is the true model, i.e. $X'_{i0}s$ are the true regressors to generate Y and $\underline{\rho} = 0$, as the cross section sample size increases, $\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}$ in (1.17) will tend to 0 if either of the following is satisfied:

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

$$f(\rho) = b(\rho) - \frac{T-1}{2} \ln \left(\rho^2 - 2\frac{b}{a}\rho + \frac{c}{a} \right) \tag{1.22}$$

and

$$b(\rho^*) + \frac{T-1}{2} \ln \frac{(T-1)\sigma^2}{d(\rho^*|M_1)} < 0 \tag{1.23}$$

where

$$\begin{aligned} d(\rho|M_i) =& \underline{a}_{|M_i}\rho^2 - 2\underline{a}_{|M_i}(\underline{\rho} + NB_{|M_i})\rho + \\ & \underline{a}_{|M_i}\underline{\rho}^2 + 2\underline{a}_{M_i}\underline{\rho} NB_{|M_i} + (T-1)\sigma^2 + h_{3|M_i}(\beta). \end{aligned} \tag{1.24}$$

2. *Though M_1 is misspecified, it can still lead to the consistent estimation of ρ, i.e. either (1.11) or (1.12) holds.*

Proposition 1.4. *When M_1 is the true model, as the cross section sample size increases, $\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}$ in (1.18) will tend to infinity if any of the following holds:*

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

$$b(\underline{\rho}) - b(\rho^*) + \frac{T-1}{2}\ln\frac{d(\rho^*|M_0)}{(T-1)\sigma^2} > 0 \tag{1.25}$$

2. *Either (1.11) or (1.12) holds.*

In addition to the Bayes factor calculated based on our parameterization of the fixed effect, we may be interested in knowing whether or not the Bayesian information criterion based on the biased MLE will lead to consistency in model selection. The results are shown in Proposition 1.5.

Proposition 1.5. *For the comparison of the two models in (1.17), when* M_1 *is the true model, BIC is consistent in model selection if the following condition is met,*

$$h_{3|M_0}(\beta) + \underline{a}_{|M_0}\underline{\rho}^2 + 2\underline{\rho}h_{2|M_0}(\beta,\underline{\rho}) - 2\underline{\rho}\sigma^2 h(\underline{\rho}) + \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}_{|M_1}} > 0 \tag{1.26}$$

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

$$\frac{\left[h_{2|M_1}(\beta,0) - \sigma^2\frac{T-1}{T}\right]^2}{\underline{a}_{|M_1}} - h_{3|M_1}(\beta) < 0 \tag{1.27}$$

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

For the comparison of the two models in (1.18), when M_1 *is the true model, BIC is consistent in model selection if the following holds*

$$\underline{a}_{|M_1}\underline{a}_{|M_0}h_{3|M_0}(\beta) + \underline{a}_{|M_0}\sigma^4 h^2(\underline{\rho}) - \underline{a}_{|M_1}\left[h_{2|M_0}(\beta,\underline{\rho}) - \sigma^2 h(\underline{\rho})\right]^2 > 0 \tag{1.28}$$

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

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

1.5 Motivations of Bayesian Model Averaging

Our method allows us to compare the goodness of fit of different model specifications. However, as Raftery and Zheng (2003) and Yuan and Yang (2005) point out, if there is substantial model uncertainty, model averaging is more preferable than model selection. In regard to our empirical application of finance and growth, the data set we have is relatively small (such as the one in Section 1.7, with cross section sample size equal to 40), which implies model uncertainty for estimation. When we want to study the relationship between economic growth and other variables from the panel data, it should be more appropriate to consider different model specifications than just drawing our conclusions based on a single model so that we can reduce the estimation risk in the presence of substantial model uncertainty. This point will be made more clear in the subsequent sections. At the moment, we will just briefly talk about the Bayesian model averaging (BMA) approach.

From different model specifications, we can have different estimates of θ.[10] Essentially, BMA consists in mixing the posterior distributions of θ from all different models according to their posterior model probabilities in (1.15). Inference about θ is drawn from its posterior distribution unconditional on the model space, which takes the following form.

$$p(\theta|Y, Y_0) = \sum_{i}^{2^{K+1}} p(\theta|Y, Y_0, M_i)p(M_i|Y, Y_0) \tag{1.29}$$

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

$$\sum_{i}^{2^{K+1}} I(\theta_j \in M_i)p(M_i|Y, Y_0). \tag{1.30}$$

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

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

1.6 Simulation Studies

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

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

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

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

In Proposition 1.2 we show that we cannot have a consistent estimate of ρ when neither (1.11) nor (1.12) holds. We want to check whether we can still select the right model asymptotically using Lancaster's transformation of the fixed effect. That is why in step 1 we want to add collinearity and serial correlation to our regressors. To achieve this, we first make each two neighboring period observations correlated with each other as follows,

$$x_{t,s} = s_{t-1}x_{t-1,s} + \bar{s}_t x_{t,ns}, \tag{1.31}$$

where $x_{t,ns}$ has no serial correlation and is generated from the i.i.d. uniform distribution U[-4,4]. We set $s_{t-1} = \frac{s'_{t-1}}{\sqrt{s'^2_{t-1}+s'^2_t}}$ and $\bar{s}_t = \frac{s'_t}{\sqrt{s'^2_{t-1}+s'^2_t}}$. For s'_{t-1}

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

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

$$S = \begin{pmatrix} 1 & s_1 & \cdots & \prod_{i=1}^{T-1} s_i \\ s_1 & 1 & \cdots & \prod_{i=2}^{T-1} s_i \\ s_2 s_1 & s_2 & \cdots & \prod_{i=3}^{T-1} s_i \\ \cdots & \cdots & \cdots & \cdots \\ \prod_{i=1}^{T-1} s_i & \prod_{i=2}^{T-1} s_i & \cdots & 1 \end{pmatrix} \tag{1.32}$$

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

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

$$X_{j,c} = \sum_{i=1}^{K} q_{j,i} X_{i,nc} \qquad j = 1, 2, \ldots, K \tag{1.33}$$

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

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

[13] We choose T to be 100 for the results to be presented later. We have also used small value of T to generate the data, all the results are similar and neither (1.21) nor (1.25) is violated. These results are available upon request from the author.

$$corr(X_{t,k}, X_{t',k'}) = S(t,t')\sum_{i=1}^{K} q_{k,i}q_{k',i} \qquad t = 1,2,\ldots,T \quad k = 1,2,\ldots,K. \tag{1.34}$$

where $S(t,t')$ denote the (t,t') element in S and K is the potential number of regressors. Through such data generating mechanism we can explicitly calculate the values of $h_2(\beta,\underline{\rho})$ and $h_3(\beta)$, $\underline{a}$ and NB in (1.14) and (1.20) respectively. Hence we can check whether condition (1.21) and (1.25) are violated or not when there is an error in our model selection based on posterior model probability.

We run the experiment for 200 times. At first we set $\eta = \frac{1}{N}$ and $\sigma^2 = 1$. The results are presented in Table 1.1. The ER (error rate) column tells us how often the model with the highest posterior model probability ends up being different from the true model. When the cross section sample size is 40 (the same as our application later), the Bayes factor criterion fails to pick up the true model by 86 out of 200 simulations. However, we can see that the error rate tends to decrease with cross section sample size, which is a sign of model selection consistency. One thing to note is that we generate β from $U[-2,2]$. When the values of some elements in β are very close to zero, it is virtually equivalent to the case when the true model does not include the corresponding regressors. In Table 1.1, the column "nest" denotes how often the top model is nested inside the true model (including the case when the top model is the true model). We can find that this number generally rises with cross section sample size. The column "nouni" checks among the errors from the Bayes factor criterion how many of them is related to the fact that either there is no solution or there are more than one solutions in (−1,1) for the equation $\underset{N\to\infty}{plim}\, f'(\rho) = 0$, where $f(\rho)$ is defined in (1.22). We show in the proof of Proposition 1.4 that when $\underset{N\to\infty}{plim}\, f(\rho)$ does not have a global maximum in the stationary region, we cannot use Laplace method to approximate the integral(s) in the Bayes factor. Hence the condition in (1.21), (1.23) and (1.25) do not hold. Under our simulating data generating mechanism, such situations do not exist.

The columns of "no(1.21)", "no(1.23)" and "no(1.25)" denote the error rates with the violation of (1.21), (1.23) and (1.25) respectively. We can see that the numbers of the columns are all zeros, which means all our errors are fixable with a large cross section sample size. The columns of "topprob" and "top10prob" are the average of the posterior model probabilities of the top model and the sum of the top ten models in the simulation. If these two numbers are far below 1, it is a sign of model uncertainty. As the cross section sample size increases, model uncertainty diminishes. If we raise the variance of the disturbance, model uncertainty will increase. The results are shown in Table 1.2 where we set the variance of the disturbance to 4. Comparing Table 1.2 to Table 1.1, we can see that the error rate is higher and the rest of the three columns are generally smaller for a particular cross section sample size. As for the model selection performance of BIC based on the biased maximum likelihood estimates, we list the results in Table 1.3 and Table 1.4. We can see that the BIC performance is much worse than our Bayes factor method. The error rates stay above 50% for different cross section sample sizes. Even for $N = 1000$, there is not much improvement. In addition to the error rate, the top model is not very often nested inside the true model as compared with the Bayes factor method. Again, it does not improve much with the cross section sample size. Moreover, the column headed with "no(1.27)" shows how many errors violate condition (1.27). Such errors are not fixable even if we have infinite cross section sample size according to Proposition 1.5. Note that around 50% of the true models do not have the lag term of the dependent variable under our simulation set-up. Also note that under our data generating scheme, we can be almost sure that $\underline{\rho} + NB = 0$ will not occur. Hence condition (1.26) will almost surely not be violated. It could be true to say that the error rate for the BIC would approach 50% in the limit since it is always possible for condition (1.27) to be violated while condition (1.26) and (1.28) hold. When the true model does not have a lag term of the dependent variable as the regressor, it is always possible to find a candidate model with both the lag term and exactly the same set of exogenous regressors as the true model such that condition (1.27) will be violated. When we compare them, we will choose

the candidate model over the true model as the cross section sample size increases. Also we could expect that the percentage under $no(1.27)$ in Table 1.3 and 1.4 should rise with cross section sample size.

Table 1.1: Simulation results when $\sigma^2 = 1$

N	ER	nest	topprob	top10prob	no(1.21)	no(1.23)	no(1.25)	nouni
40	0.40	0.83	0.38	0.85	0.00	0.00	0.00	0.00
100	0.29	0.86	0.56	0.94	0.00	0.00	0.00	0.00
200	0.31	0.88	0.62	0.96	0.00	0.00	0.00	0.00
500	0.14	0.94	0.74	0.99	0.00	0.00	0.00	0.00
1000	0.10	0.97	0.81	0.99	0.00	0.00	0.00	0.00

Table 1.2: Simulation results when $\sigma^2 = 4$

N	ER	nest	topprob	top10prob	no(1.21)	no(1.23)	no(1.25)	nouni
40	0.61	0.77	0.32	0.80	0.00	0.00	0.00	0.00
100	0.43	0.86	0.50	0.93	0.00	0.00	0.00	0.00
200	0.36	0.88	0.58	0.95	0.00	0.00	0.00	0.00
500	0.28	0.92	0.69	0.98	0.00	0.00	0.00	0.00
1000	0.16	0.96	0.78	0.99	0.00	0.00	0.00	0.00

Table 1.3: Simulation results for BIC when $\sigma^2 = 1$

N	error rate	nest	no(1.26)	no(1.27)	no(1.28)
40	0.78	0.34	0.00	0.46	0.00
100	0.69	0.42	0.00	0.60	0.00
200	0.69	0.41	0.00	0.51	0.00
500	0.54	0.51	0.00	0.65	0.00
1000	0.58	0.47	0.00	0.61	0.00

Judging from the previous simulation results, we can find that if we simply select the model with the highest model probability to provide estimates of our interest, chances are high that the model selected is not the true model. Note that the top model probability for $N = 40$ is about 32% while it is about 78% for $N = 1000$ when we set $\sigma^2 = 4$. To account for such model uncertainty, we recommend averaging the estimates from every model. We argue that BMA can reduce our estimation risk when there is substantial model uncertainty. To illustrate this, next we carry out another

Table 1.4: Simulation results for BIC when $\sigma^2 = 4$

N	**error rate**	**nest**	**no(1.26)**	**no(1.27)**	**no(1.28)**
40	0.88	0.34	0.00	0.28	0.00
100	0.77	0.41	0.00	0.36	0.00
200	0.74	0.38	0.00	0.34	0.00
500	0.65	0.43	0.00	0.40	0.00
1000	0.55	0.425	0.00	0.65	0.00

simulation, in which we set the β's to fixed values along with ρ (we set it to 0.9 as in our previous simulation). Then we use the posterior means to estimate these values. Table 1.5 shows the root mean squared errors (RMSE) from different point estimators based on 200 iterations with the cross section sample size (N) as 40. The true values of ρ and β's are shown under the column "TRUE", where the first number is the value of ρ. The column "TOP" shows the RMSE resulting from the posterior mean estimator of the top model, which has the highest posterior model probability, while the column "BMA" uses the posterior mean in (1.29). To evaluate the significance of a regressor coefficient, we calculate the sum of the posterior model probabilities of all the models which include the corresponding regressor. If the inclusion probability for a regressor is too low, we may be better off by viewing the coefficient for this regressor as zero. In Table 1.5, we try to give some ideas on how to interpret such inclusion probabilities which we will use in our application later. The RMSE in the columns headed with a percentage number are derived based on certain inclusion probability criterion. For each simulated data set, if the inclusion probability for a regressor is lower than the percentage number on top of the column, we will simply use zero as its point estimate. In the last row of Table 1.5, we sum up the RMSE in each column. We can see that the overall performances of BMA and various inclusion probability criteria are all better than that of the top model criterion in terms of smaller total RMSE. Such performance seems to fare best when we set our inclusion probability to 50%. We can also see that higher inclusion probability criterion tends to give us smaller RMSE when the true value of the parameter is exactly zero and higher RMSE when the

true value is not zero while the BMA seems to give us a safer option for almost all the parameter estimates.

Table 1.5: The RMSE of the point estimates when N=40 and $\sigma^2 = 1$

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

To add more insight into how to use inclusion probability to determine the significance of a regressor coefficient, we presents the error rates of including the wrong regressor due to different inclusion probability criteria[14] in Table 1.6. We can see that the overall error rates based on the 10% criterion is the highest. All the errors are from those parameters whose values are actually zeros. Again, the 50% criterion shows reasonably good performance, although the 60% criterion is slightly better. One thing to note is that the top model criterion has smaller overall error rate than nearly all inclusion probability criteria. Hence it seems to be a useful tool in terms of making the decision on whether to include a particular regressor or not.

Next we increase model uncertainty by increasing the variance of the disturbance to 4. Point estimate performances based on different criteria are shown in Table 1.7. When the model uncertainty is larger, the advantage of the averaging estimators becomes more obvious. Though their performances are quite alike, the 50% inclusion probability criterion still gives us the smallest overall RMSE. The error rates of whether to include a regressor

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

Table 1.6: The error rates of whether to include a regressor when N=40 and $\sigma^2 = 1$

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

are presented in Table 1.8. Now none of the inclusion probability criteria can have smaller overall error rates than that of the top model criterion. It seems that the criteria using inclusion probability above 40% (along with the top model criterion) have problems with the parameter whose value is 0.1 (close to zero).

Table 1.7: The RMSE when N=40 and $\sigma^2 = 4$

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

When we set the variance of the disturbance to 1 and the cross section sample size to 1000 such that the model uncertainty is not substantial, the performance of different criteria will be similar, which is shown in Table 1.9

Table 1.8: The error rates of whether to include a regressor when N=40 and $\sigma^2 = 4$

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

and Table 1.10. However, the averaging estimators still fare slightly better in point estimates and the top model criterion is reasonably good in deciding whether or not to include a variable.

Table 1.9: The RMSE of the point estimates when N=1000 and $\sigma^2 = 1$

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

In terms of the point estimation, BMA seems to be more preferable than simply using the estimates from the top model since it takes account of model uncertainty explicitly. Moreover, in Bayesian Econometrics we have many sensible tools to help us understand our data. As will be shown in the application later, our inference is based on the posterior distribution of the

Table 1.10: The error rates of whether to include a regressor when N=1000 and $\sigma^2 = 1$

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

parameter unconditional on the model space, which gives us information on what we are more sure of and of what we are less sure.

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

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

In our context, K is the number of potential regressors plus 1 (the lag term). We can see that Fernandez et al. (2001a) basically recommend a more non-informative prior. They argue that although the second prior is inconsistent[15], it may perform better than the first one for small samples.

[15]The inconsistency in model selection under the second prior here means the posterior model probability of the true model will not tend to 1 with increasing sample size. However, when the true model does not have a lag term as regressor, the Bayes factor under the second prior will still favour the true model, i.e. the true model still has higher model probability than the other models. For more details, see Fernandez et al. (2001a). When the model has a lag term, as long as relevant conditions in Proposition 1.3 and 1.4

In contrast to Table 1.1 and Table 1.2, Table 1.11 and Table 1.12 present the results under the second prior. It suggests that when $N = 40$ (the cross section sample size in our application), the second prior seems to do much better for smaller disturbance variance in terms of whether the true model is nested inside the top model and it also has higher posterior top model probability. However, it fares more or less the same as the first prior for bigger disturbance variance. As the sample size increases, the improvement of the second prior does not seem to be as big as that under the first prior. For large sample size (such as $N = 1000$), the first prior is more preferable than the second.

Table 1.11: Simulation results when the variance of the disturbance is 1 and under the prior $\eta = \frac{1}{K^2}$

N	ER	nest	topprob	top10prob	no(1.21)	no(1.23)	no(1.25)	nouni
40	0.43	0.9	0.45	0.90	0	0	0	0
100	0.3	0.88	0.46	0.91	0	0	0	0
200	0.27	0.9	0.48	0.92	0	0	0	0
500	0.15	0.92	0.53	0.94	0	0	0	0
1000	0.17	0.9	0.49	0.92	0	0	0	0

Table 1.12: Simulation results when the variance of the disturbance is 4 and under the prior $\eta = \frac{1}{K^2}$

N	ER	nest	topprob	top10prob	no(1.21)	no(1.23)	no(1.25)	nouni
40	0.59	0.78	0.40	0.85	0	0	0	0
100	0.48	0.8	0.42	0.88	0	0	0	0
200	0.34	0.88	0.44	0.89	0	0	0	0
500	0.39	0.77	0.47	0.92	0	0	0	0
1000	0.2	0.89	0.47	0.91	0	0	0	0

Results on the point estimation performance under the second prior are presented in Table 1.13, Table 1.14 and Table 1.15. In comparison with Table 1.5 to Table 1.9, the performance under the second prior does not seem to differ much, though when the sample size is small and model uncertainty is large, for the column of the top model criterion, the second prior seems

hold, consistency in model selection will follow. That is why we can still see improved performance under the second prior over larger sample size.

to do better than the first prior. But all the averaging estimators still tend to dominate the top model criterion for small samples. For large sample size, such dominance of averaging estimators seems to diminish and their performances are quite close to that from the top model. In Table 1.15, under the second prior, the BMA estimates even have higher RMSE than the top model criterion. Again, the first prior is more preferable than the second for large samples in terms of smaller RMSE from the averaging estimators.

Table 1.13: The RMSE of the point estimates when N=40 and $\sigma^2 = 1$ under the prior $\eta = \frac{1}{K^2}$

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

Table 1.14: The RMSE of the point estimates when N=40 and $\sigma^2 = 4$ under the prior $\eta = \frac{1}{K^2}$

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

Table 1.15: The RMSE of the point estimates when N=1000 and $\sigma^2 = 1$ under the prior $\eta = \frac{1}{K^2}$

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

1.7 An Application Example of Financial Development and Economic Growth

The model in our application is slightly different from (1.1) and it takes the following form.

$$\begin{aligned} y_{i,t} - y_{i,t-1} &= f_i + y_{i,t-1}\rho + x'_{i,t}\beta + u_{it}, \\ i &= 1 \dots N, \quad t = 1 \dots T. \end{aligned} \tag{1.35}$$

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

The data we use are taken from Beck and Levine (2004) and are available

from Levine's website. There are altogether 40 (N) countries and the panel covers the period from 1976 to 1998. Eight potential explanatory variables ($x_{i,t}$) have been proposed in the literature. Details of the variables can be found in Table 16. Here we just follow the practice of Beck and Levine (2004) on how the variables enter equation (1.35). Our focus is on the variables measuring the development of stock market and banking sector. We also include three dummy variables for each period as our potential explanatory variables.[16] Hence the total number of possible regressors is 11. Since we are studying the long run relationship between economic growth and other economic variables, we average the data over every five years. Due to missing data and the dynamic nature of our model, we can only use 143 observations in the panel. Since it is an unbalanced panel, i.e. not every country in the panel has the same number of observations (T), we have to replace some quantities that appear in the previous sections as the following,

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

There are 4,096 possible model specifications in total. Here in Table 1.17, we just present the top ten models with the highest posterior model probabilities. We can see that the top model is nested in most of the top ten models and it just accounts for around 6.4% posterior model probability while the model probability of the top 10 models in total is about 30%. The result is quite different from the simulation studies in the previous section when we have a true model to generate our data. We find that in simulation the top model alone (in many cases, the true one) usually accounts for above 30%. This confirms the fact that there is substantial model uncertainty

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

Table 1.16: Details of the explanatory variables

1. START: the per capita GDP at the starting year of the five years. It enters the equation of (1.35) in natural log.
2. PRIV: the measure of bank development, calculated from bank claims on the private sector by deposit money banks divided by GDP. It enters the equation in log.
3. PI: the inflation rate. It enters the equation as log(1+PI).
4. GOV: the ratio of government expenditure to GDP. It enters the equation in log.
5. TRADE: the shares of exports and imports to GDP. It enters the equation in log.
6. TOR: the measure of stock market development, which equals the value of traded shares on domestic exchanges divided by the total value of listed shares. It enters the equation in log.
7. BMP: the black market premium. It enters the equation as log(1+BMP).
8. SCHOOL: average years of schooling. It enters the equation as log(1+SCHOOL).

in our data. To study the relationship of economic growth and different economic variables, BMA should be a more preferable approach.

Table 1.17: Posterior Model Probabilities of the Top Ten Models

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

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

The BMA point estimates of the slope parameters from the posterior distribution in equation (1.8) are shown in Table (1.18), where we omit the results for the dummy variables. The estimates are based on 10,000 draws from the posterior model and parameter space. The column headed by "slope" presents the posterior mean of β in (1.35). The "NSE" column is the numerical standard error, which is a measure of accuracy of our simulated calculations. The true posterior means with around 95% confidence should lie in the range of [estimate$-1.96NSE$,estimate$+1.96NSE$] due to the central limit theorem. The inclusion probability is calculated as the sum of the model probabilities from the models that include the regressor. Finally, $prob < 0$ is the cumulative posterior probability of the parameter less than 0. It is based on the mixture of the models that include the regressor and can be viewed as how far away the posterior distribution is from 0. If our point estimate is negative (positive) and its posterior distribution is far away from 0, we would expect $prob < 0$ to be close to 1 (0). Not surprisingly, the regressors with the highest inclusion probability are the initial GDP and

the lagged GDP, which are closely related to our dependent variable, the per capita GDP growth. The turnover of stock market also has high inclusion probability, about 78% and it is positively related with economic growth and its posterior mean is around 1.28. This confirms the finding by Beck and Levine (2004), whose GMM point estimates of stock market turnover are significant and they range from 0.95 to 1.7 under the inclusion of different sets of exogenous variables. However, it is a surprise to see that bank credit to private sector, which is a measure of bank development, has the lowest inclusion probability among all the regressors and its point estimate is quite close to 0. Moreover, the column of $prob < 0$ tells us that the posterior distribution of stock market turnover is far away from 0 while the posterior distribution of bank credit has its center near 0. It seems that bank development is not that important for economic growth. This finding seems to contradict the results based on the GMM approach in Beck and Levine (2004).

Table 1.18: Estimates of the Slope Parameters

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

To verify our results, in Table 1.19 and Table 1.20 we present the highest (marginal) posterior probability intervals (HPDI) of bank private credit and stock market turnover respectively. Such intervals are calculated by a kernel smoothing package (ksdensity.m) in MatLab®. The package uses a normal kernel function to fit to certain number of draws from the parameter's posterior distribution. For bank private credit, the number of draws is 1,414 and the one for stock market turnover is 7,794. Note that the results are based

on the models which include the regressor. The HPDI results confirm what we found previously, i.e. the posterior distribution for stock market turnover is far different from zero while bank private credit is not. We may conclude that stock market development is more important to economic growth than bank development based on our dataset.

Table 1.19: The highest posterior density intervals for the private credit

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

Table 1.20: The highest posterior density intervals for the stock market turnover

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

Next from Table 1.21 to Table 1.24, we present the results under the g-prior $\eta = \frac{1}{K^2}$, where K is the number of potential explanatory variables plus one (the lag term of the dependent variable). As is shown in our simulation, this prior may have better performance when the cross section sample size is as small as in our application. We can see that the second prior mainly reconfirms our previous results. First there is substantial model uncertainty as shown by the top model probability.[17] Second the stock market development is more significant and the bank private credit is more insignificant under the second prior than the first prior. One difference under the second prior is that trade seems more important. The top model now consists of stock market development and trade.

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

Table 1.21: Posterior Model Probabilities of the Top Ten Models under the prior $\eta = \frac{1}{K^2}$

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

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

Table 1.22: Estimates of the Slope Parameters under the prior $\eta = \frac{1}{K^2}$

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

Table 1.23: The highest posterior density intervals for the private credit under the prior $\eta = \frac{1}{K^2}$

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

Table 1.24: The highest posterior density intervals for the stock market turnover under the prior $\eta = \frac{1}{K^2}$

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

1.8 Conclusion

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

with the highest posterior model probability. Using the method developed, we study the connection of stock market and bank development with economic growth. Consistent with the results from the classical approach, our finding suggests that stock market development is positively correlated with economic growth. However, the effect of bank development on economic growth is not significant, which contradicts the classical results. In our framework, we have restricted our attention to stationary data and strictly exogenous explanatory variables. Future research to relax such restrictions could be promising.

1.9 Appendix

1.9.1 The Informatation Orthogonal Method

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

$$\begin{aligned} & E\left(\frac{\partial^2 \ln p(y_i|\theta, f_i)}{\partial f_i \partial \theta}\right)\Bigg|_{f_i=f_{i,true},\theta=\theta_{true}} \\ & = \int \frac{\partial^2 \ln p(y_i|\theta, f_i)}{\partial f_i \partial \theta} p\left(y_i|\theta, f_i\right)\, dY \Bigg|_{i=f_{i,true},\theta=\theta_{true}} = 0 \end{aligned} \tag{1.36}$$

Lancaster (2002) showed that if (1.36) evaluated at the true values of f_i and θ is satisfied, the mode of the marginal posterior of θ ($p(\theta|y) \propto \int \prod_{i=1}^{N} p(y_i|\theta, f_i) p(f_i|\theta) d\, f$), which is obtained by integrating out f_i with repect to the flat prior, $p(f_i\theta) \propto 1$[18], is a consistent estimator. When the the original incidental parameter is not information orthogonal to the common parameter, Lancaster (2002) suggested we reparameterize $f_i = (g_i, \theta)$ such that the new incidental pa-

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

rameter g_i is information orthogonal to θ. To find the new parameterization is equivalent to finding the solution for the following differential equation,

$$\frac{\partial f_i}{\partial \theta} = -\left(E_Y(\frac{\partial^2 \ln p(y_i|\theta, f_i)}{\partial f_i^2})\right)^{-1} E_Y\left(\frac{\partial^2 \ln p(y_i|\theta, f_i)}{\partial f_i \partial \theta}\right) \tag{1.37}$$

However, when θ is a vector, say $\theta = (\theta_1, \theta_2)$, there is no guarantee that (1.37) has a solution since the compatability condition $\frac{\partial^2 f_i}{\partial \theta_1 \partial \theta_2} = \frac{\partial^2 f_i}{\partial \theta_2 \partial \theta_1}$ may not be satisfied. For the linear AR(1) panel model, Lancaster (2002) showed that an information orthogonal parameterization can not be found.

1.9.2 Proof of Proposition 1.1

Denote y_i as $[y_{i,1}, y_{i,2}, \ldots, y_{i,T}]'$ and $y_{i\_}$ as $[y_{i,0}, y_{i,1}, \ldots, y_{i,T-1}]'$. We can rewrite model (1.1) as

$$y_i = f_i \iota + y_{i\_}\rho + X_i\beta + u_i \tag{1.38}$$

$$y_{i\_} = f_i c_1 + y_{i,0} c_2 + C X_i \beta + C u_i \tag{1.39}$$

where

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

and ι is a $T \times 1$ vector of ones. Note that $h(\rho)$ defined in 1.13 is equal to $\frac{1}{T}\iota' c_1 = -trace(C'H)$.

Lancaster (2002) finds the appropriate reparameterization of the fixed effect parameter by drawing analogy from two simpler cases, i.e. when the model only contains either the lag term of the dependent variable or the exogenous regressors. Here we provide a slightly different derivation of the reparameterization. In brief, we attempt to find a correction function attached to the marginal posterior density of ρ such that the mode of the

marginal posterior is a consistent estimator for ρ. The solution turns out to be the same as Lancaster's. The derivation strategy adopted here is also needed for the proof of Proposition 1.2. To obtain such a correction function, let us first reparameterize the fixed effect as

$$f_i = g_i \underline{r}(\rho) - \frac{1}{T}\iota' X_i \beta \tag{1.40}$$

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

$$y_i = g_i \underline{r}(\rho)\iota + y_{i\_}\rho + HX_i\beta + u_i, \tag{1.41}$$

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

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

$$\begin{aligned} p(\theta)p\left(Y|\theta, Y_0\right) = p(\beta|\sigma^2)\frac{1}{2}I(-1 \leq \rho \leq 1)(2\pi)^{-\frac{TN}{2}}\sigma^{2(-\frac{NT+2}{2})} \\ \prod_{i=1}^{N}\exp\left\{-\frac{1}{2\sigma^2}[w_i - g_i\underline{r}(\rho)\iota - HX_i\beta]'\left[w_i - g_i\underline{r}(\rho)\iota - HX_i\beta\right]\right\}, \end{aligned} \tag{1.42}$$

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

Next we derive the posterior distribution of g_i. Note that $\iota' H = 0$ so we can rewrite equation (1.42) as

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

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

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

or equivalently

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

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

$$\begin{aligned} p(\theta)p\left(Y|\theta, Y_0\right) = p(\beta|\sigma^2)\frac{1}{2}I(-1 \leq \rho \leq 1)(2\pi)^{-\frac{TN}{2}}\sigma^{2(-\frac{NT+2}{2})} \\ \prod_{i=1}^{N}\exp\left\{-\frac{\underline{r}^2(\rho)}{2\frac{\sigma^2}{T}}[g_i - \frac{\iota'(y_i - y_{i\_}\rho)}{T\underline{r}(\rho)}]^2\right\} \\ \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - y_{i\_}\rho - X_i\beta)'H(y_i - y_{i\_}\rho - X_i\beta)\right] \end{aligned} \tag{1.43}$$

Remember $p(\beta|\sigma^2)$ does not involve parameters other than σ^2. Moreover, since we ignore the distribution of Y_0 and assume the prior of θ is independent of it, from (1.43) it is clear that the posterior distribution of g_i conditional on $y_{i,0}$, σ^2 and ρ is i.i.d. normal as in (1.7).

Next we go on to derive the posterior distributions for β and σ^2. First

we can integrate out g in equation (1.43) to obtain

$$\begin{aligned} p(\rho,\beta,\sigma^2,Y|Y_0) &= p(\rho,\beta,\sigma^2|Y,Y_0)p(Y|Y_0) \\ &= p(\beta|\sigma^2)\frac{1}{2}I(-1<\rho\leq 1)T^{-\frac{N}{2}}(2\pi)^{-\frac{N(T-1)}{2}}\sigma^{2\left[-\frac{N(T-1)+2}{2}\right]} \\ &\quad \underline{r}^{-N}(\rho)\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-y_{i\_}\rho-X_i\beta)'H(y_i-y_{i\_}\rho-X_i\beta)\right]. \end{aligned} \tag{1.44}$$

Let us define a new function $r(\rho) = \underline{r}^{-N}(\rho)$, $\tilde{w}_i = H(y_i - y_{i\_}\rho)$ and $\tilde{X}_i = HX_i$. Incorporating the prior of β in (1.6) we can rewrite equation (1.44) as

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

Then completing the square of β yields

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

We can see that the posterior kernel for β is normal as in (1.8) and hence we can integrate it out. The posterior distribution for ρ and σ^2 is

$$
\begin{aligned}
p(\rho,\sigma^2|Y,Y_0)p(Y|Y_0) =&\frac{1}{2}I(-1\leq\rho\leq 1)\left(\frac{\eta}{\eta+1}\right)^{\frac{k}{2}}T^{-\frac{N}{2}}(2\pi)^{-\frac{N(T-1)}{2}}\\
&\sigma^{2\left[-\frac{N(T-1)+2}{2}\right]}r(\rho)\exp\left\{-\frac{1}{2\sigma^2}[\sum_{i=1}^{N}\tilde{w}_i{}'\tilde{w}_i\right.\\
&\left.-\frac{1}{\eta+1}\sum_{i=1}^{N}\tilde{w}_i'\tilde{X}_i\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i]\right\}
\end{aligned}
\tag{1.45}
$$

It is also clear from equation (1.45) that conditional on ρ, σ^2 follows an inverted gamma distribution with mean $\frac{A}{N(T-1)-2}$ and degrees of freedom $N(T-1)$ as in (1.9).

Now we can integrate out σ^2 to obtain the posterior distribution of ρ as in (1.10). Another way to write the posterior of ρ is as follows

$$
p(\rho|Y,Y_0)\propto I(-1<\rho<1)r(\rho)t(\frac{b}{a},\frac{c}{av}-\frac{b^2}{a^2v},v) \tag{1.46}
$$

where

$$
\begin{aligned}
a &= \sum_{i=1}^{N} y_{i\_}' H y_{i\_} - \frac{1}{\eta+1} \sum_{i=1}^{N} \left(y_{i\_}' H X_i\right) \left(\sum_{i=1}^{N} X_i' H X_i\right)^{-1} \sum_{i=1}^{N} \left(X_i' H y_{i\_}\right) \\
b &= \sum_{i=1}^{N} y_{i\_}' H y_i - \frac{1}{\eta+1} \sum_{i=1}^{N} \left(y_{i\_}' H X_i\right) \left(\sum_{i=1}^{N} X_i' H X_i\right)^{-1} \sum_{i=1}^{N} \left(X_i' H y_i\right) \qquad (1.47) \\
c &= \sum_{i=1}^{N} y_i' H y_i - \frac{1}{\eta+1} \sum_{i=1}^{N} \left(y_i' H X_i\right) \left(\sum_{i=1}^{N} X_i' H X_i\right)^{-1} \sum_{i=1}^{N} \left(X_i' H y_i\right).
\end{aligned}
$$

Equation (1.46) tells us that when ρ is in the stationary region, its kernel of the posterior distribution can be viewed as the product of $r(\rho)$ and the t distribution with the mean parameter $\frac{b}{a}$ and the variance parameter $\frac{c}{av} - \frac{b^2}{a^2 v}$, where $v = N(T-1)-1$ is the degrees of freedom. Note that $\frac{b}{a}$ is the within-group estimator, which we could obtain if we operate on the first difference data and adopt a non-informative prior for ρ by assuming our model is stationary ($|\rho| < 1$) and the regressors are exogenous. This estimator is inconsistent and the bias is a function of the true value of ρ. If our posterior estimate of ρ is consistent, $r(\rho)$ should act as the correction term to the bias. Let us denote NB as the bias and $\underline{\rho}$ as the true value of the parameter. We

will have the following[19]

$$
\begin{aligned}
&\underset{N\to\infty}{plim}\frac{b}{a} = \underline{\rho} + NB \\
&\underset{N\to\infty}{plim}\frac{1}{N}a = \underline{a} \\
&NB = \underset{N\to\infty}{plim}\frac{\sum_{i=1}^{N} y_{i\_}' H u_i - \sum_{i=1}^{N} y_{i\_}' H X_i \left(\sum_{i=1}^{N} X_i' H X_i\right)^{-1} \sum_{i=1}^{N} X_i' H u_i}{\sum_{i=1}^{N} y_{i\_}' H y_{i\_} - \sum_{i=1}^{N} y_{i\_}' H X_i \left(\sum_{i=1}^{N} X_i' H X_i\right)^{-1} \sum_{i=1}^{N} X_i' H y_{i\_}} \\
&\qquad = -\frac{\sigma^2 h(\underline{\rho})}{\underline{a}}, \\
&\underset{N\to\infty}{plim}\frac{1}{N}\left[\sum_{i=1}^{N} y_{i\_}' H u_i - \sum_{i=1}^{N} y_{i\_}' H X_i \left(\sum_{i=1}^{N} X_i' H X_i\right)^{-1} \sum_{i=1}^{N} X_i' H u_i\right] = -\sigma^2 h(\underline{\rho}), \\
&\underset{N\to\infty}{plim}\frac{1}{N}\left[\sum_{i=1}^{N} u_i' H u_i - \sum_{i=1}^{N} u_i' H X_i \left(\sum_{i=1}^{N} X_i' H X_i\right)^{-1} \sum_{i=1}^{N} X_i' H u_i\right] = (T-1)\sigma^2,
\end{aligned}
\tag{1.48}
$$

where the function $h(\cdot)$ is given in (1.13). So we can obtain

$$
\underset{N\to\infty}{plim}\frac{c}{a} = cta = \underline{\rho}^2 + 2\underline{\rho} NB + \frac{(T-1)\sigma^2}{\underline{a}} \tag{1.49}
$$

.

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

$$
p(\rho|Y, Y_0) \propto I(-1 < \rho < 1) r(\rho) t(\underline{\rho} + NB, \frac{1}{v}\left(cta - (\underline{\rho} + NB)^2\right), v) \tag{1.50}
$$

Recall that $v = N(T-1) - 1$. If our estimate from the above kernel is consistent, the posterior distribution of ρ should become a spike at the true value of ρ (the mode of the kernel). The mode of the kernel in (1.50) can be obtained from the following first order condition,

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

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

So we will have

$$\frac{1}{N}\frac{d\,r(\rho)}{d\,\rho}=(T-1)\frac{\rho-\underline{\rho}-NB}{cta-(\underline{\rho}+NB)^2+(\rho-\underline{\rho}-NB)^2}. \tag{1.51}$$

If our specification of $r(\rho)$ leads to consistent estimator, the true value $\underline{\rho}$ should be a solution for the above differential equation. By using (1.48), we can obtain

$$Nh(\underline{\rho})d\,\underline{\rho}=\frac{1}{r(\underline{\rho})}d\,r. \tag{1.52}$$

Finally by using (1.13), we will have

$$\begin{aligned} r(\rho)&=exp\,(Nb(\rho))\\ \underline{r}(\rho)&=exp\,(-b(\rho))\,, \end{aligned} \tag{1.53}$$

where $b(\rho)$ is given in (1.3). By inserting (1.53) into (1.40), we will get the transformation in (1.2). By replacing $\underline{r}(\rho)$ and $r(\rho)$ in our derivation, we will have exactly the same results as those from (1.7) to (1.10).

1.9.3 Proof of Proposition 1.2

When the regressors under the candidate model are neither perfectly correlated nor perfectly uncorrelated with those under the true model, we can define $h_2(\beta,\underline{\rho})$ and $h_3(\beta)$ as in (1.14) whereX_i and $\underline{X}_i$ denote the regressors under the true and the candidate model respectively. We can also rewrite (1.47) as (1.19) and in the limit we will have (1.20). We can still have (1.50), but the differential equation in (1.51) has now become

$$\frac{-N(T-1)\left[h_2(\beta,\underline{\rho})-\sigma^2h(\underline{\rho})\right]}{h_3(\beta)+(T-1)\sigma^2}d\,\underline{\rho}=\frac{1}{r(\underline{\rho})}d\,r \tag{1.54}$$

If the solution in (1.53) is still valid, we can insert (1.52) into (1.54) to obtain

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

It is obvious that unless we have either $\frac{-(T-1)h_2(\beta,\underline{\rho})}{h_3(\beta)}=h(\underline{\rho})$ or $h_2(\beta,\underline{\rho})=h_3(\beta)=0$, (1.53) is not a solution for (1.54). In other words, the reparameterization of the fixed effect in (1.2) cannot lead to consistent estimation of ρ.[20] Generally speaking, if the candidate model does not nest the true model, it is likely that the reparameterization that will enable us to estimate ρ consistently will involve the true values of the common parameters (β, σ^2 and ρ).

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

1.9.4 Proof of Proposition 1.3

To prove Proposition 1.3 and 1.4, essentially we need to simplify the integral(s) which appears in the Bayes factor. One way to do it is Laplace's method, the details of which can be found in Tierney and Kadane (1986) and Kass et al. (1990). To apply the method, we can first multiply both the numerator and the denominator by $\left(\frac{1}{N}\right)^{-\frac{N(T-1)}{2}}$. The integral appearing in

[20] The inconsistency of the estimator for σ^2 follows since σ^2 is not independent from ρ (asymptotically) as can be seen from (1.9).

the Bayes factor can be simplified as

$$\begin{aligned}
&\left(\frac{1}{N}\right)^{-\frac{N(T-1)}{2}} \int_{-1}^{1} \exp\left[Nb(\rho)\right] \\
&\left[\sum_{i=1}^{N} w_i' H w_i - \frac{1}{\eta+1}\sum_{i=1}^{N} w_i' H \underline{X}_i \left(\sum_{i=1}^{N} \underline{X}_i' H \underline{X}_i\right)^{-1} \sum_{i=1}^{N} \underline{X}_i' H w_i\right]^{-\frac{N(T-1)}{2}} d\rho \\
&= \left(\frac{1}{N}\right)^{-\frac{N(T-1)}{2}} \int_{-1}^{1} \exp\left[Nb(\rho)\right] \left(a\rho^2 - 2b\rho + c\right)^{-\frac{N(T-1)}{2}} \, d\rho \\
&= \left(\frac{a}{N}\right)^{-\frac{N(T-1)}{2}} \int_{-1}^{1} \exp\left(N\left[b(\rho) - \frac{T-1}{2}\ln(\rho^2 - 2\frac{b}{a}\rho + \frac{c}{a})\right]\right) d\rho \\
&= \left(\frac{a}{N}\right)^{-\frac{N(T-1)}{2}} \int_{-1}^{1} \exp\left[Nf(\rho)\right] \, d\rho
\end{aligned} \tag{1.55}$$

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

$$\begin{aligned}
f(\rho) &= b(\rho) - \frac{T-1}{2}\ln(\rho^2 - 2\frac{b}{a}\rho + \frac{c}{a}), \\
f'(\rho) &= h(\rho) - \frac{(T-1)(\rho - \frac{b}{a})}{\rho^2 - 2\frac{b}{a}\rho + \frac{c}{a}}, \\
f''(\rho) &= h'(\rho) - \frac{(T-1)(\rho^2 - 2\frac{b}{a}\rho + \frac{c}{a}) - 2(T-1)(\rho - \frac{b}{a})^2}{(\rho^2 - 2\frac{b}{a}\rho + \frac{c}{a})^2},
\end{aligned} \tag{1.56}$$

where $h'(\rho) = \sum_{i=1}^{T-2} \frac{i(T-i-1)}{T}\rho^{i-1} = \frac{1}{(1-\rho)^2} - \frac{(T+2)\rho^{T+1} - 2\rho^T - T\rho^{T-1} - 2\rho + 2}{T(1-\rho)^4}$. Based on (1.20), if we take the probability limit of (1.56), we can arrive at (1.57)

as follows,

$$\begin{aligned}
&\underset{N\to\infty}{plim}\, f(\rho) = b(\rho) - \frac{T-1}{2}\ln\left[\rho^2 - 2(\underline{\rho}+NB)\rho + \underline{\rho}^2 + 2\underline{\rho}NB + \frac{(T-1)\sigma^2 + h_3(\beta)}{\underline{a}}\right],\\
&\underset{N\to\infty}{plim}\, f'(\rho) = h(\rho) - \frac{(T-1)(\rho - \underline{\rho} - NB)}{\rho^2 - 2\rho(\underline{\rho}+NB) + \underline{\rho}^2 + 2\underline{\rho}NB + \frac{(T-1)\sigma^2+h_3(\beta)}{\underline{a}}},\\
&\underset{N\to\infty}{plim}\, f''(\rho) = h'(\rho) -\\
&\frac{(T-1)\left[\rho^2 - 2\rho(\underline{\rho}+NB) + \underline{\rho}^2 + 2\underline{\rho}NB + \frac{(T-1)\sigma^2+h_3(\beta)}{\underline{a}} - 2(\rho - \underline{\rho} - NB)^2\right]}{\left[\rho^2 - 2\rho(\underline{\rho}+NB) + \underline{\rho}^2 + 2\underline{\rho}NB + \frac{(T-1)\sigma^2+h_3(\beta)}{\underline{a}}\right]^2}.
\end{aligned} \tag{1.57}$$

Now we can use Laplace's method to approximate the integral. Suppose for the equation $\underset{N\to\infty}{plim}\, f'(\rho) = 0$, there exists only one solution ρ^* in (−1,1) and $\underset{N\to\infty}{plim}\, f''(\rho^*) < 0$. For large N, the integral in (1.55) can be approximated by

$$\begin{aligned}
&\left(\frac{1}{N}\right)^{-\frac{N(T-1)}{2}} \int_{-1}^{1} \exp\left[Nb(\rho)\right]\\
&\left[\sum_{i=1}^{N} w_i' H w_i - \frac{1}{\eta+1}\sum_{i=1}^{N} w_i' H \underline{X}_i \left(\sum_{i=1}^{N} \underline{X}_i' H \underline{X}_i\right)^{-1} \sum_{i=1}^{N} \underline{X}_i' H w_i\right]^{-\frac{N(T-1)}{2}} d\rho\\
&\approx \underline{a}^{-\frac{N(T-1)}{2}} \sqrt{\frac{2\pi}{N\left|f''(\rho^*)\right|}} \exp\left[Nf(\rho^*)\right]\\
&= \sqrt{\frac{2\pi}{N\left|f''(\rho^*)\right|}} \exp\left[Nb(\rho^*) - \frac{N(T-1)}{2}\ln d(\rho^*)\right],
\end{aligned} \tag{1.58}$$

where $d(\rho)$ is defined in (1.24).

Moreover, if our choice of the set of regressors included can lead to consistent estimation of ρ, i.e. either (1.11) or (1.12) is satisfied, by substituting the true value of ρ (i.e. $\underline{\rho}$) into (1.57) we can obtain

$$\begin{aligned}
&\underset{N\to\infty}{plim}\, f(\underline{\rho}) = b(\underline{\rho}) - \frac{T-1}{2}\ln\frac{(T-1)\sigma^2 + h_3(\beta)}{\underline{a}},\\
&\underset{N\to\infty}{plim}\, f'(\underline{\rho}) = 0,\\
&\underset{N\to\infty}{plim}\, f''(\underline{\rho}) = h'(\underline{\rho}) - \frac{\underline{a}(T-1)}{(T-1)\sigma^2 + h_3(\beta)} + \frac{2h^2(\underline{\rho})}{T-1}.
\end{aligned} \tag{1.59}$$

For large value of N, the integral in (1.55) can now be approximated by

$$\begin{aligned}
&\left(\frac{1}{N}\right)^{-\frac{N(T-1)}{2}} \int_{-1}^{1} \exp\left[Nb(\rho)\right]\\
&\left[\sum_{i=1}^{N} w_i' H w_i - \frac{1}{\eta+1}\sum_{i=1}^{N} w_i' H \underline{X}_i \left(\sum_{i=1}^{N} \underline{X}_i' H \underline{X}_i\right)^{-1} \sum_{i=1}^{N} \underline{X}_i' H w_i\right]^{-\frac{N(T-1)}{2}} d\rho\\
&\approx \sqrt{\frac{2\pi}{N\left|h'(\underline{\rho}) - \frac{\underline{a}(T-1)}{(T-1)\sigma^2+h_3(\beta)} + \frac{2h^2(\underline{\rho})}{T-1}\right|}}.\\
&\quad \exp\left[Nb(\underline{\rho}) - \frac{N(T-1)}{2}\ln\left((T-1)\sigma^2 + h_3(\beta)\right)\right]
\end{aligned} \tag{1.60}$$

Considering (1.17), if $X_{i1}'s$ are the true regressors to generate Y (so $h_2(\beta, \underline{\rho}) = h_3(\beta) = 0$), in the probability limit (1.17) can be approximated by

$$
\begin{aligned}
&\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}\\
\approx&\frac{1}{2}\left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}}\left[\underline{a}_{|M_0}\underline{\rho}^2-2\underline{\rho}\sigma^2h(\underline{\rho})+2\underline{\rho}h_{2|M_0}(\beta,\underline{\rho})+h_{3|M_0}(\beta)+(T-1)\sigma^2\right]^{\frac{N(T-1)}{2}}\\
&\sqrt{\frac{2\pi}{N\left|h'(\underline{\rho})-\frac{a}{\sigma^2}+\frac{2h^2(\underline{\rho})}{T-1}\right|}}\exp\left[Nb(\underline{\rho})-\frac{N(T-1)}{2}\ln(T-1)\sigma^2\right]\\
=&\frac{1}{2}\left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}}\sqrt{\frac{2\pi}{N\left|h'(\underline{\rho})-\frac{a}{\sigma^2}+\frac{2h^2(\underline{\rho})}{T-1}\right|}}\exp\Bigg\{Nb(\underline{\rho})+\\
&\frac{N(T-1)}{2}\ln\left[\frac{\underline{a}_{|M_0}\underline{\rho}^2-2\underline{\rho}\sigma^2h(\underline{\rho})+2\underline{\rho}h_{2|M_0}(\beta,\underline{\rho})+h_{3|M_0}(\beta)+(T-1)\sigma^2}{(T-1)\sigma^2}\right]\Bigg\}.
\end{aligned}
\tag{1.61}
$$

So we can guarantee $\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}=\infty$ ($\underline{\rho}\neq 0$) as long as (1.21) holds. It does not matter whether we choose η to be $O(\frac{1}{N})$ or $\frac{1}{K^2}$ as used in the simulation studies.

Now let us consider the case when the true model is M_0 in (1.17), i.e. the true value of ρ is 0 and X_{i0} are the right regressors. Given the assumptions in Proposition 1.3, the probability limit of the Bayes factor in (1.17) takes the following form,

$$
\begin{aligned}
&\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}\\
\approx&\frac{1}{2}\left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}}\left[(T-1)\sigma^2\right]^{\frac{N(T-1)}{2}}\\
&\sqrt{\frac{2\pi}{N\left|f''(\rho^*|M_1)\right|}}\exp\left[Nb(\rho^*)-\frac{N(T-1)}{2}\ln d(\rho^*)\right]\\
=&\frac{1}{2}\left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}}\sqrt{\frac{2\pi}{N\left|f''(\rho^*|M_1)\right|}}\exp\left[Nb(\rho^*)+\frac{N(T-1)}{2}\ln\left[\frac{(T-1)\sigma^2}{d(\rho^*|M_1)}\right]\right].
\end{aligned}
\tag{1.62}
$$

If (1.23) holds, then the Bayes factor in (1.17) will tend to 0 for large sample size. If M_1 is misspecified but it can still give consistent estimates of ρ, i.e. $\rho^* = 0$ (either (1.11) or (1.12) holds), we can simplify (1.62) as

$$\begin{aligned}&\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}\\ =&\frac{1}{2}\left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}}\sqrt{\frac{2\pi}{N\left|f''(0|M_1)\right|}}\exp\left[\frac{N(T-1)}{2}\ln\left[\frac{(T-1)\sigma^2}{(T-1)\sigma^2+h_{3|M_1}(\beta)}\right]\right].\end{aligned} \tag{1.63}$$

If $h_{3|M_1}(\beta) > 0$, the Bayes factor in (1.63) will be 0 when N tends to infinity. If $h_{3|M_1}(\beta) = 0$, we should have $k_1 - k_0 > 0$. Once again, the choice of η between $O(\frac{1}{N})$ and $\frac{1}{K^2}$ are not important here.

1.9.5 Proof of Proposition 1.4

For (1.18), suppose the true model is M_1 and M_0 despite being misspecified can still lead to consistent estimation of ρ, (1.18) can be approximated as

$$\begin{aligned}&\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}\\ \approx&\left(\frac{\eta}{\eta+1}\right)^{\frac{k_1-k_0}{2}}\sqrt{\left|1+\frac{\frac{a}{\sigma^2}-\frac{a(T-1)}{(T-1)\sigma^2+h_3(\beta)}}{h'(\underline{\rho})-\frac{a}{\sigma^2}+\frac{2h^2(\rho)}{T-1}}\right|}\left\{\frac{(T-1)\sigma^2+h_3(\beta)}{(T-1)\sigma^2}\right\}^{\frac{N(T-1)}{2}}.\end{aligned} \tag{1.64}$$

Since $h_3(\beta)$ is a semi-positive definite quadratic form of β, it should be greater than or equal to 0. It is 0 when M_0 nests M_1 ($k_1 < k_0$). It is not hard to see that $\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)} = \infty$ when M_1 is the true model and we set η to be $O(\frac{1}{N})$. Under the choice of $\eta = \frac{1}{K^2}$, the Bayes factor in the limit will not tend to infinity, but rathe a constant, which is still possible to be greater than 1 and favours the true model.

If we are comparing the true model to a model under which we cannot obtain consistent estimate of ρ using the transformation of the fixed

effect, we cannot use (1.60) to approximate the marginal likelihood of the misspecified model. In fact we may not be able to use Laplace's method to approximate the integral since $\underset{N\to\infty}{plim} f(\rho)$ may not have a unique maximum point in (-1,1). However, if $\underset{N\to\infty}{plim} f(\rho)$ has a nice bell shape in the stationary region, we can prove that when using the reparameterization of the fixed effect, Bayes factor can lead to the selection of the true model asymptotically under certain circumstances. To see this, we continue to suppose M_1 is the true model in (1.18) and denote ρ^* as our estimate of ρ under M_0. The Bayes factor (1.18) can be approximated by

$$\begin{aligned}&\underset{N\to\infty}{plim}\frac{p(Y|Y_0,M_1)}{p(Y|Y_0,M_0)}\\ &\approx\left(\frac{\eta}{\eta+1}\right)^{\frac{k-k_0}{2}}\sqrt{\left|\frac{f''(\rho^*)}{f''(\underline{\rho})}\right|}\exp\left\{N\left[b(\underline{\rho})-b(\rho^*)+\frac{(T-1)}{2}\ln\left(\frac{d(\rho^*)}{d(\underline{\rho})}\right)\right]\right\}\end{aligned} \tag{1.65}$$

Note that $d(\underline{\rho})=(T-1)\sigma^2$. So if (1.25) is satisfied, the Bayes factor is consistent in selecting the true model, as claimed by Proposition 1.4. It is difficult to interpret under what circumstances our data can satisfy (1.25). Note that the equation $\underset{N\to\infty}{plim} f'(\rho)=0$ generally do not have analytical solution when our model is misspecified and it does not nest the true model. Therefore it is hard to check (1.25) and we have to rely on simulation studies to shed some light on this issue.

1.9.6 Proof of Proposition 1.5

The likelihood function takes the following form,

$$\begin{aligned}p\left(Y|\theta,Y_0\right)=&(2\pi)^{-\frac{TN}{2}}\sigma^{2(-\frac{NT}{2})}\\ &\prod_{i=1}^{N}\exp\{-\frac{1}{2\sigma^2}[y_i-y_{i\_}\rho-\iota f_i-X_i\beta]'\left[y_i-y_{i\_}\rho-\iota f_i-X_i\beta\right]\}.\end{aligned} \tag{1.66}$$

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

$$\begin{aligned}
\sigma^2 &= \frac{1}{NT}\sum_{i=1}^{N}[y_i - y_{i_}\rho - \iota f_i - X_i\beta]'\,[y_i - y_{i_}\rho - \iota f_i - X_i\beta]\,, \\
f_i &= \frac{\iota'(y_i - y_{i_}\rho - X_i\beta)}{T}, \\
\beta &= \sum_{i=1}^{N}(X_i' H X_i)^{-1}\sum_{i=1}^{N}X_i' H(y_i - yi_{_}\rho), \\
\rho &= \frac{b}{a},
\end{aligned} \tag{1.67}$$

where a and b are defined in (1.19) with $\eta = 0$. Based on the MLE, we can find the Bayesian information criterion (BIC) as the following,

$$BIC = NT\left(\ln\frac{c - \frac{b^2}{a}}{NT} + \ln 2\pi + 1\right) + (1 + k + N)\ln(NT). \tag{1.68}$$

A BIC value close to zero calculated under a model indicates evidence in favor of the model. Using (1.20), we can find the probability limit of BIC as

$$\begin{aligned}
\underset{N\to\infty}{plim}\, BIC =& NT\left(\ln\frac{\underset{N\to\infty}{plim}\frac{1}{N}c - (\underset{N\to\infty}{plim}\frac{1}{N}b)^2(\underset{N\to\infty}{plim}\frac{1}{N}a)^{-1}}{T} + \ln(2\pi) + 1\right) \\
&+ (1 + k + N)\ln(NT) \\
=& NT\left(\ln\frac{(T-1)\sigma^2 + h_3(\beta) - \underline{a}NB^2}{T} + \ln(2\pi) + 1\right) \\
&+ (1 + k + N)\ln(NT) \\
=& NT\left\{\ln\frac{(T-1)\sigma^2 + h_3(\beta) - \frac{\left[h_2(\beta,\underline{\rho}) - \sigma^2 h(\underline{\rho})\right]^2}{\underline{a}}}{T} + \ln(2\pi) + 1\right\} \\
&+ (1 + k + N)\ln(NT).
\end{aligned} \tag{1.69}$$

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

$$\underset{N\to\infty}{plim}\, BIC = NT\left(\ln\frac{(T-1)\sigma^2-\frac{\sigma^4h^2(\underline{\rho})}{\underline{a}}}{T}+\ln(2\pi)+1\right)+(1+k+N)\ln(NT). \tag{1.70}$$

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

$$\begin{aligned}\underset{N\to\infty}{plim}\, BIC &= NT\left(\ln\frac{\underset{N\to\infty}{plim}\frac{1}{N}c}{T}+\ln(2\pi)+1\right)+(k+N)\ln(NT)\\ &= NT\left(\ln\frac{(T-1)\sigma^2+h_3(\beta)+\underline{a}\underline{\rho}^2+2\underline{\rho}h_2(\beta,\underline{\rho})-2\underline{\rho}\sigma^2h(\underline{\rho})}{T}+\ln(2\pi)+1\right)\\ &\quad+(k+N)\ln(NT).\end{aligned} \tag{1.71}$$

Let us now look at the case of (1.17). When X_{i1} are the true regressors to generate Y_i, the difference between the BIC under M_0 and M_1 is

$$\begin{aligned}BIC_{|M_0}-BIC_{|M_1} =&\\ &NT\ln\frac{(T-1)\sigma^2+h_{3|M_0}(\beta)+\underline{a}_{|M_0}\underline{\rho}^2+2\underline{\rho}h_{2|M_0}(\beta,\underline{\rho})-2\underline{\rho}\sigma^2h(\underline{\rho})}{(T-1)\sigma^2-\frac{\sigma^4h^2(\underline{\rho})}{\underline{a}_{|M_1}}}\\ &+(k_0-k_1-1)\ln(NT)\end{aligned} \tag{1.72}$$

Clearly if we have $BIC_{|M_0}-BIC_{|M_1}>0$ for large N, which means M_1 is the preferred model, inside the natural log on the right hand side of the equation, the numerator should be larger than the denominator. In other words, we should have (1.26) stated in Proposition 1.5. If $\underline{\rho}=0$, it is clear that (1.26) can be satisfied and model selection is consistent. However, if $X_{i1}=X_{i0}$, we can have $\underline{a}_{|M_0}=\underline{a}_{|M_1}=\underline{a}$, $k_1=k_0$ and $h_{2|M_0}(\beta,\underline{\rho})=h_{3|M_0}(\beta)=0$. Hence

we can simplify (1.72) as

$$\begin{aligned} BIC_{|M_0} - BIC_{|M_1} =& NT \ln \frac{(T-1)\sigma^2 - \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}} + \underline{a}\underline{\rho}^2 - 2\underline{\rho}\sigma^2 h(\underline{\rho}) + \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}}}{(T-1)\sigma^2 - \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}}} \\ &+ (k_0 - k_1 - 1)\ln(NT) \\ =& NT \ln \left[1 + \frac{\left(\underline{a}\underline{\rho} - \underline{\rho}\sigma^2 h(\underline{\rho})\right)^2}{(T-1)\underline{a}\sigma^2 - \sigma^4 h^2(\underline{\rho})} \right] - \ln(NT) \end{aligned} \tag{1.73}$$

If $\underline{a}\underline{\rho} - \underline{\rho}\sigma^2 h(\underline{\rho}) = 0$, i.e. $\underline{\rho} + NB = 0$, we will always have $BIC_{|M_0} - BIC_{|M_1} < 0$, which means we will always prefer M_0 over M_1 even if $\underline{\rho} \neq 0$. In a situation like this, model selection is not consistent.

The problem with BIC also arises when M_0 is the true model. Now the difference between the two BICs is

$$\begin{aligned} BIC_{|M_0} - BIC_{|M_1} =& NT \ln \frac{(T-1)\sigma^2}{(T-1)\sigma^2 + h_{3|M_1}(\beta) - \frac{\left[h_{2|M_1}(\beta,0) - \sigma^2 \frac{T-1}{T}\right]^2}{\underline{a}_{|M_1}}} \\ &+ (k_0 - k_1 - 1)\ln(NT). \end{aligned} \tag{1.74}$$

If we want to have M_0 preferred by BIC, we should have $BIC_{|M_0} - BIC_{|M_1} < 0$, which means we should have (1.27) claimed in Proposition 1.5. However, if we have $h_{3|M_1}(\beta) = 0$, which implies $k_1 \geq k_0$, (1.27) cannot be satisfied since $\frac{[h_{2|M_1}(\beta,0) - \sigma^2 \frac{(T-1)}{T}]^2}{\underline{a}_{|M_1}} \geq 0$. Again, this implies inconsistency in model selection.

For the case of (1.18), suppose M_1 is the true model, the difference

between the BICs calculated under M_0 and M_1 is

$$\begin{aligned} BIC_{|M_0} - BIC_{|M_1} =& NT \ln \frac{(T-1)\sigma^2 + h_{3|M_0}(\beta) - \frac{\left[h_{2|M_0}(\beta,\underline{\rho}) - \sigma^2 h(\underline{\rho})\right]^2}{\underline{a}_{|M_0}}}{(T-1)\sigma^2 - \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}_{|M_1}}} \\ &+ (k_0 - k_1)\ln(NT) \\ =& NT \ln \left\{ 1 + \frac{h_{3|M_0}(\beta) + \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}_{|M_1}} - \frac{\left[h_{2|M_0}(\beta,\underline{\rho}) - \sigma^2 h(\underline{\rho})\right]^2}{\underline{a}_{|M_0}}}{(T-1)\sigma^2 - \frac{\sigma^4 h^2(\underline{\rho})}{\underline{a}_{|M_1}}} \right\} \\ &+ (k_0 - k_1)\ln(NT). \end{aligned} \tag{1.75}$$

If M_1 is the true model, (1.28) stated in Proposition 1.5 should hold. If X_{i0} nests the true set of regressors, i.e. $h_{2|M_0}(\beta,\underline{\rho}) = h_{3|M_0}(\beta) = 0$ and $\underline{a}_{|M_1} = \underline{a}_{|M_0}$, (1.75) is reduced to

$$BIC_{|M_0} - BIC_{|M_1} = (k_0 - k_1)\ln(NT) \tag{1.76}$$

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

Chapter 2

A Correction Function Approach to Solve the Incidental Parameter Problem

2.1 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 chapter, 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 the previous chapter 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 chapter is as follows. Section 2.2 gives a Bayesian perspective on the incidental parameter problem and our strategy to find the correction function to solve the problem. Section 2.3 demonstrates how our strategy is applied to the linear panel AR model of order more than one to derive the correction function. Section 2.3.2 and Section 2.3.3 discuss the algorithms to carry out point estimation and model comparison, while Section 2.3.4 and Section 2.3.5 give the respective examples using simulated datasets before Section 2.4 concludes.

2.2 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,

$$\begin{aligned} p(\theta|Y) &\propto \int_F p(\theta, f)p(Y|\theta, f)\, df \\ &\propto \int_F p(\theta)p(f|\theta)p(Y|\theta, f)\, df. \end{aligned} \tag{2.1}$$

Here we use Y to stand for the collection of the dependent variable(s) and $p(f|\theta)$ is a permissible prior function with support F[1]. The problem with the Bayesian method is that there is no guarantee for us to obtain consistent

[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).

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,

$$\nu(\Omega) = \int_{\theta \in \Omega} p(\theta|Y)\, d\theta. \tag{2.2}$$

The incidental parameter problem now can be interpreted as

$$\underset{N\to\infty}{plim}\, \nu(\Omega) = I(\theta_b \in \Omega) \tag{2.3}$$

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 (2.1) 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 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

[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.

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

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 (2.3). 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

$$r(\theta) = \frac{\int_F p_r(f|\theta)p(Y|\theta,f)d\,f}{\int_F p_b(f|\theta)p(Y|f,\theta)d\,f}. \tag{2.4}$$

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,

$$\nu^r(\Omega) = \int_{\theta\in\Omega} k \cdot r(\theta)p_b(\theta|Y)\,d\theta = \int_{\theta\in\Omega} k \cdot r(\theta)\,d\nu. \tag{2.5}$$

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

$$\underset{N\to\infty}{plim}\,\nu^r(\Omega) = I(\theta_{true} \in \Omega). \tag{2.6}$$

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.

$$E_Y\left(\frac{\partial^2 \ln p(Y|\theta,f)}{\partial f\partial\theta}\right) = \int \frac{\partial^2 \ln p(Y|\theta,f)}{\partial f\partial\theta} p\left(Y|\theta,f\right)\,dY = 0 \tag{2.7}$$

we can just use a flat prior $p(f|\theta) \propto 1$[8] to integrate out the incidental

[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.

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 profile likelihood in Cox and Reid (1987), see also Sweeting (1995). The modified profile likelhood 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

$$\frac{\partial f}{\partial \theta} = -\left(E_Y(\frac{\partial^2 \ln p(Y|\theta,f)}{\partial f \partial f'})\right)^{-1} E_Y\left(\frac{\partial^2 \ln p(Y|\theta,f)}{\partial f \partial \theta}\right) \tag{2.8}$$

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 (2.8) as

$$\frac{\partial f_i}{\partial \theta} = \chi(f_i, \theta) \tag{2.9}$$

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

$$\chi(f_i, \theta) = -\left(E_y(\frac{\partial^2 \ln p(y_i|\theta, f_i)}{\partial f_i^2})\right)^{-1} E_y\left(\frac{\partial^2 \ln p(y_i|\theta, f_i)}{\partial f_i \partial \theta}\right). \tag{2.10}$$

Since f_i is defined implicitly as a one-one function of g_i, we can differentiate both sides of (2.9) 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

$$-\frac{\partial \ln |\frac{\partial g_i}{\partial f_i}|}{\partial \theta} = \frac{\partial \ln |\frac{\partial f_i}{\partial g_i}|}{\partial \theta} = \frac{\partial^2 f_i}{\partial \theta \partial g_i}\left(\frac{\partial f_i}{\partial g_i}\right)^{-1} = \frac{\partial \chi(f_i, \theta)}{\partial f_i}. \tag{2.11}$$

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 (2.11) to solve the incidental parameter problem.

Example 2.1. *Let us consider a simple panel Poisson count model:* $y_{i,t} \sim$ *i.i.d.Poisson* $(f_i \exp(x_{i,t}\theta))$ *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*

$$l_i(f_i, \theta) = p(y_i|f_i, \theta) \propto e^{-f_i \sum_t \exp(x_{it\theta})} f_i^{\sum_t y_{it}} e^{\theta \sum_t y_{it} x_{it}} \tag{2.12}$$

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. $l_i(f_i(g,\theta),\theta) = l_{i1}(g_i)l_{i2}(\theta)$,

$$l_i(f_i(g,\theta),\theta) \propto e^{-g_i} g_i^{\sum_t y_{it}} \times \frac{e^{\theta \sum_t y_{it} x_{it}}}{(\sum_t \exp(x_{it}\theta))^{\sum_t y_{it}}} \tag{2.13}$$

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

$$p(y_i|\theta) \propto \frac{e^{\theta \sum_t y_{it} x_{it}}}{(\sum_t \exp(x_{it}\theta))^{1+\sum_t y_{it}}}, \tag{2.14}$$

which is different from $l_{i2}(\theta)$ in (2.13) 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 (2.14) 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

$$\begin{aligned} E_y\left(\frac{\partial^2 l_i(f_i,\theta)}{\partial f_i \partial \theta}\right) &= -\sum_t x_{it}\exp(x_{it}\theta) \neq 0 \\ E_y\left(\frac{\partial^2 l_i(f_i,\theta)}{\partial f_i^2}\right) &= E_y\left(-\frac{\sum_t y_{it}}{f_i^2}\right) = -\frac{\sum_t \exp(x_{it}\theta)}{f_i} \\ \chi(f_i,\theta) &= -\frac{f_i \sum_t x_{it} exp(x_{it}\theta)}{\sum_t exp(x_{it}\theta)} \end{aligned} \tag{2.15}$$

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 (2.11) 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

gurantee that we can find $\lambda(f_i, \theta)$ from the differential equation (2.11) 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 (2.11). 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

$$p(f_i|\theta) \propto \exp\left[\lambda_1(f_i, \theta) + \lambda_2(f_i, \theta) + \dots\right]. \tag{2.16}$$

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.

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

2.3.1 The Bias Reducing Prior and the Posterior Results

Suppose our model has p lags and can be written as

$$y_i = \iota f_i + Y_{i\_}\rho + X_i\beta + u_i \tag{2.17}$$

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 the 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 (2.11) 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:

$$p(f_i|\theta) \propto \exp\left[\lambda_\rho(f_i,\theta)+\lambda_\beta(f_i,\theta)+\lambda_{\sigma^2}(f_i,\theta)\right]. \tag{2.18}$$

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,

$$l_i = \ln p(y_i|f_i,\theta,y_{i,-p}) \propto -\frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left(y_i-\iota f_i - Y_{i\_}\rho - X_i\beta\right)'\left(y_i-\iota f_i - Y_{i\_}\rho - X_i\beta\right). \tag{2.19}$$

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

$$E_y\left(\frac{\partial^2 l_i}{\partial f_i^2}\right) = -\frac{T}{\sigma^2}, \tag{2.20}$$

$$E_y\left(\frac{\partial^2 l_i}{\partial f_i \partial \beta}\right) = -\frac{T}{\sigma^2}X_i'\iota, \tag{2.21}$$

$$E_y\left(\frac{\partial^2 l_i}{\partial f_i \partial \sigma^2}\right) = -E_y\left[\frac{\left(y_i-\iota f_i - Y_i\rho - X_i\beta\right)'\iota}{(\sigma^2)^2}\right] = 0, \tag{2.22}$$

$$\begin{aligned} E_y\left(\frac{\partial^2 l_i}{\partial f_i \partial \rho}\right) &= -\frac{T}{\sigma^2}E_y(Y_{i\_}'\iota) \\ &= -\frac{T}{\sigma^2}\left[Th(\rho)f_i + \omega_1(X_i\beta,\rho) + \omega_2(y_{i,-p},\rho)\right], \end{aligned} \tag{2.23}$$

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 (2.22), we can see that f_i is information orthogonal to σ^2. The right hand side of (2.21) does not involve f_i. Hence we can have $\lambda_\beta(f_i,\theta) = 0_{K\times 1}$ and $\lambda_{\sigma^2}(f_i,\theta) = 0_{1\times 1}$,

[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.

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,

$$\frac{\partial \lambda_\rho(f_i,\theta)}{\partial \rho} = h(\rho). \tag{2.24}$$

We show in the appendix that (2.24) 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 2.1 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 chapter 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

Table 2.1: 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	$\frac{1}{2}\rho_1$	$\frac{1}{3}\sum_{t=1}^{2}\frac{3-t}{t}\rho_1^t + \frac{1}{3}\rho_2$	$\frac{1}{4}\sum_{t=1}^{3}\frac{4-t}{t}\rho_1^t + \frac{1}{4}\rho_1\rho_2 + \frac{1}{2}\rho_2 + \frac{1}{4}\rho_3$

the bias reducing prior, $p(f_i|\theta)$ under our strategy in (2.18) is

$$p(f_i|\theta) = p(f_i|\rho) \propto exp(\tau(\rho)). \tag{2.25}$$

[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.

Note that this prior involves ρ only. The correction function defined in (2.4) is therefore

$$r(\theta) = r(\rho) = exp[N\tau(\rho)]. \tag{2.26}$$

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 (2.4). That is why the correction function in (2.26) has closed form. It is possible that the bias reducing prior defined in (2.16) 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,

$$\begin{aligned} p(f,\theta) &= p(f,\rho,\beta,\sigma^2) = p(f_1|\rho)\dots p(f_N|\rho)p(\rho)p(\sigma^2)p(\beta|\sigma^2) \\ &\propto r(\rho)\frac{1}{\sigma^2}I(\rho \in S)\frac{1}{m(S)}p(\beta|\sigma^2) \end{aligned} \tag{2.27}$$

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],

$$\beta|\sigma^2 \sim N\left(0, \sigma^2(\eta\sum_{i=1}^{N} X_i' H X_i)^{-1}\right), \tag{2.28}$$

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

Proposition 2.1. *Conditional on the initial p observations of the dependent variable, using the bias reducing prior (2.25) and the priors described in*

[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.

(2.27) and (2.28), we can obtain the following posterior distributions,

$$f_i|Y, y_{i,0}, \sigma^2, \rho, \beta \sim N\left(\frac{\iota'(y_i - Y_{i\_}\rho - X_i\beta)}{T}, \frac{\sigma^2}{T}\right), \tag{2.29}$$

$$\beta|Y, Y_0, \sigma^2, \rho \sim$$
$$N\Big(\frac{1}{\eta+1}\left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1}\sum_{i=1}^{N} X_i'H(y_i - Y_{i\_}\rho), \sigma^2\left((\eta+1)\sum_{i=1}^{N} X_i'HX_i\right)^{-1}\Big), \tag{2.30}$$

$$\sigma^2|\rho, Y, Y_0 \sim IG(N(T-1), \Delta), \tag{2.31}$$

where

$$\begin{aligned}\Delta =& \sum_{i=1}^{N}(y_i - Y_{i\_}\rho)'H(y_i - Y_{i\_}\rho) - \\ & \frac{1}{\eta+1}\sum_{i=1}^{N}(y_i - Y_{i\_}\rho)'HX_i\left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1}\sum_{i=1}^{N} X_i'H(y_i - Y_{i\_}\rho).\end{aligned} \tag{2.32}$$

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

$$\rho|Y, Y_0 \propto I(\rho \in S)r(\rho)t(A^{-1}b, \frac{1}{N(T-1)-p}(c - b'A^{-1}b)A^{-1}, N(T-1)-p) \tag{2.33}$$

where

$$\begin{aligned}\underset{p\times p}{A} &= \sum_{i=1}^{N} Y_{i\_}'HY_{i\_} - \frac{1}{\eta+1}\sum_{i=1}^{N}\left(Y_{i\_}'HX_i\right)\left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1}\sum_{i=1}^{N}\left(X_i'HY_{i\_}\right) \\ \underset{p\times 1}{b} &= \sum_{i=1}^{N} Y_{i\_}'Hy_i - \frac{1}{\eta+1}\sum_{i=1}^{N}\left(Y_{i\_}'HX_i\right)\left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1}\sum_{i=1}^{N}\left(X_i'Hy_i\right) \\ \underset{1\times 1}{c} &= \sum_{i=1}^{N} y_i'Hy_i - \frac{1}{\eta+1}\sum_{i=1}^{N}\left(y_i'HX_i\right)\left(\sum_{i=1}^{N} X_i'HX_i\right)^{-1}\sum_{i=1}^{N}\left(X_i'Hy_i\right).\end{aligned} \tag{2.34}$$

Equation (2.33) 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.

2.3.2 Estimation Algorithm

Our estimation is based on the draws of the parameters from their posterior distributions. From (2.29), (2.30) and (2.31) 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 (2.33) 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 (2.27). 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

$$\kappa_i^{(k)} = \kappa_i^{(k-1)} - \pi_k \kappa_{k-i}^{(k-1)}, \quad i = 1, \ldots, k-1, \tag{2.35}$$

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

$$J(\pi) = \prod_{k=2}^{p} (1-\pi_k)^{[\frac{k}{2}]} (1+\pi_k)^{[\frac{k-1}{2}]} \tag{2.36}$$

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

$$\kappa_i^{(k-1)} = \frac{\kappa_i^{(k)} + \kappa_k^{(k)} \kappa_{k-i}^{(k)}}{1 - \left(\kappa_k^{(k)}\right)^2} \tag{2.37}$$

As Jones (1987) showed, if ρ follows a uniform distribution in the stationary

region, PAC will be related to a beta distribution as follows,

$$\frac{\pi_k + 1}{2} \quad \sim \quad i.i.d. Beta\left([\frac{1}{2}(k+1)], [\frac{1}{2}k] + 1\right) \tag{2.38}$$

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 2.1[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 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 (2.33), which can be rewritten as,

$$\begin{aligned} p(\rho|Y, Y_0) &\propto I(\rho \in S) \exp\left\{N\left[\tau(\rho) - \frac{T-1}{2}\ln(\rho' A\rho - 2\rho' b + c)\right]\right\} \\ &\propto I(\rho \in S) \exp\left[N\vartheta(\rho)\right] \end{aligned} \tag{2.39}$$

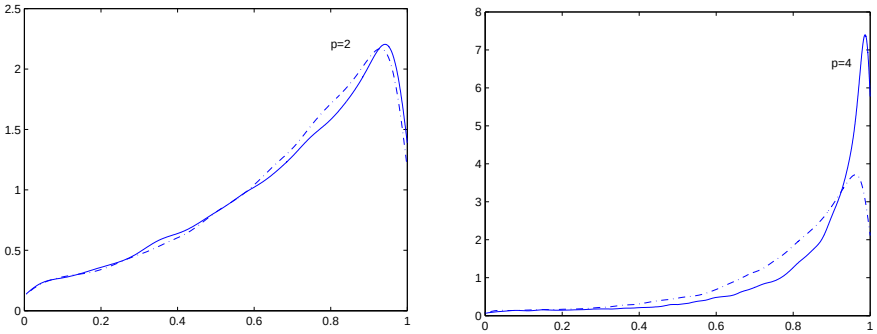
where

$$\vartheta(\rho) = \tau(\rho) - \frac{T-1}{2}\ln(\rho' A\rho - 2\rho' b + c). \tag{2.40}$$

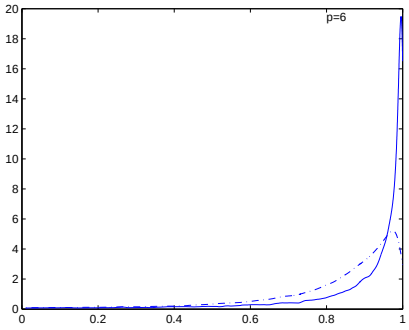
Since the mode of the posterior distribution is a consistent estimator, we can expect $\vartheta(\rho)$ has a unique global maximum in the stationary region when N tends to infinite. Under certain regularity conditions, the posterior distribution will converge to a normal distribution as the sample size N increases,

[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 2.1: 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.

see Bernardo and Smith (section 5.3 1994). It is sensible to use the following truncated normal distribution to approximate the posterior:

$$\rho|Y, Y_0 \quad \overset{a}{\sim} \quad I(\rho \in S)N\left(\hat{\rho}, \frac{1}{N}\left[-\vartheta''(\hat{\rho})\right]^{-1}\right). \tag{2.41}$$

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 2.1 in the following is a Metropolis-Hastings (MH) algorithm, which makes draws from (2.39) using (2.41) 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 (2.41), 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 2.1. *Starting from the current value of $\rho_0 \in S$, we repeat the following steps.*

1. *We propose a draw ρ_c from (2.41).*
2. *We accept ρ_c as a draw from the posterior distribution (2.39) with the probability*
$$\alpha(\rho_0, \rho_c) = \min\left(1, \frac{\exp\left[N\vartheta(\rho_c)\right] q(\rho_0)}{\exp\left[N\vartheta(\rho_0)\right] q(\rho_c)}\right) \tag{2.42}$$
where $q(\cdot)$ is the density function of the truncated normal distribution (2.41).
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 2.1 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 2.2 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 (2.41) has a negative (definite) variance. Although such situation does not always arise, it is not hard to imagine that 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 (2.41) by a positive definite variance matrix, such as $\frac{1}{N(T-1)-p}(c - b'A^{-1}b)A^{-1}$ in (2.33). 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.

2.3.3 Comparison of Different Model Specifications

In the last chapter, we noticed that when our model is misspecified, such as the case when we include the wrong set of exogenous regressors, the solution for (2.24) may not enable us to obtain consistent estimate of ρ under the AR(1) panel model. Therefore we suggested comparing different model specifications using the Bayes factor and showed certain regularity conditions

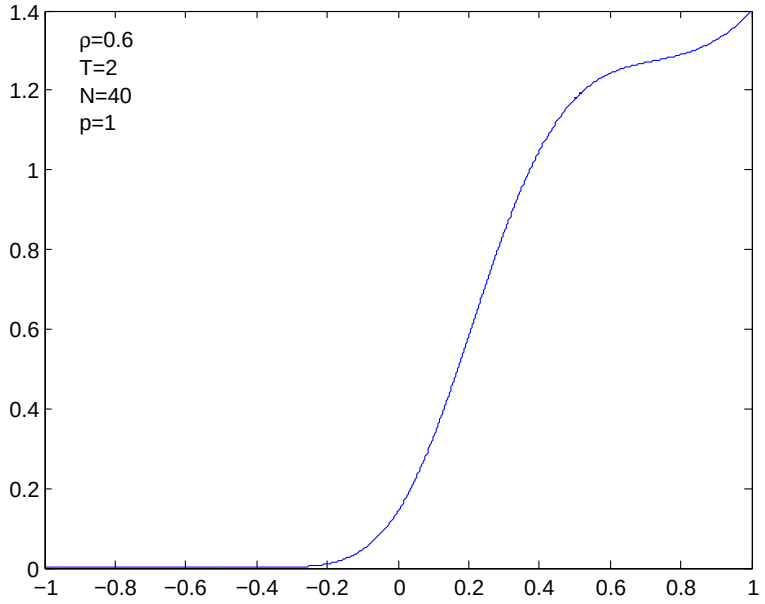


Figure 2.2: The plot of a non-bell shape posterior density function of ρ

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 (2.17). 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 (2.17) 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

[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 $A=\sum_{i=1}^{N} Y_{i\_}' H Y_{i\_}$, $b=\sum_{i=1}^{N} Y_{i\_}' H y_i$ and $c=\sum_{i=1}^{N} y_i' H y_i$.

posterior model probability of model i is defined as

$$\begin{aligned} p\left(M_i|Y,Y_0\right) &= \frac{p\left(M_i\right)p\left(Y|Y_0,M_i\right)}{p\left(Y|Y_0\right)} \\ &= \frac{p\left(M_i\right)p\left(Y|Y_0,M_i\right)}{\sum_{j=1}^{(P+1)2^K} p\left(M_j\right)p\left(Y|Y_0,M_j\right)}. \end{aligned} \tag{2.43}$$

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.

$$\begin{aligned} & p\left(Y|Y_0,M_i\right) \\ &= \int p(g,\theta|Y_0,M_i)p(Y|g,\theta,Y_0,M_i)dg\,d\theta \\ &= \int\limits_{\rho\in S} p(\rho|Y_0,M_i)p(Y|\rho,Y_0,M_i)d\rho \end{aligned} \tag{2.44}$$

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,

$$p(M_i|Y,Y_0) = \frac{p(\rho^*|M_i)p(Y|\rho^*,Y_0,M_i)}{p(\rho^*|M_i,Y,Y_0)} \tag{2.45}$$

For ρ^* we can choose arbitrary value in the stationary region, but for estimation efficiency, the estimated mode of ρ from (2.41) is preferred. According to Chib and Jeliazkov (2001), $p(\rho|M_i,Y,Y_0)$ can be estimated by

$$\hat{p}(\rho|M_i,Y,Y_0) = \frac{K^{-1}\sum_{k=1}^{K}\alpha(\rho^{(k)},\rho^*)q(\rho^{(k)},\rho^*)}{J^{-1}\sum_{j=1}^{J}\alpha(\rho^*,\rho^{(j)})} \tag{2.46}$$

As in Algorithm 2.1 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 2.2) which samples the parameter space and the model space at the same time.

Algorithm 2.2. *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 (2.37) 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 (2.35). The acceptance probability is calculated as*

$$\min\left(1, \left(\frac{\eta}{\eta+1}\right)^{(k^{(c)}-k^{(0)})} \cdot \frac{m\left(S^{(0)}\right)\exp\left[N\vartheta(\rho^{(c)}|ix^{(c)})\right]q(c,0)}{m\left(S^{(c)}\right)\exp\left[N\vartheta(\rho^{(0)}|ix^{(0)})\right]2^{p^{(0)}-p^{(c)}}q(0,c)}\left|\frac{\partial\rho^{(c)}}{\partial(\rho^{(0)\prime}, u')}\right|\right), \tag{2.47}$$

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

where $\vartheta(\cdot)$ is defined in (2.40). $q(x,y)$ 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

$$\left|\frac{\partial\rho^{(c)}}{\partial(\rho^{(0)\prime},u')}\right| = \prod_{i=1}^{p^{(c)}-p^{(0)}}(1+u_i)^{[\frac{p^{(0)}+i-1}{2}]}(1-u_i)^{[\frac{p^{(0)}+i}{2}]}, \quad (2.48)$$

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*

$$\min\Bigg(1, \left(\frac{\eta}{\eta+1}\right)^{(k^{(c)}-k^{(0)})} \frac{m\left(S^{(0)}\right)\exp\left[N\vartheta(\rho^{(c)}|ix^{(c)})\right]2^{p^{(c)}-p^{(0)}}q(c,0)}{m\left(S^{(c)}\right)\exp\left[N\vartheta(\rho^{(0)}|ix^{(0)})\right]q(0,c)} \cdot \left|\frac{\partial\rho^{(0)}}{\partial(\rho^{(c)\prime},\pi_{p^{(c)}+1}^{(0)},\ldots,\pi_{p^{(0)}}^{(0)})}\right|^{-1}\Bigg). \quad (2.49)$$

where the Jacobian takes the following form

$$\left|\frac{\partial\rho^{(0)}}{\partial(\rho^{(c)\prime},\pi_{p^{(c)}+1}^{(0)},\ldots,\pi_{p^{(0)}}^{(0)})}\right| = \prod_{i=p^{(c)}+1}^{p^{(0)}}(1+\pi_i^{(0)})^{[\frac{i-1}{2}]}(1-\pi_i^{(0)})^{[\frac{i}{2}]}. \quad (2.50)$$

- *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*

$$\min\left(1, \frac{\exp\left[N\vartheta(\rho^{(0)}|ix^{(c)})\right]}{\exp\left[N\vartheta(\rho^{(0)}|ix^{(0)})\right]}\right). \quad (2.51)$$

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 2.1 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 2.2 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

$$q(p^{(0)}, p^{(c)}) \propto \exp\left(-\varsigma|p^{(c)} - p^{(0)}|\right), \quad p^{(c)}, p^{(0)} \in [1, \ldots, P], \tag{2.52}$$

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

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

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 $q(k)$,

$$q(k) = \binom{K}{k} \gamma^k (1-\gamma)^{K-k} \left(1 - (1-\gamma)^K\right)^{-1} \tag{2.53}$$

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

2.3.4 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 2.3.2 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 2.1. 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 2.2). 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 2.3. 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 2.2 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 (2.41), 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, though for $N = 1000$ the marginal improvement compared to $N = 200$ is not that much. 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 $N = 1000$.

Table 2.2: 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

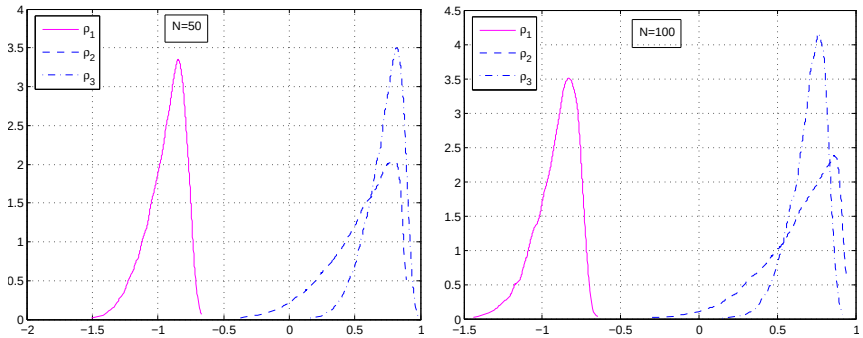
Though point estimates could be important, sometimes we may be more interested in knowing the uncertainty surrounding our estimators. Figure 2.3 shows the posterior marginal density plots for ρ_1, ρ_2 and ρ_3 under different

Table 2.3: 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

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 under $N = 1000$. Table 2.4 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



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

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

Figure 2.3: The marginal density plots of the posterior draws of ρ for $T = 2$

Table 2.4: 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

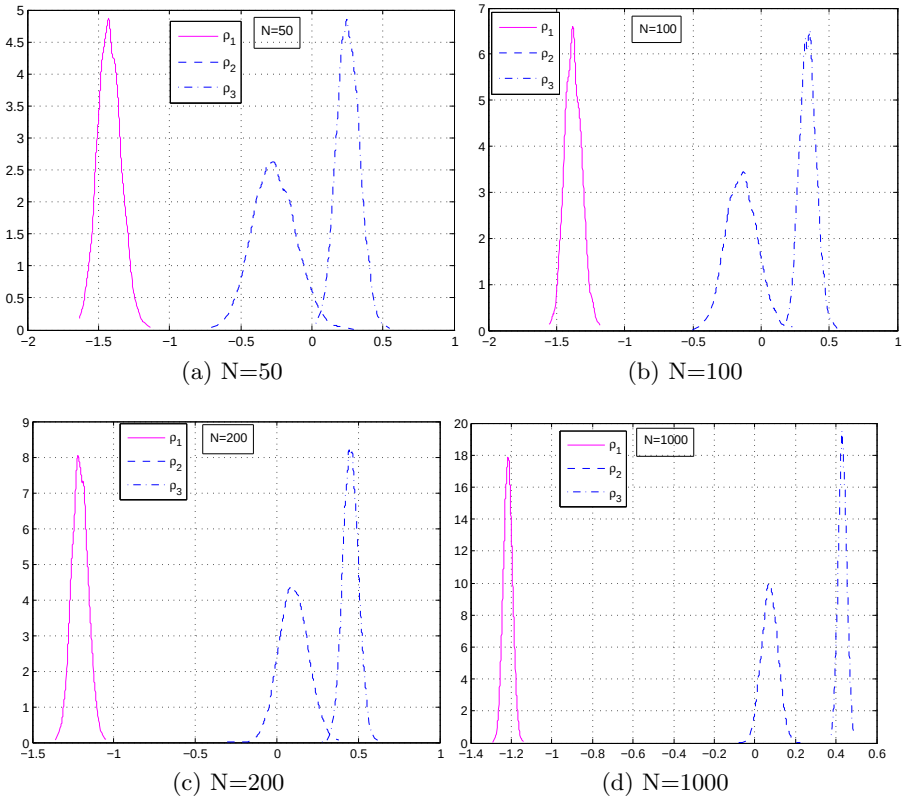
confidence intervals are closer to the true values, none of them can have the true values inside for different cross section sample sizes. As for our correction function method, under $T = 4$, even for $N = 50$, the mode of the posterior distribution for ρ is no longer on the unit circle as before and 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 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 2.5: 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

[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

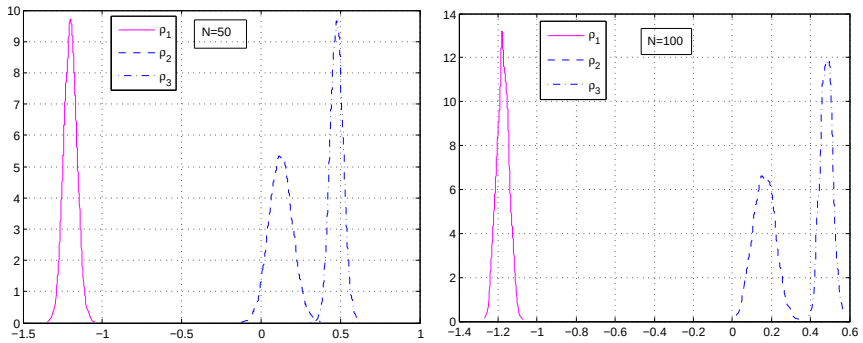
Figure 2.4: The marginal density plots of the posterior draws of ρ for $T = 4$

Table 2.6: 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

Table 2.7: 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 2.5: The marginal density plots of the posterior draws of ρ for $T = 10$

Table 2.8: 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

2.3.5 Demonstration Examples for Model Comparison

In this section, we show how well the algorithms developed in Section 2.3.3 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. As in the previous chapter[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 2.9 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 con-

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

spicuous 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., 2001b). 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 2.10 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. This implies we include regressor 1, 3, 4 and 6 into our model. The column under "inclp" is the inclusion probability obtained from the reversible jump method, while the column "inclpC" is calculated based on the Chib's method. Both the Chib's method and the reversible jump give us similar estimates. Except for regressor 3, the coefficients of other true regressors 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

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

Table 2.9: 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

can conclude that our method can not only achieve consistent estimates for ρ, but also consistent for β.

Table 2.10: 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 2.11, 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 case with 6 potential regressors, four of which are included in the true model. Table 2.12 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 2.11: 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

Table 2.12: 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

2.4 Conclusion

In this chapter, 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 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 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.

2.5 Appendix

2.5.1 Solution for (2.24)

By repetitive substitution, we can rewrite the model in (2.17) as the following,

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

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

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

$$\underset{(T-1+p)\times T}{C} = \begin{pmatrix} 0_{p\times 1} & 0_{p\times 1} & \ldots & 0_{p\times 1} & 0_{p\times 1} \\ 1 & 0 & \ldots & 0 & 0 \\ P_{(1,1)} & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ P^{T-2}_{(1,1)} & P^{T-3}_{(1,1)} & \ldots & 1 & 0 \end{pmatrix}. \tag{2.54}$$

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 (2.54). For the convenience of the 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

$$
\begin{aligned}
\underset{p\times 1}{h}\left(\underset{p\times 1}{\rho}\right) &= \frac{1}{T}\begin{pmatrix} \iota' c_{1(p:T+p-1)} \\ \iota' c_{1(p-1:T+p-2)} \\ \cdots \\ \iota' c_{1(1:T)} \end{pmatrix} = -\begin{pmatrix} trace\left(HC_{(p:T+p-1,:)}\right) \\ trace\left(HC_{(p-1:T+p-2,:)}\right) \\ \cdots \\ trace\left(HC_{(1:T,:)}\right) \end{pmatrix} \\
\underset{p\times 1}{\omega_1}\left(\underset{T\times 1}{X_i\beta}, \underset{p\times 1}{\rho}\right) &= \begin{pmatrix} \iota'(CX_i\beta)_{(p:T+p-1)} \\ \iota'(CX_i\beta)_{(p-1:T+p-2)} \\ \cdots \\ \iota'(CX_i\beta)_{(1:T)} \end{pmatrix} \\
\underset{p\times 1}{\omega_2}\left(\underset{p\times 1}{y_{i,-p}}, \underset{p\times 1}{\rho}\right) &= \begin{pmatrix} \iota'(I_{T-1+p}\otimes y'_{i,-p}c_2)_{1(p:T+p-1)} \\ \iota'(I_{T-1+p}\otimes y'_{i,-p}c_2)_{1(p-1:T+p-2)} \\ \cdots \\ \iota'(I_{T-1+p}\otimes y'_{i,-p}c_2)_{1(1:T)} \end{pmatrix}
\end{aligned} \tag{2.55}
$$

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 (2.23).

Since the right hand side of (2.24) 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,

$$
d\,\tau(\rho) = \sum_{k=1}^{p} h_k(\rho)\, d\rho_k. \tag{2.56}
$$

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 2.1.

[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 (2.10), involves all the common parameters and the incidental parameter, what strategy is required for consistent estimation. It should depend on the specific problems.

Lemma 2.1.

$$\frac{\partial P_{(1,1)}^{i+j}}{\partial \rho_i} = \frac{\partial P_{(1,1)}^{i'+j}}{\partial \rho_{i'}} \tag{2.57}$$

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]

$$P_{(1,1)}^{n} = \sum_{k=1}^{p} \rho_k P_{(1,1)}^{n-k}. \tag{2.58}$$

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 (2.57) by mathematical induction, which involves the following three steps:

1. We assume that for any integer less than j equation (2.57) holds. The left and right hand side of (2.57) can be rewritten as

$$\frac{\partial P_{(1,1)}^{i+j}}{\partial \rho_i} = \rho_1 \frac{\partial P_{(1,1)}^{i+j-1}}{\partial \rho_i} + \cdots + \frac{\partial \left(\rho_i P_{(1,1)}^{i+j-i} \right)}{\partial \rho_i} + \cdots + \rho_{i'} \frac{\partial P_{(1,1)}^{i+j-i'}}{\partial \rho_i} + \cdots + \rho_p \frac{\partial P_{(1,1)}^{i+j-p}}{\partial \rho_i} \tag{2.59}$$

$$\frac{\partial P_{(1,1)}^{i'+j}}{\partial \rho_{i'}} = \rho_1 \frac{\partial P_{(1,1)}^{i'+j-1}}{\partial \rho_{i'}} + \cdots + \rho_i \frac{\partial P_{(1,1)}^{i'+j-i}}{\partial \rho_{i'}} + \cdots + \frac{\partial \left(\rho_{i'} P_{(1,1)}^{i'+j-i'} \right)}{\partial \rho_{i'}} + \cdots + \rho_p \frac{\partial P_{(1,1)}^{i'+j-p}}{\partial \rho_{i'}} \tag{2.60}$$

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

$$\rho_n \frac{\partial P_{(1,1)}^{i+j-n}}{\partial \rho_i} = \rho_n \frac{\partial P_{(1,1)}^{i'+j-n}}{\partial \rho_{i'}}, \tag{2.61}$$

   where $n \in \{1, 2, \ldots, p\} \setminus \{i, i'\}$. Now to prove (2.59) and (2.60) are

[25] It is obvious that if $i = i'$, equation (2.57) 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 $j - n < j$.

equal to each other is reduced to proving

$$\frac{\partial\left(\rho_i P_{(1,1)}^{i+j-i}\right)}{\partial\rho_i}+\rho_{i'}\frac{\partial P_{(1,1)}^{i+j-i'}}{\partial\rho_i}=\rho_i\frac{\partial P_{(1,1)}^{i'+j-i}}{\partial\rho_{i'}}+\frac{\partial\left(\rho_{i'}P_{(1,1)}^{i'+j-i'}\right)}{\partial\rho_{i'}},\quad(2.62)$$

which is equivalent to

$$P_{(1,1)}^{j}+\rho_i\frac{\partial P_{(1,1)}^{i+j-i}}{\partial\rho_i}+\rho_{i'}\frac{\partial P_{(1,1)}^{i+j-i'}}{\partial\rho_i}=P_{(1,1)}^{j}+\rho_i\frac{\partial P_{(1,1)}^{i'+j-i}}{\partial\rho_{i'}}+\rho_{i'}\frac{\partial P_{(1,1)}^{i'+j-i'}}{\partial\rho_{i'}}.\quad(2.63)$$

It is not hard to see that (2.63) is true due to our assumption. Finally we know that if (2.57) 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 (2.57) are equal to 1. So (2.57) holds.

3. From the above two points, we know that Lemma 2.1 is true.

□

Now we are ready to prove that there exists a solution for the partial differential equation system (2.56).

Proof. It can be seen from (2.56) that if the system has a solution, the differential of $\tau(\rho)$ must be exact, which implies the following must be satisfied,

$$\frac{\partial h_i(\rho)}{\rho_{i'}}=\frac{\partial h_{i'}(\rho)}{\rho_i}\quad(2.64)$$

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 (2.64), we need to have

$$\frac{T-i-i'}{T}\frac{\partial P^{i'}_{(1,1)}}{\partial \rho_{i'}}+\cdots+\frac{1}{T}\frac{P^{i'+T-i-i'-1}_{(1,1)}}{\partial \rho_{i'}}=\frac{T-i'-i}{T}\frac{\partial P^{i}_{(1,1)}}{\partial \rho_{i}}+\cdots+\frac{1}{T}\frac{P^{i+T-i-i'-1}_{(1,1)}}{\partial \rho_{i}} \tag{2.65}$$

By Lemma 2.1, we know that (2.65) is true. Hence (2.64) is true and $d\,\tau(\rho)$ is exact. So we can conculde that $\tau(\rho)$ exists and (2.56) has a solution. □

Next we go on to solve (2.56). A solution for $\tau(\rho)$ can take the following form,

$$\tau(\rho) = R_1(\rho) + \phi_1(\rho_{2:p}) \tag{2.66}$$

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

$$\frac{\partial \tau(\rho)}{\partial \rho_2} = h_2(\rho) = \frac{\partial R_1(\rho)}{\partial \rho_2} + \frac{\partial \phi_1(\rho_{2:p})}{\partial \rho_2}. \tag{2.67}$$

Hence

$$\phi_1(\rho_{2:p}) = \int \left(h_2(\rho) - \frac{\partial R_1(\rho)}{\partial \rho_2} \right) d\rho_2 + \phi_2(\rho_{3:p}). \tag{2.68}$$

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

$$\tau(\rho) = \sum_{i=1}^{p} R_i(\rho_{i:p}) + k \tag{2.69}$$

where k is an arbitrary constant not depending on ρ and

$$R_i(\rho_{i:p}) = \int \left(h_i(\rho) - \sum_{j=1}^{i-1} \frac{\partial R_j(\rho_{j:p})}{\partial \rho_i} \right) d\rho_i \qquad for \quad i = 2, \ldots, p \tag{2.70}$$

with $R_1(\rho) = \int h_1(\rho) d\rho_1$. If we look at (2.69) 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).

2.5.2 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),

$$\begin{aligned}\ln p(Y|r,b,s^2) &\propto Q_N(r,b,s^2) \\ &= -\frac{1}{2s^2}\sum_i (y_i - Y_{i\_}r - X_i b)'H(y_i - Y_{i\_}r - X_i b) - \frac{N(T-1)}{2}\ln s^2 + N\tau(r).\end{aligned} \tag{2.71}$$

where r, b and s^2 are the specific values that θ takes. Subsituting (2.17) into (2.71) yields

$$\begin{aligned}\ln p(Y|r,b,s^2) &\propto Q_N(r,b,s^2) \\ &= -\frac{1}{2s^2}\Bigg\{(\rho-r)'\sum_i Y_{i\_}'HY_{i\_}(\rho-r) + (\beta-b)'\sum_i X_i'HX_i(\beta-b) \\ &+ \sum_i u_i H' u_i + 2(\rho-r)'\sum_i Y_{i\_}'Hu_i + 2(\rho-r)'\sum_i Y_{i\_}'HX_i(\beta-b) \\ &+ 2\sum_i u_i'HX_i(\beta-b)\Bigg\} - \frac{N(T-1)}{2}\ln s^2 + N\tau(r).\end{aligned} \tag{2.72}$$

Next we assume the following probability limits exist:

$$\begin{aligned}
&\underset{N\to\infty}{plim}\frac{1}{N}\sum_{i}^{N}Y_{i_}'HY_{i_}=\underset{p\times p}{Y_Y_}\\
&\underset{N\to\infty}{plim}\frac{1}{N}\sum_{i}^{N}Y_{i_}'Hu_i=\sigma^2\begin{pmatrix}trace\left(HC_{(p:T+p-1,:)}\right)\\ trace\left(HC_{(p-1:T+p-2,:)}\right)\\ \cdots\\ trace\left(HC_{(1:T,:)}\right)\end{pmatrix}=-\sigma^2h(\rho)\\
&\underset{N\to\infty}{plim}\frac{1}{N}\sum_{i}^{N}Y_{i_}'HX_i=\underset{p\times K}{Y_X}\\
&\underset{N\to\infty}{plim}\frac{1}{N}\sum_{i}^{N}X_i'HX_i=\underset{K\times K}{XX}
\end{aligned}\tag{2.73}$$

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

$$\begin{aligned}
\underset{N\to\infty}{plim}\frac{1}{N}Q_N(r,b,s^2)&=Q(r,b,s^2)\\
&=-\frac{1}{2s^2}\Big\{(\rho-r)'\left(Y_Y_\right)(\rho-r)+(\beta-b)'\left(XX\right)(\beta-b)+(T-1)\sigma^2\\
&\quad-2\sigma^2(\rho-r)'h(\rho)+2(\rho-r)'\left(Y_X\right)(\beta-b)\Big\}-\frac{T-1}{2}\ln s^2+\tau(r).
\end{aligned}\tag{2.74}$$

[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:

$$\begin{aligned}
\frac{\partial Q}{\partial r} &= \frac{1}{s^2}\left[(\rho - r)'\left(Y_{-}Y_{-}\right) - \sigma^2 h(\rho) + \left(Y_{-}X\right)(\beta - b)\right] + h(r) \\
\frac{\partial Q}{\partial b} &= \frac{1}{s^2}\left[XX(\beta - b) + (\rho - r)'(Y_{-}X)\right] \\
\frac{\partial Q}{\partial s^2} &= \frac{1}{2s^2}\Big\{(\rho - r)'\left(Y_{-}Y_{-}\right)(\rho - r) + (\beta - b)'\left(XX\right)(\beta - b) + (T-1)\sigma^2 \\
&\quad - 2\sigma^2(\rho - r)'h(\rho) + 2(\rho - r)'\left(Y_{-}X\right)(\beta - b)\Big\} - \frac{(T-1)}{2s^2}.
\end{aligned} \tag{2.75}$$

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.

2.5.3 Proof of Proposition 2.1

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

$$\begin{aligned}
p(\theta)p(Y|\theta, Y_0) = \frac{1}{m(S)} I(\rho \in S) p(\beta|\sigma^2)(2\pi)^{-\frac{TN}{2}} \sigma^{2(-\frac{NT+2}{2})} r(\rho) \\
\prod_{i=1}^{N} \exp\left\{-\frac{1}{2\sigma^2}\left[w_i - f_i\iota - X_i\beta\right]'\left[w_i - f_i\iota - X_i\beta\right]\right\},
\end{aligned} \tag{2.76}$$

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 (2.76) as

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

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

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

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

$$\begin{aligned} p(\theta)p\left(Y|\theta, Y_0\right) = p(\beta|\sigma^2)\frac{1}{m(S)}I(\rho \in S)(2\pi)^{-\frac{TN}{2}}\sigma^{2(-\frac{NT+2}{2})}r(\rho) \\ \prod_{i=1}^{N}\exp\left\{-\frac{1}{2\frac{\sigma^2}{T}}[f_i - \frac{\iota'(y_i - Y_{i\_}\rho - X_i\beta)}{T}]^2\right\} \\ \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - Y_{i\_}\rho - X_i\beta)'H(y_i - Y_{i\_}\rho - X_i\beta)\right] \end{aligned} \tag{2.77}$$

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 (2.77) 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 (2.29).

Next we go on to derive the posterior distributions for β and σ^2. First

we can integrate out g in equation (2.77) to obtain

$$\begin{aligned} p(\rho,\beta,\sigma^2,Y|Y_0) &= p(\rho,\beta,\sigma^2|Y,Y_0)p(Y|Y_0) \\ &= p(\beta|\sigma^2)\frac{1}{m(S)}I(\rho\in S)T^{-\frac{N}{2}}(2\pi)^{-\frac{N(T-1)}{2}}\sigma^{2\left[-\frac{N(T-1)+2}{2}\right]} \\ &\quad r(\rho)\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-Y_{i\_}\rho-X_i\beta)'H(y_i-Y_{i\_}\rho-X_i\beta)\right]. \end{aligned} \tag{2.78}$$

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 (2.28) we can rewrite equation (2.78) 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

$$\begin{aligned} & p(\rho,\beta,\sigma^2|Y,Y_0)p(Y|Y_0) \\ =& \frac{1}{m(S)}I(\rho\in S)T^{-\frac{N}{2}}(2\pi)^{-\frac{N(T-1)+k}{2}}\sigma^{2\left[-\frac{N(T-1)+2+k}{2}\right]}r(\rho)\left|\eta\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right|^{\frac{1}{2}} \\ & \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}\tilde{w}_i'\tilde{w}_i-\frac{1}{\eta+1}\sum_{i=1}^{N}\tilde{w}_i'\tilde{X}_i\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i\right]\right\} \\ & \exp\left\{-\frac{1}{2\sigma^2}\left[\beta-\frac{1}{\eta+1}\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i\right]'\right. \\ & \left.\left(\sum_{i=1}^{N}(\eta+1)\tilde{X}_i'\tilde{X}_i\right)\left[\beta-\frac{1}{\eta+1}\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i\right]\right\} \end{aligned} \tag{2.79}$$

We can see that the conditional posterior of β follows a normal distribution as in (2.30). Now we can integrate out β in (2.79) to obtain the posterior distribution for ρ and σ^2 as the following,

$$p(\rho,\sigma^2|Y,Y_0)p(Y|Y_0) = \frac{1}{m(S)} I(\rho\in S)\left(\frac{\eta}{\eta+1}\right)^{\frac{k}{2}} T^{-\frac{N}{2}}(2\pi)^{-\frac{N(T-1)}{2}} \sigma^{2\left[-\frac{N(T-1)+2}{2}\right]} r(\rho)\exp\Bigg\{-\frac{1}{2\sigma^2}[\sum_{i=1}^{N}\tilde{w}_i'\tilde{w}_i - \frac{1}{\eta+1}\sum_{i=1}^{N}\tilde{w}_i'\tilde{X}_i\left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i'\tilde{w}_i]\Bigg\}. \quad (2.80)$$

It is clear from equation (2.80) 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 (2.31),

Now we can integrate out σ^2 to obtain the posterior distribution of ρ as in (2.82).

$$p(\rho|Y,Y_0)p(Y|Y_0) = \frac{1}{m(S)} I(\rho\in S)\left(\frac{\eta}{\eta+1}\right)^{\frac{k}{2}} (\Delta)^{-\frac{N(T-1)}{2}} \Gamma\left[\frac{N(T-1)}{2}\right] T^{-\frac{N}{2}}(\pi)^{-\frac{N(T-1)}{2}} r(\rho) \quad (2.81)$$

$$p(\rho|Y,Y_0) \propto I(\rho\in S) r(\rho)\,(\Delta)^{-\frac{N(T-1)}{2}}, \quad (2.82)$$

Another way to interpret the posterior of ρ is given in (2.33) under Proposition 2.1.

2.5.4 Proof of Equation (2.48)

Proof. Note that there is a differentiable mapping from $(\pi^{(0)\prime}, u')'$ to $\rho^{(c)}$, whose Jacobian is given in (2.36), and also from $\rho^{(0)}$ to $\pi^{(0)}$. Hence we can

obtain

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

□

2.5.5 Data Generating Process for the Exogenous Regressors in Section 2.3.5

We go through the following steps to generate the exogenous regressors used in Section 2.3.5:

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,

$$x_{t,s} = s_{t-1}x_{t-1,s} + \bar{s}_t x_{t,ns}, \tag{2.83}$$

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

$$S = \begin{pmatrix} 1 & s_1 & \cdots & \prod_{i=1}^{T-1} s_i \\ s_1 & 1 & \cdots & \prod_{i=2}^{T-1} s_i \\ s_2 s_1 & s_2 & \cdots & \prod_{i=3}^{T-1} s_i \\ \cdots & \cdots & \cdots & \cdots \\ \prod_{i=1}^{T-1} s_i & \prod_{i=2}^{T-1} s_i & \cdots & 1 \end{pmatrix} \tag{2.84}$$

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.

$$X_{j,c} = \sum_{i=1}^{K} q_{j,i} X_{i,nc} \qquad j = 1, 2, \ldots, K \tag{2.85}$$

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

[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.

same across different individuals and can be calculated as

$$corr(X_{t,k}, X_{t',k'}) = S(t,t')\sum_{i=1}^{K} q_{k,i}q_{k',i} \qquad t = 1,2,\ldots,T \quad k = 1,2,\ldots,K. \tag{2.86}$$

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

Chapter 3

The Horizon Effect of Stock Return Predictability and Model Uncertainty on Portfolio Choice: UK Evidence

3.1 Introduction

Finance advisors often tell people with long investment horizon to invest more into stocks than bonds. Fund managers will recommend different portfolios to investors with different investment horizons. For example, they may recommend some stock shares for long term investment and some others just for short term. Such ideas to allocate wealth according to the length of investment horizon have been challenged by academics. Early work about horizon effect can be seen in Samuelson (1969) and Merton (1969), in which they prove that if the return of a risky asset is unpredictable, rational investors should choose the same portfolio regardless of the length of their investment. Later work by Samuelson (1989) and Samuelson (1990)

readdressed the irrelevance of the length of investment horizon in portfolio management.

The absence of horizon effect primarily relies on the assumption that the return of the risky asset is unpredictable. However, there are also studies showing that return predictability can affect investor's optimal portfolio decision, see, for example, Kandel and Stambaugh (1996), Barberis (2000) and Xia (2001). To add more valuable insight into this debate, it is important to understand the nature of stock market inefficiencies, which is closely related to the question of whether stock return is predictable or not. Though most studies using daily or weekly data find very little evidence of predictability in terms of low R-squares or low p-values, many academic investigations into monthly data suggest some variables may have the ability to explain the movements in stock expected return. Fama and French (1988) reported that apart from dividend yields, past stock return in the US market can predict 40 percent of future return over the long run. Fama and French (1993) then identified five common risk factors in explaining the return of stocks and bonds. Consistent with Fama and French's results, Kothari and Shanken (1997) also found that book-to-market ratio (B/M) has predictive power. However, these studies have invited criticisms from other scholars. Hodrick (1992) and Goetzmann and Jorion (1993) argued that many findings based on long-horizon return regressions may be inappropriate due to problems such as data snooping[1], nonrobustness of test statistics and poor small sample properties of the inference method.

Such controversy about stock return predictability can be better explained from two aspects. First, though there are many articles addressing the issue of stock return predictors, there is little consensus on what the important conditioning variables are. This issue can be regarded as model uncertainty, which in general refers to our uncertainty about the underlying data generating process (DGP) of stock return, see, for example, Brennan and Xia (1999). Secondly, even if one believes to have found the correct set of predictors, the predictive relationship between stock return and the predictors cannot be estimated with certainty due to limited sample size.

[1] It implies that such patterns in the data may happen by chance.

In other words, it is not possible for us to identify the true values of the parameters for our model in real life application. Parameter uncertainty or estimation risk also has an important effect on investor's optimal portfolio choice, see Bawa et al. (1979) and Barberis (2000). By taking into account both parameter and model uncertainty, one could better answer the question of whether stock return is predictable or not. Cremers (2002) and Avramov (2002) both used Bayesian model averaging (BMA) to consider such uncertainty and found that the BMA method, which averages all the potential models according to their posterior probabilities, can provide better forecasts of stock return than those selected based on certain criterion. The above studies are based on the US stock market. Relevant research on the UK market can be seen in Pesaran and Timmermann (1995), in which they employed recursive regression method to select a best single model based on certain information criterion to make out-of-sample forecasts. Though they acknowledged there was uncertainty about which model best forecasted stock returns over time, they did not address this issue explicitly in their method.

In this chapter we study the stock return predictability in the UK market by accounting for both parameter and model uncertainty. We then investigate the effect of such predictability on a rational investor's portfolio choice given different lengths of investment horizons. We find that the stock return predictability in the UK market is weak if we allow for model uncertainty. Many explanatory variables are not as strong predictors as classical results suggest. Moreover, if we take account of the data generating processes of the explanatory variables and allow them to be correlated with that of the stock return, the predicting power of these explanatory variables will decrease further. As for the horizon effect, we propose a computationally convenient statistic that can be used as a reference for how a rational buy-and-hold investor should adjust her optimal portfolio given different lengths of investment. We find that although the return predictability is weak, it still has a considerable effect on a rational buy-and-hold investor's portfolio choice as evidenced by different allocation proportion of wealth to risky asset given different initial investment conditions over time.

The chapter proceeds as follows. Section 3.2 explains the asset allocation problem and the computation techniques used to solve it. Section 3.3 investigates the horizon effect when the risky asset's return is unpredictable. We look into the cases with and without parameter uncertainty and then propose a measure to capture the horizon effect. Section 3.4 studies the stock return predictability in the UK market by considering model uncertainty. Section 3.5 then examines the horizon effect of stock return predictability and model uncertainty. Finally Section 3.6 concludes.

3.2 The Asset Allocation Problem and the Calculation of the Optimal Portfolio

The basic economic model of the analysis consists of a risk averse investor, who allocates her wealth to either riskless (e.g. treasury bond) or risky asset (e.g. stock share) in order to maximize her utility function. This model has been studied by Kandel and Stambaugh (1996), Barberis (2000) and Avramov (2002) with a focus on the time horizon effects, i.e. how the investor will allocate her wealth given different lengths of investment horizons. Different from their studies, we will look into the horizon effect based on the UK data. Compared to Avramov (2002), we will take into account not only the effects of parameter and model uncertainty, but also the interactions between the data generating process (DGP) of the return of the risky asset and those of its explanatory variables. Moreover, we will propose a computationally convenient statistic, which may shed some light on the behavior of a rational investor when she has to choose between risky and riskless asset.

The investor's preference over terminal wealth is described by the constant relative risk-aversion power utility function (v) with the following form.

$$v(W) = \begin{cases} \frac{W^{1-A}}{1-A} & for \quad A > 0 \quad and \quad A \neq 1 \\ \ln W & for \quad A = 1 \end{cases} \tag{3.1}$$

where A is commonly referred to as the investor's coefficient of relative risk

aversion and W denotes the investor's wealth. Without loss of generality, we assume the initial wealth of the investor is equal to one. Let us denote the rate of return of the riskless asset by r_f and the excess return of the risky asset over the riskless by r[2]. For simplicity, we further assume that r_f is non-stochastic and only r is a random variable. Suppose the investor is going to hold the portfolio of the two assets from period T till period $T+\hat{T}$. At the end of her investment horizon, her cumulative excess return will be $R_{T+\hat{T}} = r_{T+1} + r_{T+2} + ... + r_{T+\hat{T}}$, which will also follow a certain distribution. If we assume the returns are continuously compounded and the investor allocates ω of her wealth to the risky asset, her total wealth at the end of the investment will be $(1-\omega)\exp(\hat{T}r_f) + \omega\exp(\hat{T}r_f + R_{T+\hat{T}})$. The asset allocation problem for the investor is to solve

$$\max_{\omega} \int_{R_{T+\hat{T}}} \frac{\left[(1-\omega)\exp(\hat{T}r_f) + \omega\exp(\hat{T}r_f + R_{T+\hat{T}})\right]^{1-A}}{1-A} \quad p(R_{T+\hat{T}})dR_{T+\hat{T}} \tag{3.2}$$

That is, given a period of time, which is $\hat{T}$ periods long, the problem facing the investor is to choose ω to maximize her expected utility at the start of her investment, i.e. period T. Our study will focus on the investment horizon effect, i.e. the relationship between ω and $\hat{T}$. For the moment, we just assume the integral in (3.2) exists and will leave more detailed discussion of this issue to later section. Note that it is generally impossible to obtain a closed form solution for (3.2) even if $p(R_{T+\hat{T}})$ is some standard density function. To solve the problem, Barberis (2000) restricted ω to $[0,1]$ and performed a grid search after simulating draws from $p(R_{T+\hat{T}})$ to integrate $R_{T+\hat{T}}$ out. Here we use a relatively convenient and possibly more efficient numerical method to tackle this problem. First we use Taylor expansion to approximate the power utility function around the mean of $R_{T+\hat{T}}$ to produce a polynomial of $R_{T+\hat{T}}$. We can choose the order of Taylor expansion to control the approximation accuracy. Then we obtain the moments of $R_{T+\hat{T}}$ analytically or by simulation and insert them into the polynomial to

[2]That is the difference of the rate of return between the two assets.

obtain a function of only ω. Finally we use a numerical routine to maximize the function.[3] Our method relies heavily on the existence of the high order moments of $R_{T+\hat{T}}$. However, we argue that if the moments of $R_{T+\hat{T}}$ do not exist to certain orders, we should cast doubt on the existence of the integral in (3.2). In our application in later sections, we find that Taylor approximation with order around 10 could give us reasonably accurate results when $R_{T+\hat{T}}$ follows a normal or t distribution with high degrees of freedom.

Next we discuss the force that may drive the horizon effect. Note that the demand for the risky asset in the investor's portfolio clearly hinges on how we set up the maximization problem and the constraints confronting the investor. However, it should be no surprise that the risky asset's return and its level of risk are the key factors. In other words, the first and the second moments of $R_{T+\hat{T}}$ should have an important role in determining the horizon effect. Note that the density function $p(R_{T+\hat{T}})$ will change with $\hat{T}$. Hence both the first and the second moments of $R_{T+\hat{T}}$ are functions of $\hat{T}$. We may be interested in knowing how fast the return changes relative to the change of risk. For example, if the expected risk of an asset increases with time, will the asset's expected return increase fast enough to offset such effect so that the asset will still remain attractive to a rational investor? Here we propose the following expression which may help to answer this question.

$$MtoS = \frac{\frac{\partial \mu_{\hat{T}}}{\partial \hat{T}} \times \sigma_{\hat{T}}}{\frac{\partial \sigma_{\hat{T}}}{\partial \hat{T}} \times \mu_{\hat{T}}} \tag{3.3}$$

where $\mu_{\hat{T}}$ and $\sigma_{\hat{T}}$ denote the mean and standard deviation of $R_{T+\hat{T}}$ respectively. The expression in (3.3) is no more than the ratio between the percentage rate of change of $R_{T+\hat{T}}$'s mean and standard deviation. It could be interpreted as the Sharpe ratio in a time context and provides a measure of the value of risk (in terms of the mean return) over time. In the following sections, we will illustrate how expression (3.3) is related to the investment horizon effect under different probability density functions of $R_{T+\hat{T}}$.

[3] All these procedures could be easily implemented in Maple once we obtain the moments of $R_{T+\hat{T}}$.

3.3 When the Excess Return is Unpredictable

Samuelson (1969) and Merton (1969) showed that when stock's return is not predictable, the optimal portfolio will be independent of wealth and all consumption-saving decisions in a multi-period portfolio rebalancing model. Different from Samuelson (1969), the distribution of the excess return will change with time in this chapter. Barberis (2000) used the US data and shows that the optimal portfolio is insensitive to investment time horizon if $R_{T+\hat{T}}$ is unpredictable and follows a normal distribution with mean and variance increasing linearly with time. In our empirical study, we will use the UK 3 month treasury bill rate as r_f. The excess return of the risky asset (r) is calculated as the return difference between the FTSE All-Share Index and r_f. Now we assume the excess return is unpredictable and follows a normal distribution as below,

$$r_t = \mu + \epsilon_t, \qquad \epsilon_t \sim IIDN\left(0, \sigma^2\right) \tag{3.4}$$

where μ is the mean of the stock excess return and σ^2 is the variance in the normal distribution. The cumulative excess return $R_{T+\hat{T}}$ will also be normal as the following,

$$R_{T+\hat{T}}|\mu, \sigma^2 \sim N(\mu_{\hat{T}}, \sigma^2_{\hat{T}}) \tag{3.5}$$

where $\mu_{\hat{T}} = \hat{T}\mu$, and $\sigma^2_{\hat{T}} = \hat{T}\sigma^2$. As pointed out by Barberis (2000), if the investor ignores parameter uncertainty, i.e.taking the estimates of μ and σ from the past data as the true values of these parameters, the optimal holding proportion of the risky asset (ω) will not change with time. It is easy to see that under such setup, $MtoS$ defined in (3.3) is equal to 2 and also independent of $\hat{T}$.

Our data sample is from November 1978 up to September 2003, which includes 299 (T) observations of the FTSE All-Share Index. The mean of the excess rate of return (μ) of the FTSE index over T-bill in our sample is 0.4772% , while the sample standard deviation is 4.88%(σ). When calculating the optimal ω, Barberis (2000) restricted ω to be within 0 and 1 to

ensure the integral in (3.2) exists. The case corresponding to $\omega < 0$ happens if the investor sells short the risky asset and the case for $\omega > 1$ means the investor sells short the riskless asset. The nonexistence of the integral in (3.2) arises since we allow infinite disutility for the power utility function. To understand this point, first note that A is the investor's coefficient of relative risk aversion in (3.1) and usually takes the value of a positive integer. If the investor's wealth is equal to zero, [4], we can easily find that the investor will suffer infinite disutility[5] if $R_{T+\hat{T}} = \ln(\frac{\omega-1}{\omega})$. It is clear that when $0 \leq \omega \leq 1$, $R_{T+\hat{T}} \neq \ln(\frac{\omega-1}{\omega})$ will always hold[6] and the expectation in equation (3.2) will exist for certain distribution of $R_{T+\hat{T}}$. But if we allow for short sale of either the risky or riskless asset, the existence of the expectation may be in question with A bigger than 1. On the other hand, it is reasonable to expect the investor to sell short the risky (riskless) asset when the expected return of the risky asset is a large negative (positive) percentage number. So to allow for a bit more generality, we may restrict ω to be within $[-0.3, 1.3]$ such that the real values of $\ln \frac{\omega-1}{\omega}$ will range outside $(-1.46, 1.46)$. Since it is rare to observe a return or a loss of more than 150% for $R_{T+\hat{T}}$ in our data given that $\hat{T}$ is a moderate number, we could expect our calculation algorithms to work well.

In our following studies we set $A = 5$ and $r_f = 0.3\%$, which is the last observation of the monthly rate of return of the 3-month T-bill. Here we study the horizon effect from one month to 5 years, i.e. 60 months. By using the method described in the previous section[7], the optimal holding proportion of the risky asset and the $MtoS$ defined in (3.3) are shown in the left column of Figure 3.1 for different investment lengths. We can see that ω is about 0.5 while $MtoS$ is 2. Both of them do not change with time. These results confirm the empirical findings of Barberis (2000) using the US data. We have just added $MtoS$ to analyze the relative change of the return and risk of the risky asset over time.

[4] That is $(1-\omega)\exp(\hat{T}r_f) + \omega\exp(\hat{T}r_f + R_{T+\hat{T}}) = 0$

[5] That is $[(1-\omega)\exp(\hat{T}r_f) + \omega\exp(\hat{T}r_f + R_{T+\hat{T}})]^{1-A} = -\infty$ for $A \geq 1$.

[6] Note that $R_{T+\hat{T}}$ must be real.

[7] Note that it is easy to derive the analytic moments of different orders for a normal distribution once its first and second moments are known.

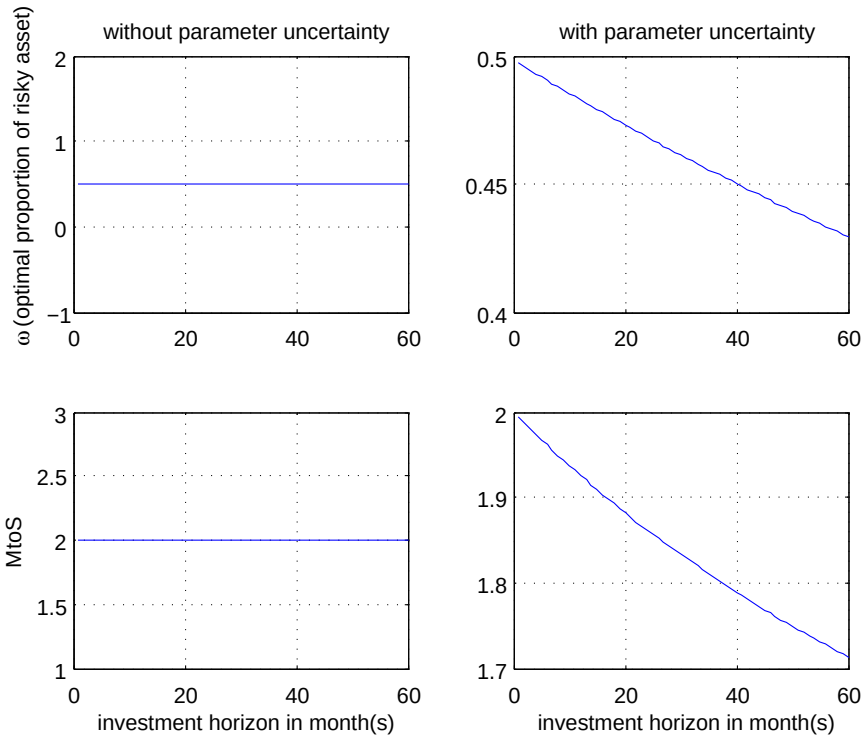


Figure 3.1: Optimal holding of stock and $MtoS$ with respect to time horizon when excess return is unpredictable

Next we turn our attention to the case when the investor no longer treats the estimates of μ and σ as their true values, i.e. the investor is now taking parameter uncertainty into account as termed by Barberis (2000). To model the parameters μ and σ in (3.5) as random variables, we adopt the Bayesian inference framework by assuming the joint distribution of μ and σ^2 follows a noninformative prior and hence their posterior follows a normal-gamma distribution.

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \tag{3.6}$$

$$\mu|\sigma^2, D \sim N(\bar{r}, \frac{\sigma^2}{T}) \tag{3.7}$$

$$\sigma^2|D \sim IG\left((T-1)s^2, T-1\right) \tag{3.8}$$

where $\bar{r} = \frac{1}{T}\sum_{1}^{T} r_t = 0.4772\%$ and $s^2 = \frac{\sum_{t=1}^{T}(r_t-\bar{r})^2}{T-1} = 0.0023797$. Here σ^2 follows an inverted gamma distribution with degrees of freedom $T-1$. Equation (3.7) and (3.8) show the posterior distributions of μ and σ^2, which are conditional on the data[8], denoted by D. The posterior mean and variance of σ^2 are

$$E(\sigma^2|D) = \frac{\sum_{t=1}^{T}(r_t - \bar{r})^2}{T-3} = 2.40 \times 10^{-3} \tag{3.9}$$

$$Var(\sigma^2|D) = \frac{2(\sum_{t=1}^{T}(r_t - \bar{r})^2)^2}{(T-3)^2(T-5)} = 3.91 \times 10^{-8} \tag{3.10}$$

The conditional distribution of $R_{T+\hat{T}}$ is still given by (3.5). However, the posterior distribution of $R_{T+\hat{T}}$ unconditional of μ and σ^2 now becomes

$$R_{T+\hat{T}}|D \sim t(\hat{\mu}_{\hat{T}}, \sigma^2_{\hat{T}}, T-1) \tag{3.11}$$

where $\hat{\mu}_{\hat{T}}$ denotes the mean parameter, which is equal to $\bar{r}\hat{T}$ and $\sigma^2_{\hat{T}}$ is the variance parameter equal to $s^2\hat{T}(1+\frac{\hat{T}}{T})$. With parameter uncertainty of μ and σ^2, it is equivalent as saying that $R_{T+\hat{T}}$ in (3.2) has a t density function

[8] Here the data are the observed excess returns.

with parameters described in (3.11). It can be seen that the mean and variance of $R_{T+\hat{T}}$ grow at different speeds as compared to (3.5). The ratio of the percentage rate of change between them ($MtoS$) is now the following,

$$MtoS = \frac{\frac{\partial \hat{\mu}_{\hat{T}}}{\partial \hat{T}} \times \hat{\sigma}_{\hat{T}}}{\frac{\partial \hat{\sigma}_{\hat{T}}}{\partial \hat{T}} \times \hat{\mu}_{\hat{T}}} = 2 - \frac{2\hat{T}}{T + 2\hat{T}}. \tag{3.12}$$

Unlike the case of no parameter uncertainty, this ratio depends on $\hat{T}$ and is a decreasing function of $\hat{T}$, whose value is less than 2 unless $\hat{T}$ is 0. Barberis (2000) shows that the optimal holding proportion of the risky asset under parameter uncertainty will no longer be insensitive to the length of investment horizon. It can be interpreted as the rational investor constantly updating her estimation of the risky asset's mean and standard deviation. As there are more samples to be included for estimation in the long run, she will become more doubtful about her initial estimation made at period T. Therefore her expected risk of the asset will grow faster with time than the case with no parameter uncertainty. To calculate the optimal ω, we will still use the same numerical method as for no parameter uncertainty, since for t-distribution it is also easy to derive the moments of different orders. The plots of the optimal portfolio and $MtoS$ are shown in the right column of Figure 3.1. We can see that both ω and $MtoS$ drop with the length of investment horizon. For ω, it falls by around 7% while $Mtos$ drops from 1.99 to 1.71. If we look at the expression of $MtoS$ in (3.12), we could find that it also involves T, i.e. the original sample size. We may conjecture that the optimal portfolio may also be related to the sample size. Indeed, if we keep the values of all other parameters the same and just change T to its one tenth, the optimal ω will fall from 46% to 23% while $MtoS$ is from 1.93 to 1.2. This can be due to the fact that as the investor's estimation is based on smaller sample size, she will have less confidence in it and will feel that the asset is more risky in the long run. Hence the size of the drop of ω is much larger and the horizon effect is more pronounced. The interesting point here is that the ratio between the percentage rate of change of the excess return's mean and standard deviation ($MtoS$) seems to be able to

tell how a rational investor will behave given different lengths of investment horizons. However, in many applications, such as the one in the following sections, the first and second moments of $R_{T+\hat{T}}$ may not have closed forms, needless to say $MtoS$. Here we propose the following statistic, which can circumvent this problem and approximate the $MtoS$,

$$\hat{MtoS} = \frac{\ln(\mu_{\hat{T}}/\mu_{\hat{T}-1})}{\ln(\sigma_{\hat{T}}/\sigma_{\hat{T}-1})} \approx MtoS, \tag{3.13}$$

which is the ratio of the log differences between the contemporaneous expected mean and the one of one period earlier to its standard deviation counterpart[9]. The statistic approximates the instantaneous relative percentage change of expected return to that of risk. As we argued before, the $MtoS$ could be viewed as a measure of the economic value of risk in terms of return over time. While the optimal holding proportion of the risky asset is hard to calculate and depends on the setup of the maximization problem, such as what form the utility function takes and how risk averse is the investor, the statistic defined in (3.13) is easy to calculate and may provide a reference for the investor as to how attractive a particular portfolio is over time. We will apply this statistic to the subsequent sections where the analytical forms of the first and second moments of $R_{T+\hat{T}}$ are not available.

3.4 Whether Stock Return is Predictable or Not

3.4.1 Data and Summary of Statistical Results

All our data, except HGSC Index, are from DataStream, covering the period from November 1978 to September 2003, altogether 299 observations (T). As before, we use r to denote the FTSE All-Share Index excess return, which is our dependent variable. Its explanatory variables along with their short forms used in the analysis are shown in the following list. Consistent

[9] Here we assume $\mu_{\hat{T}}$ is a monotonic function of $\hat{T}$ and $\mu_{\hat{T}}$ and $\mu_{\hat{T-1}}$ have the same sign. Note that with no parameter uncertainty (3.13) is equal to 2 just the same as the result calculated in (3.3) given $\hat{T} > 1$. For the case with parameter uncertainty, the behaviour of (3.13) is similar to that of (3.3) and they have the same limit as $\hat{T}$ tends to infinity.

Table 3.1: Details of the explanatory variables

1. January Dummy (Jan), which captures the January effect in the stock market
2. monthly return of the three-month Treasury bill (Tb)
3. the first difference of Treasury bill ($Tbchng$), which is calculated as $Tb(t) - Tb(t-1)$
4. the difference of return between small market capitalization companies and big ones (Smb), which is the difference between the total returns of Hoare Govett Smaller Companies index (HGSC) and FTSE 100 Index
5. dividend yield, the ratio of dividend over stock price(Dy)
6. the difference between monthly returns of 20 year UK government gilt and the 3 month T-bill ($TERM$)
7. monthly industrial production ($Indp$)
8. money supply, seasonally adjusted ($M0$)
9. monthly percentage change of industrial production ($Indp\%ch$)
10. monthly percentage change of monetary supply ($M0\%ch$)
11. Monthly inflation (Inf)
12. monthly oil price ($Oilp$)
13. monthly percentage change of oil price ($Oil\%ch$)
14. the difference between returns of high book-to-market ratio company index and low ones (HML), which is calculated as the difference between the total returns of MSCI value index and growth index
15. monthly change of inflation rate($Infch$), which is calculated as $Inf(t) - Inf(t-1)$

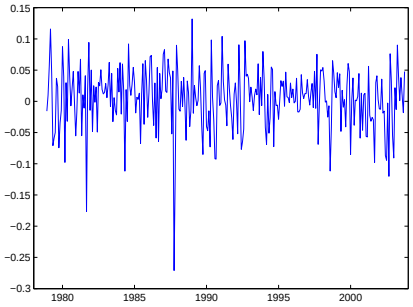


Figure 3.2: Monthly excess return on Financial Times All-Share Index

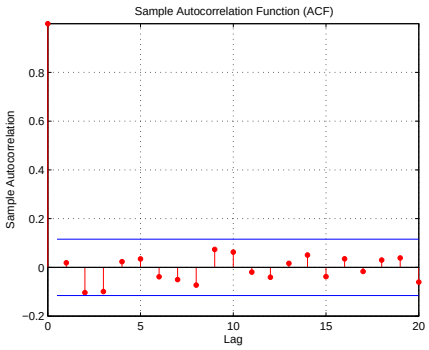


Figure 3.3: Sample autocorrelation of FTSE All-Share Index excess return

with the study by Pesaran and Timmermann (1995), we do not include the observation in October 1987[10], which is an outlier[11]. Figure 3.2 displays the monthly excess returns of the FTSE All-Share Index over our sample range. First sight suggests there do not seem to be any obvious patterns, such as autocorrelation. This can be confirmed in Figure 3.3[12]. The two parallel horizontal lines indicate the 95% confidence interval. We can see that all the autocorrelation coefficients up to twenty lags are well within the 95% confidence lines.

[10] We achieve this by putting a dummy variable (1987Oct) in the linear regression later.
[11] In that month, there was a stock market crash. The index dropped by around 27%.
[12] The results are obtained from the MatLab routine autocor.m.

Table 3.2: The OLS estimated excess return equation including all regressors

Regressors	Coefficients	T-statistic	P-value
Intercept	0.007	0.088	0.465
Jan	**0.0219**	**1.614**	**0.054**
Tb	**-9.1271**	**-1.7349**	**0.041**
Tbchng	**7.45**	**1.3325**	**0.092**
Smb	-0.0864	-1.0134	0.156
Dy	**2.265**	**3.1529**	**8.96e-4**
TERM	**-9.998**	**-2.018**	**0.0223**
Indp%ch	0.04985	0.225	0.411
M0%ch	-0.106	-0.795	0.214
Inf	-0.174	-1.093	0.138
Oil%ch	-0.0881	-0.734	0.232
HML	-0.0689	-0.746	0.228
Oilp	**0.00033**	**1.303**	**0.0968**
Indprd	-1.41e-5	-0.019	0.492
M0	**-0.0031**	**-2.04**	**0.0213**
Infch	0.0605	0.104	0.4588
R^2	0.17778		
Adjusted R^2	0.12804		

Apart from the excess returns, all other variables are either business cycle variables or financial market variables, which may possess explanatory power for excess returns. A rough idea about the extent to which the excess return can be predicted using different variables can be seen in the OLS regression results obtained by regressing the excess return on all other variables as summarized in Table 3.2. The numbers in bold indicate they are significant at 10% level of significance. The signs of the coefficients and R^2 statistics are close to the findings for the US market, see Granger (1992). However, such practice by regressing the excess return on all other variables could be subject to criticism such as data snooping and model misspecification. In the next subsection, we will use the BMA method to look into this issue from another perspective.

3.4.2 Bayesian Model Averaging in a Univariate Linear Model

The Efficient Market Hypothesis (EMH, e.g. Fama (1970)) states that in an efficient capital market, stock return is not predictable. Numerous empirical

work has shown that the capital market is not efficient in this sense. While traditional asset pricing models, like CAPM, precludes the use of predictors in determining return, the literature of style investment, which bases investment on certain economic or accounting variables, has prospered in the past decades. Banz (1981) documented that small-cap stocks have historically outperformed large-cap stocks in the US by a margin that could not be explained by conventional measures of risk. Hence the capital size of a stock may help predict its return. Later influential work can be seen in Fama and French (1993), who documented five common risk factors in the returns on stocks and bonds, i.e. the whole market returns, firm size, book-to-market ratio, maturity risk and default risk. For the UK market, Pesaran and Timmermann (1995) found that in addition to dividend yield, several business cycle variables (see Table 3.1) help to predict the excess return. Different variables can be seen in predicting returns in numerous other papers. The variables presented in the last subsection are based primarily on these studies. Though there are many articles mentioning possible predictors, there is little consensus on what the most important conditioning variables are.

Here we apply BMA techniques to a linear model to identify the most important predictors using the UK data. We assume the predictors and the dependent variable have a linear relationship with no serial correlation and heteroscedasticity in the disturbance, i.e.

$$r_t = \underline{a}_p + B_p' x_{t-1,p} + \epsilon_{t,p},\ \epsilon_{t,p} \sim i.i.d.N(0, \sigma_p^2) \tag{3.14}$$

where r stands for the excess return, and x stands for the set of predictors used, which does not include any lag terms of r due to the weak autocorrelation of excess returns as shown before. The subscript p is a model specific parameter, which implies the parameters are different for different models. There are altogether 15 (K) possible predictors which may enter x to explain the excess return. The total number of different models with different regressors, is $P = 2^{15}$. Each model, M_p, is described by a $K \times 1$ binary vector $\gamma = (\gamma_1, ..., \gamma_K)'$, where a one (zero) indicates the inclusion (exclusion) of a variable. We denote the sum of all elements in γ by k_p, which is the

dimension of the column vector $x_{t-1,p}$. If we stack up all the observations for equation (3.14), then it can be written as,

$$\mathbf{r} = \underline{a}_p\iota + \mathbf{X}_p B_p + \epsilon_p \qquad \epsilon_p \sim N(0, \sigma_p^2\mathbf{I}), \tag{3.15}$$

where ι is a vector of ones, $\mathbf{X}_p = [x_{0,p}, x_{1,p}, x_{2,p}, ..., x_{T-1,p}]'$, and $\mathbf{r}$ =[r_1, r_2, r_3,...r_T]'.

The following analysis relies heavily on the bench mark prior developed by Fernandez et al. (2001a). To implement their approach, we first reparameterize the intercept term ($\underline{a}_p$) in the regression such that the new intercept term (a_p) is orthogonal to the slope (B_p) in the likelihood function, i.e. $\underline{a}_p = a_p - \frac{\iota' X_p B_p}{T}$. In doing so, we have changed (3.15) into

$$\mathbf{r} = a_p\iota + H\mathbf{X}_p B_p + \epsilon_p,\ \epsilon_p \sim N(0, \sigma_p^2\mathbf{I}) \tag{3.16}$$

where $H = I_T - \frac{\iota\iota'}{T}$ is the demean matrix. The bench mark prior proposed by Fernandez et al. (2001a) looks like the following.

$$p(a_p, \sigma_p^2) \propto \frac{1}{\sigma_p^2} \tag{3.17}$$

$$B_p|\sigma_p^2, a_p \sim N(0, \sigma_p^2(g\mathbf{X}_p'H\mathbf{X}_p)^{-1}) \tag{3.18}$$

Here we use noninformative prior for the equation variance and the constant. For the slope vector B_p, we use the g prior designed by Zellner (1986), which uses the explanatory variables to specify the prior variance. It substantially reduces the trouble of eliciting too many hyperparameters. Now the strength of the prior only depends on g. Fernandez et al. (2001a) elicit g based on truth searching. They first generate hypothetical datasets in a linear model and then try different values for g to find the one which can identify the true model under different circumstances. After extensive experiments, they recommend choosing

$$g = \begin{cases} \frac{1}{T} & \text{if } T > K^2 \\ \frac{1}{K^2} & \text{if } T \leq K^2 \end{cases} \tag{3.19}$$

where T stands for the sample size and K stands for the number of potential predictors. Note that g appears in the prior variance of the slope vector, which controls our confidence in the prior. The choice of g in (3.19) means we always prefer a more noninformative prior such that the variances for the slopes in (3.18) are bigger than the alternative. It can be shown that the posterior of B_p follows a multivariate t distribution with mean:

$$E(B_p|D, M_p) = \bar{B}_p = \frac{1}{g+1}(\mathbf{X}_p' H \mathbf{X}_p)^{-1}\mathbf{X}_p' H \mathbf{r} \tag{3.20}$$

and covariance matrix:

$$Var(B_p|D, M_p) = \frac{\bar{v}\bar{s}_p^2}{\bar{v}-2}\bar{V}_p \tag{3.21}$$

where $\bar{v} = T$ is the degree of freedom, $\bar{v}\bar{s}_p^2 = \mathbf{r}'H\mathbf{r} - \frac{1}{1+g}\mathbf{r}'H\mathbf{X}_p(\mathbf{X}_p'H\mathbf{X}_p)^{-1}\mathbf{X}_p'H\mathbf{r}$ and D denotes the data. Note that σ_p^2 and a_p enter into all models and their dimensions will not change. It is acceptable to use uninformative priors for them when we want to compare different models using posterior odds ratios, which will be discussed later, see Koop (2003). The marginal likelihood takes the following form:

$$p(D|M_p) \propto (\frac{g}{1+g})^{\frac{k_p}{2}} \left(\bar{v}\bar{s}_p^2\right)^{-\frac{T-1}{2}} \tag{3.22}$$

We can see that the marginal likelihood penalizes the models with a large number of regressors (k_p) since $\frac{g}{1+g}$ is less than 1. For our case there are $P = 2^{15}$ models. Given this model space, there is uncertainty about what is the correct model. Hence it makes sense to consider the parameters unconditional of the model space. This requires us to calculate the posterior model probability as shown in equation (3.23).

$$p(M_p|D) = \frac{p(D|M_p)p(M_p)}{p(D)} = \frac{p(D|M_p)p(M_p)}{\sum_{p=1}^{P} p(D|M_p)p(M_p)} \tag{3.23}$$

To specify the model prior, $p(M_p)$, it is possible to use a uniform prior,

which gives every model the same prior probability($\frac{1}{2^{15}}$ in our case). However, we may want to penalize models with regressors which are highly correlated with each other. George (2001) suggests the following model prior to achieve this,

$$p(M_p) \propto |\mathbf{R}| \prod_{i=1}^{K} \pi^{\gamma_i} (1-\pi)^{\gamma_i} \tag{3.24}$$

where $\mathbf{R}$ is the correlation matrix of regressors included and π is the probability of including a variable. Here we set $\pi = \frac{1}{2}$ to allocate equal prior probability to each model[13]. The determinant of the correlation matrix in the model prior serves to penalize the models with redundant regressors. We can see this by noting that $|\mathbf{R}| = \mathbf{1}$ when the regressors are orthogonal and $|\mathbf{R}|$ approaches 0 when the regressors become more collinear. By putting $|\mathbf{R}|$ into the prior, we can downweigh the models with similar regressors. Table 3.3 displays the parts of the correlation matrix of both dependent and independent variables. Here we just report those elements with absolute values greater than 0.1. As we can see from the table, the use of the model prior in (3.23) is justifiable since certain variables in our regression exhibit high degree of correlation such as oil price, industrial production, monetary supply, treasury bill rate and dividend yield.

Next we turn to how to construct the BMA estimates. Let β denote the intercept and the slopes for the predictors. Though we choose different regressors for different models, we can still think that we are running the usual linear regression model where all possible regressors are included, but for different models different elements of β are set to zeros with probability one. As mentioned in Poirier (1985) we always condition on the full set of available regressors. After model uncertainty is accounted for, Leamer (1978) showed that the mean and variance of the elements in β can be

[13]It is possible to choose other values of π: a smaller value of π will favour more parsimonious moodel, while a bigger value of π will prefer models with more explanatory variables. In terms of inclusion probabilities, the order (from high to low) of the most robust explanatory variables, which are picked up for the subsequent SUR analysis, does not vary much under different values of π's. The results are available upon request.

calculated as[14]:

$$E(\beta_i|D) = \sum_{p=1}^{P} I(\gamma_i = 1|M_p, D)p(M_p|D)E(\beta_i|M_p, D) \tag{3.25}$$

$$Var(\beta_i|D) = E(\beta_i^2|D) - E^2(\beta_i|D) \tag{3.26}$$

where $E(\beta_i^2|D) = \sum_{p=1}^{P} I(\gamma_i = 1|M_p, D)p(M_p|D)E(\beta_i^2|M_p, D)$. An investor may be interested in knowing how important the variables are in explaining the excess return. We therefore need to have a measure of the importance of the included regressor i unconditional of the model space. The following posterior inclusion probability of variable i serves this purpose.

$$p(\gamma_i = 1|D) = \sum_{p=1}^{P} I(\gamma_i = 1|D, M_p)p(M_p|D) \tag{3.27}$$

[14]In fact when we apply the BMA method, the formulae can also be used for other parameters of our interest such as the predicted stock return.

Table 3.3: The correlation matrix of the dependent and explanatory variables

	r	Jan	Tb	$Tbchng$	Smb	Dy	$TERM$	$Indp\%ch$	$M0\%ch$	Inf	$Oil\%ch$	HML	$Oilp$	$Indprd$	$M0$
r	1					0.14									
Jan		1													
Tb			1	0.10	-0.11	0.77	-0.58			0.84	0.11		-0.81	-0.72	-0.79
$Tbchng$			0.10	1			-0.13								
Smb			-0.11		1	-0.10	0.14					0.14			
Dy	0.14		0.77		-0.10	1				0.79	0.12		-0.83	-0.83	-0.79
$TERM$			-0.58	-0.13	0.14		1			-0.40			0.13		0.10
$Indp\%ch$								1		-0.13					
$M0\%ch$		0.69							1		0.13				
Inf			0.84			0.79	-0.40	-0.13		1	0.20	-0.11	-0.66	-0.64	-0.62
$Oil\%ch$			0.11			0.12			0.13	0.20	1				-0.10
HML					0.14					-0.11		1			
$Oilp$			-0.81			-0.83	0.13			-0.66			1	0.87	0.96
$Indprd$			-0.72			-0.83				-0.64			0.87	1	0.88
$M0$			-0.79			-0.79	0.10			-0.62	-0.10		0.96	0.88	1

There are altogether $2^{15} = 32768$ (P) models to be analyzed. The availability of the closed forms of posterior moments and marginal likelihood substantially ease the computation, which enables us to evaluate all the models. It takes about 30 seconds to evaluate all the models and carry out the model averaging exercise in a computer with Pentium IV 3 GHz CPU. It can be estimated that it takes around one hour to estimate models with 22 explanatory variables. For models with more predictors, which the current computing technology cannot handle, we can rely on algorithms such as Markov Chain Monte Carlo Model Composition (MC3) developed by Madigan and York (1995). In our case, the cost to analyse all the models is relatively small. Table 3.4 shows the estimation results for the slope parameters. Table 3.5 lists the 10 models with the highest posterior model probabilities. If the posterior distribution of a slope parameter can be approximated by a normal distribution (e.g. when we have a large sample size), we could characterize the distribution by only its mean and standard deviation and we could know its distance from zero. However, under model uncertainty, the coefficients' posterior distributions are now mixtures of the posterior distributions from all models, which makes the standard deviation hard to interpret. Here we use the posterior inclusion probability defined in (3.27) as a measure of the coefficient's significance. Additionally, we also report the probability of the coefficient being less than zero unconditional on the model space. This probability and the statistic obtained through dividing it by the inclusion probability (we will call this ratio statistic henceforth) are designed to tell what the posterior distribution of a slope parameter looks like. First note that the point zero has a probability mass for each parameter, which is the probability of not including the parameter conditional on the data. Except for this point, the posterior distribution is continuous anywhere in the real line. If the posterior distribution of a slope parameter is different from zero, we should expect the point zero receives little probability mass and the continuous part of the distribution is far away from zero. For example, if a parameter's posterior mean is positive (negative), we should expect $p(slope < 0)$ be close to zero (one). The ratio statistic, defined by $p(slope < 0)$ divided by the inclusion probability, is the probability

of the slope less than zero unconditional of the models with the regressor included. In other words, it does not consider the point zero and only takes into account those models with the parameter. We have put the variables with more than 10% inclusion probabilities in bold. These are the relatively powerful explanatory variables in our results. As we can see such models' ratio statistics are also significant at 10%[15]. The variables with the highest inclusion probabilities in descending order are Dy, Inf, Smb, Jan and $M0$. If we compare the results to those in Table 3.2, we can see that only Dy, Jan and $M0$ are robust for both Bayesian and classical approaches while the variables of oil price, Tb and $Tbchng$ have relatively lower inclusion probability in contrast to their significant results without model averaging. Hence one should be more cautious of the significance of the latter set of explanatory variables.

Table 3.4: Univariate Posterior Estimation of the Slope Parameters

	slope	standard deviation	incl prob	p(slope<0)	$\frac{\text{p(slope<0)}}{\text{incl prob}}$
Jan	**1.62e-3**	**5.455e-3**	**0.1257**	0.01122	**0.089**
Tb	-0.01926	0.43324	5.1e-2	0.02585	0.51
$Tbchng$	0.223	1.6047	0.06596	0.01735	0.263
Smb	**-0.0175**	**0.0522**	**0.14752**	0.137	**0.93**
Dy	**0.311**	**0.489**	**0.391**	0.0039	**0.01**
$TERM$	0.0512	0.5828	0.06325	0.0202	0.319
$Indp\%ch$	2.47e-3	5.17e-2	0.05564	0.02328	0.418
$M0\%ch$	-0.0289	0.1649	0.07454	0.0591	0.793
Inf	**-0.0362**	**0.0998**	**0.1644**	0.15	**0.913**
$Oil\%ch$	-0.0098	0.044	0.090	0.0766	0.852
HML	-0.0059	0.0325	0.076	0.061	0.80
$Oilp$	-6.1e-6	3.01e-5	0.081	0.071	0.881
$Indp$	-3.33e-5	1.56e-4	0.085	0.075	0.884
$M0$	**-6.21e-5**	**2.3e-4**	**0.11**	0.099	**0.925**
$Infch$	-0.0197	0.1466	0.065	0.0476	0.732

Table 3.5 lists the top 10 models with the highest posterior model probabilities. The column headed by "model" list the regressors included for the particular model. The explanation of the variables can be found in Table

[15]That is if the posterior mean of the slope is positive (negative), the ratio statistic is more than 90% (less than 10%).

Table 3.5: Univariate Posterior Model Probabilities

ranking	model	model prob
1	0	0.129
2	Dy	0.098
3	Dy,Inf	0.056
4	M0	0.036
5	Indp	0.028
6	Oilp	0.027
7	Smb	0.025
8	Jan	0.019
9	Smb,Dy	0.014
10	Jan,Dy	0.0137

3.1. We can see that the model with the highest posterior probability is the one without any explanatory variables. Moreover, the top 10 models are all parsimonious models with at most 3 regressors. Their posterior probabilities sum up to 0.49, while for the top 100 models out of 32768, the sum is 84%. All of the top 100 models have no more than 4 regressors, with 82 of them with less than 3. A point to note is that although the model without any explanatory variables has the highest posterior model probability, its posterior probability is not much higher than those of other top models. Their model probabilities sum up to around 87%. Another point to note is that the variable with the highest inclusion probability (around 40%) is the dividend yield. The inclusion probabilities of other variables are at most between 10% and 15%. Remember that our prior for the inclusion probability of each variable is 50%. However, the data do not seem to confirm our prior. This reveals a substantial amount of model uncertainty. It seems that none of the explanatory variables are overwhelmingly strong predictors of the stock return, although the models supporting predictability have higher posterior probability than the model supporting no predictability.

3.4.3 Bayesian Model Averaging in a SUR Model

The previous subsection reveals that the excess return seems to be predictable and some explanatory variables have relatively high posterior inclusion probabilities. Judging from these results, we cannot rule out the possibility that the stock return is predictable. However, as Holmes et al. (2001) point out, if one can incorporate the data generating processes of not only the dependent variable but also the explanatory variable into estimation, the true model may receive higher posterior model probability since different data generating processes can borrow strength from each other. In this subsection, we will implement this idea in a seemingly unrelated regression (SUR) model to investigate more closely the predictability of excess return.

We assume that the explanatory variables have their own data generating processes and that such processes could be correlated with each other and that of the excess return.

$$r_t = a_{0p} + B'_{0p} x_{t-1,p} + F'_{0p} y_{t,p} + \epsilon_{0t,p} \tag{3.28}$$

$$x_{i,t} = a_i + B'_i w_{i,t-1} + \epsilon_{it} \tag{3.29}$$

As before p is the model specific subscript. Here we separate the explanatory variables into dummy variables y and non-dummy variables x, which have their own generating processes described in equation (3.29). The regressors for the predictor equations (w) may include the lag of the excess stock return and those of the explanatory variables. To ease the computational burden in estimation[16], we wish to reduce the number of equations and the parameters to be estimated. We only pick up the five variables with the highest inclusion probabilities calculated in the previous subsection, which consist of one dummy variable, Jan, two financial variables, Smb and Dy, and two business cycle variables, Inf and $M0$. All the equations in (3.29) ($i = 1, 2, 3, 4$) have an intercept term. Table 3.6 describes the regressors

[16]The author agrees that model uncertainty should be considered for all equations at the same time. However, the current computation power does not allow such practice. Moreover we should place our focus on the first equation about the stock return.

we choose for each predictor equation. Holmes et al. (2001) suggests a full search of potential regressors for each equation under SUR framework. To make it simpler, here we just use the univariate BMA method as described before to pick up the variables with the highest inclusion probabilities for each predictor equation. The potential explanatory variables for each equation are the predictor variables of the stock return and the stock return itself [17], i.e. $r, Smb, Dy, Infl, M0$ and Jan.

Table 3.6: Regressors for Equation (3.29)

Equation (for $x's$)	Regressors (w)				
$Smb(x_1)$	r				
$Dy(x_2)$	r		Dy		
$Inf(x_3)$				Inf	
$M0(x_4)$					$M0$

Since our focus is still on the excess return, for different models we assume only the regressors in equation (3.28) will change and the predictor equations will stay the same for different models. Our total number of potential models is $P = 2^5 = 32$. Let us define $\epsilon_t = [\epsilon_{0t}, \epsilon_{1t}, \epsilon_{2t}, \epsilon_{3t}, \epsilon_{4t}]'$ and we assume

$$\epsilon_t \sim N(0, \Sigma) \text{ and } E(\epsilon_j \epsilon_k') = 0 \text{ for } j \neq k \tag{3.30}$$

We will estimate equation (3.28), (3.29) and (3.30) in an SUR framework. Koop (2003) illustrates how to estimate SUR model in a Bayesian framework. Our analysis partly relies on it. First we need to write equation (3.28) and (3.29) into matrix form by defining the following notations.

$$\underset{(m\times 1)}{z_t} = \begin{bmatrix} r_t \\ x_{1t} \\ \dots \\ x_{4t} \end{bmatrix}, \underset{(m\times k_p)}{\tilde{X}_t} = \begin{bmatrix} x'_{t-1,p} & y'_t & 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & w'_{1,t-1} & 1 & 0 & \dots & 0 & 0 \\ \dots & & & \dots & & \dots & & \dots & \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & w'_{4,t-1} & 1 \end{bmatrix}$$

$$\underset{(k_p\times 1)}{C} = \begin{bmatrix} B'_{0p} & F'_{0p} & a_{0p} & B'_1 & a_1 & \dots & B'_4 & a_4 \end{bmatrix}'$$

[17] All explanatory variables enter the predictor equation in the form of one period lag. Detailed results are available from the author upon request.

So equation (3.28) and (3.29) can be rewritten as

$$z_t = \tilde{X}_t C + \epsilon_t \tag{3.31}$$

We adopt the independent Normal Wishart prior for C and Σ, which looks like

$$p(C, \Sigma) = p(C)p(\Sigma) = f_N(C|\underline{C}, \underline{V}) f_{IW}(\Sigma|\underline{\Sigma}, \underline{v}) \tag{3.32}$$

where the prior parameters $\underline{C}$ and $\underline{V}$ denote the mean and variance in the normal distribution and $\underline{\Sigma}$ and $\underline{v}$ denote respectively the matrix and degrees of freedom in the inverted Wishart distribution. Although we have tried to limit the number of our parameters, we still end up with 31 parameters to estimate when we include all regressors into equation (3.28), which means the specification of the hyperparameters could be a huge task. Here we try to be as least subjective as possible. Koop (2003) recommends a general rule of thumb for doing BMA which suggests it is acceptable to use a noninformative prior for parameters which are common to all models and informative proper priors for parameters changing over models. Since for different models we only change the regressors in equation (3.28), only the dimension of B_{0p} and F_{0p} will change across models. For these parameters, we use a proper prior. We will use a noninformative prior for the other parameters. The prior for Σ looks like the following.

$$f_{IW}(\Sigma|\underline{\Sigma}, \underline{v}) \propto |\Sigma|^{-\frac{1}{2}(5+1)} \tag{3.33}$$

Let us denote $\mathbf{c}_p = \begin{bmatrix} B'_{0p} & F'_{0p} \end{bmatrix}'$. For parameters $[a_{0p}\ B'_1\ a_1 ... B'_4\ a_4]'$, we set their covariance elements in $\underline{V}$ and the diagonal elements in $\underline{V}^{-1}$ to zero so that the corresponding values of the hyperparameters in $\underline{C}$ are irrelevant. We will leave the prior for $\mathbf{c}_p$ to later discussion. For the moment we just assume we have a proper prior for it.

The posterior distributions for C and Σ have no analytical forms since the stock return equation and the predictor equations have different regressors. We have to use the following Gibbs sampler algorithm (see Geweke, 2005)

to evaluate them.

$$C|D,\Sigma \sim N(\bar{V}(\underline{V}^{-1}\underline{C} + \sum_{t=2}^{T}\tilde{X}_t\Sigma^{-1}z_t, \bar{V}) \tag{3.34}$$

$$\Sigma|D,C \sim IW(\sum_{t=2}^{T}(z_t - \tilde{X}_t C)(z_t - \tilde{X}_t C)', T) \tag{3.35}$$

where $\bar{V} = \underline{(V}^{-1} + \sum_{t=2}^{T}\tilde{X}_t'\Sigma^{-1}\tilde{X}_t)^{-1}$.

We first choose some arbitrary values for C and draw Σ from equation (3.35) and then plug the draw of Σ into (3.34) to make a new draw of C. Repeating this process will give us a chain of draws. We discard a certain number of the initial draws as burn-in and only retain the remaining draws. The sample average of such draws can give us the estimates of the posterior means for C and Σ.

Our system is comprised of equation (3.28) and (3.29) (described in Table 3.6), where there are altogether 5 equations. The marginal likelihood for a model in our case should be based on all the equations. Though our focus is on the excess return equation, we should not ignore the DGPs of other variables when calculating the marginal likelihood for a model. In this sense our work differs from the previous researchers such as Avramov (2002). To make different models comparable, we keep the specifications for the explanatory variables the same across different models and only change that of the excess return equation. Unlike the univariate case, an SUR model like ours has no closed form for the marginal likelihood. We use Savage-Dickey density ratio discussed in Verdinelli and Wasserman (1995) to calculate the Bayes factors of all restricted models relative to the model with all regressors included in equation (3.28). We denote the model with all regressors included by subscript *all*. We can view different models as fixing different parts of the elements in $\mathbf{c}_{all}$, which we call η, to 0 with probability 1. Again we attach a model specific subscript p to η for all restricted models[18].

[18]Note that the other unrestricted elements in $\mathbf{c}_{all}$ form $\mathbf{c}_p$

Then the Savage-Dickey density ratio (Bayes factor) could be evaluated as

$$BF_{p,all} = \frac{p(D|M_p)}{p(D|M_{all})} = \frac{p(\eta_p = 0|D, M_{all})}{p(\eta_p = 0|M_{all})} \tag{3.36}$$

Though it is straightforward to evaluate the denominator from the marginal prior distribution, there is no direct way to evaluate the numerator since we do not know the analytical form of the posterior distribution for η_p. However, we know the marginal posterior distribution of η_p conditional on Σ and we can have posterior draws of C and Σ from the Gibbs sampler. If we denote the number of draws from the Gibbs sampler by N, we can evaluate the numerator in (3.36) as

$$p(\eta_p = 0|D, M_{all}) = \frac{1}{N}\sum_{i=1}^{N} p(\eta_p = 0|\Sigma_i, D, M_{all}) \tag{3.37}$$

For us to use the above Savage-Dickey density ratio to calculate the Bayes factor, the following condition must hold (see Verdinelli and Wasserman, 1995).

$$p(\mathbf{c}_p|\eta_p = 0, M_{all}) = p(\mathbf{c}_p|M_p) \tag{3.38}$$

To guarantee the above condition to hold, we must choose the prior for $\mathbf{c}_p$ carefully. We first specify the prior for $\mathbf{c}_{all}$ using the g prior like that in equation (3.18) without the term σ_p^2. We choose g as in (3.19). Here we use Ω to denote the variance hyperparameter for $\mathbf{c}_{all}$ and break it into blocks corresponding to $\mathbf{c}_p$ and η_p.

$$\mathbf{c}_{all} = \begin{bmatrix} \mathbf{c}_p \\ \eta_p \end{bmatrix} \Bigg| M_{all} \sim N\left(0, \Omega = \begin{bmatrix} \Omega_{11,p} & \Omega_{12,p} \\ \Omega_{21,p} & \Omega_{22,p} \end{bmatrix}\right) \tag{3.39}$$

where Ω takes the form of a g prior in (3.18). It can be proved that the prior for $\mathbf{c}_p$ should have the following form for condition (3.38) to be satisfied,

$$p(\mathbf{c}_p|M_p) \sim N(0, \Omega_{11,p} - \Omega_{12,p}\Omega_{22,p}^{-1}\Omega_{21,p}) \tag{3.40}$$

which means for models with restriction $\eta_p = 0$, we have more confidence in

$\mathbf{c}_p = 0$ a priori compared to the all inclusive model.

With the Bayes factor we are able to calculate the posterior odds ratio as

$$PO_{p,all} = \frac{p(M_p|D)}{p(M_{all}|D)} = \frac{p(D|M_p)p(M_p)}{p(D|M_{all})p(M_{all})} \tag{3.41}$$

The prior model probability is calculated as before in (3.24) with $\pi = \frac{1}{2}$. Finally we can calculate the posterior model probability for model p using the following,

$$p(M_p|D) = \frac{PO_{p,all}}{\sum_{p=1}^{P} PO_{p,all}}. \tag{3.42}$$

The mean and variance estimates of C and Σ unconditional on the model space can be obtained in a similar way as in equation (3.25). The advantage of using Savage-Dickey density ratio is that to obtain posterior model probabilities for different models, we only need to estimate the model with all the regressors and do not have to calculate the marginal likelihoods of different models one by one, which substantially reduces the computation time.

There are altogether 31 parameters to be estimated in the model described by (3.28) and (3.29). Table 3.7 lists all the models along with their posterior probabilities in descending order. We obtain the results after 1 million draws in the Gibbs sampler. Remember we only change the regressors of the excess return equation in (3.28) to form different models, while the regressors for other equations of (3.29) remain the same in the BMA exercise. Different from the univariate BMA case, the model without any regressor has much higher probability while the posterior model probabilities of most of the other top ten models in the univariate framework fall substantially. The sum of the model probabilities of all the models supporting stock return predictability is now only around 26%. This indicates under the SUR model we find less favorable evidence for stock return's predictability. A point to note is that the posterior model probability of the one with only January effect jumps from 0.02 to about 0.135, which accounts for around half of the posterior model probability of the models supporting

Table 3.7: Posterior model probability under dynamic context

Ranking	Model Probability	Regressors in (3.28)	Ranking	Model Probability	Regressors in (3.28)
1	0.736	0	17	5.1966e-5	Smb Dy Jan
2	0.135	Jan	18	3.8916e-5	Infl M0 Jan
3	0.070567	Dy	19	3.4636e-5	Smb M0 Jan
4	0.028658	M0	20	2.8122e-5	Dy M0 Jan
5	0.01023	Dy Jan	21	2.2804e-5	Smb Infl
6	0.0055782	M0 Jan	22	1.3884e-5	Smb Dy Infl
7	0.0052467	Smb	23	5.672e-6	Dy Infl M0
8	0.0032275	Infl	24	3.6299e-6	Smb Infl Jan
9	0.0025978	Dy Infl	25	1.4882e-6	Smb Infl M0
10	0.00086235	Smb Jan	26	1.4777e-6	Smb Dy Infl Jan
11	0.0005714	Infl Jan	27	1.0426e-6	Smb Dy M0
12	0.00039074	Smb Dy	28	6.725e-7	Dy Infl M0 Jan
13	0.00029996	Dy Infl Jan	29	2.7876e-7	Smb Infl M0 Jan
14	0.0001986	Smb M0	30	1.494e-7	Smb Dy M0 Jan
15	0.00018582	Infl M0	31	3.0519e-8	Smb Dy Infl M0
16	0.00018445	Dy M0	32	3.4185e-9	Smb Dy Infl M0 Jan

stock return predictability. It seems that if we incorporate the DGPs of the explanatory variables from the excess return equation and allow such DGPs to be correlated with each other, we find much weaker support for stock return predictability compared with the univariate case. However, further analysis needs to be carried out to see whether such weak predictability has an impact on the investor's portfolio strategy.

Table 3.8 shows the estimation results of all the parameters for the excess return equation (3.28) after model averaging, where numerical standard error (NSE) is equal to $\frac{\text{standard deviation}}{\sqrt{\text{number of draws}}}$, which is a measure of accuracy for the mean estimates, see Koop (2003). When the true population mean has no closed form, the numerical method we use implies that it should lie in the region of (estimated mean-1.96NSE, estimated mean+1.96NSE) with about 95% probability. Compared with Table 3.4, the slopes of Smb, Dy and Inf have decreased in scale (in absolute value). We can also see a huge drop in their inclusion probabilities, while the inclusion probability of Jan has risen moderately. The estimates for Σ are shown in the lower triangle of Table 3.9 with standard deviations in brackets. The correlation coefficients of different equations are in the upper triangle. Note that the correlation between some equations can be as high as 0.32. The fact that some equations of (3.28) and

Table 3.8: BMA estimation results for the excess return equation under dynamic context

return equation	Smb	Dy	$Infl$	$M0$	Jan	$1987Oct$	$const$
slope	-0.00075	0.043	-0.00079	-0.0022	0.0052	-0.2685	0.0038
std	0.0112	0.163	0.01611	0.01315	0.013	0.04569	0.0077
NSE	3.53e-5	0.0005	5.09e-5	4.16e-5	4.14e-5	0.00014	2.44e-5
incl prob	0.00683	0.0843	0.00697	0.03492	0.1527	1	1

Table 3.9: BMA estimate for error variance matrix under dynamic context

Equation of	r	Smb	Dy	$Infl$	$M0$
r	0.0021775 (0.00018281)	*-0.14287*	*0.013314*	*-0.095545*	*0.077462*
Smb	-0.00020629 ($8.645e-5$)	0.0009575 ($7.9883e-5$)	*0.16883*	*0.090462*	*-0.063564*
Dy	$4.9834e-7$ ($2.252e-6$)	$4.1904e-6$ ($1.4901e-6$)	$6.4337e-7$ ($5.3803e-8$)	*0.32002*	*0.16179*
$Infl$	$-2.4844e-5$ ($1.5398e-5$)	$1.5598e-5$ ($1.0253e-5$)	$1.4303e-6$ ($2.7633e-7$)	$3.1049e-5$ ($2.5901e-6$)	*0.015963*
$M0$	$2.2598e-5$ ($2.2496e-5$)	$-1.2296e-5$ ($1.1387e-5$)	$8.1132e-7$ ($2.9989e-7$)	$5.561e-7$ ($2.048e-6$)	$3.9085e-5$ ($3.2603e-6$)

(3.29) are correlated could imply that the use of SUR model should lead to improved estimation.

3.5 The Horizon Effect of Stock Return Predictability and Model Uncertainty

Equation (3.28) and (3.29) provide us a framework to make forecasts of more than one period ahead based on the information of current period. To simplify the illustration, we need to write equation (3.28) and (3.29) into the form of vector autoregression (VAR) model. Actually the SUR model can be viewed as the restricted form of the VAR model. First let us define

the following[19],

$$\underset{5\times1}{z_t} = \begin{bmatrix} r_t \\ x_{1t} \\ ... \\ x_{4t} \end{bmatrix}, \underset{5\times5}{B} = \begin{bmatrix} 0 & & B'_{0,p} & & \\ B_1 & 0 & 0 & 0 & 0 \\ B_{21} & 0 & B_{22} & 0 & 0 \\ 0 & 0 & 0 & B_3 & 0 \\ 0 & 0 & 0 & 0 & B_4 \end{bmatrix}$$

$$\underset{5\times1}{A} = \begin{bmatrix} a_{0p} & a_1 & a_2 & a_3 & a_4 \end{bmatrix}', F = \begin{bmatrix} F'_{0,p} \\ 0 \end{bmatrix}$$

For the DGPs of the explanatory variables, we use the regressors described in Table 3.6. So equation (3.28) and equation (3.29) can be written as

$$z_t = B \cdot z_{t-1} + A + H \cdot y_{t,p} + \epsilon_t \tag{3.43}$$

Now we can use the following to estimate the mean and variance of z_{t+h} periods ahead conditional on a particular model and the parameters in the excess return and the predictor equations.

$$E(z_{T+h}|C, \Sigma, D, M_p) = B^h \cdot z_T + \sum_{i=0}^{h-1} B^i \cdot A + \sum_{i=0}^{h-1} B^i \cdot H \cdot y_{T+h-i,p} \tag{3.44}$$

$$Var(z_{T+h}|C, \Sigma, D, M_p) = \Sigma + B\Sigma B' + ... + B^{h-1}\Sigma(B^{h-1})' \tag{3.45}$$

Note that we assume ϵ_t has no heteroscedasticity and no serial correlation as in (3.30). This is the base to obtain the conditional forecast variance in (3.45). Also note that y is the dummy variable, i.e. Jan in our case, which captures the periodic phenomenon. In our evaluation of the moments of the cumulative excess return (i.e.$R_{T+\hat{T}} = r_{T+1} + r_{T+2} + ... + r_{T+\hat{T}}$), we set $y = 0$ since we are more interested in the relationship between the stock excess return and the economic fundamentals over time. Note that the cumulative excess return $R_{T+\hat{T}}$ is the first element in the vector $\sum_{h=1}^{\hat{T}} z_{T+h}$, whose mean

[19] Here B_i $(i = 1, 3, 4)$ denote the slope parameters in equation (3.29), which are all scalars, while B_{21} and B_{22} are the slope parameters in the second predictor equation (the equation for dividend yield). See Table 3.6.

and variance can be calculated as

$$\mu_{cum} = B(B^{\hat{T}} - I)(B-I)^{-1} z_T + [B(B^{\hat{T}} - I)(B-I)^{-2} - \hat{T}(B-I)^{-1}]A \quad (3.46)$$

$$Var_{cum} = \sum_{h=1}^{\hat{T}} \delta(h)\Sigma\delta(h)', \ \delta(h) = (B^h - I)(B-I)^{-1} \quad (3.47)$$

The predictive distribution of the cumulative excess return conditional on C, Σ and the model is

$$R_{T+\hat{T}}|C, \Sigma, D, M_p \sim N(\mu_{cum}^{(1)}, Var_{cum}^{(1,1)}) \quad (3.48)$$

where $\mu_{cum}^{(1)}$ stands for the first element in μ_{cum} and $Var_{cum}^{(1,1)}$ is the (1,1) element of the covariance matrix.

To summarize, when we make $\hat{T}$ periods ahead forecast of the stock excess return, we will go through the following steps:

1. We sort the posterior model probabilities of the 32 models calculated from (3.42) descendingly as shown in Table 3.7.

2. Conditional on each model, we draw C and Σ from (3.34) and (3.35). The number of draws for each model corresponds to its posterior model probability[20].

3. Conditional on the draws of C and Σ, we take draws of $R_{T+\hat{T}}$ from (3.48) and calculate the moments.

Note that for the models with explanatory variables other than the dummy in the excess return equation, our results are sensitive to the initial condition of the variables included, i.e. z_T. Predictability in the context of equation (3.44) and (3.45) means that investors use the dynamic model to predict the future based on the current information. The estimated mean and variance of $R_{T+\hat{T}}$ from (3.48) should be viewed as the investor's subjective belief of the cumulative stock excess return and risk accordingly.

[20] If we denote the total number of draws in the algorithm by TN, then the number of draws for model p will be around $TN \times p(M_p|D)$.

Another reason for incorporating the DGPs for the explanatory variables of the excess return equation is that we want to make forecast of excess return $\hat{T}$ periods ahead. Barberis (2000) showed that when the excess return can be predicted only by dividend yield, the optimal stock holding proportion will be very sensitive to the initial value of dividend yield while less sensitive if the investor takes into account parameter uncertainty[21]. However, note that in Table 3.7 the top two models receiving large amount of posterior model probabilities include no explanatory variables and only the dummy variable respectively[22]. These regressors do not appear in our forecast exercise in equation (3.46). In the situations like these, we are virtually saying that the stock return is unpredictable. However, our BMA results are based on the average of all the potential models. Whether the weak predictability will lead to any conspicuous horizon effect requires further analysis.

First we will use the sample mean of all the explanatory variables concerned to form the investor's initial condition. The final results shown in Figure 3.4 are obtained after 8 million draws. The solid line represents the forecast path of the mean and standard deviation of $R_{T+\hat{T}}$ over time. We can see that the mean of the excess return is positive throughout our investment horizon and like the standard deviation, it rises in scale as the investment horizon lengthens. In addition to our forecast, we have also included the evolution paths of the mean and standard deviation of $R_{T+\hat{T}}$ when the excess return is unpredictable with and without parameter uncertainty (in dotted and dashed line respectively), as indicated in (3.5) and (3.11). We can see that $R_{T+\hat{T}}$'s forecast mean does not rise as fast as the one under no predictability in the long run, while its standard deviation is above the one without parameter uncertainty and slightly below the one with parameter uncertainty. Given z_T is the mean of the predictors, the evolution paths of the mean and standard deviation of the excess return from the BMA results are very similar to those under no predictability and

[21] Note that the marginal effect of dividend yield on stock return in most applications is positive. Therefore given that a rational investor's initial value of dividend yield is positive, she should have more position in stock if she has longer investment horizon.

[22] The sum of their model probabilities is around 87%.

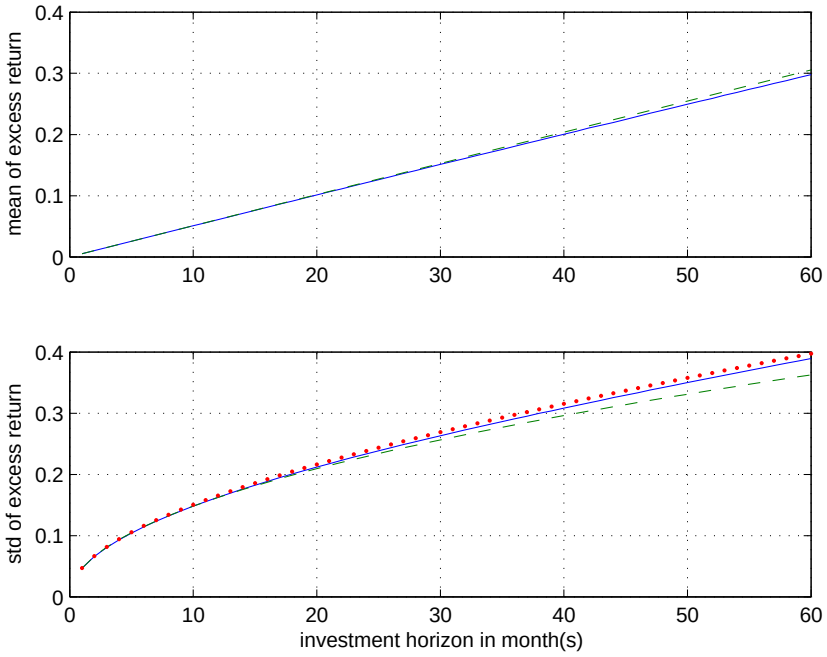


Figure 3.4: The mean and standard deviation of $R_{T+\hat{T}}$ (solid line)

with parameter uncertainty. Therefore we may conjecture that the optimal holding proportion of stock should decrease with time in the long run. This should make intuitive sense since our framework does not only take into account parameter uncertainty but also model uncertainty. When the initial condition for the investor is formed by taking the sample mean of the predictors, it is close to the case with no predictability since in our sample we find not much evidence supporting predictability.

To confirm our guess, we can calculate the $M\hat{t}oS$ statistic defined in (3.13) and the optimal holding proportion of stock. We use the algorithm mentioned in Section 3.2 to search for the optimal ω. Figure 3.5 shows the results in solid lines. Except for the initial tiny rise, the optimal holding proportion of stock falls consistently. As for the $M\hat{t}oS$, although it has some

zigzag movements[23], it clearly demonstrates a downward sloping trend over the long run. Hence we have reason to believe that the $\hat{MtoS}$ statistic captures investor's willingness to hold a risky asset over time to some degree. Under our initial condition, the weak predictability and model uncertainty lead to relatively slow increase of the mean of the excess return compared to the risk, which makes the FTSE rather unattractive in the long run. A rational investor under our utility maximization setting hence should decrease her holding of the FTSE index asset over time. While it is difficult to calculate the optimal holding proportion of the risky asset, it is very convenient to calculate the $\hat{MtoS}$ statistic as long as we can simulate draws from the predictive distribution of the excess return. Although the statistic does not depend on how the utility maximization problem is set up, it may still provide a reference for the investor in regard to how attractive an asset is over time.

Next we will turn to the question of whether the weak predictability of stock return will induce any horizon effect. As mentioned before, predictability should imply that the investor use the present information to predict the future. If there is no horizon effect caused by predictability, the investor should be insensitive to different values of z_T (the initial condition). Here we try two more values in addition to the mean of the predictors: zero and twice the mean of the predictors. The results are also shown in Figure 3.5. The dashed line is obtained from the initial value zero while the dashed-star line is from twice the predictors' mean. We can see that the three paths of ω from three initial values look quite different, though all of them are downward sloping over time. The dashed-star line (from twice the predictors' mean) falls faster than the other two and its $MtoS$ line is below those of the other two. If we set the initial condition to a zero vector, the starting ω is much less than the other two cases. Over time, its optimal holding proportion seems to be parallel to the one obtained by setting z_T to the mean of the predictors. We can see that its $MtoS$ line is initially below the z_T-mean line and the two get intertwined over time. To summarize, it

[23]Such movements could be due to the numerical error during the simulation. As the number of draws increases in the Gibbs sampler, the range of oscillation should be reduced.

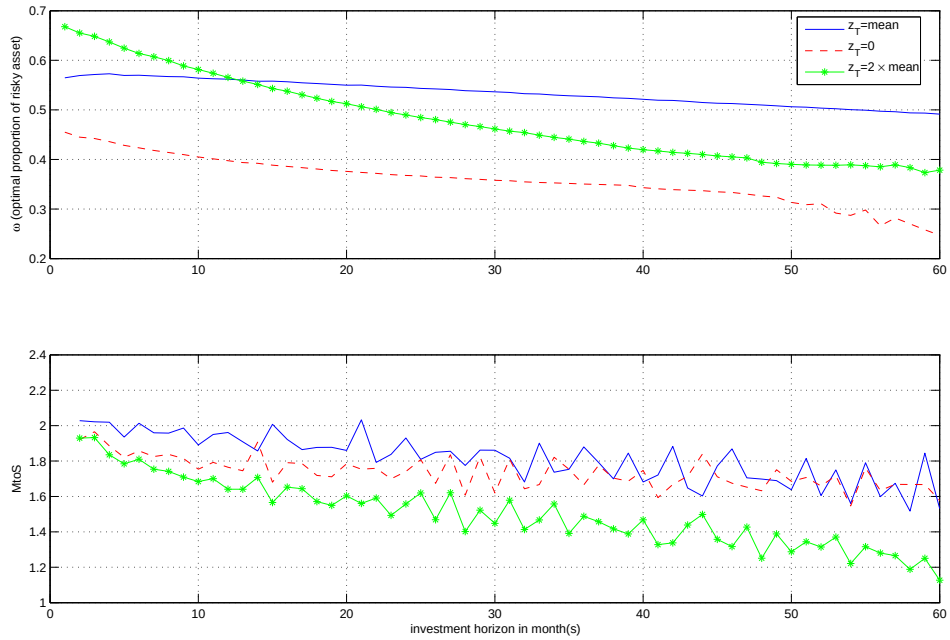


Figure 3.5: The optimal holding proportion of stock and the MtoS statistic

appears that although the stock return predictability is weak, it still has a considerable effect on the investor's optimal portfolio decision over time.

3.6 Conclusion

In this chapter, we study the horizon effect of stock return's predictability, that is, for different lengths of investment horizons how a rational investor should allocate between risky and risk free asset. We show that the investor's portfolio choice for different investment horizons can be linked to the relative time variation of stock's expected return and its expected risk. We propose a computationally convenient statistic to capture such horizon effect and show that it could be related to an investors' optimal holding proportion of a risky asset. We also study the stock return's predictability for the UK market, i.e. what variables may be useful in predicting stock excess return. We argue that Bayesian model averaging is more preferable than simply focusing on a particular model in terms of picking up the variables truly useful in predicting the return. By using BMA, we can avoid the problem of data snooping and take into account parameter and model uncertainty. We have studied the potential useful predictors under both univariate and multivariate frameworks. Our univariate results show that for the UK market, the most powerful predictors are dividend yield, January effect, monetary supply, inflation rate and company size effect. However, if we allow the data generating processes of stock excess return to be correlated with those of its explanatory variables, the predicting power decreases for most variables. Only January effect still remains relatively robust. Though the evidence for stock return predictability is rather weak, it can still lead to considerable horizon effect. It is possible to extend our framework to consider several risky assets. With regard to stock's predictability, we have just considered the predictability in return. It could be fruitful to study the case when the same set of explanatory variables can predict stock's conditional volatility.

Bibliography

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.

Avramov, D. (2002): "Stock-Return Predictability and Model Uncertainty," *Journal of Financial Economics*, 64, 423–458.

Banz, R. (1981): "The Relation between Return and Market Value of Common Stocks," *Journal of Financial Economics*, 9, 3–18.

Barberis, N. (2000): "Investing for the Long Run when Returns Are Predictable," *Journal of Finance*, 55, 225–264.

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

Bawa, V. S., S. J. Brown, and R. W. Klein (1979): *Estimation Risk and Optimal Portfolio Choice*, North-Holland, New York.

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

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.

Brennan, M. J. and Y. Xia (1999): "Assessing Asset Pricing Anomalies," working paper, University of California, Los Angeles.

Brooks, S. P., P. Giudici, and G. Roberts (2003): "Efficient Construction of Reversible Jump Markov Chain Monte Carlo Proposal Distributions," *Journal of the Royal Statistical Society B*, 65, 3–55.

Brooks, S. P. and B. J. Morgan (1995): "Optimization Using Simulated Annealing," *The Satistician*, 44, 241–257.

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.

Cremers, K. J. M. (2002): "Stock Return Predictability: A Bayesian Model Selection Perspective," *The Review of Financial Studies*, 15, 1223–1249.

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.

Fama, E. F. (1970): "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance*, 25, 383–417.

Fama, E. F. and K. R. French (1988): "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics*, 23, 3–25.

——— (1993): "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33, 3–56.

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

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

George, E. I. (2001): "Dilution Priors for Model Uncertainty," Notes for msri workshop on nonlinear estimation and classification, Unicersity of California, Berkeley.

Geweke, J. (2005): *Contemporary Bayesian Econometrics and Statistics*, Wiley-Interscience.

Goetzmann, W. N. and P. Jorion (1993): "Testing the Predictive Power of Dividend Yields," *Journal of Finance*, 48, 663–679.

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

Granger, C. W. J. (1992): "Forecasting Stock Market Prices: Lessons for Forecasters," *International Journal of Forecasting*, 8, 3–13.

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.

Hodrick, R. (1992): "Dividend Yields and Expected Stock Returns," *Review of Financial Studies*, 5, 357–386.

Holmes, C., B. Mallick, and D. Denison (2001): "Bayesian model order determination and basis selection for seemingly unrelated regressions," Technical report, Oxford Centre for Gene Function.

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.

Kandel, S. and R. F. Stambaugh (1996): "On the Predictability of Stock Returns: An Asset-Allocation Perspective," *Journal of Finance*, 51, 385–424.

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

Koop, G. (2003): *Bayesian Econometrics*, Chichester, England: John Wiley & Sons.

Kothari, S. P. and J. Shanken (1997): "Book-to-Market, Dividend Yield, and Expected Market Returns: A Time-Series Analysis," *Journal of Financial Economics*, 44, 168–203.

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.

Leamer, E. E. (1978): *Specification Searches: Ad Hoc Inference with Non-Experimental Data*, New York: Wiley.

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

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

Merton, R. C. (1969): "Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case," *The Review of Economics and Statistics*, 51, 247–57.

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.

Pesaran, M. H. and A. Timmermann (1995): "Predictability of Stock Returns: Robustness and Economic Significance," *Journal of Finance*, 50, 1201–28.

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.

Poirier, D. (1985): "Bayesian Hypothesis Tesing in Linear Models with Continuously Induced Conjugate Priors Across Hypotheses," in *Bayesian Statistics 2*, ed. by J. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, New York: Elsevier, 711–722.

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

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

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

Samuelson, P. A. (1969): "Lifetime Portfolio Selection by Dynamic Stochastic Programming," *The Review of Economics and Statistics*, 51, 239–46.

——— (1989): "The Judgment of Economic Science on Rational Portfolio Management: Indexing, Timing, and Long-Horizon Effects," *Journal of Portfolio Management*, 4–12.

——— (1990): "Asset Allocation could be Dangerous to Your Health," *Journal of Portfolio Management*, 5–8.

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.

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

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

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

Verdinelli, I. and L. Wasserman (1995): "Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio," *Journal of the American Statistical Association*, 90, 614–618.

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.

Xia, Y. (2001): "Variable Selection for Portfolio Choice," *Journal of Finance*, 56, 205–246.

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

Zellner, A. (1986): "On Assessing Prior Distributions and Bayesian Regression Analysis with G-prior Distribution," in *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, ed. by P. K. Goel and A. Zellner, Amsterdam: North-Holland, 233–243.