ESSAYS ON MODEL SELECTION USING BAYESIAN INFERENCE

by

GUO CHEN

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Economics

Written under the direction of

Professor Hiroki Tsurumi

And approved by

____

____

____

____

New Brunswick, New Jersey

October, 2009

ABSTRACT OF THE DISSERTATION

Essays on Model Selection Using Bayesian Inference

By GUO CHEN

Dissertation Director:

Professor Hiroki Tsurumi

This dissertation is composed of three essays evaluating Bayesian model selection criteria in various models, and whenever necessary, the Bayesian criteria are compared with sampling theory criteria. In chapter two, I compare the 2-regime threshold ARMA model (TARMA) and 2-state Markov switching model (MSM). Bayesian Markov Chain Monte Carlo (MCMC) algorithms are devised to obtain coefficient estimates, conditional and unconditional predictive densities. Posterior densities and cumulative densities of the mean square error of forecast (MSEF) of two competing models are generated. The main finding is that for one-day conditional prediction, the 2-regime TARMA model predicts the interest rate better than the MSM. Under the unconditional prediction, however, MSM has less prediction error than TARMA.

In chapter three, I compare the MSEF and Pseudo Bayes Factor (PSBF) obtained by 10-fold CV method and those from an out of sample prediction for fixed points. The MSEF suggests there is a slightly superior performance for the CV method in model selection over traditional out-of-sample forecast in the i.i.d sample. However, the same result is not obtained by PSBF. By excluding forecasted data in constructing coefficients within MCMC, the out-of-sample method is further improved by yielding higher probability to select the true model.

In chapter four, I evaluate logit and probit binary choice models. Monte Carlo experiments are conducted to compare the following five criteria in choosing the univariate probit and logit models: the deviance information criterion (DIC), predictive DIC, Akaike information criterion (AIC), weighted and unweighted sums of squared errors. The results show that if data are balanced no model selection criterion can distinguish the probit and logit models. If data are unbalanced and the sample size is large the DIC and AIC choose the correct models better than the other criteria. If unbalanced binary data are generated by a leptokurtic distribution the logit model is preferred over the probit model. The probit model is preferred if unbalanced data are generated by a platykurtic distribution.

Acknowledgements

This thesis could not be finished without the help and support of many people who are gratefully acknowledged here.

At the very first, I am deeply indebted to my advisor, Professor Hiroki Tsurumi. Without his meticulous guidance, I could not have completed my dissertation. He has offered me valuable ideas, suggestions and comments in every step of the way. His patience and kindness are greatly appreciated. Besides, he is always willing to discuss with me anytime he is available. I have learnt from him a lot not only about economics research, but also the professional ethics.

I'm also extremely grateful to my dissertation committee. Professor Norman Swanson offered invaluable suggestions to my first essay, which enabled me to expand my future research. I enjoyed and benefitted from my discussions with professor John Landon-Lane on improving my third assay. I also thank him for reading over my presentation slides. Professor Ayse Akincigil was kind enough to sit on my dissertation committee. She has worked extensively on qualitative choice models such as probit, logit, and multinomial logit models

I extend my gratitude to the faculty and my fellow students of Rutgers University, for their priceless comments and suggestions on the earlier drafts of my dissertation. I also thank all my friends and the administrative assistants, who made my time at Rutgers enjoyable, especially Dorothy Rinaldi for her enormous assistance.

I would like to give special thanks to my husband Gang for his past, present and future support. Finally, I would like to thank my parents, who always believe in me.

Dedication

To My Parents

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Research Questions, Background and Significance

In empirical studies we need to select a model that performs best among competing models. To select the best model we rely on model selection criteria. The model selection criteria are complex depending on the class of models as well as the characteristics of data sets. In my dissertation I evaluate Bayesian and sample theory selection criteria in various models, specifically (i) a non-linear threshold autoregressive ARMA model versus a Markov switching model, (ii) Non-nested linear models. (iii) Binary choice logit versus probit models.

In recent years, non-linear models have been studied and applied to analyze many macro economic and financial time series data capturing the discrete regimes of the time series. Both TARMA and MSM describe the economics system shifting from one state to another and can yield similar results. Although there are many papers studying either a TARMA model or MSM, only a handful of studies have examined the relative performance of these two models. In particular, the comparison under Bayesian framework is not done yet.

The aim of my first essay is to test which model performs better or which model more accurately captures the characteristics of the data under some common model selection criteria. I use the distribution of the forecasted MSE as the comparison criteria. My results show that both models capture the nonlinearity of the data, however, the prediction ability is different. TARMA is better in conditional prediction than MSM, while MSM outperforms TARMA in unconditional prediction in which data are used twice.

Except for popular model selection criterion MSE, there is another criterion, predictive density combined with Cross Validation (CV). CV measures the predictive density of a given model and chooses the best model which provides the lowest CV error for the given data distribution across several candidate models. The major advantage of CV approach is that it reduces the estimation risks due to the double use of the same dataset for both estimation and validation. Bekara and Fleury (2003) show that CV is a consistent and robust model selection criterion. However, CV has its drawback as it is computationally demanding. That could be one reason that the application of CV in economic models is quite limited yet.

In chapter three, I test the model selection ability for CV method by comparing it with the traditional out-of-sample forecast method. My research questions in this chapter are: (1) whether CV outperform the traditional out-of-sample method; (2) test under which condition, these two methods have different ability in selecting correct model among certain model space.

I develop an algorithm for 10-fold CV for non-nested regression model. Using MSE of forecasted (MSEF) and Pseudo Bayes Factor (PSBF) as the comparison criteria, I found that CV performs slightly better based on MSEF. However, the PSBF yield opposite results. In addition, the out-of-sample method is improved if the predicted value is not used in constructing coefficients in the MCMC loop. My results provide implication for the use of CV in economic research.

In my forth chapter, I worked on the dichotomous model. Logit and probit models are two popular qualitative choice models, where the endogenous random variables take only discrete values. There are numerous applications of these two models in economics, biometrics, marketing research, credit analysis, etc. In prior literature, the univariate logit and probit are considered identical in application as they have similar statistical distributions. They cannot be discriminated unless the sample size is large. My results show that these two models can be discriminated if we have a large sample size and unbalanced data. In addition, I demonstrate that we can choose from a model, either logit or probit to better fit the data if the kurtosis of the distribution is available.

My essay contributes to this literature in the following way: First, I introduced a more general form of distribution: the exponential power distribution, into my study. It incorporates both normal and logistic distribution and makes the results applicable to a broader range of data sets. The next contribution is that I use a new model selection criterion Predictive Deviance Information Criteria (PDIC) based on the predictive density in comparing the univariate probit and logit model. In addition, I incorporate model selection criteria in both Bayesian inference and sampling theory to identify the conditions under which these two models can/cannot be distinguished. A large portion of the past studies are focused on the similarity of statistical distribution, hypothesis testing, and the comparison of the coefficient estimates between these two distributions. The Last contribution is that I devised a Metropolis-Hasting (M-H) algorithm with random walk for logit and probit model, which has less computation time compared with the available Bayesian estimation method, such as Gibb sampler with data augmentation.

1.2 General Methodology

In all my three essays, a MCMC algorithm is applied in estimating coefficients, obtaining the predictive density, and calculating the model selection criteria. Different models induce different priors and likelihood function, thus leads to different posterior distribution. The general outline of the MCMC procedure is as follows;

1) Generate parameter matrix A from the proposal density.

2) Apply accept-reject algorithms to generate the sequence (chain) of parameters. The probability of acceptance is given by:

$$\alpha\left(A,\hat{A}\right)=\min\left\{\frac{P\left(A|data\right)q(\hat{A})}{P\left(\hat{A}|data\right)q(A)}\right\} \tag{1}$$

where $P\left(\cdot\right)$ is the posterior density and $q\left(\cdot\right)$ is proposal density. A is the updated (current) draw of parameter matrix and $\hat{A}$ is previous draw of parameter matrix.

3) Repeat above two steps until each sequence converges.

There are different ways to choose the proposal density for the convergence purpose, thus yield different accept-reject algorithm. I use are M-H with random walk, hybrid M-H and M-H with efficient jump in my three essays.

2 Comparison of Threshold ARMA Model and Markov Switching Model Using MCMC

2.1 Introduction

In recent years, non-linear models have been studied and applied to analyze many macro economic and financial time series data. These models capture the interesting properties of the data, such as frequency dependence, cycling behavior and jump phenomena. Among the various discrete time non-linear models, two popular classes are Threshold Autoregressive or Threshold Autoregressive ARMA model (TAR or TARMA) and Markov Switching Models (MSM).

The popularity of TAR class models is due to the fact that they are simple to specify, estimate and interpret compared to other non-linear time-series models. The MSM has its advantage in capturing the ups and downs of the state. The intuitive appeal of both models lies in the fact that the behavior of economic time series often exhibit distinct phases. For instance, the national economy shifts between a recession and expansion, government regimes change over time, and financial markets experience bubbles and crashes. Since both models describe the economics system shifting from one state to another, they can be applied to the same time series, and can yield similar results. For example, Garcia and Perron (1996) used MSM to investigate the threshold behavior of interest rate, and they find that the interest rate follow a unit root in low regime and mean-reversion process in upper regime. The same result is obtained by Goldman and Agebeyegbe (2005), Gospodinov (2005) using TARMA-GARCH model.

However, besides these similarities, both models also have their own distinct features. First, MSM captures the probabilistic state transitions over time, and the optimal probabilistic inference based on the observed behavior of the dataset; whereas the TAR models treat the shifts in regimes as directly observable. Secondly, MSM is used in a cyclical process, such as bull and bear market states in the stock market, the recession and expansion characterized by GDP growth change, but TAR models estimate the threshold and the dynamics of data in different regimes, not necessarily cyclical process. Henneke, Rachev and Fabozzi (2001) also point out that in MSM, the process can leave a state and returns with a positive probability; while TAR models only use the data between changes in the regimes and disregard the rest of the data set. Thus, MSM tends to yield better estimate for the "normal state" as it bases on much larger data set, therefore it may provide better forecasting than TAR models.

Observing the similarities and difference of the both models, one interesting question would be to compare these two models and test which one perform better under some common model selection criteria. Thus, the purpose of this essay is to compare the two-regime TARMA model with the two-state MSM. In particular, we try to answer following questions, when applying to the same dataset with distinct phase: which model will perform better or which model will more accurately capture the characteristics of the data. I use the short term interest rate data. One reason is that the short-term interest rate is an essential component of monetary policy formulation and asset valuation. It plays a key role in the yield curve, inflation expectation and aggregate demand in macro economy. In addition, it has important implications for the dynamics of long rate, the pricing of other underline derivatives. Therefore, a good understanding the characteristics of short-term interest rate data is crucial. Another important reason is that there are already many papers which use either TAR or MSM to estimate the short term interest rate and the results obtained by using these two models are similar. However, there has been quite limited studies in comparing these models. Thus, the short term interest rate would be the ideal observation. Furthermore, we can evaluate our results in the traditional literature of nonlinear modeling for short term interest rate.

Immediately, we encounter two problems. First, how to measure the performance of two different models and make it comparable. We need to set common model selection criteria. It is a complex matter as there are many criteria. A natural way to assess the goodness of the model is to estimate its predictive ability. Thus, my model selection is based on the predictive density. I do the comparison within the Bayesian modeling framework. In addition, a single summary number also is a choice for model selection, as one number is straightforward to compare and cause no ambiguity. Following the standard literature on Bayesian model choices, we use the MSEF as a model choice criteria, which make the comparison on the aggregated level, instead of case level diagnostics used in Geisser (1988), Pettit and Young (1990) and Gelfand et al. (1992). The second issue is the model specification. There are many sub-categories in these two broad classes of models. To make the comparison applicable and reliable, we will conduct our comparison between two-regime TARMA model and two-state MSM with fixed parameter dimension. The error structure of the model is the ARMA model.

This essay is organized as follows. In Sections 2.2 we review the literature of TAR/TARMA models and MSM. Description of the two-regime TARMA model and two-state MSM are provided in section 2.3. Also, a brief summary of the model selection criteria are outlined in this section. In section 2.4, Bayesian estimation procedures are specified, including the prior setting, maximum likelihood and the posterior distribution for each model. Metropolis-Hasting (M-H) algorithm is also presented briefly in this section. Section 2.5 discusses the data set and the descriptive statistics. Analyses of empirical results are given in Section 2.6. Concluding remarks and extension are made in Section 2.7.

2.2 Literature Review

The TAR model is first introduced by Tong and Lim (1980). Tong (1990) gives more detailed review of this model. The basic idea is to divide the data into a small number of regimes according to the different values of threshold variables. Thus, the threshold model describe a process of piecewise linear in the threshold space. After Tong's study, a large number of papers have been devoted to this model both in discussion of statistical inference and testing procedure. Tsay (1989, 1998) constructs a test using predictive residuals to detect threshold nonlinearity in a vector time series. Petruccelli and Davis (1986) propose a CUSUM type test for TAR nonlinearity. Chan (1986) develops a conditional likelihood ratio test statistic. The TAR models are also applied to real macro and financial data. Hansen (1997) reports a significant threshold effects in U.S. unemployment rate. Koop and Potter (1999) use a two-regime threshold model to show the dynamic asymmetry between unemployment rate rising and falling. Gospodinov (2005) estimates conditional mean and variance for short–term rate with threshold nonlinearity, and also finds its better forecast performance than single regime model. Other applications of TAR model and its extension, see Forbes, Kalb, and Kofman (1999), Lanne and Saikkonen (2002).

The Bayesian analyses of TAR models have been made as well. These analyses effectively estimates multiple threshold simultaneously without the "curse of dimension". Geweke and Terui (1993) use Monte Carlo integrations for two-regime TAR model. Phann, Schotman and Tscherig (1996) use Griddy Gibbs sampler within MCMC. Chen and Lee (1995) apply Metropolis algorithm within a Gibbs sampler. Goldman and Agebeyegbe (2005) used M-H algorithm with efficient jump for ARMA and ARMA-GARCH model to estimate the U.S. short term interest rates. The same approach is used for multivariate threshold model in Tsay (1998). The Vector Error Correction Threshold models for future index is studied by Huang (2004), and the procedure provides the reliable estimates of parameters and a test criteria for detecting threshold nonlinearity in a vector model.

The MSM is introduced and developed by Hamilton (1989, 1990) for modeling U.S. business cycle using U.S. post war GNP data. This model also considers the change of a state of time series. Generally speaking, a process is governed by different states, where switch between them are based on a probabilistic process. In addition, this probability is unique. Essentially, the process modeled here subject to discrete shifts in state; and the parameters of the process change over time driven by a Markov state variable, which is assumed to be unobserved or latent. Similar to TAR model, MSM has been applied for many markets, the successful ones including the application in short term interest rates, foreign exchange rate, stock return, and stock return volatilities.

The MSM is seen in modeling heteroskedasticity and duration dependence of stock returns. ARCH or GARCH type models generate high persistence of stock volatility to match the fat tails and volatility clustering, but they usually provide poor forecast. Perron (1989) has pointed out that if data follows a process with structure break, then it tends to reveal unit root although it is actually stationary. Since GARCH model incorporating a unit root and structural break usually accounts for part of the high persistence, if the process is stationary with structural breaks, GARCH model with a single regime yields poor forecast. Based on this idea, Hamilton and Susmel (1994) suggest an ARCH model allowing parameter changing. This SWARCH specification offers better fit to the data and better forecast. Cai (1994), Derker (1997), Haas, Mittnik, and Paolella (2004) also build their empirical work upon these combination of ARCH/GARCH and regime switching. The other application of MSM is found in exchange rate and inflation rate. Engel and Hamilton (1990) propose Markov switching model in Foreign exchange rates and claim that exchange rate follow a model of long swing instead of random walk. Rapach and Strauss (2005) construct GARCH models for exchange rate volatility. Evans and Wachtel (1993); Ricketts and

Rose (1995) use MSM to estimate inflation rate. One extension to Hamilton's MSM is to specify the transition probabilities to be a function of underling variable instead of a constant. Applied to U.S. post war real GNP, Durland and MacCurdy (1994) obtain the evidence of duration dependence for recession, but not for expansion. They also find that economy is more likely to transit out of its current state the longer it has been in it, which is not implied by Hamilton's model. Lunde and Timmermann (2000) find asymmetry in stock prices.

There are many studies of the MSM using Bayesian inference. This due to the computational tractability and reliability of Bayesian approach. In estimating MSM with many unknown parameters, such as ARMA-GARCH parameters, ML estimation becomes computationally unfeasible if the data are few. The inference based on the unobserved state variable is also problematic. To avoid the high dimension problem, Hamilton and Susmel (1994) use low order of ARCH with Markov switching. More favorable method adopted by many authors is Markov Chain Monte Carlo methods (MCMC). Among them are Carlin, Polson and Stoffer (1992), Francq and Zakoian (2001), Henneke, Rachev and Fabozzi(2006), and Hark Yoo (2006).

Interest rate is another most applied area for Markov switching model. As the early example, Hamilton (1988) uses MSM for real interest rates. Gray (1996) develop a generalized regime-switching (GRS) model for the short-term interest based on the first-order Markov process with state-dependent transition probabilities. GRS has a better forecasting performance than single regime model for out-of-sample forecasting. Garcia and Perron (1996) detect the random behavior in both mean and variance of ex-post real interest rates with three regimes. Ang and Bekaert (2002a) show that the regime switching outperforms single regime in out-of-sample forecasting, and corresponds well with the business cycles. Other short term interest rate model based on MSM include Evans and Lewis (1994), Smith (2000), Bekaert, Hodrick and Marshall (2001). Except MSM, there are other nonlinear models proposed to address the high persistence and conditional heteroskedasticity of interest rate data, among them are Tsay (1989, 1998), Lanne and Saikkonen (2003), Gospodinov (2005) Ait-Sahalia(1996) and Stanton (1997). In addition, Waston (1999) argue that the existence of regime switching in the conditional mean implies the relationship between the short-rate persistence and the long-rate variability.

2.3 Model and Model Selection Criteria

2.3.1 Two-Regime Threshold ARMA Model

In general, a time series y_t follows $k+1$ regime stochastic threshold model with ARMA error component and the threshold variable z_{t-d} if it satisfies the following model,

$$y_t = x_t\gamma^{(j)} + u_t \tag{2}$$

$$u_t = \frac{\Theta^{(j)}(B)}{\Phi^{(j)}(B)}\epsilon_t \tag{3}$$

$$\Theta^{(j)}(B) = 1 + \theta_1^{(j)}B + \cdots + \theta_{q^{(j)}}^{(j)}B^{q(j)}$$

$$\Phi^{(j)}(B) = 1 - \phi_1^{(j)}B + \cdots + \phi_{p^{(j)}}^{(j)}B^{p(j)}$$

$$\epsilon_t \;\sim\; N\left(0,\; \sigma_t^2\right) \tag{4}$$

where y_t belongs to regime j if $r_{r-1}\dot{\leq}z_{t-d} < r_j$, $j = 1, \ldots, k$ are non-negative integers, and r is the threshold on real line. Thus, if we have k thresholds, then there are $k+1$ regimes, and $\{r_1, \ldots, r_k\}$ forms a partition of the real line. d is the delay variable, and we assumed $d \in \{1, \ldots, d_0\}$. The $p^{(j)}$ and $q^{(j)}$ are the orders of AR and MA processes, and they can be different for different regimes. In our model, we assume they are the same across regimes. B is the backward shift operator. The parameters $\left\{\gamma^{(j)},\ \phi^{(j)},\ \theta^{(j)}\right\}$ have different values for different regimes, i.e., they take $k+1$ values depending on the regime j which z_{t-d} belongs to. Following Hansen (1997) and Gospodinov (2005), I choose the stationary lag difference $\Delta z_{t-d} = |y_{t-d} - y_{t-d-1}|$ as the threshold variable. There are other candidates for the threshold variable, such as moving average $\sum_{i=1}^{d} \frac{|y_{t-i}|}{d+1}$ or stationary lag value of dependent variable y_{t-d} use by Tsay (1998).

As discussed earlier, the ARMA order p and q for different regimes are the same. I choose p and q to be $(1, 1)$. This specification is accepted for modeling short term interest rate, see Gospodinov (2005) and among others. Other diagnostic check, such as autocorrelation function (ACF) and the partial autocorrelation functions (PACF) also indicate that the right ARMA order is $(1, 1)$ for our data. Table 1 summarizes the result of ACF and PACF of the data with the first ten lags.

2.3.2 Two State Markov-Switching Model with ARMA Errors

In this essay, I consider the MSM in mean with ARMA (p, q) error, and it is defined as,

$$y_t = \gamma_0 + \gamma_1 S_t + u_t \tag{5}$$

$$u_t = \sum_{i=1}^{p} \phi_i u_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \epsilon_t \tag{6}$$

$$\epsilon_t \text{ ~ } N\left(0, \sigma_t^2\right) \tag{7}$$

so, y_t has two mean values depending on the state variable S_t. If $S_t = 0$, then mean is γ_0; if $S_t = 1$, then mean is $\gamma_0 + \gamma_1$, $\gamma_1 > 0$. The state variable S_t evolves according to a two-state first order Markov switching process with following transition probabilities:

$$\Pr\left[S_t = 0 | S_{t-1} = 0\right] = p_{00},\ \Pr\left[S_t = 0 | S_{t-1} = 1\right] = q_{00}$$

The functions (6) and (7) specifying ARMA error are similar to those in TARMA model, except the former one is not grouped by regimes. We choose two-state MSM corresponding to the two-regime TARMA model, which make the comparison feasible.

2.3.3 Model Selection Criteria

Predictive density has been proposed to make model selection in application. I use MCMC algorithm to draw predict value $\tilde{y}^i_{t+j}$ based on the posterior means of the estimated parameters. Then, I compare the shape of the predictive density curve of the two models. The kernel density of the predicted $\tilde{y}^i_{t+j}$ is drawn accordingly. Ideally, the model better capturing the characteristics of the data has the distribution close to normal and relative smaller probability at tails. The better model also has less variance, which indicates that the curve is tight around its mean.

The MSEF is used as another criteria for model selection. It is given by:

$$MSEF^{(i)} = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{y}^{(i)}_{t+j} - y_{t+j}\right)^2 \tag{8}$$

where $\hat{y}^i_{t+j}$ is the i-th draw of the predictive value of y_{t+j}, $j = 1, \ldots, n$. $\hat{y}^i_{t+j}$ is the estimated value based on the posterior mean of the parameter, and y_{t+j} is the realized value. The model with the smaller MSEF is chosen to be the best model. This approach has been applied in Anderson, Bollerslev, Diebold and Labys (2003), Goldman, Nam and Wang (2005), and among others. The posterior density of MSEF and the cumulative density of MSEF are also drawn for further comparison. The posterior density with tighter shape and smaller tail is the better one. For cumulative density, the one more quickly attaining the highest level and having higher level at each point is considered to be the one leading model.

2.4 MCMC

2.4.1 Two-regimes TARMA Model

Prior, Likelihood Function and Posterior Let $\pi\left(\gamma, \phi, \theta, r\right)$ to be the proper prior given by,

$$\pi\left(\gamma, \phi, \theta, r\right) = \prod_{i=1}^{k} N\left(\gamma_{0i,} \sum_{\gamma_i}\right) \times N\left(\phi_{0i,} \sum_{\phi_i}\right) \times N\left(\theta_{0i,} \sum_{\theta_i}\right) \times I\left(r \in \left[r_{low}^{(i)}, r_{up}^{(i)}\right]\right) \tag{9}$$

where the threshold parameter r has uniform prior. In order to guarantee sufficient sample size for each regime, we impose a restriction of $\delta = 20\%$ of total sample size as a minimum observation number in each regime. Thus, the threshold r_i is restrained in the interval $\left[r\_low^{(i)},\ r\_up^{(i)}\right]$. The hyperparameters $\left\{\gamma_{0i,} \sum_{\gamma_i}; \phi_{0i,} \sum_{\phi_i}; \theta_{0i,} \sum_{\theta_i}\right\}$ are assumed to be known. The likelihood function of the TARMA model is given by,

$$\ell\left(y|x, \gamma, \phi, \theta, \sigma^2, r\right) = \prod_{j-1}^{k+1} \prod_{t \in T_J} \frac{1}{\sigma_t \sqrt{2\pi}} \phi\left(\frac{y_t^{(j)} - g(Z_t)}{\sqrt{2}\sigma_t}\right) \tag{10}$$

where for each $t \in T_j = \{t :\ r_{j-1} \leq z_{t-d} \leq r_j\}$,

$$e_t = y_t^{(j)} - x_t^{(j)} \gamma^{(j)}$$

$$\epsilon_t = y_t^{(j)} - g(Z_t)$$

$$g(Z_t) = x_t^{(j)} \gamma^{(J)} - \sum_{i=1}^{k} \phi_i^{(j)} e_{t-j} - \sum_{i=1}^{q} \theta_i^{(j)} \epsilon_{t-i} \tag{11}$$

Thus, the posterior distribution is given by,

$$P\left(\gamma, \phi, \theta, \sigma^2, r|data\right) \propto \pi\left(\gamma, \phi, \theta, r\right) \prod_{j-1}^{k+1} \prod_{t \in T_J} \frac{1}{\sigma_t \sqrt{2\pi}} \phi\left(\frac{y_t^{(j)} - g(Z_t)}{\sigma_t}\right) \tag{12}$$

MCMC Procedure Given the number of regimes and the order of p and q, the parameters to be estimated are $(\gamma, \phi, \theta, \sigma^2, r)$. The M-H algorithm for ARMA component has been explained by Nakatsuma (2000). Instead of using constrained nonlinear maximization algorithm (CML) in MA block for independent chain, I used random-walk Markov chain. The random walk is more efficient than the CML in reducing the computational time, whereas "not losing much of the acceptance rate of Metropolis-Hasting algorithm", as done in Goldman and Tsurumi (2005).

The conditional distribution of the threshold is non-standard. The choice of the spread or scale, and the candidate-generating density is critical and has implications for the efficiency of the algorithm. Here, I applied efficient jump method developed by Goldman and Agbeyegbe (2005). The efficient jump algorithm goes as follows: Threshold r is generated by normal distribution, $N\left(r_j^{(i-1)},\ stdr_j^{(i-1)}\right)$. The standard deviation is initially selected as a constant C_0, which is equal to the half-distance between the upper and lower bound for each regime. After some number of draws, the standard deviation of the sample of the accepted draws is multiplied by a scaling constant C.

$$stdr_j^{(i-1)} = C \times stdr\left(\{r_m\}\right),\ m = n_0, \ldots, i-1 \tag{13}$$

$\left\{r_j^{(m)},\ m = 1, \ldots, i-1\right\}$ is the sample of accepted draws in the regime j. We can control the accept rate by changing the scaling constant C.

I estimate the parameters in block: 1) regression parameters γ, 2) AR coefficients ϕ, 3) MA coefficients θ, 4) σ^2. 5) threshold parameter r. The proposal distributions for these parameters are based on the original ARMA model:

$$y_t = x_t\gamma^{(j)} + \sum_{i=1}^{p} \phi_i^{(j)} \left(y_{t-i} - x_{t-i}\gamma^{(j)}\right) + \epsilon_t + \sum_{i=1}^{q} \theta_i^{(j)}\epsilon_{t-i},\ \epsilon_t\ \tilde{}\ N\left(0, \sigma_t^2\right) \tag{14}$$

Using above equation, I generate $(\gamma, \phi, \theta, \sigma^2, r)$ from their proposal density, the details are provided in Appendix II.A. The initial value for γ is obtained from the OLS estimate. The initial value of (ϕ, θ) is set arbitrarily. For threshold r, I use the mean value of $(r\_up + r\_low)$ for the two-regime model.

Then, I sort data according to the increasing order of the threshold variable Δz_{t-d}. Thus, all the observations in each regime follow the same ARMA model. The OLS estimates of gamma is used for each regime as the starting value of $\gamma_1^{(0)}, \gamma_2^{(0)}$. When the data are sorted into regimes based on threshold r, I transform the original model into the arranged ARMA model. The outline of the MCMC procedure is as follows;

1. Generate $(\gamma, \phi, \theta, \sigma^2, r)$ from the proposal density based on the equation (14) block by block.

2. Apply M-H algorithm after each parameter is generated, i.e., generate a value of $\left(\hat{\gamma}, \hat{\phi}, \hat{\theta}, \hat{\sigma}^2, \hat{r}\right)$ from proposal and accept the proposal value with probability:

$$\lambda(A, \hat{A}) = \min\left\{\frac{P(\hat{A}|data)/h(\hat{A})}{P(A|data)/h(A)},\ 1\right\} \tag{15}$$

where $A = (\gamma, \phi, \theta, \sigma^2, r)$, $h(A)$ is proposal density defined in equation (14), see Appendix II.B of Metropolis-Hasting for each block in details.

3. Repeat above two steps until each sequence converges.

Here, several remarks need to be made. First, parameters in each block are drawn for all regimes separately, but accepted or rejected jointly in one block. The acceptance rates are controlled by multiplying the variance of the proposal density with a scaling constant. Second, regression coefficient and ARMA coefficients are drawn from independent multivariate normal distribution. The threshold r follows efficient jump algorithm as I discussed above. Third, I make N draws of the parameters in each of the five blocks, and burn the first M draws. Out of the remaining $N-M$ draws, I keep every h-th draw. N, M and h are chosen optimally according to the convergence test.

2.4.2 Two-state Markov Switching Model with ARMA Error

Prior, Likelihood Function and Posterior Following the Bayesian rule and conditional distribution, we derive the posterior distribution of the parameter,

$$P\left(\Theta, S|Y\right) \propto P(\Theta, S)L(Y|\Theta, S) \propto \pi(\Theta)P(S|\Theta)P\left(Y|\Theta, S\right) \tag{16}$$

where $\Theta = (\gamma, \phi, \theta, p_{00}, p_{11})$, $S = (S_{1,} \ldots, S_T)$, and $Y = (y_{t,} \ldots, y_T)$.

The prior distribution of parameters are defined as,

$$\pi\left(\gamma, \phi, \theta, p_{00}, p_{11}\right) = N\left(\mu_\gamma, \sum_\gamma\right) \times N\left(\mu_\phi, \sum_\phi\right) \times N\left(\mu_\theta, \sum_\theta\right)$$

$$\times Beta(u_{00,}u_{01}) \times Beta(u_{11,}u_{10}) \tag{17}$$

Where $Beta(.)$ is the beta function.

The conditional distribution of the state variable is $P(S|\Theta)$, which only depends on $(p_{00,}p_{11})$, is defined as,

$$P(S|\Theta) = P\left(S|(p_{00,}p_{11})\right) = \prod_{i=1}^{T} P(S_{t+1}|S_t, p_{00,}p_{11}) = p_{00}^{\eta_{00}}(1-p_{00})^{\eta_{01}}p_{11}^{\eta_{10}}(1-p_{11})^{\eta_{10}} \tag{18}$$

where η_{ij} is the number of the transitions from state i to state j. The last term $P(Y|\Theta, S)$ is likelihood function assuming that the states and the parameters are known. Therefore, it is a full information likelihood function.

$$P(Y|\Theta, S) = \prod_{i=1}^{T-1} L(y_i|Y_{i-1}, S_t, \dots, S_1, \Theta) = \prod_{i=1}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left[-\frac{\epsilon_t^2}{2\sigma_t^2}\right] \qquad (19)$$

MCMC Procedure Following Hark Yoo (2006), I conduct the MCMC procedure as follows,

Step 1: Generate each S_t from $P(S_t|S_{\neq t}, Y, \Theta)$ for $t = 1, \dots, T$ by single move approach.

Step 2: Generate the transition probabilities, p and q from Beta distribution.

Step 3: Generate regression coefficients and ARMA parameters from $P(\gamma, \phi, \theta|S, p, q, Y)$.

Use single move developed by Albert and Chib(1993), I draw state one by one from each of the following conditional Distributions,

$$P(S_t|S_{\neq t}, Y, \Theta) \propto P(S_{t+1}|S_t, \Theta_1)P(S_t|S_{t-1}, \Theta_1)P(y_t|Y_{t-1}, S_t, \Theta) \qquad (20)$$

The detailed derivation of above equation is provided in Appendix II.C. The first and second term in the RHS of above equation can be obtained by transition probabilities and the last term is given by likelihood function. After calculating $P(S_1|S_{\neq 1}, Y, \Theta)$, we can generate S_t using a uniform distribution. For example, we generate a random number from uniform distribution between 0 and 1, if the generated number is less than or equal to the calculated value of $P(S_1|\ S_{\neq 1}, Y, \Theta)$, we set $S_t = 1$, otherwise 0.

Conditioning on S_T, the transition probabilities p and q are independent of data Y and other parameters in the model. I use beta distribution as conjugate priors, and the likelihood function for p and q is given by,

$$L(p, q|S_T) = p^{\eta_{00}}(1-p)^{\eta_{01}}q^{\eta 11}(1-q)^{\eta 10} \qquad (21)$$

Thus, the conditional distribution of (p, q) is given by ,

$$\begin{aligned} P(p,q|S_T) &= \pi(p,q)L(p,q|S_T) \\ &\propto p_{00}^{u_{00}-1}(1-p_{00})^{u_{01}-1}q_{11}^{u_{11}-1}(1-q_{00})^{u_{10}-1}p_{00}^{\eta_{00}}(1-p_{00})^{\eta_{01}}q_{11}^{\eta_{11}}(1-q_{11})^{\eta_{10}} \\ &\propto p_{00}^{u_{00}+\eta_{00}-1}(1-P_{00})^{u_{01}+\eta_{01}-1}q_{11}^{u_{11}+\eta_{11}-1}(1-q_{11})^{u_{10}+\eta_{10}-1} \qquad (22) \end{aligned}$$

Then, p_{00} and q_{11} are generated by Gibbs sampler from the following independent beta distribution:

$$p_{00}|S_T \sim Beta(u_{00}+\eta_{00}, u_{01}+\eta_{01}) \tag{23}$$

$$q_{11}|S_T \sim Beta(u_{11}+\eta_{11}, u_{10}+\eta_{10}) \tag{24}$$

The regression and ARMA type parameters are generated by M-H algorithm in a way similar to those describe in TARMA model. To save space, I do not list all the details here. The only difference is that in generating MA coefficient θ_t, we use the approach proposed by Chib and Greenberg, i.e. linearizing ϵ_t by first-order Taylor expansion:

$$\epsilon_t(\theta) \approx \epsilon_t(\theta^*) + \psi_t(\theta - \theta^*) \tag{25}$$

where $\epsilon_t(\theta^*) = y_t^*(\theta^*) - x_t^*(\theta^*)\mu$, $\psi_t = \left[\psi_{1t}, \ldots \psi_{qt}\right]$ is the first-order derivative of $\epsilon_t(\theta)$ evaluated at θ^* given by the following equation,

$$\psi_{1t} = \epsilon_{t-i}(\theta^*) - \sum_{j=1}^{q} \theta_j^* \psi_{it-j}, \psi_{it} \text{ for } t \leq 0 \tag{26}$$

and the non-linear least square estimate of θ_t is given by

$$\theta^* = \arg\min_{\theta} \sum_{t=1}^{n} \left\{\epsilon_t(\theta)\right\}^2 / \sigma_t^2 \tag{27}$$

The initial value of γ is given by the OLS estimates, and the starting values of other parameters are arbitrary.

2.4.3 Drawing Predictive Value $\tilde{y}_{T+j}^{(i)}$

There are two ways to obtain the i-th draw of the predictive value $\tilde{y}_{T+j}^{(i)}$. The first and the easier way is to draw $\tilde{y}_{T+j}^{(i)}$ from the conditional density of $f(\tilde{y}_{T+1}|\hat{\Theta}, \, data)$, where $\tilde{\Theta}$ is the posterior mean of the parameters. And we keep randomly drawing for n times. For convergence purpose, I get rid of the first m values, and choose every $j-$th value as the final MCMC series. Then, the density is the final sample produced by the iteration.

The second way is to obtain the unconditional density of the predictive $\tilde{y}_{T+1}$ given by, $f(\tilde{y}_{T+1}|data) = \int f(y_{T+1}|\Theta)P(\Theta|data)d\Theta$. So, one complete cycle of MCMC procedure actually consists of 3 steps. First I use method one described in previous paragraph to get the predictive value of $y_{t=1}^{(i-1)}$ conditional on $\hat{\Theta}^{(i-1)}$. Then draw $\hat{\Theta}^{(i)}$ based on the generated $y_{t+1}^{(i-1)}$and the original data, and finally I draw the predictive $y_{t+1}^{(i)}$based on the $\hat{\Theta}^{(i)}$. I iterate this process by n times, and the final sample is produced by the iteration after burning the initial values and choose every h-th value. In this essay, I use the both ways to obtain the predictive value of the one-period ahead $\tilde{y}_{T+1}$.

2.5 Data

The data for my empirical analysis is the monthly UK 3-month Treasury Bill from the Global Financial website. It covers the horizons from January 1950 to December 2005, total 672 observations. Figure 1 and Figure 2 plot the dynamics of the level and absolute change of the 3-month interest rate over time. There are presence of frequent jumps that can be characterized by high level of volatility and mean revision. For example, the interest rate for the period of 1973-1991 has high level and high volatility. In particular, during the late 1970s and early 1980s, macroeconomic conditions were dominated by the oil price shocks and other rises in agricultural and mineral products. These were considered to play a big role in the stagflation of that decade. The recession is characterized by significantly higher interest rate and somewhat more variable interest rate. The return also changes significantly in economic recession.

Table 2 summarizes the descriptive statistics. We find that the interest rate is slightly right skewed; however, it does not show excess kurtosis. Jarque-Bera statistics indicates that it is not normally distributed. The series are highly auto correlated even at lag 30. Table 3 present the ADF and KPSS unit root test, and we find that we cannot reject the unit root for the level at 5% significance level, but can reject the unit root for first difference at 1% level.

2.6 Results

The threshold-like behavior of interest rate can arise from following aspects. The first one is the transaction cost [see Anderson (1997)]. Because the transaction is not continuous, the adjustment of the deviation from the arbitrage condition is infrequent, and happens after passing through a threshold. Another reason comes from the policy aspect. Short-term interest rates are generally determined by the Federal Reserve in U.S. or central bank in other countries. The small deviation from the target zone is admissible for policy maker, only the deviation exceed its market expectation, i.e. beyond certain critical value, then central bank will re-evaluate the rate, thus the nonlinearity of the interest rate will be induced.

The estimation of TARMA model is summarized in Table 4. The threshold variable is the absolute change of the short rate $\Delta z_{t-1} = |y_{t-1} - y_{t-2}|$. The returns are used as a proxy for latent monthly volatility process, thus the threshold variable divide the data into low- and high- volatility regimes. The threshold parameter is selected to be 0.178. There are 390 observations falling in the high-volatility regime and 279 observations classified into the lower volatility regime. Except the persistence coefficient, the estimated parameters for two regimes are quite different. All parameters lie within their 95% HPDI. Most parameters are statistically significant except the MA coefficient and AR coefficient in regime 1. This indicates that the data do exhibit some regime changing.

The estimated γ_2 is 0.9848 and 0.9855 for regime 1 and regime 2 respectively. Thus the persistence of the low- and high-regimes is both strong, which means the probability of interest rate remaining in the previous regime for next period is more persistent. We cannot reject the unit root for both regimes as the persistence parameter lies in the 95% HPDI. This results are different from the observations made by some former regime switching literature [Gray (1996), Ang and Bekaet (2002a), Bansal and Zhou (2002), Goldman and Agebeyegbe (2005)]. They use US data and find that interest rate follows a unit root in lower regimes and the mean-reversion process become stronger after pass certain threshold.

The possible reason for the difference is the model specification. According to Brenner (1996), the interest rate volatility has both level and GARCH effects. If GARCH component is omitted in the model, the estimated level effect is biased. Since conditional variance could also be random and they are modeled as GARCH process in the literatures I mentioned above, the high conditional volatility in upper regime is attributed to the GARCH estimation, and the lagged level effect plays the mean reverting role. In our model, variance σ_t^2 is assumed to be conditional heteroskedasticity, but this is not modeled explicitly. Thus it is included in the persistence coefficient, which makes the upper regime also exhibit unit root process. The variance in both regimes exhibit large difference, which indicates that the conditional variance also follows regime switching process. The GARCH component can be added in modeling conditional variance to improve the estimation accuracy or forecast performance. The stationery of non-linear time series implies that the stationary two regime TAR model is characterized by a unit root in lower regimes and stationary in upper regimes, which is not the case we find here. Therefore, the interest rate is not stationary here.

I apply the same data to the 2-state Markov switching model, which is defined in section 3. I perform MCMC iteration 9000 times, and the first 6000 draws are discarded. The estimate results are summarized in Table 5. Except for the intercept coefficient, all the other coefficients are statistically significant. We also find that the all estimated parameters are in the 95% HPDI.

Figure 3 illustrate the effectiveness of the model. The black line stands for the true value of interest rate, and the fitted value is illustrated by the red line. We find that the model captures the trend of peaks and troughs of the real data. Further more, we can distinguish the difference between the peaks after the estimation; the second peak is due to the change of state while the first one is just an error process.

We obtained the mean value and the 95% HPDI of one-day ahead forecast for both models use the two methods described in section 4.3, and compare them with the realized value. The results in table 6 indicate that the mean values of the predictor for both models are close to the realized value no matter which method we choose. Except for the conditional predictive value by the MSM, all the realized values fall within the 95% HPDI in both models. However, the unconditional predictive mean is slightly closer to the true value than the conditional predictive mean for TARMA model. The same patter is found for MSM. Under unconditional case, the mean value of the predictor is better for MSM than for TARMA model as the former one is close to the true value. In addition, the prediction by MSM has much less standard deviation and tighter 95% HDPI compared with TARMA model.

The above numbers are not so clear for deciding which model is better. I use the predictive density of Y_{n+1} to illustrate the case. Figure 4 displays the posterior density of one-day forecast by both models using conditional predictive density. There is no distinctive superiority of TARMA or MSM. The shape of the density by TARMA model looks more close to normal distribution, and centered at 5.2.; however, it has more probability at the right tail. The density of MSM has two modes and relative large probability at the left tail. It's hard to choose one model to be the leading one based on these results. If we draw the unconditional predictive density, the results change in the way that MSM is relatively more accurate. Figure 5 shows the posterior density for unconditional predictive by both models. Clearly, the MSM still has two modes, but it has much less probability at the tails, and tightens around its peak.

I then calculate the mean, standard deviation; mode and HPDI for the MSEF of the one-day forecast for both models, and the results are presented in Table 7. The one-day forecast by the MSM model has less mean, standard deviation and mode for unconditional predictive value. If we use these as the criteria for model selection, the MSM is preferred to TARMA model. The conditional case is decisive now. The mean, standard deviation and mode of MSEF for MSM is large than for TARMA model, therefore, TARMA model is superior than MSM under conditional case.

To make further comparison, I graph the posterior density of MSEF for both models, and Figure 6 and Figure 8 illustrate the results. For conditional case, the posterior density of TARMA is tighter than that of MSM, However, it peaks at the bigger value, which cannot demonstrate that TARMA beats MSM. The unconditional case is also ambiguous. On the one hand, MSEF by MSM has less probability at extreme values, on the other hand, it has bigger peak value than TARMA model does. Thus, I turn to another more accurate comparison criteria: the cumulative density of MSEF. I graph the cumulative density for both conditional and unconditional predictive values in Figure 7 and Figure 9. Clearly, the TARMA model is superior to the MSM in conditional case as it rises quickly and smoothly to the highest level. For the unconditional predictive MSEF, it shows that MSM is the better one. Therefore, the TARMA model is better in predicting future short-term interest rate in conditional case while the MSM is superior for unconditional case.

2.7 Conclusions

In this essay, I use both conditional and unconditional predictive density to compare the TARMA model and MSM. The results are mixed. For conditional predictor, the MSEF and the posterior density graph both indicate that the TARMA model is superior to MSM, with less MSEF and less probability at tails for posterior density of MSEF. But this conclusion is just opposite if we use unconditional prediction. The posterior density of predictor is much tight, and MSEF is smaller for MSM than for TARMA model. Under both methods, the posterior density of prediction for MSM has turned out to be two modes, which makes the comparison not so straightforward. Since unconditional density uses data twice, we may consider conditonal density is more precise in predict future value of interest rate than unconditional method. However, it leaves to the researcher's judgemental call which model is better.

In extension, we can use the alternative model selection criteria, such as cross validation combining predictive density. The cross-validation method dates back to Geisser and Eddy (1979). Gelfand (1996) provide further discussion on it. The underlying idea is to divide dataset into two parts. One proportion of the observations is used for estimation and the other proportion is used for validation. However, there are different approaches to partition the observations. In this way, it reduces the estimation risks due to the double use of the same dataset for both estimation and validation, which pointed out by Hastie, Tibshirani and Friedman (2001).

Appendices

II.A: Proposal Density for TARMA Model

The ARMA algorithm was suggested by Chib and Greenberg (1994). For the TARMA model with efficient jump, we follow Goldman (2005). The multivariate Normal distribution is used as the proposal density for each block of the parameters. The TARMA model is given by

$$y_t = \epsilon_t + g(Z_t) = x_t\gamma^{(j)} - \sum_{i=1}^{p} \phi_i^{(j)} e_{t-j} + \epsilon_t + \sum_{i=1}^{q} \theta_i^{(j)} \epsilon_{t-j} \tag{28}$$

where, $\epsilon_t \backsim N(0, \sigma^2)$ for every $t \in T_j = \{r_{j-1} \leq z_{t-d} \leq r_j\}$, $j = 1, \ldots, k+1$. The likelihood function is given by,

$$\ell\left(y|x, \gamma, \phi, \theta, r\right) = \prod_{j-1}^{k+1} \prod_{t \in T_J} \frac{1}{\sigma_t\sqrt{2\pi}} \phi\left(\frac{y_t^{(j)} - g(Z_t)}{\sqrt{2}\sigma_t}\right) \tag{29}$$

where $\sigma_t$is fixed and known. Given the i-th draws of $r^{(i)}$, we classify the sample of y, and x_t into regimes. Then estimating the parameters for each regime.

1) Proposal density for $\gamma^{(i)}$. For the observations in regime $j(t \in T_j = \{r_{j-1} \leq z_{t-d} \leq r_j\})$,equatio (28) is re-written as

$$y_t^* = x_t^*\gamma^{(j)} + \epsilon_t \tag{30}$$

where y_t^* and x_t^* are calculated by the following transformation,

$$y_t^* = y_t - \sum_{i=1}^{p_j} \phi y_{t-j} - \sum_{i=1}^{q_j} \theta_i^{(j)} y_{t-j}^* \tag{31}$$

$$x_t^* = x_t - \sum_{i=1}^{p_j} \phi x_{t-j} - \sum_{i=1}^{q_j} \theta_i^{(j)} x_{t-j}^* \tag{32}$$

$y_t = y_t^* = 0$, $x_t = x_t^* = 0$ for $t \leq 0$. Let $X_{\gamma_j^{(i-1)}} = (x_{j_1}^{*'}, \ldots, x_{j_{nj}}^{*'})'$ be the matrix of

x_t^* belonging to regime j, and $\left(\sum_{\gamma^{(j)}}^{(i-1)}\right)^{-1}$ be a diagonal $n_j \times n_j$ variance-covariance matrix of $\epsilon_t(t \in T_j = \{r_{j-1} \le z_{t-d} \le r_j\})$. We have following proposal density of $\gamma^{(j)}$,

$$\gamma_j^{(i)\sim}N\left(\gamma_j^{(i-1)}, \sum\limits_{\gamma_j^{(i-1)}}\right), \sum\limits_{\gamma_j^{(i-1)}} = s_1^2\left(X'_{\gamma_j^{(i-1)}}(\sum\limits_{j}^{(i-1)}) + \sum\limits_{r_j}^{-1}\right)^{-1} \tag{33}$$

We draw each $\gamma_j^{(i)}$ separately because the posterior density of γ_j is independent of γ_l for $l \neq j$, and accept or reject them jointly for the whole vector $\gamma^{(i)} = \left(\gamma_1^{(i)}, \ldots, \gamma_{k+1}^{(i)}\right)$ with probability λ_1 given in the text. We set $s_t = 1.0$ if the accept rate greater than 40%.

2) Proposal density for AR coefficient $\phi^{(i)}$. For the observations in regime $j(t \in T_j = \{r_{j-1} \le z_{t-d} \le r_j\})$, equation (28) is re-written as,

$$\tilde{y}_t = \tilde{x}_t\phi^{(j)} + \epsilon_t \tag{34}$$

where y_t^{**} and x_t^* are calculated by the following transformation,

$$\tilde{y}_t = y_t^{**} - x_t^{**}\gamma^{(j)} - \sum_{i=1}^{q_j}\theta_i^{(j)}y_{t-j}^{**} \tag{35}$$

$\tilde{x}_t = \left[\tilde{y}_{t-1}, \ldots, \tilde{y}_{t-p_j}\right]$ and $\tilde{y}_t = y_t^{**} = 0$ for $t \le 0$. Let $X_{\phi_j} = \left(\tilde{x}'_{j_1}, \ldots, \tilde{x}'_{j_{nj}}\right)'$ be the matrix of $\tilde{x}_t$ belonging to regime j, we have following proposal density of $\phi^{(i)}$,

$$\phi_j^{(i)\sim}N\left(\phi_j^{(i-1)}, \sum\limits_{\phi_j^{(i-1)}}\right), \sum\limits_{\phi_j^{(i-1)}} = s_2^2\left(X'_{\phi_j^{(i-1)}}(\sum\limits_{j}^{(i-1)}) + \sum\limits_{\phi_j}^{-1}\right)^{-1} \tag{36}$$

We draw each $\phi_j^{(i)}$ separately and accept or reject them jointly for whole vector $\phi^{(i)} = \left(\phi_1^{(i)}, \ldots, \phi_{k+1}^{(i)}\right)$ with probability λ_2 given in the text.

3) Proposal density for MA coefficient $\theta^{(i)}$, For the observations in regime $j(t \in T_j = \{r_{j-1} \le z_{t-d} \le r_j\})$, equation (28) is re-written as,

$$\hat{y}_t = \hat{x}_t \phi^{(j)} + \epsilon_t \tag{37}$$

where y_t^{**} and x_t^{*} are calculated by the following transformation,

$$\hat{y}_t = y_t^{**} - x_t^{**}\gamma - \sum_{i=1}^{p} \phi_i^{(j)} \left(y_{t-j}^{**} - x_t^{**}\gamma^{(j)} \right) \sum_{i=1}^{q_j} \theta_i^{(j)} y_{t-j}^{**} \tag{38}$$

and $\hat{y}_t = y_t^{**} = 0$ for $t \leq 0$. Let $X_{\theta_j} = \left(\hat{x}'_{j_1}, \ldots, \hat{x}'_{j_{nj}} \right)'$ be the matrix of $\hat{x}_t$ belonging to regime $j, \hat{x}_t = \left[\hat{y}_{t-1}, \ldots, \hat{y}_{t-p_j} \right]$. We have following proposal density of $\theta^{(i)}$,

$$\theta_j^{(i)} \sim N \left(\theta_j^{(i-1)}, \sum_{\theta_j^{(i-1)}} \right), \sum_{\theta_j^{(i-1)}} = s_3^2 \left(X'_{\theta_j^{(i-1)}} (\sum_{j}^{(i-1)}) + \sum_{\theta_j}^{-1} \right)^{-1} \tag{39}$$

We draw each $\phi_j^{(i)}$ separately and accept or reject them jointly for whole vector $\phi^{(i)} = \left(\phi_1^{(i)}, \ldots, \phi_{k+1}^{(i)} \right)$ with probability λ_3 given in text.

II.B: Metropolis-Hasting Accept or Reject Rule for Each Block

1) Regression coefficient block γ_j. For each regime, $j = 1, 2,$, generate $\gamma_j^{(i)}$ from the proposal density $N\left(\gamma_j^{(i-1)}, \sum_{\gamma_j^{(i-1)}}\right)$, which is defined previously. Let $\gamma^{(i)} = \left(\gamma_1^{(i)}, \gamma_2^{(i)}, \gamma_3^{(i)}\right)$. Accept or reject $\gamma^{(i)}$ with probability:

$$\lambda_1 = \min\left\{\frac{P(\gamma^{(i)}, \phi^{(i-1)}, \theta^{(i-1)}, \sigma^{2(i)}, r^{(i-1)}|data}{P(\gamma^{(i-1)}, \phi^{(i-1)}, \theta^{(i-1)}, \sigma^{2(i)}, r^{(i-1)}|data}, 1\right\}, \tag{40}$$

otherwise $\gamma^{(i)} = \gamma^{(i-1)}$.The proposal density in numerator and denominator has been canceled out as we take random walk draw.

2) AR coefficient block ϕ_j. For each regime, $j = 1, 2,$, generate $\phi_j^{(i)}$ from the proposal density $N\left(\phi_j^{(i-1)}, \sum_{\phi_j^{(i-1)}}\right)$, which is defined previously. Let $\phi^{(i)} = \left(\phi_1^{(i)}, \phi_2^{(i)}, \phi_3^{(i)}\right)$. Accept or reject $\phi^{(i)}$ with probability:

$$\lambda_2 = \min\left\{\frac{P(\gamma^{(i)}, \phi^{(i)}, \theta^{(i-1)}, \sigma^{2(i-1)}, r^{(i-1)}|data}{P(\gamma^{(i)}, \phi^{(i-1)}, \theta^{(i-1)}, \sigma^{2(i-1)}, r^{(i-1)}|data}, 1\right\}, \tag{41}$$

otherwise $\phi^{(i)} = \phi^{(i-1)}$.

3) AR coefficient block θ_j. For each regime, $j = 1$ and 2, generate $\theta_j^{(i)}$ from the proposal density $N\left(\theta_j^{(i-1)}, \sum_{\theta_j^{(i-1)}}\right)$, which is defined previously. Let $\theta^{(i)} = \left(\theta_1^{(i)}, \theta_2^{(i)}, \theta_3^{(i)}\right)$. Accept or reject $\theta^{(i)}$ with probability:

$$\lambda_3 = \min\left\{\frac{P(\gamma^{(i)}, \phi^{(i)}, \theta^{(i)}, \sigma^{2(i-1)}, r^{(i-1)}|data}{P(\gamma^{(i)}, \phi^{(i)}, \theta^{(i-1)}, \sigma^{2(i-1)}, r^{(i-1)}|data}, 1\right\}, \tag{42}$$

otherwise$\theta^{(i)} = \theta^{(i-1)}$.

4) σ^2 block. For each regime j with n_j observations, we generate $\sigma_j^{2(i)}$ from Inverted Gamma distribution,

$$g = IG(v, d) \tag{43}$$

where $v = \dfrac{n_j + v_0}{2}$, $d = \epsilon'\epsilon + \delta_0,\ v_0 = \delta = 0$, and $\epsilon_t = y_t^* - x_t^*\gamma, y_t^*, x_t^*$ are given in proposal density part in appendix. Let $\sigma^{2(i)} = \left(\sigma_1^{2(i)}, \sigma2_2^{(i)}, \sigma_3^{2(i)}\right)$. Accept or reject $\sigma^{2(i)}$ with probability:

$$\lambda_4 = \min\left\{\frac{P(\gamma^{(i)}, \phi^{(i)}, \theta^{(i)}, \sigma^{2(i)}, r^{(i-1)}|data}{P(\gamma^{(i)}, \phi^{(i)}, \theta^{(i)}, \sigma^{2(i-1)}, r^{(i-1)}|data}, 1\right\}, \tag{44}$$

otherwise $\sigma^{2(i)} = \sigma^{2(i-1)}$.

II.C: Derivation of Conditional Density of S_t

The conditional distribution of S_t is given as follows,

$$P(S_t|S_{\neq t}, Y, \Theta) = \frac{P(Y|S)P(S_t|S_{\neq t})}{P(Y|S_{\neq t})} \propto P(Y|S)P(S_t|S_{\neq t}) \tag{45}$$

The first term in above equation is generated by

$$\begin{aligned} P(Y|S) &= P(y_1|S)P(y_2|y_1, S), \ldots P(y_T|y_{T-1}, S) = P(y_1|S)P(y_2|y_1, S_1, S_2), \ldots, P(y_t|Y_{T-1,}S) \\ &\propto P(y_t|Y_{T-1,}S_1, \ldots, S_t), \ldots, P(y_t|y_{T-1}, S) \end{aligned} \tag{46}$$

In above equation, step 2 to step 3 is given by the fact that the likelihood function of $y_t = (t = 1, \ldots, T)$ is independent of all past history of state and Y_{t-1}.

The second term is derived by,

$$\begin{aligned} P(S_t|S_{\neq t}) &= P(S_t|S_{1,} \ldots, S_{t-1,}S_{t+1,}S_T) = \frac{P(S_{t+1,} \ldots, S_T|S_{1,} \ldots, S_t)P(S_t|S_{1,} \ldots, S_{t-1})}{P(S_{t+1,} \ldots, S_T|S_{1,} \ldots, S_{t-1})} \qquad (47) \\ &\propto P(S_{t+1,} \ldots, S_T|S_{1,} \ldots, S_t)P(S_t|S_{1,} \ldots, S_{t-1}) \\ &= P(S_{t+1}|S_{1,} \ldots, S_t)P(S_{t+2}|S_{1,} \ldots, S_{t+1}), \ldots, P(S_T|S_{1,} \ldots, S_T - 1)P(S_t|S_{1,} \ldots, S_{t-1}) \\ &= P(S_{t+1}|S_t)P(S_{t+2}|S_{t+1}), \ldots, P(S_T|S_{T-1})P(S_{t+1}|S_t)P \\ &\propto P(S_{t+1}|S_t)P(S_t|S_{t-1}) \end{aligned}$$

Combine above two equations, we have,

$$P(S_t|S_{\neq t}, Y, \Theta) \propto P(S_{t+1}|S_t, \Theta_1)P(S_t|S_{t-1}, \Theta_1)P(y_t|Y_{t-1}, S_{t,}\Theta) \tag{48}$$

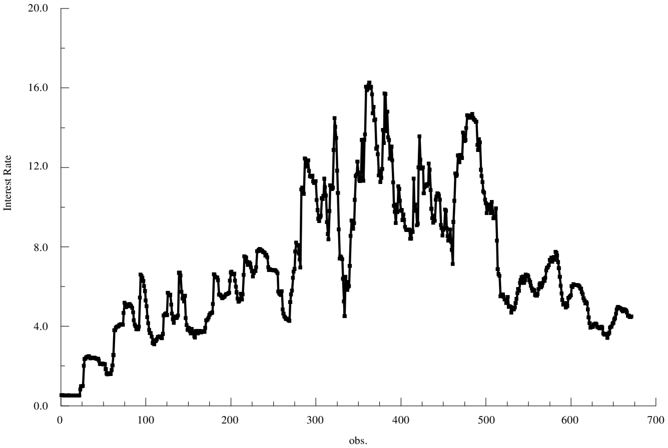


Figure 1: UK 3-Month T-bill Rate

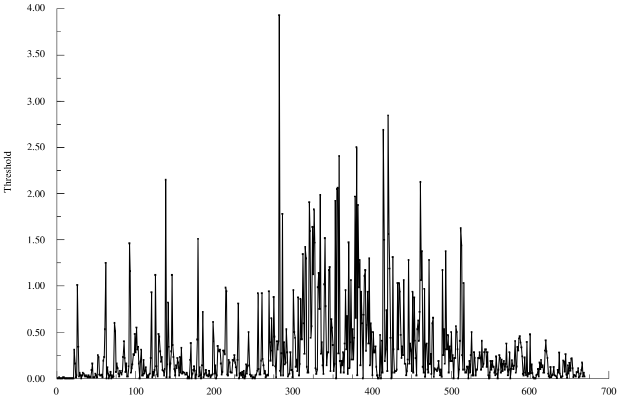


Figure 2: Threshold; $\Delta z_{t-1} = |y_{t-1} - y_{t-2}|$

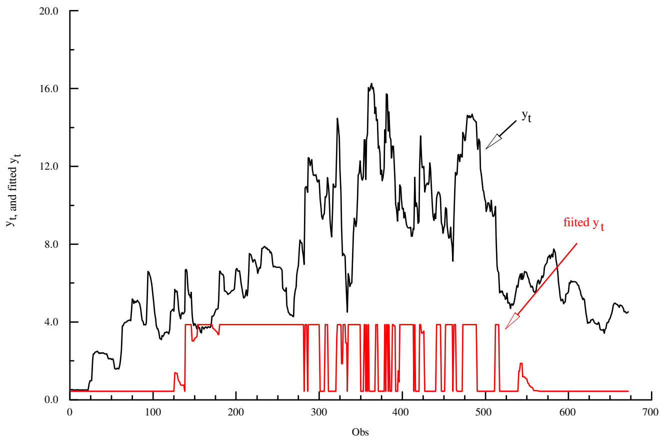


Figure 3: y_t **and** $\hat{y}_t = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{S}_t$

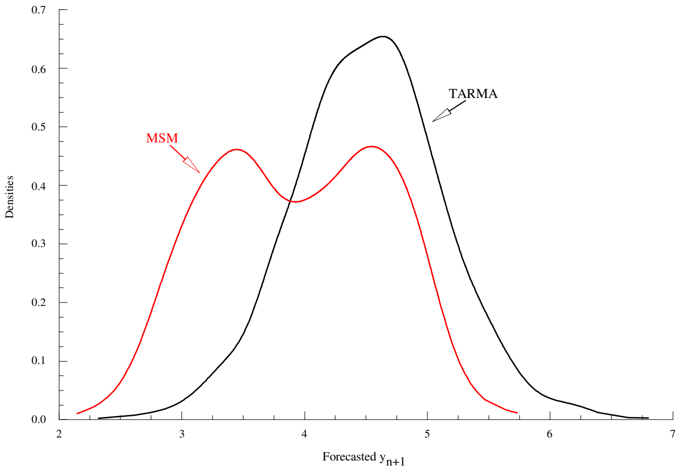


Figure 4: Conditional Predictive Density of Y_{n+1}

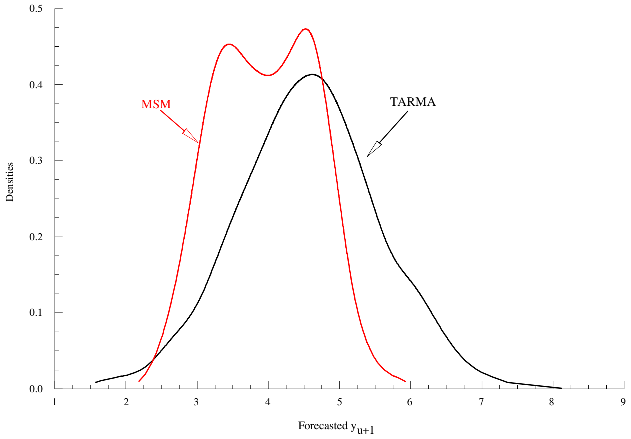


Figure 5: Unconditional Predictive Density of Y_{n+1}

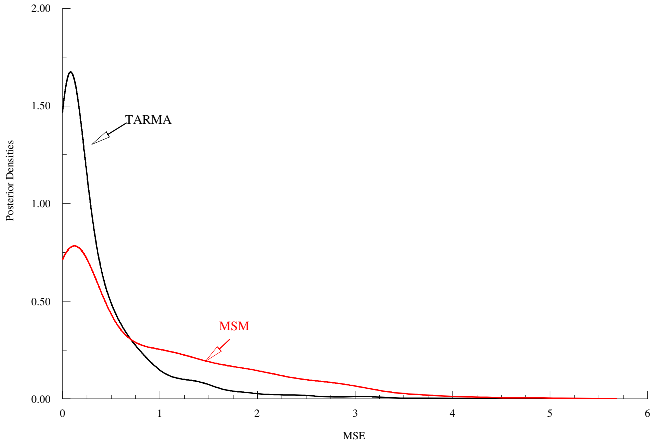


Figure 6: Posterior Density of MSE for Condtional Predictive Y_{n+1}

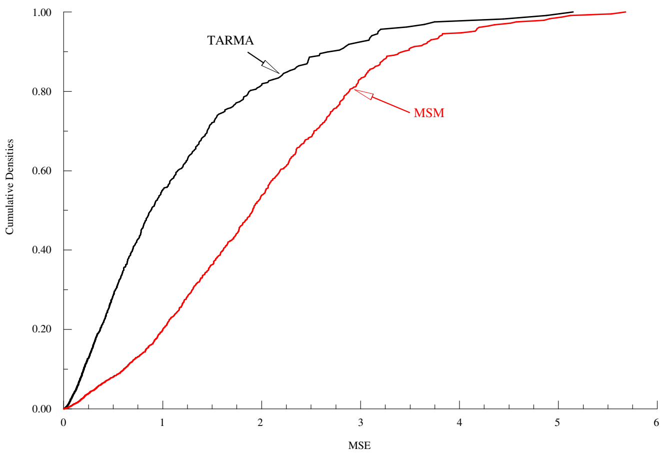


Figure 7: Cumulative Density of MSE for Uncond. Predictive Y_{n+1}

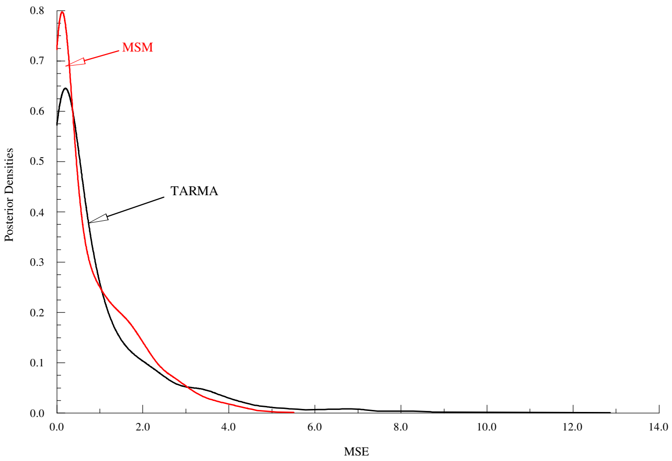


Figure 8: Posterior Density of MSE for Unconditional Predictive Y_{n+1}

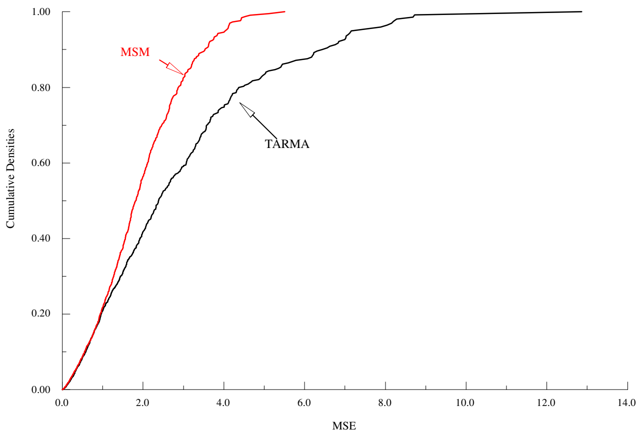


Figure 9: Cumulative Density of MSE for Unconditional Predictive Y_{n+1}

Table 1: AC and PAC for Interest Rate

	interest		Δinterest	
lag	AC	PAC	AC	PAC
1	.99	.99	.13	.13
2	.98	-.03	.05	.03
3	.97	.02	-.04	-.55
4	.96	-.02	.03	.04
5	.95	-.01	.05	.04
6	.94	-.02	-.00	-.21
7	.93	-.01	.03	.03
8	.92	-.04	-.96	-.10
9	.91	-.01	-.41	-.23
10	.90	-.00	-.10	-.85

Table 2: Descriptive Statistics

	Mean	Std. Dev.	Skewness	Kurtosis	J-B	L-BQ(30)
y	6.98	3.58	.53	2.61	34.86	119.30

Table 3: Unit Root Tests

Variables	Test	Deterministic Terms	Lags	Test Value
interest	ADF	Drift	25	-2.60
	ADF	Drift&Trend	25	-2.40
	KPSS	drift	25	1.02^{***}
	KPSS	drift & trend	25	0.60^{***}
Δinterest	ADF	none	24	-22.76^{***}
	ADF	drift	24	-21.72^{***}
	KPSS	drift	24	0.12

Note: *** indicate rejection the null hypothesis at 1% level.

Table 4: Estimates of TARMA Model for Short-Term Interest Rate

		Two-Regime TARMA		
Parameters	Mean	St dev.	Corr.	HPDI at 95%
γ regime 1	.1100	.0275	.0530	$(.0536, .1818)$
	.9848	.0040	.0456	$(.9740, 09917)$
γ regime 2	.1085	.0393	-0.019	$(.0277, .2137)$
	.9855	.0044	.0228	$(.9745, .9953)$
ϕ regime 1	.1523	.1457	.0245	$(-0.1256, .4807)$
ϕ regime 2	.0911	.0215	-0.0144	$(0.0394, 0.1303)$
θ regime 1	.1414	.1169	-0.0493	$(-0.0778, 0.4265)$
θ regime 2	.0918	.0205	.0216	$(.0438, .1290)$
s^2 regime 1	.1780	.0138	0.289	$(.1526, .2043)$
s^2 regime 2	.1618	.0422	.0211	$(.3849, .5475)$
Threshold r	.1780	.0144	.0390	$(.1470, .2072)$
Max AR root	.1669	.1286	.0175	$(.0002, .4273)$
	.0911	.0214	-0.0118	$(.0384, .1303)$
% obs in regime 1	58.30			
% obs in regime 2	41.70			
MBIC	-139.52			

Table 5: Estimates of MSM

Two-State MSM				
Parameters	Mean	St. dev.	Corr	95% HPDI
γ_0	0.4326	0.3298	0.0078	$(-0.2263, 1.0657)$
γ_1	1.7062	0.3392	0.0240	$(1.0234, 2.3656)$
ϕ	0.9958	0.0026	-0.0144	$(0.9909, 1.0000)$
θ	0.4663	0.0362	0.0666	$(0.3940, 0.5339)$
s^2	0.1557	0.0088	0.0803	$(0.1388, 0.1733)$
p_{00}	0.9380	0.0128	0.0793	$(0.9130, 0.9626)$
p_{11}	0.0912	0.0181	0.0283	$(0.0556, 0.1268)$

Table 6: Forecast of TARMA and MSM

		Cond.	Uncond.
Realized	y_{T+1}	4.5300	4.5300
TARMA	$\tilde{y}_{T+1}$	4.5218 (.6038)	4.5357 (.3987)
	95% $HPDI$	$3.3158, 5.7081$	$3.7914, 5.2870$
MSM	$\tilde{y}_{T+1}$	4.5528 (.9832)	4.5218 (.4063)
	95% $HPDI$	$.6395, 6.4956$	$3.7331, 5.3272$
Note:	The numbers in brakect are standard deviations		

Table 7: MSE of Forecast for TARMA and MSM

	TARMA		MSM	
	Cond.	Uncond.	Cond.	Uncond.
Mean	0.3644	0.9667	0.8455	0.8199
Std. Dev	0.5549	1.3868	0.9879	0.9620
Mode	0.0833	0.2012	0.1221	0.1241

3 Model Selection: Comparison of 10-Fold Cross Validation and Out-of-Sample Forecast

3.1 Introduction

Model selection is an important data analysis task and it has its application in many scientific research fields. The model selection criteria are complex depending on the class of models as well as the characteristics of data sets. There are various model selection criteria. One popular approach is using mean squared errors (MSE). More accurate criteria, the predictive density has been proposed to make model selection in application. Cross Validation (CV) is an accepted method to measure the predictive density of a given model. CV will choose the best model which provides the lowest CV error for the given data distribution across several candidate models. The studies for CV method have been focused on using CV in model selection within some model spaces or compare the properties of different CV tests.

Theoretically, CV has its advantages as it reduces the estimation risks due to the double use of the same dataset for both estimation and validation, which pointed out by Hastie, Tibshirani and Friedman (2001). In field of machine learning, it avoids the danger of over-fitting . However, it also has disadvantages as it requires cumbersome computation, accurate choice of partition data point, and unavailable routine for dynamic models. The aim of this essay is to test the model selection ability for CV method by comparing it with the traditional out-of-sample forecast method. If CV method does not outperform the relative standard and less-computational burden out-of-sample method, there is no need to use it. In this essay, we develop the algorithm for 10-fold CV for simple linear non-nested regression model using the generated data. We perform two comparison experiments with true model included in our model space. Using MSEF and Pseudo Bayes Factor (PSBF) as the comparison criteria, we first compare the prediction results from CV and those from the traditional out of sample forecast for fixed points. Normally, we would expect that CV exhibits a better model selection ability for i.i.d sample as it uses the most information of the data. Our results suggest that based on MSEF, randomized CV method and the fixed out-of-sample method can choose the true model with high probability, however none of them has 100% probability to choose the true model. However, the PSBF shows the fixed point method seems to beat the CV method in choosing the true model. We also conduct the second experiment using revised Markov Chain Monte Carlo (MCMC). As pointed out by Robert & Titterington (2002), there is double use of data in applying MCMC to draw the parameters, which has over-fit risk. We then exclude the forecasted data in drawing coefficients, the results for CV improve by having higher PSBF and higher probability of choosing the true model.

This essay is organized as follows. In Sections 3.2 we review the literature of Cross Validation method. Model design and model selection criteria are in section 3.3. In section 3.4, CV algorithm and Bayesian estimation procedures are specified for both experiments, including the prior setting, maximum likelihood and the posterior distribution for each model. Section 3.5 discusses the results. Concluding remarks and extension are discussed in Section 3.6.

3.2 Related Literature

3.2.1 Leave-one-out CV

Based on the size of the validation dataset, there are several methods of CV test. The first is the Leave-one-out CV. It uses all the data points except rth data point to do estimation. The benefit for leave-one-out CV is that it uses the most information of the dataset, so it is accurate to some extend. However, it has the drawback of huge computation load. Stone (1977a), Efron (1983), and Shao (1993) point out that leave-one-out CV is asymptotically equivalent to Akaike information criterion (AIC), C_p and bootstrap. Therefore, it suffers the same inconsistency in model selection as for those criteria. This inconsistency is that the probability of choosing the model with the best predictive ability is not converge to 1 as sample size n tends to infinity. Shao (1993) prove this result based on the classic linear model.

In applying leave-one-out CV, importance sample approach is used in drawing parameters. The quality of the importance sampling depends on the variability of the importance sampling weights. For example, if we leave one dominant data point out, it will change the posterior substantially and the variance of the weights could become infinite. So, leave-one-out CV is not accurate in this case. Another drawback of leave-one-out CV is for high dimensional models. It could fail because of the large variance of importance sampling weight. The examples can be found in Vehtari and Lampinen (2002). If the liability of importance sampling is questionable, the better choice is to use k-fold CV.

3.2.2 K-fold CV

In k-fold CV, the dataset is randomly partitioned into k groups. For each k, leaving the k-th group for validation and the rest $(k-1)$ groups to form the set data T for estimation. Then, report the mean errors over all k validation sets. Using simulated data, Bekara and Fleury (2003) show that the consistency of CV can be achieved by changing the estimation sample size. The ratio of $\dfrac{k}{n-k} = \frac{1}{3}$ is suggested by Hastie, Tibshirani and Friedman (2001). Vehtari and Lampinen (2002) perform simulations and found the cases with k-fold CV works well while leave-one-out CV fails.

Chakrabarti and Ghosh (2006) give the comprehensive discussion on sample partition of CV in model selection. They study how much of the sample should be used as estimation and how much should be used as validation for fixed parameter dimension p under both M-closed (True model is in the model space) and M-open (True model is not in the model space) case; and the infinite parameter p dimension for M-closed case, where $p \in R^p$. Their simulation results suggest that if the parameter dimension is small (fixed), under regularity condition, the better discriminating power of Bayes factor is found for larger size of validation group k. To be exact, as $k \longrightarrow \infty, n-k \longrightarrow \infty$ in a way such that $\frac{k}{n-k} \longrightarrow \infty$, then the Bayes factor will be better to discriminate between models. If the parameter dimension go to infinite, for the nested linear model with normal error, and the more complex model as the true model, then the true model is chosen as $k \longrightarrow \infty, n-k \longrightarrow \infty$ such that $\frac{k}{n}$ is bounded away from zero in the limit. Chakrabarti and Ghosh (2006) also disagree with Stone (1977a) and Shao (1993) in that AIC and CV are not equivalent. They argue that Stone and Shao's conclusions only hold when the considered model is the correct model.

3.2.3 CV in Bayesian Context

In Bayesian framework, the combination of CV and Bayes Factor is popular in model selection. It uses predictive density, normally the posterior density of the validation data conditioned on the estimation data set and the true candidate model. The cross validation Bayes factor, also called Pseudo-Bayes factor (PSBF) (Geisser and Eddy, 1979), which is originally aroused from Stone (1974), and Geisser (1975). Gelfand (1996) provides further discussion on it. Compared with formal Bayes factor, CV Bayes factor avoids the Lindley's paradox[1] inherent in former one. The underlying idea of Bayes factor is to adopt a broader notion of predictive distribution and densities. The predictive density is obtained by averaging a density arising from the likelihood with respect to a distribution arising from the data-based updating of the prior. Using Laplace approximations, Gelfand and Dey (1994) obtained the asymptotic behavior of the predictive density for the Bayes factors under fixed dimensional

[1] Lindley's paradox shows the conflict between Bayesian and frequentist evidences in hypothesis testing. Even if sample sizes increases to infinity, Bayesian methods accept the point null hypothesis for values where the frequentist method leads to rejection. It is a result of the prior having a sharp feature at H_0 and no sharp features anywhere else. See Lindley, Dennis V. (1957). "A Statistical Paradox".

parameters case. In particular, they proved that Lindley's paradox disappears in PSBF. Vehtari and Lampinen (2002) suggests the combination of posterior predictive density with CV method as the posterior predictive density alone is not good in model selection. It generally prefers the overfitted model. Only when the lower dimensional models, the posterior predictive densities are good approximation of CV predictive density. Therefore, CV method is a good supplement to posterior predictive density in model assessment and model comparison. This has been applied by Gelfand et al. (1992) , Gelfand (1996).

The merit of CV comparing with other model selection criteria is shown in Bekara and Fleury (2003). They found that CV combined with predictive density (CVBPD) outperformed the popular AIC and Minimum Description Length (MDL) as the probability of selecting the correct model is consistently higher using their CVBPD measure. As $N \rightarrow \infty$, the probability of choosing the correct model for CVBPD approaches 1. In addition, it also achieves faster convergence than Minimum Description Length (MDL) for small samples. Therefore, CV could be a consistent and robust model selection criterion. Chakrabarti and Ghosh (2006) also show that CV perform equivalent or better job in model selection compared with AIC under certain condition.

The recent applications of CV method are found in Alqallaf and Gustafson (2000), Bekara and Fleury (2003). Kárny, Nedoma, and Šmí¿dl (2005), Chakrabarti and Ghosh (2006). Vehtari and lampinen (2002) list some difficulties in using CV in a Bayesian context. In sum, there is no standard way of applying CV in a Bayesian context.

3.3 Model Specification and Model Selection Criteria

3.3.1 Choice of K

There are challenges in applying CV method in Bayesian context. First, we face the computational challenges. When applying leave-one-out CV, for each re-fit process, Markov Chain Monte Carlo (MCMC) requires many iterations. If we have large dataset, the computation is cumbersome. In this essay, we focus on k-fold CV. One problem associated with the k-fold CV is that it is biased since the estimation data set may not be a good proxy for the full data. However, this bias can be ignored in model comparison because biases are canceled out for both models. But if we assess the model performance, this biased cannot be neglected. There is no agreed principle or theoretical justification to determine the optimal value for k. To some extend, CV is highly depend on how "luckily" or "unluckily" the validation dataset chosen. The appropriate k depends on the models and the dataset. Simulation studies show that value of k between 8 and 16 seems to balance well between the increased accuracy and increased computational load. Therefore, we set k to be 10 here.

3.3.2 Model Specification

Normally, there are two cases for model selection in applications: true model is among the considered group of models or none of the models is the true model. For the first case, researchers need to compare those models to find out the true or optimal model. For the second case, they need to find out which model is closer to true model. Based on the relation among candidate models, there are also two design strategies. One is nested models and the other is non-nested models. For each strategy, we can set the parameters in models fixed, meaning the dimension of parameters in both models is much less than the size of data, i.e, $p << n$. Or, we can set one model with infinite parameter dimension and the others have fixed parameter dimension. However, this scenario may not have much appeal in empirical study. For non-nested models, typically, both models have fixed parameter dimension. If considering the model type, there is linear model and dynamic model. Most paper, including Chakrabarti and Ghosh (2006), Shao (1993), and Bekara and Fleury (2003) choose M-closed case with fixed parameters and nested linear models, we follow their rule of including the true model in our model space, but consider the non-nested models. The reason of choosing non-nested model is its wide application in economic studies.

In our simulation case, we specify two simple linear regression models, one is the correct model and the other is the wrong model. The correct model is specified as follows,

$$\text{Model One: } y = x\beta + \epsilon_1, \text{with } x_2 \sim U\left(0, 5\right), \qquad \beta = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \epsilon_1 \sim N(0, 3) \tag{49}$$

Here x is a 100×2 matrix, with x_1 constant and x_2 uniform distributed with range from zero to five. The error term is normally distributed with zero mean and variance equal to three. Therefore, the data is y_1 generated by the correct model with sample size of 100 and it is an i.i.d sample. The second model is defined as follows:

$$\text{Model Two: } y = z\theta + \epsilon_2, \text{with } z \sim N(2.5, \frac{1}{12}) \tag{50}$$

where z is a 100×2 matrix, with z_1 constant and z_2 is normally distributed with mean of 2.5 and variance equal to 1/12. Based on our data generating process, the cross-validation hopefully lead us to choose model one.

3.3.3 Model Selection Criteria

Predictive density has been proposed to make model selection. Here, we combine the predictive density and CV method to evaluate the two models. Two measures have been applied in model selection. The first one is the MSEF. We draw the predict value $\tilde{y}_j^i$ based on the MCMC draws of the estimated parameters. The MSEF for one validation dataset is given by

$$MSEF_v = \frac{1}{mm}\sum_{i=1}^{mm}\frac{1}{r}\sum_{j=1}^{r}\left(\tilde{y}_j^i - y_j\right)^2 \tag{51}$$

Where $\tilde{y}_j^i$ is the i-th draw of the predictive value of y_j, $j = 1, 2, \cdots, r$. y_j is the realized value. r is the number of observations in the validation data set. mm is the number of MCMC draws. The overall MSEF cross all validation dataset is given by,

$$MSEF = \frac{r}{n}\sum_{v=1}^{n/r} MSEF_v \tag{52}$$

The model with the smaller mean squared errors is chosen as the best model. The posterior density of MSEF and the cumulative density of MSEF are also drawn for further comparison. The posterior density with mass close to the origin and smaller tail is the better one. For cumulative density, the one which more quickly attain the highest level and has higher level at each point is considered the one for the best model. In addition, we compare the shape of the predictive density curve of the two models. So, we draw the kernel density of the predicted $\tilde{y}_j$. Ideally, the model, which better captures the characteristics of the data, has the distribution close to normal and has relative smaller probability at tails. The better model also has smaller variance.

The second criteria are the CV ratio and its log value. According to Gelfand and Dey (1994), assume $y_i, i = 1, 2, ..., n$ be a sequence of independent observations which has density $f(y_j|\theta_{k,}M_k), k = 1, 2$ under model k. Let J_n denote the set $\{1, 2, ..., n\}, \{y_s = (y_j, j \in S)\}$ with size of $r, \{y_{s^c} = (y_{j,}j \notin S)\}$. The conditional density of y_s is given by,

$$\begin{aligned} f(y_s|y_{s^c}, M_k) &= \int L(\theta_k; y_s, M_k)\pi(\theta_k|y_{s^c})d\theta_k \\ &= \frac{\int L(\theta_k; y_s, M_k)L(\theta_k; y_{s^c}, M_k)\pi(\theta_k)d\theta_k}{\int L(\theta_k; y_s, M_k)(\theta_k)d\theta_k}, k = 1, 2 \end{aligned} \tag{53}$$

where $L(\cdot)$ is the likelihood function over the sample space of y_s, and $\pi(\cdot)$ is the prior updated by the sample space of y_{s^c}. Therefore, equation (53) defined a predictive density by averaging the joint density of y_s with respect to the prior for θ_k. If $S = (r), S^c = J_n - \{r\}$, i.e, S^c is the set of $(n-1)$ observations with the rth observation deleted, the CV predictive odds ratio or Pseudo-Bayes Factor (PSBF) is given by

$$\prod_r f(\tilde{y}_r|y_{(r)}, M_1)/f(\tilde{y}_r|y_{(r)}, M_2) \tag{54}$$

where $f(\tilde{y}|y_{(r)}, M_k)$ is the posterior density of the predicted $\tilde{y}$ evaluated at realized y based on the estimation set $y_{(r)}$ and model k. It is defined as follows,

$$f(\tilde{y}_r|y_{(r)}, M_k) = \int L(y_r|\theta|y_{(r)})\pi(\theta|y_{(r)})d\theta \tag{55}$$

To obtain this posterior density, we first draw the kernel density of the predicted $\tilde{y}_r$ based on the MCMC sequence, then obtain the corresponding density $f(\tilde{y}_r^m)$ where $\tilde{y}_r^m = y_r$. Finally, take the average of $f(\tilde{y}_r^m)$ and $f(\tilde{y}_r^{m-1})$ to get the posterior density at the realized y_r. If Model one is the true model or the better model, this ratio is greater than one. We can also average the ratio by taking its log value as follows,

$$\frac{1}{n}\sum_{r=1}^{n} \log \frac{f(\tilde{y}_r|y_{(r)}, M_1)}{f(\tilde{y}_r, |y_{(r)}, M_2)} \tag{56}$$

The result is expected to be positive.

3.4 Simulation

3.4.1 Procedure

There are three major steps in our algorithm. The first step is partitioning the data. CV approach partitions data into two parts, with one part to do the estimation and the other part used as validation set. Since we use 10-fold, which means we leave 10 data points out as the validation set and the remaining 90 data points for estimation. We use two ways to do the partition for comparison purpose. One way, the non-random approach or fixed draw approach, is to fix those 10 point. Let sample size $n = 100$. Choosing the last 10 data points in the sample for validation and use the first 90 data points for estimation. The other way is the actual CV method or random partition, in which we randomly leave 10 points out each time and use the other 90 points to do the estimation. If we have 100 data points, we do the partition 10 times. It ends up with 20 different partition dataset, 10 of them are with 10 data points and the other 10 dataset with 90 data points. For non-random method, there is only two partition datasets.

The second step is applying MCMC algorithm to draw the parameters and predicted $\tilde{y}$. The estimated coefficient β or θ is obtained using estimation sample $y_{(10)}$, where $y_{(10)}$ is the sample space leaving 10 observations out. Then we use the coefficient to obtain the predicted 10 data points. Both coefficients and $\tilde{y}$ are drawn by Metropolis-Hasting (M-H) algorithms. For non-random method, we only need to refit model once. For random draw, we need to refit model 10 times as the sample size is equal to 100. For each refit, we need to do the Markov Chain Monte Carlo (MCMC) iteration m times, with the first n iterations burned out. In our case, we do the iteration 2100 times (repeating step 2 for 2100 times) with first 100 iterations burned; then get an matrix with its row of 2000 and its column of 20 for each model.

The last step is to do the validation. We obtain the MSEF of forecasted $\tilde{y}$ and its distribution. In addition, the CV Bayes factor is also calculated using equation (54) - equation (56), where $r = 10$.

3.4.2 MCMC

Prior, Likelihood Function and Posterior Let $\pi(\theta, \sigma)$ to be the proper prior given by,

$$\pi(\theta, \sigma) = \pi(\theta|\sigma)\pi(\sigma) \propto \pi(\sigma) \propto \sigma^{-1} \tag{57}$$

The likelihood function for simple linear regression model is given by,

$$\ell(\theta, \sigma^2|y, x) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(y - X\theta)'(y - X\theta)\right\} \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(y - X\theta)'(y - X\theta)\right\} \tag{58}$$

The joint posterior pdf of θ and σ is given by,

$$f(\theta, \sigma|y, X) = \pi(\theta, \sigma)\ell(\theta, \sigma^2|y, x) \propto \sigma^{-(n-k)-1} \exp\left\{-\frac{1}{2\sigma^2}\nu s^2\right\} \sigma^{-k} \exp\left\{-\frac{1}{2\sigma^2}(\theta - \hat{\theta})'X'X(\theta - \hat{\theta})\right\} \tag{59}$$

where $\nu s^2 = y'My$, $\nu = n - k$.

By integrating σ out, we obtain the marginal posterior pdf of θ as follows,

$$f(\theta|\sigma, y) = (2\pi)^{-\frac{k}{2}} \sigma^{-k} \exp\left\{-\frac{1}{2\sigma^2}(\theta - \hat{\theta})'X'X(\theta - \hat{\theta})\right\} \tag{60}$$

The joint posterior density of $\tilde{y}$ given θ, σ and $\tilde{X}$ is,

$$\ell(\tilde{y}|\theta, \sigma, \tilde{X}) \propto \sigma^{-m} \exp\left\{-\frac{1}{2\sigma^2}(\tilde{y} - \tilde{X}\theta)'(\tilde{y} - \tilde{X}\theta)\right\} \tag{61}$$

Then the joint posterior density for $\tilde{y}$, θ and σ is given by,

$$f(\tilde{y}, \theta, \sigma | y, X, \tilde{X}) = \ell(\tilde{y} | \theta, \sigma, \tilde{X}) f(\theta, \sigma | y, X) \tag{62}$$

MCMC Procedure We use the M-H algorithm to draw the coefficients and predicted $\tilde{y}$. The parameters are estimated in block: 1) regression parameters θ, 2) variance σ^2, 3) predicted $\tilde{y}$. We can generate θ and $\tilde{y}$ from the proposal density. The initial value of θ is the OLS estimates using the estimation data set with size equal to 90.

The outline of the MCMC procedure is as follows;

1) θ block: set initial value of $\tilde{y}$ and σ^2. Draw $\theta^{(1)}$ using the proposal density given by the following distribution,

$$\begin{bmatrix} \theta_1^{(1)} \\ \theta_2^{(1)} \end{bmatrix} \sim N \left\{ \begin{bmatrix} \begin{bmatrix} \theta_1^{(0)} \\ \theta_2^{(0)} \end{bmatrix}, & \sigma^{2(0)} (x'x)^{-1} \end{bmatrix} \right\} \tag{63}$$

$$\text{Set } \theta^{(1)} = \begin{cases} \theta^{(1)} \text{ with probability } \alpha \left(\theta^{(0),} \, \theta^{(1)} \right) \\ \theta^{(0)} \text{with probability } 1 - \alpha \left(\theta^{(0),} \, \theta^{(1)} \right) \end{cases} \tag{64}$$

where $\alpha = \min \left\{ \dfrac{f(\theta^{(1)}) q(\theta^{(1)})}{f(\theta^{(0)}) q(\theta^{(0)})}, \ 1 \right\}$, $f(.)$ is the posterior density defined by function (62), and $q(.)$ is the proposal density defined in function (63).

2) σ block: draw $\sigma^{(1)}$ by inverted gamma using obtained $\theta^{(1)}$ in regression block, then apply M-H algorithm and accept the proposal value $\sigma^{(1)}$ with probability:

$$\alpha = \min \left\{ \frac{f(\sigma^{(1)} \, , \ \theta^{(1)}) q(\sigma^{(1)}, \theta^{(1)})}{f(\sigma^{(0)} \, , \ \theta^{(1)}) q(\sigma^{(0)}, \theta^{(1)})}, \ 1 \right\} \tag{65}$$

3) Prediction block: draw $\tilde{y}$ based on the coefficients, then apply M-H algorithm to choose. Use function (62) as the posterior density and accept or reject with probability:

$$\alpha = \min\left\{\frac{f(\sigma^{(1)}\ ,\ \theta^{(1),},\ y^{(1)})q(\sigma^{(1)}\ ,\ \theta^{(1),},\ y^{(1)})}{f(\sigma^{(1)}\ ,\ \theta^{(1),},\ y^{(0)})q(\sigma^{(1)}\ ,\ \theta^{(1),},\ y^{(0)})},\ 1\right\} \tag{66}$$

4) Repeat above three steps until each sequence converges.

For the second experiment without forecasted $\tilde{y}$ involved, we use the above steps with slight difference. After drawing θ and σ blocks, we use posterior density defined as equation (61) to draw $\tilde{y}$. Then the information of forecasted $\tilde{y}$ is not used to draw the coefficient. Thus, the joint posterior density of parameters (θ, σ) is based on equation (59). In this way, it avoids double using the data.

3.5 Results

3.5.1 Experiment One: include $\tilde{y}$ to draw θ

The expected results based on our data generating is that model one is the best model under both fixed partition method and random partition method, i.e., CV approach. The mean, standard deviation and median of MSEF for each forecasted $\tilde{y}$ are expected to be lower for model one. However, the results for fixed partition presented in table 8 are not consistent. If we leave the last ten data points out, we got a mixed results: with the 92^{th} and 97^{th} observations have bigger mean ,standard deviation and median for MSEF under model one.

To do the further check, we graph the posterior density and cumulative density of MSEF for each predicted $\tilde{y}$ for both models. Figure 10 to Figure 19 present the corresponding densities for the last ten data points. We find that the graphs are consistent with the results from Table 8 . For data points $\tilde{y}_{92}$ and $\tilde{y}_{97}$, Model two totally beat model one as the former one yield a higher and tighter posterior density for MSEF, and its MSEF reach the highest level faster than model two does. So, the traditional out-of-sample prediction does not always lead to the true model.

Now, we turn to the results of randomized CV method. Table 9 gives the mean, standard deviation and median for forecasted $\tilde{y}_s$ for each partition under both models. The results show that in all cases, randomized CV method chooses the correct model with only one exception. Model one is not chosen as the better model in first partition dataset since it has slightly bigger mean, standard deviation and median of MSEF than that of model two.

The posterior density and cumulative density of MSEF for the first partition dataset shown by figure 20 conform to the MSEF result. Model two appears to be superior than model one with tighter tail in posterior density and quicker convergence for cumulative density. All the figures (Figure 21-Figure 29) for other partition sets indicate that model one is chosen over model two. Therefore, the graphs show that randomized CV method could lead to the correct model selection based on predictive density.

The second measure is PSBF and its log value and the results are shown in table 10. For fixed partition, we got PSBF greater than one and its log value positive, which is consistent with our previous results in MSEF. If we count the times that the posterior density of predicted $\tilde{y}$ evaluated at the realized y for model one is greater than that of model two, we got 60%, greater than 50%. For the random partition, we got PSBF less than one and log of PSBF negative for seven out of ten partition sets, which shows that predictive density combined with randomized CV is less accurate in model selection for this case. It only has probability of 48% that model one yield higher posterior density than model two. The MSEF result is not consistent with the PSBF. If we only check MSEF, randomized CV method is slightly better in choosing the true model than out-of-sample method. However, it is not superior in model selection based on Bayes factors.

In addition, we compare the MSEF from two approaches under more aggregated level by calculating the mean and standard deviation cross all predicted $\tilde{y}_s$. The results in table 11 show that overall, both methods capture the data property with lower MSEF for model one. However, for random partition method, it needs to points out that for some particular data points, apply only predictive density cannot always choose the true model. Therefore, it is not a robust in model selection.

3.5.2 Experiment Two: exclude $\tilde{y}_s$

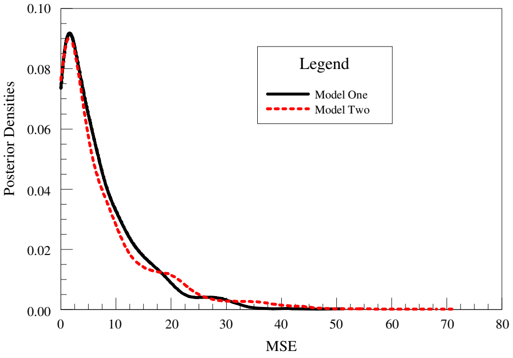
In this experiment, forecasted $\tilde{y}_s$ is not used in constructing coefficients. The MSEF results presented in table 12 show that there is an improvement for fixed partition method because the standard deviation of MSEF for the $97th$ data is smaller for model one, and the mean of MSEF is almost the same for both models. So, the only exception is the $92nd$ data.

The posterior density and cumulative density of MSEF for the predicted $\tilde{y}_{92}$ and $\tilde{y}_{97}$ are shown in figures 30-31. The graph shows that model two beats model one for $\tilde{y}_{92}$ only at certain range as their cumulative density curve cross each other. For $\tilde{y}_{97}$, model one outperforms model two with quicker converge in cumulative density.

The PSBF is 13.20 and its log value is 0.11, which are higher than PSBF if we use the predicted $\tilde{y}$ to obtain the coefficients. 80% of the posterior density evaluated at realized y for model one is higher than posterior density of model two. The probability is also higher than the probability we obtained without using the information of predicted $\tilde{y}$.

3.6 Conclusions

The simulation results are mixed. Randomized CV combined with Bayes Factor does not reveal superior performance in model selection than traditional out-of-sample forecast. MSEF obtained under randomized CV method is slightly better in choosing the true model, but the dominance over out-of-sample method is not noticeable. For i.i.d sample, excluding the forecasted data in constructing coefficients in MCMC improves the model selection ability by fixed partition method.



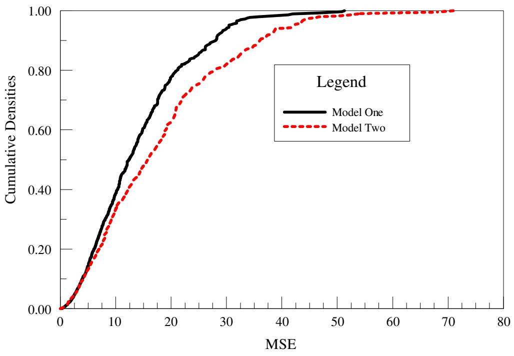
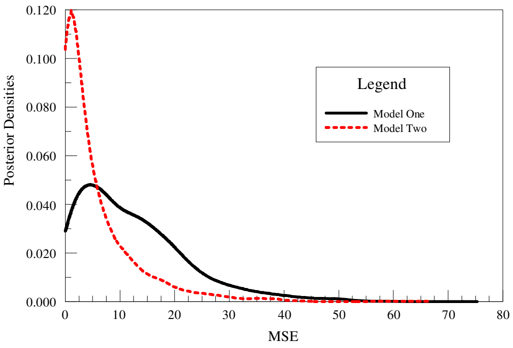


Figure 10: Posterior Density of MSE for Predictive $y_{\tilde{9}1}$.

Cumulative Density of MSE for Predictive $\tilde{y}_{91}$



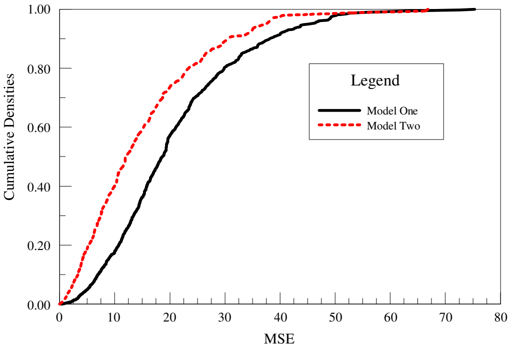
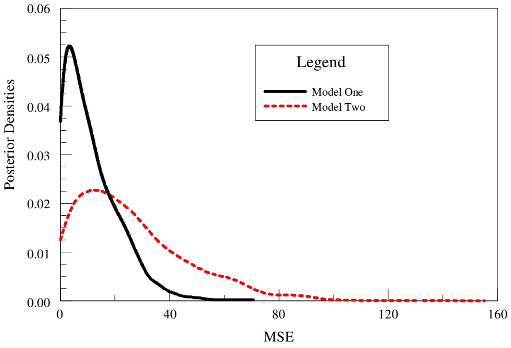


Figure 11: Posterior Density of MSE for Predictive $y_{\tilde{9}2}$.

Cumulative Density of MSE for Predictive $\tilde{y}_{92}$



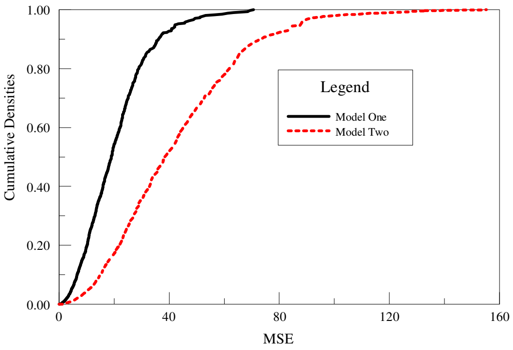
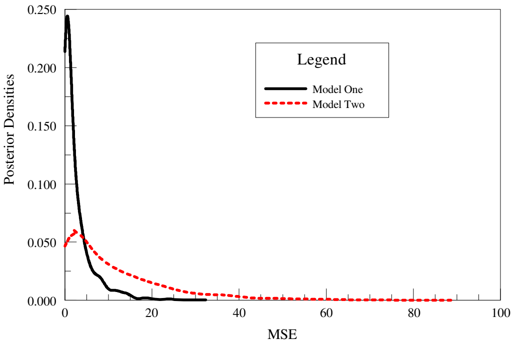


Figure 12: Posterior Density of MSE for Predictive $y_{\tilde{9}3}$.

Cumulative Density of MSE for Predictive $\tilde{y}_{93}$



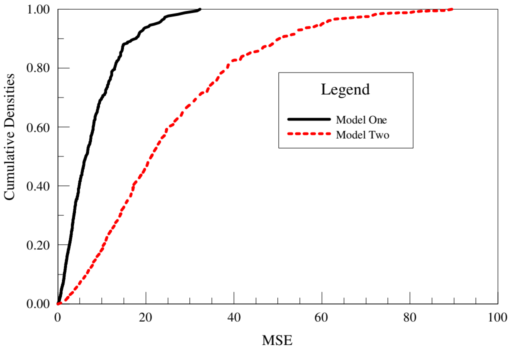


Figure 13: Posterior Density of MSE for Predictive $\tilde{y}_{94}$.

Cumulative Density of MSE for Predictive $\tilde{y}_{94}$

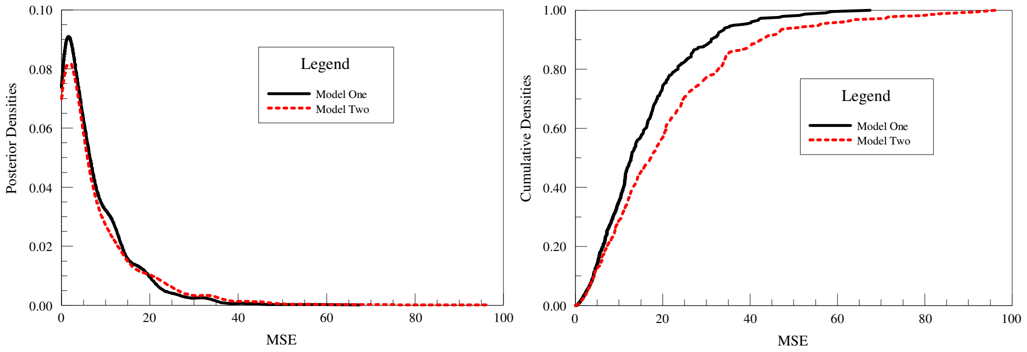


Figure 14: Posterior Density of MSE for Predictive $\tilde{y}_{95}$.

Cumulative Density of MSE for Predictive $\tilde{y}_{95}$

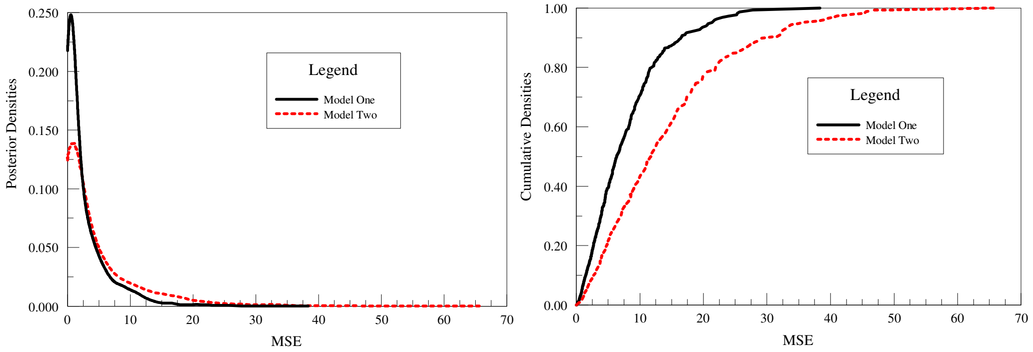


Figure 15: Posterior Density of MSE for Predictive $\tilde{y}_{96}$.

Cumulative Density of MSE for Predictive $\tilde{y}_{96}$

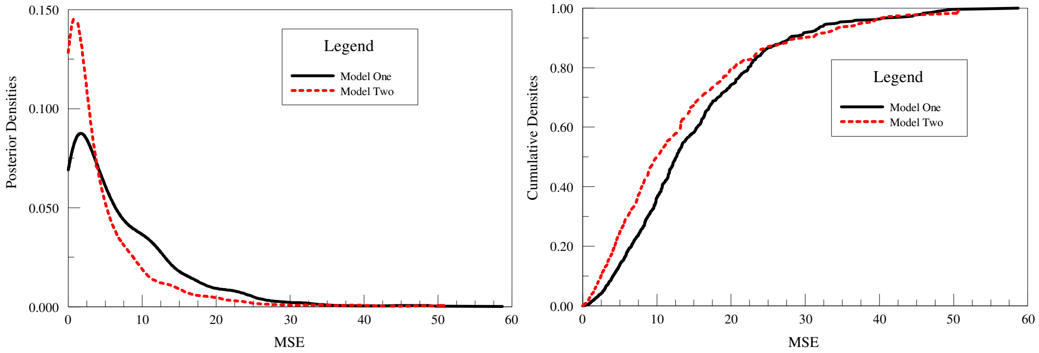


Figure 16: Posterior Density of MSE for Predictive $\tilde{y}_{97}$.

Cumulative Density of MSE for Predictive $\tilde{y}_{97}$.

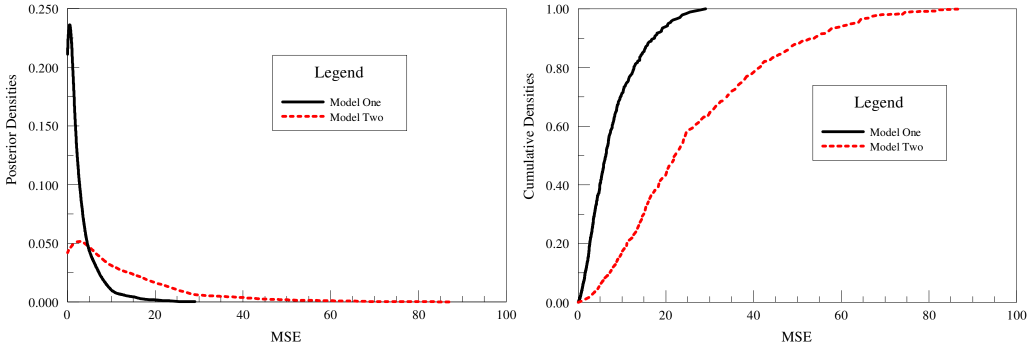


Figure 17: Posterior Density of MSE for Predictive $\tilde{y}_{98}$.

Cumulative Density of MSE for Predictive $\tilde{y}_{98}$.

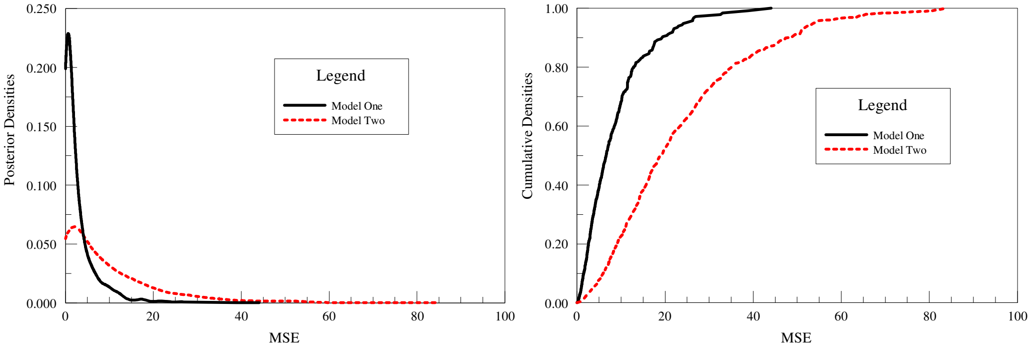


Figure 18: Posterior Density of MSE for Predictive $\tilde{y}_{99}$.

Cumulative Density of MSE for Predictive $\tilde{y}_{99}$

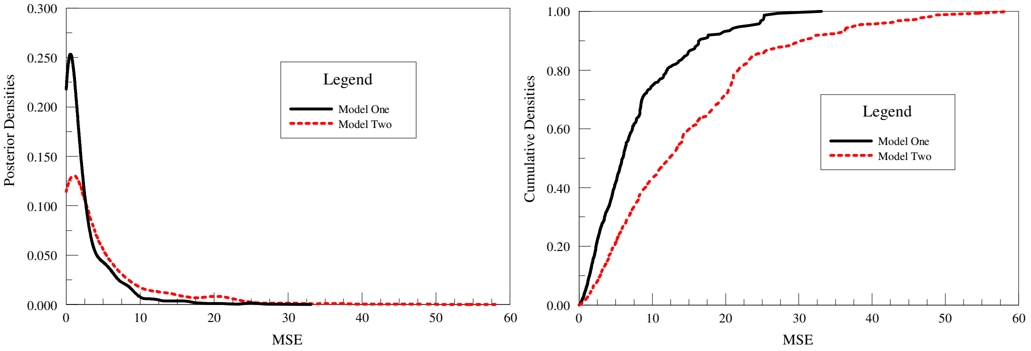


Figure 19: Posterior Density of MSE for Predictive $\tilde{y}_{100}$.

Cumulative Density of MSE for Predictive $\tilde{y}_{100}$

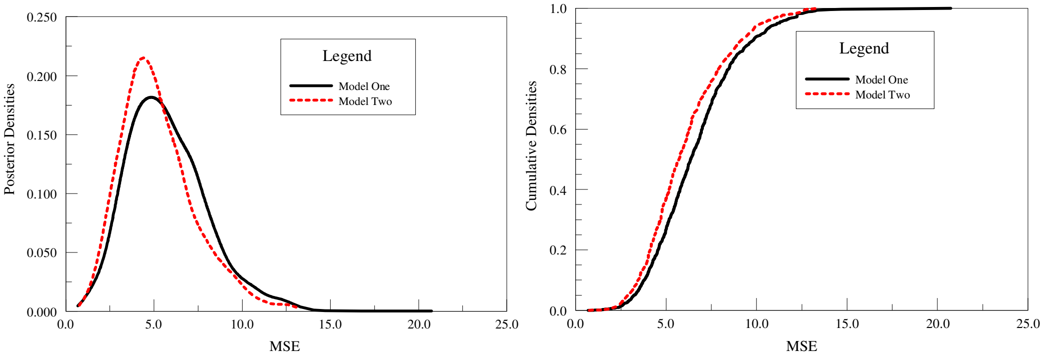


Figure 20: Posterior Density of MSE for Partition 1.

Cumulative Density of MSE for Partition 1

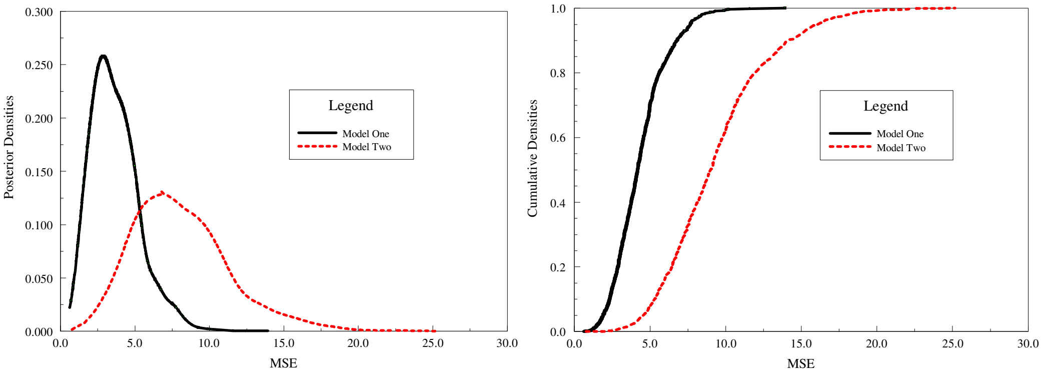
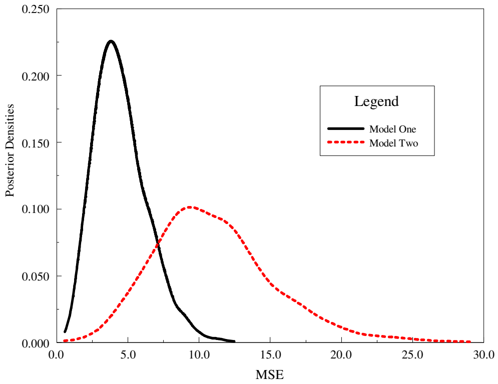


Figure 21: Posterior Density of MSE for Partition 2.

Cumulative Density of MSE for Partition 2



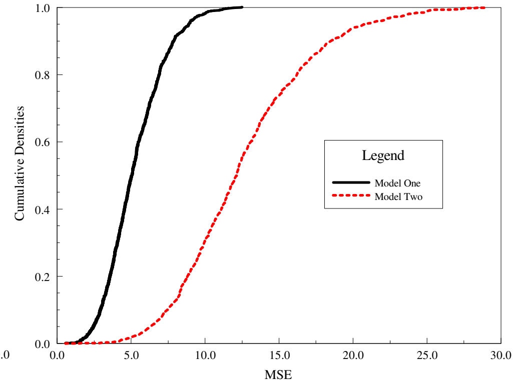
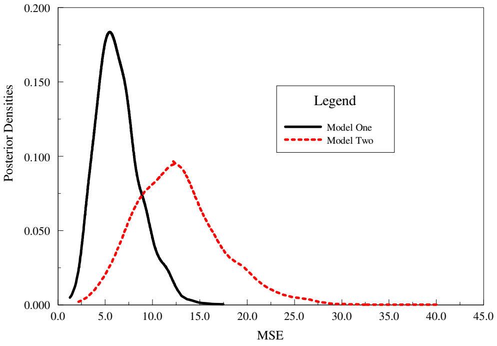


Figure 22: Posterior Density of MSE for Partition 3.

Cumulative Density of MSE for Partition 3



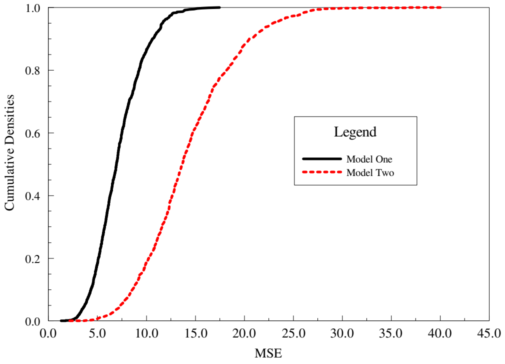
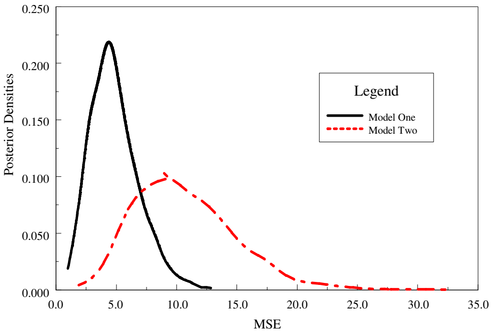


Figure 23: Posterior Density of MSE for Partition 4.

Cumulative Density of MSE for Partition 4



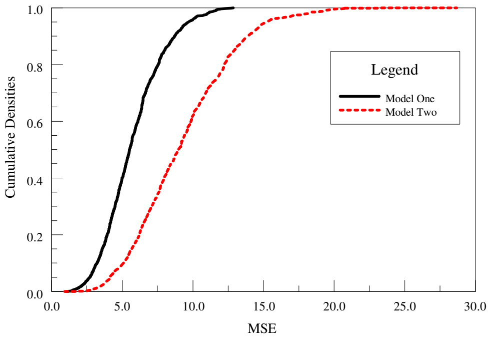
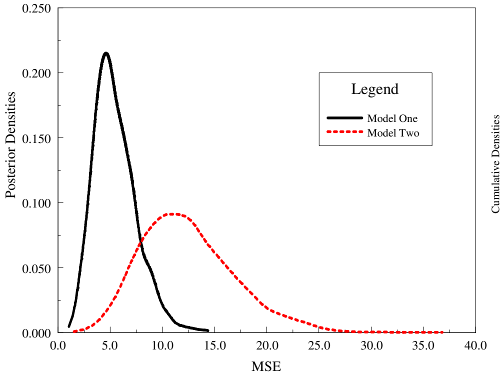


Figure 24: Posterior Density of MSE for Partition 5.

Cumulative Density of MSE for Partition 5



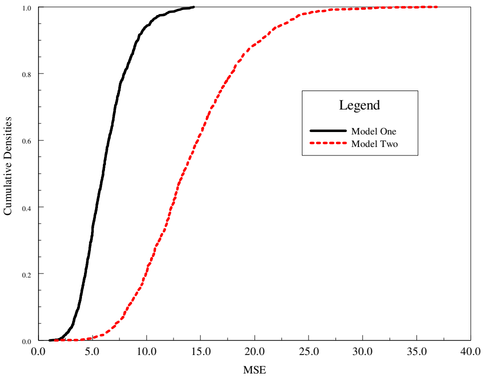
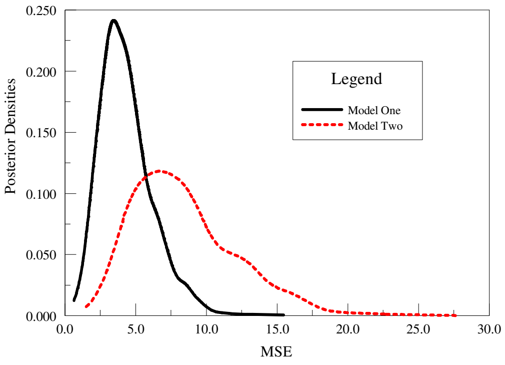


Figure 25: Posterior Density of MSE for Partition 6.

Cumulative Density of MSE for Partition 6



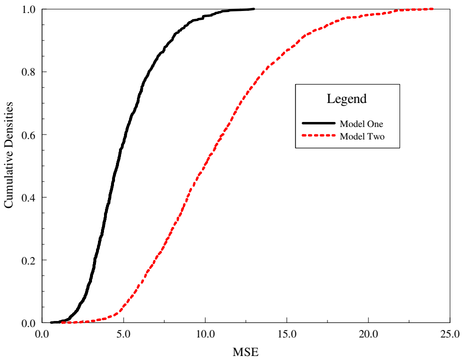
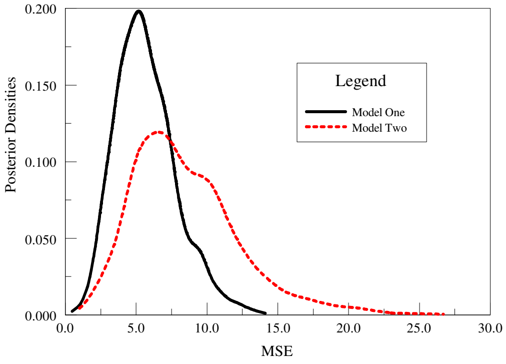


Figure 26: Posterior Density of MSE for Partition 7.

Cumulative Density of MSE for Partition 7



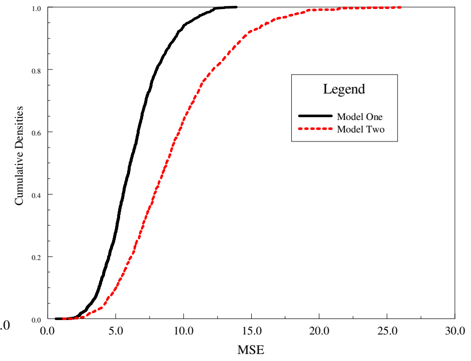
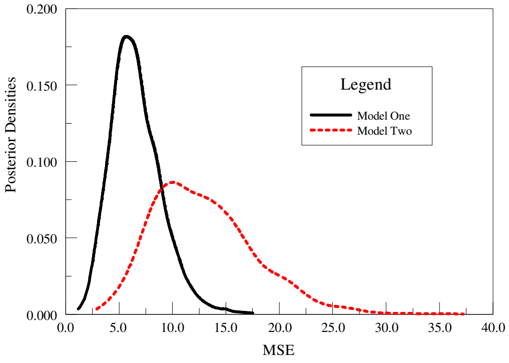


Figure 27: Posterior Density of MSE for Partition 8.

Cumulative Density of MSE for Partition 8



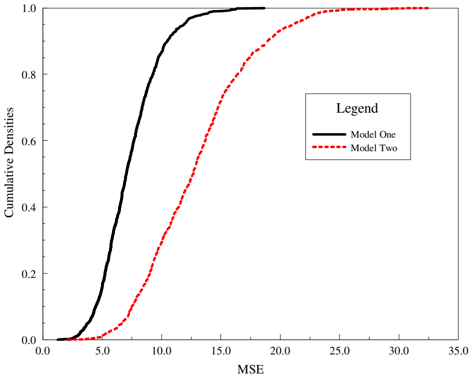
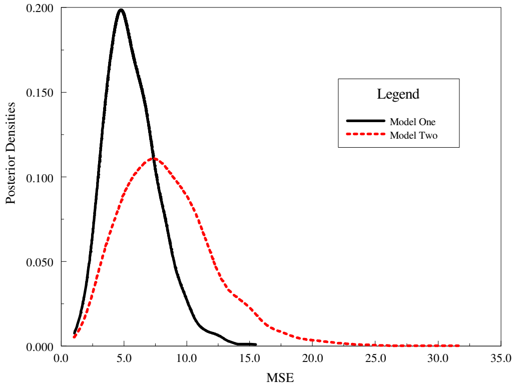


Figure 28: Posterior Density of MSE for Partition 9.

Cumulative Density of MSE for Partition 9



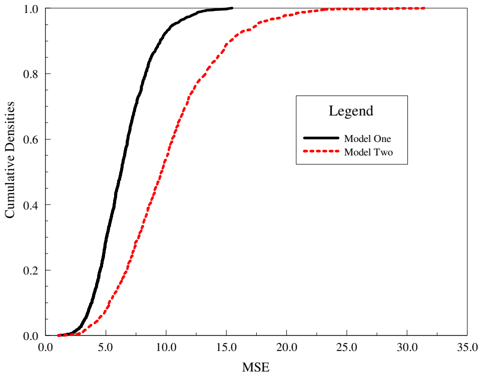


Figure 29: Posterior Density of MSE for Partition 10.

Cumulative Density of MSE for Partition 10

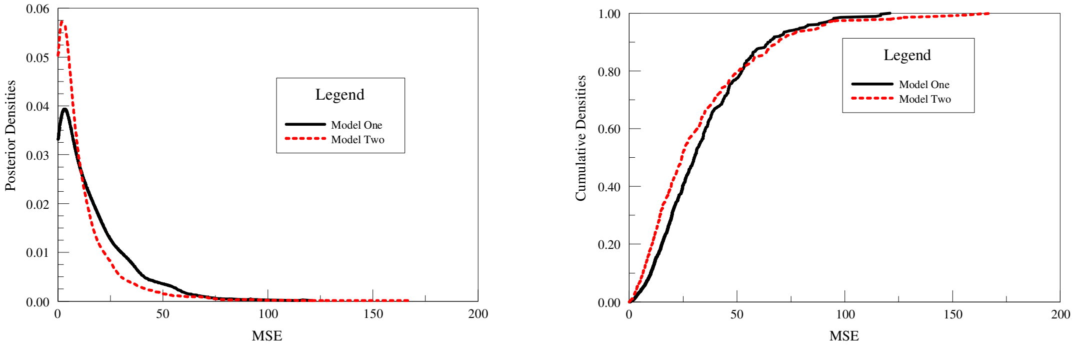


Figure 30: Posterior Density of MSE for Predictive $\tilde{y}_{92}$ by Exluding $\tilde{y}_s$.Cumulative Density of MSE for Predictive $\tilde{y}_{92}$ by Exluding $\tilde{y}_s$

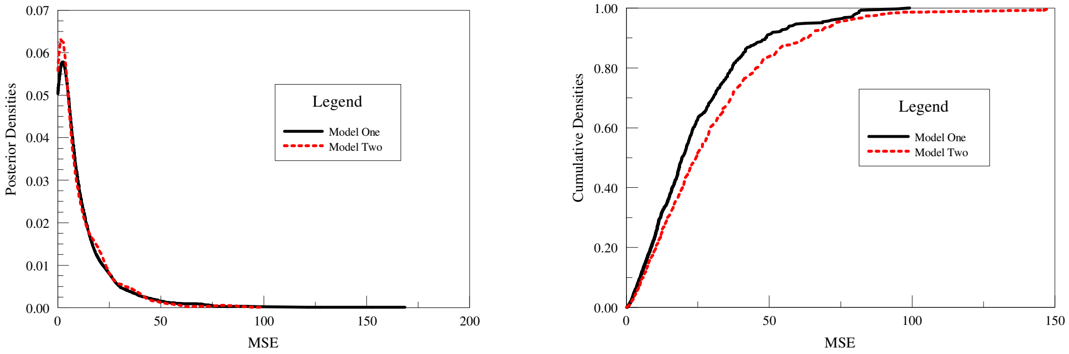


Figure 31: Posterior Density of MSE for Predictive $\tilde{y}_{97}$ by Exluding $\tilde{y}_s$.Cumulative Density of MSE for Predictive $\tilde{y}_{97}$ by Exluding $\tilde{y}_s$

Table 8: MSE of Forecasted $\tilde{y}_s$ (every 10 $\tilde{y}_s$) for Both Models under Fixed Partition

	MSEF of Model One			MSEF of Model Two		
	Mean	Std. Dev	Median	Mean	Std.Dev	Median
y_{91}	6.73	7.10	4.63	7.15	8.84	3.79
y_{92}	12.33	10.26	10.26	5.53	7.26	2.59
y_{93}	11.43	10.24	8.93	26.11	20.69	21.82
y_{94}	2.59	3.79	1.18	11.17	12.64	6.91
y_{95}	6.88	7.78	4.38	7.67	10.22	3.97
y_{96}	2.60	3.75	1.07	4.58	6.66	1.85
y_{97}	7.12	7.52	4.76	4.29	6.26	1.96
y_{98}	2.65	3.75	1.22	12.07	13.26	7.82
y_{99}	2.72	4.08	1.18	9.70	11.47	5.93
y_{100}	2.49	3.61	1.17	4.85	6.92	2.12

Table 9: MSE of Forecasted $\tilde{y}_s$ (every 10 $\tilde{y}_s$) for Both Models under Randomized CV method

	MSEF of Model One			MSEF of Model Two		
	Mean	Std.Dev	Median	Mean	Std.Dev	Median
Partition 1	5.75	2.26	5.49	5.22	2.09	4.85
Partition 2	3.82	1.75	3.57	5.00	2.20	4.72
Partition 3	4.66	1.99	4.35	7.68	3.23	7.32
Partition 4	6.34	2.31	6.11	10.67	3.91	10.18
Partition 5	4.95	2.02	4.71	7.92	3.31	7.48
Partition 6	5.66	2.23	5.25	9.98	3.91	9.54
Partition 7	4.32	1.86	4.95	8.38	3.65	7.83
Partition 8	5.69	2.14	5.43	8.47	3.72	7.85
Partition 9	6.59	2.34	6.31	12.88	4.73	12.35
Partition 10	5.83	2.22	5.58	61.57	51.25	45.24

Table 10: PSBF and log Value of PSBF for Both Methods

	Fiexed		Randomized CV	
	PSBF	log(PSBF)	PSBF	log(PSBF)
Partition 1	6.61	0.08	0.83	−0.01
Partition 2	N/A	N/A	1.20	0.01
Partition 3	N/A	N/A	1.74	0.02
Partition 4	N/A	N/A	0.60	−0.02
Partition 5	N/A	N/A	0.43	−0.04
Partition 6	N/A	N/A	0.36	−0.04
Partition 7	N/A	N/A	0.72	−0.01
Partition 8	N/A	N/A	0.02	−0.16
Partition 9	N/A	N/A	0.58	−0.02
Partition 10	N/A	N/A	42.09	0.16

Table 11: MSE of Cross All Forecasted Ys for Both Methods

	MSEF of Fixed		MSEF of Randomized CV	
	Model One	Model Two	Model One	Model Two
Mean	5.75	9.29	5.36	13.78
Std. Dev	2.73	4.42	2.40	19.35
Median	2.80	3.88	1.92	3.08

Table 12: MSE of Forecasted $\tilde{y}_s$ Without Using $\tilde{y}_s$ to Draw Coefficients

	MSEF of Model One			MSEF of Model Two		
	Mean	Std.Dev	Median	Mean	Std.Dev	Median
y_{91}	9.55	12.48	4.67	12.60	17.58	5.46
y_{92}	15.80	17.27	10.48	10.74	15.42	5.31
y_{93}	15.02	16.17	9.79	31.02	31.09	22.91
y_{94}	5.88	8.14	2.53	16.52	21.27	8.40
y_{95}	10.41	13.22	5.07	13.96	17.99	6.86
y_{96}	5.87	8.34	2.48	9.77	13.98	4.34
y_{97}	10.11	13.00	5.11	10.07	14.23	4.59
y_{98}	5.58	7.60	2.62	17.61	21.99	9.45
y_{99}	6.35	8.96	2.98	16.47	22.88	7.93
y_{100}	5.77	7.98	2.84	9.90	13.01	4.84

4 Logit and Probit Model Selection

4.1 Introduction

Probit and logit models have been used to model binary or polychotomous choice. Observed data of the dependent variable Y takes the values of 0 or 1. In the regression framework, we are interested in estimating the coefficients β of the regressors X. One way to achieve this goal is to use the linear probability model (LPM) . However, there exists some unattractive features for this LPM. First, it produces heteroscedascitiy for ordinary-least squares estimation when Y approaches to 0 or 1. The more serious problem is that the linear specification cannot confine the estimated of Y to the unit interval [0,1]. It also yields constant marginal effects. To overcome all those problems associated with the LPM, one way is to introduce a positive monotone function to transform the linear indicator function $Y^* = X\beta$, where Y^* is unobservable or latent variable and the relationship between Y and Y^* is given by,

$$Y = \begin{cases} 0 \text{ if } Y^* < 0 \\ 1 \text{ if } Y^* \geq 0 \end{cases} \tag{67}$$

thus,

$$P(Y = 1) = P\left(Y^* \geq 0\right) = P(\epsilon \geq -X\beta) = F(X\beta) \tag{68}$$

where $F\left(\cdot\right)$ is the cumulative distribution function (cdf) of ϵ. For probit model, this cdf is normal distribution, that is $\epsilon \sim N(0, 1)$. Therefore,

$$Y = \Phi\left(X\beta\right) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{X\beta} e^{-\frac{1}{2}z^2} dz \tag{69}$$

The logit model is similar to probit model with the cdf as logistic distribution, thus,

$$P(Y=1) = P\left(Y^* \geq 0\right) = P(\epsilon \geq -X\beta) = F(X\beta) = \frac{e^{X\beta}}{1+e^{X\beta}} \qquad (70)$$

In the mathematical sense, the probit and logit model is almost indistinguishable. By taking the inverse of the logistic function defined in equation (70) we have,

$$\text{logit }(F) = \log \frac{F}{1-F} = X\beta \qquad (71)$$

In model selection, the conventional perception is that there is no difference to choose either of them. However, some studies demonstrate that there are similarities as well as differences between the probit and logit model. No universal agreement is reached on which model is superior, probit or logit. The aim of this essay is to compare the probit and logit model in terms of Bayesian model selection criteria and sampling model selection criteria. Compared to other researches, the contribution of our study lie in three aspects. First, we propose the Bayesian model selection criteria based on the predictive probability. A large portion of the past studies concentrate on the comparison of the coefficient estimates. Our study differs from these traditional comparisons by studying the various model selection criteria based on the predictive probability. Then, we use Exponential Power Distribution (EPD) to further check the performances of the probit and logit models. EPD permits a range of kurtosis values, and provides a relatively complete investigation of the distribution of the error term. The third feature is that we devise Metropolis-Hasting (M-H) algorithm with random walk, which convergences to steady state faster than other popular Bayesian estimation method, such as data augmentation. Using generated data, we find the probit and logit model perform almost identically when the data is balanced, i.e, the percentage of $y = 1$ is $40\% - 60\%$. However, they can be discriminated when the data is unbalanced, that is the % of $y = 1$ is around $10\% - 20\%$. Thus, if there is difference between these two models, they must reside in the extreme values at the tails. In addition, the forth moment ,i.e., the Kurtosis of the error distribution also contributes to the differentiation of these two models., which is revealed by the EPD model.

This essay is organized as follows. In Sections 4.2 we review the literature of probit and logit models. Model selection criteria are discussed in section 4.3. In section 4.4, sampling experiment specification, M-H algorithm with random walk and Bayesian estimation procedures are constructed, including the prior setting, maximum likelihood and the posterior distribution for each model. Section 4.5 discusses the results. EPD experiments are analyzed in section 4.6. Applications to real data are discussed in section 4.7. Concluding remarks are made in section 4.8.

4.2 Literature Review

The logit model is introduced and first named by Verhulst (1845) in his study of human population growth in 19th century. The rediscovery and wide acceptance of this model seems to have been motivated by Pearl and Reed (1920, 1922, 1923). It is also used in autocatalystic or chain reaction analysis by Reed and Berkson (1925), Yule (1925) and Wilson (1925). Since then, the logit model has been widely applied in the market demand study of new products and technologies. The probit model was first introduced by Gaddum (1933) and Bliss (1934a, 1934b) in the field of bio-assay. It is then used for modeling discrete binary outcome in economics and market research. At the early days and for quite a long time, the comparison of logit and probit models is concentrated in bio-assay area. Although many researchers demonstrate the resemblance of logistic to normal distribution function (see Wilson (1925), Winsor (1932), etc.), the logit model is regarded inferior to probit model as it lacks the specific underlying process required in bio-assay. However, the logit model has advantage in its computational ease compared to the probit especially in the days of less developed computer technology. This is also the reason that logit model is used frequently in empirical study. In recent 30 years, the logit model has gained much more popularity than before. This is attributed to the direct interpretation of odds, odds ratios and log-odds in the logit model whereas the coefficients from the probit model are difficult to interpret. In addition, the theoretical foundation for logistic regression was built in 1970's by McFadden (see McFadden (2000)). The complete discussion of the origin and the development of probit and logit models can be found in Cramer, J.S. (2003).

The resemblance of these two models is summarized in many studies. Mainly, the researchers prove the similarity of the statistical distribution of the normal and logistic distribution. Amemiya (1981) is one of the first to show the cdf's of the logistic distribution and normal distribution can be made as close as possible. Both normal distribution and logistic distribution are symmetric around zero. Their distribution functions are bounded between 0 and 1. The transformed logistic distribution[2] is given by

$$L_\lambda(w) = \frac{e^{\lambda w}}{1 + e^{\lambda w}} \tag{72}$$

The above function can be made approximate to the normal distribution by choosing an appropriate value of λ. Since the variance of logistic distribution is $\frac{\pi^2}{3}$ of the variance of the standard normal distribution, scaling the logit estimates by $\frac{\pi}{\sqrt{3}}$ results in probit estimates. By trial and error, Amemiya find the approximation can be even closer when setting the scaling parameter to be 1.6. By choosing different w from 0.0 to 3.0, he lists the cdf of normal, $\Phi(w)$ and 1.6 times of the cdf of logistic distribution, $L_{1.6}(w)$. The values of $\Phi(w)$ and $L_{1.6}(w)$ are very close in the mid-range. For example, $\Phi(w) = L_{1.6}(w) = .5$ when $w = 0.0$. $\Phi(w) = L_{1.6}(w) = .5793$ when $w = .2$. However, if $w = 3.0$, $\Phi(w) = .9987$ while $L_{1.6}(w) = .9918$. Thus,

[2] Amemiya uses "transformed logistic distribution" to describe the realtionship between logistic distribution and normal distribution. The logistic distribution can be transformed to approximate normal distribution by setting the proper scaling parameter λ.

$1 - L_{1.6}(w) = .0082$, which is greater than .0013 obtained by $1 - \Phi(w)$.This is due to the relative heavier tails of logistic distribution. However in general, it is hard to differentiate the univariate probit and univariate logit model. Relying on Amemiya's results, Davidson & MacKinnon (1993) state that the probit and logit models are similar. Their scaling parameter is different from Amemiya's. They find if the average of $X\beta$ is around zero, the coefficients of logit model must be roughly 1.6 times of those of probit model in order to get the same prediction for the marginal effect of X. However, if the average value of the P_i is extremely away from .5, this approximation is less effective.

Similar results are also found in Long (1997). He also propose a scaling parameter which makes the cdf's of logistic and normal distribution as close as possible. Long illustrates three distributions: standard normal, standard logistic and standardized logistic distribution[3]. The standard logistic distribution has mean zero and variance of $\frac{\pi^2}{3}$. If scaling the variance of standard logistic distribution to be one, the standard logistic distribution turns to be standardized logistic distribution. The pdf and cdf of the standardized logistic distribution is given by,

$$f(x) = \frac{\exp(\gamma x)}{\gamma \left[1 + \exp(\gamma x)\right]}, \; P(x) = \frac{\exp(\gamma x)}{1 + \exp(\gamma x)} \tag{73}$$

The scaling parameter γ is still equal to $\frac{\pi}{\sqrt{3}}$ in theory. Long shows that the standardized logistic distribution has identical cdf as that of normal distribution. The pdf of standardized logistic distribution also approximates the normal pdf. If there are two regressions:

$$y_L^* = x\beta_L + \epsilon_L, \qquad y_P^* = x\beta_P + \epsilon_P \tag{74}$$

[3] Long uses "standardized logistic distribution" in describing the relationship between logistic distribution and normal distribution. The scaling parameter is chosen to make the standardized logistic distribution with variance equal to one. The crux of the matter is that the regression coefficients of the univariatelogit and probit models are related by this scaling parameter.

where the first model is the logit regression and the latter one is the probit model. y_L^* and y_P^* are latent, thus their variances are undefined. β_L and β_P are unidentified consequently. It can be shown that $Var(\epsilon_L|x) = \frac{\pi^2}{3}Var(\epsilon_P|x)$. Since ϵ_L and ϵ_P are not equal, but approximately equal, $\beta_L \approx \sqrt{Var(\epsilon_L|x)}\beta_P \approx 1.81\beta_P$. Long points out the the parameter of 1.81 is based only on equalizing variance. However, Amemiya's scaling parameter of 1.6 is to approximate the cdf's of probit and logit, not just the variance of logistic and normal distribution. Long gives his own estimation of $\beta_L \approx 1.7\beta_P$. He uses the labor force participation model as an example and obtains the identical likelihood and the statistics for both logit and probit model. Albert and Chib (1993) demonstrate that by setting the scaling parameter to be $\frac{2}{\pi}$, the logistic distribution has approximately identical distribution as the t distribution with degree of freedom of 8. By exploring exponential family, Gill (2001) shows that probit and logit function are theoretically similar and conclude that "In general, with social science data any of these link functions can be used and will provide identical substantive conclusions." Hardin and Hilbe (2001) point out that probit model and logit model yield similar results for binary or grouped binomial data. However, if the underlying process is normal, or the researchers are interested in the prediction or classification instead of odds ratio, then the probit model is preferred to logit model when the Bayesian Information Criteria (BIC) is smaller. Greene (2003) also concludes that there is no much difference in applications for both models by saying that "the logit and probit models results are nearly identical". He uses the study of Spector and Mazzeo (1980) as an example and shows that both model provide comparable estimate coefficients and standard errors. Similar statements are also found in Powers and Xie (2000), Fahrmeir and Tutz (2001). All these studies show that the logistic distribution can be excellent fit of normal distribution; therefore, the logit and probit models are essentially the same. In addition, the logit estimates can be scaled to yield probit estimates.

One of the pioneers in studying the difference of these two models are Chambers and Cox (1967). They analyze the dose-response data by univariate probit and logit model and test the conditions under which these two models can be distinguished. They formulate the experiment by assuming either logit or probit to be the true model and specify this true model as the null hypothesis and the other one as the alternative hypothesis. Let n_i be the observations made at does levels $x_i (i = 1, 2, 3)$, $x_1 < x_2 < x_3$. R_i is the number of deaths at level x_i, $P_i = R_i/n_i$.is the proportion of death and θ_i is the corresponding probability. Asymptotically for fixed θ_i's, all test statistics can be treated as the linear functions of the empirical logistic transforms as $n_i's \to \infty$.,

$$Z_i = \log\left(\frac{P_i}{1-P_i}\right) = \log\left(\frac{R_i}{1-R_i}\right) \tag{75}$$

and they are asymptotically normally distributed with mean

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \frac{x_i - \mu}{\lambda} \tag{76}$$

Then, they obtained the test statistics when logit model is the null hypothesis, which is

$$T_l = \frac{U\sqrt{n}}{A} \tag{77}$$

$$A = \left\{ \frac{(x_2 - x_3)^2}{\rho_1 P_1(1-P_1)} + \frac{(x_3 - x_1)^2}{\rho_2 P_2(1-P_2)} + \frac{(x_1 - x_2)^2}{\rho_3 P_3(1-P_3)} \right\}$$

Under the null hypothesis, equation (77) has standard normal distribution as $n \to \infty$ and holding the levels x_1, x_2, x_3 and ρ_1, ρ_2, ρ_3 fixed. Under alternative hypothesis, T_l is asymptotically normally distributed with unit variance and expectation equal to

$$E(U)\sqrt{n}/E(A) \tag{78}$$

where U is the linear combination of the $Z's$ in equation (75) with asymptotic expectation independent of the nuisance parameters,

$$U = (x_2 - x_3)Z_1 + (x_3 - x_1)Z_2 + (x_1 - x_3)Z_3 \tag{79}$$

Asymptotically

$$nvar(U) = \frac{(x_2 - x_3)^2}{\rho_1\theta_1(1-\theta_1)} + \frac{(x_3 - x_1)^2}{\rho_2\theta_2(1-\theta_2)} + \frac{(x_1 - x_2)^2}{\rho_3\theta_3(1-\theta_3)} \tag{80}$$

where,

$$E(U)/\sigma \sim (t_2 - t_3)\log\left\{\frac{\Phi(t_1)}{1-\Phi(-t_1)}\right\} + (t_3 - t_1)\log\left\{\frac{\Phi(t_2)}{1-\Phi(-t_2)}\right\} + (t_1 - t_2)\log\left\{\frac{\Phi(t_3)}{1-\Phi(-t_3)}\right\} \tag{81}$$

$$E(A)/\sigma \sim \left\{\frac{(t_2 - t_3)^2}{\rho_1\Phi(t_1)\Phi(-t_1)} + \frac{(t_3 - t_1)^2}{\rho_2\Phi(t_2)\Phi(-t_2)} + \frac{(t_1 - t_2)^2}{\rho_3\Phi(t_3)\Phi(-t_3)}\right\} \tag{82}$$

If the null hypothesis is probit and the logit is the alternative, the analog for equation (79) and (80) are obtained as follows,

$$V = (x_2 - x_3)\Phi^{-1}(P_1) + (x_3 - x_1)\Phi^{-1}(P_2) + (x_1 - x_3)\Phi^{-1}(P_3) \tag{83}$$

$$nvar(V) = 2\pi\,(x_2 - x_3)^2\,\theta_1(1-\theta_1)\exp[\{\Phi^{-1}(\theta_1)\}^2]/\rho_1 + \cdots \tag{84}$$

and the test statistics is defined as

$$T_p = \frac{V\sqrt{n}}{B} \tag{85}$$

where B is the estimate of the right-hand side of equation (84) by replacing the θ_i with P_i. Chambers and Cox find that large sample size and extreme independent variable levels are two condition making the probit and logit different. They choose three independent variable values, 1, 2, 3.2, and the extreme independent variable level occurs at $x = 1$ or $x = 3.2$. The second thing is that there is large proportion of the total sample size at that extreme level; and finally the probability of success at that level is extreme, say greater than 99%.

Most papers after Chambers and Cox follow their approach, but refine their model specifications. For example, Hahn and Soyer (2005) follow the approach of Chambers and Cox and test the in-sample model fit and out-of-sample predictive performance of bivariate probit and bivariate logit model. They adopt the bivariate model from Ashford and Sowden (1970):

$$P(Y_{ij} = 1|X_{ij}) = \Phi(Y_{ij}^*), \ j = 1, 2$$

$$P(Y_{i1} = 1, Y_{i2} = 1|X_{ij}) = \Phi_2(Y_{1,}^* Y_{2,}^* \rho) \tag{86}$$

The model selection criteria are Bayes factors proposed by Kass and Raftery (1995) and Deviation Information Criterion (DIC) introduced in Spiegelhalter et al. (2002). They use Monte Carlo study for both small sample size ($n = 90$) and large sample size ($n = 450$). Their results show that under small sample size, for non-extreme independent variable case, the probit model performs slightly better than logit model for both moderate and high dependent correlation in bivariate case. For extreme independent variable cases, the logit model beats probit model. As the dependent variable correlation moves from moderate to high, the model fitting and out-of-sample prediction difference become even more pronounced. When sample size become larger, the difference between the two link functions are increasingly distinctive. The logit model becomes more preferable. Dow and Endersby (2004) compare Multinomial Probit (MNP) and Multinomial logit (MNL) models in voting research. The MNP and MNL differ in the error structure. The MNP assumes the errors are distributed multivariate normal, with mean zero and covariance matrix $\sum$. The probability is calculated by,

$$P\left(Y_i = j|\beta, X_{ij}, \sum\right) = \int_{-\infty}^{\beta^* X_1^*} \cdots \int_{-\infty}^{\beta^* X_{j-1}^*} f(\epsilon_{i1,}^*, \cdots, \epsilon_{ij-1,}^*)\partial\epsilon_{i1,}^*, \cdots, \partial\epsilon_{ij-1} \quad (87)$$

where $f(\cdot)$ is the density function of the multivariate normal distribution. The MNL model is specified as.

$$P(Y_i = j) = \frac{e^{X_{ij}\beta_j}}{\sum_{j=0}^{p} e^{\left(X_{ij}\beta_j\right)}} \quad (88)$$

Dow and Endersby's work is based on the studies by Alvarez and Nagler (2001), Quinne et al. (1999). Alvarez and Nagler use Monte Carlo analysis to compare MNP and the independent MNP (IMNP). The independent MNL is the MNL with the off-diagonal error covariances constrained to zeros. They use IMNP as a substitute for MNL. Dow and Endersby question this proxy and conclude that its inference is limited under the presence of correlated error term. They estimate the parameter values and variances, assess the sensitivity of the probability change to the change of independent variables, and evaluate the accuracy of optimization for both MNL and MNP. They find that MNP and MNL estimator are remarkably similar in their consistency, normality and efficiency for large sample size. However, for small sample size, neither model seems to produce the observed data. MNP is superior to MNL as it incorporates the correlated errors. In voting problem, the correlated error can be interpreted as the dependence of irrelevant alternatives property of voter choice. Although the MNL is criticized by its independent error specification, Dow and Endersby pointed out the independence of irrelevant alternative property on voter choice is irrelevant and unrestricted.

In sum, the probit and logit model are similar. In particular, it's hard to distinguish them for univariate case. Only under some special conditions shown by Chambers and Cox's test, these two models can be distinguished; however, it usually require large observations for each level of independent variable. For multivariate or multinomial cases, these two models differ substantially even for small sample size.

The Bayesian analyses of probit and logit models concentrate on the estimation algorithm and Bayesian inference. It can be found in Ritter and Tanner (1992), McCulloch and Rossi (1994), Chib and Greenberg (1998), Liu and Wu (1999), McCulloch et al. (2000), Webb and Forester (2006), Liu and Daniels (2006), and others.

4.3 Model Selection Criteria

4.3.1 Bayesian Model Selection Criteria

The Bayesian model selection criteria we use are DIC, Predictive DIC (PDIC) and Akaike Information Criteria (AIC). The first measure is the DIC which has been applied in Hahn and Soyer. It is originally given in Spiegelhalter *et. al.* (2003) as follows,

$$\text{DIC} = D(\bar{\theta}) + 2p_D = \bar{D} + p_D \tag{89}$$

where $D(\bar{\theta})$ is the deviance evaluated at the posterior mean, $\bar{D}$ is the posterior mean of the deviance, and $p_D = \bar{D} - D(\bar{\theta})$. The DIC combines measurement for model fit and model complexity. p_D[4] measures the model complexity, such as the number of parameters. The Bayesian deviance, $D(\theta)$, in general is defined as,

[4] In defining DIC, p_D appears in the equation (89) as it stands for the model complexity. In calculating DIC, the equation (89) is algebraically simplified as,
DIC= $D(\bar{\theta}) + 2p_D = D(\bar{\theta}) + 2\left[\bar{D} - D(\bar{\theta})\right] = 2\bar{D} - D(\bar{\theta})$

$$D(\theta) = -2\ln\{(L(y|\theta)\} + 2\ln\{(f(y)\} \tag{90}$$

Following Spiegelhalter *et. al.,* we define $\bar{D}$ and $D(\bar{\theta})$ as follows,

$$\bar{D} = E_{\theta|y}[-2\ln\{L(y|\theta\}] = -2\int \ln L\,(y|\theta)\,p\,(\theta|y)\,d\theta \tag{91}$$

$$D(\bar{\theta}) = -2\ln L\left(y|\bar{\theta}\right) \tag{92}$$

where $L(y|\theta)$ is the likelihood function and $p(\theta|y)$ is the posterior pdf. The constant term $\ln\{(f(y)\}$ is cancelled out when calculating p_D. In MCMC, equation (91) can be computed as

$$-2\left[\frac{1}{m}\sum_{i=1}^{m}\ln L\left(y|\theta^{(i)}\right)\right] \tag{93}$$

where $\theta^{(i)}$ is the i-th MCMC draw of θ. We obtain the DIC for both probit and logit models, namely DIC_{probit} and $\text{DIC}_{\log it}$.

Normally, the model with the smaller DIC is the preferred model. However, like other model selection criteria, the cutoff values or critical values need to be specified. In Hahn and Soyer (2005), they calculate the difference of DIC between the probit and logit model and use 3-7 as the 'significant' difference. For example, if ($\text{DIC}_{\log it}$-DIC_{probit}) is greater than 3, they select probit model as the superior model. Hahn and Soyer take these cutoff values from Burnham and Anderson (1998). Burnham and Anderson discuss the "AIC differences" rather than the "DIC difference". The AIC difference is defined as,

$$\Delta_i = \text{AIC}_i - \min\text{AIC} = E_{\hat{\theta}}[\hat{I}(f, g_i)] - \min E_{\hat{\theta}}[\hat{I}(f, g_i)] \tag{94}$$

over all candidate models in the set. Here, f is the true model and g_i refer to the candidate models. AIC is defined as,

$$AIC = -2log(L(\hat{\theta}|y)) + 2K \tag{95}$$

The best model is the model with minimum AIC. Thus, minAIC is the AIC from the true model. or the best model f. Δ_i measure the relative expected Kullback-Liebler (K-L) difference between f and $g_i(x|\theta)$. The larger the Δ_i is, the less plausible is the candidate model g_i to fit the data. Burnham and Anderson compare four models: Weibull distribution, lognormal distribution, inverse Gaussian and F distribution. They call these models "approximating models", The true model f is the gamma distribution, which is called "generating model" by Burnham and Anderson. It means that the data is generated by this model. They put the rough cutoff value of Δ_i as follows:

1. If $\Delta_i \leq 2$, the candidate model can be considered as a feasible model in making inference.

2. If $4 \leq \Delta_i \leq 7$, then the candidate model is considered less support.

3. If $10 \leq \Delta_i$, the candidate model has virtually not plausible for the data and can be omitted from further consideration. However, they also point out the above guideline really depend on some conditions: the assumption of independent observations, large sample size and model selection situations (nesting or non-nesting). Their cutoff values are for the "simple situation", i.e., independent observation, large sample, nesting models and several candidate models (at least five). If these conditions change, the cutoff points should be revised. For example, if the sample size is smaller, and the parameters for models are large relative to the data size, then Δ_i need to be larger. They emphasize if selecting from only two models with no nesting of one in the other, a simple Δ_i value may not exist from a frequentist sampling viewpoint. To obtain those cutoff point, Burham and Anderson use Monte-Carlo method. Following them, we define the DIC difference and AIC difference as our selection criteria. The

DIC difference is calculated as,

$$\Delta DIC = \text{DIC}_{probit} - \text{DIC}_{\log it}, \text{where DIC} = \bar{D} + p_D \tag{96}$$

The AIC difference is defined as follows,

$$\begin{aligned}\Delta AIC &= \text{AIC}_{probit} - \text{AIC}_{\log it} = \left[-2log(L(\hat{\theta}_{probit}|y)) + 2K\right] - \left[-2log(L(\hat{\theta}_{\log it}|y)) + 2K\right] \\ &= 2log(L(\hat{\theta}_{\log it}|y)) - 2log(L(\hat{\theta}_{probit}|y)) \end{aligned} \tag{97}$$

where $L((\hat{\theta}|y))$ is the likelihood evaluated at Maximum Likelihood estimator (MLE) of θ. K is the number of parameters.

The last model selection criteria is PDIC which is proposed by Ando (2007). Pointed by Robert & Titterington (2002), the same data were used twice in constructing of p_D for calculating DIC. Thus, DIC overfits the observed data. Ando's PDIC is based on the DIC but it corrects the asymptotic bias of the posterior mean of the log likelihood. We follow Ando's approach and define the PDIC by,

$$\text{PDIC} = \tilde{D} + 2\tilde{p}_D \tag{98}$$

where $\tilde{D} = -2\int \ln L\left(\tilde{y}|\theta\right) p\left(\theta|y\right) d\theta, \tilde{p}_D = \tilde{D} - \tilde{D}(\bar{\theta}), \tilde{D}(\bar{\theta}) = -2\ln L\left(\tilde{y}|\bar{\theta}\right)$ and $\tilde{y}$ is the realized post-sample y's. The selection criteria is similar to that of DIC, except the critical values are different. Similarly, we are interested in the PDIC difference, which is

$$\Delta PDIC = \text{PDIC}_{probit} - \text{PDIC}_{\log it} \tag{99}$$

We compute all the Δ's. $\Delta > 0$ indicates the above criteria chooses the logit model, and $\Delta < 0$ indicates that the probit model is preferred to logit model based on these criteria. We also use the Mean Square Error (MSE) of θ_j as the criterion. It is defined as,

$$\mathrm{MSE}_{\theta_j} = E((\hat{\theta}_j^i - \theta)^2) \tag{100}$$

where $\hat{\theta}_j^i$ is the i-th draw of θ_j, θ is the true parameter.

4.3.2 Model Selection Criteria in Sampling Theory

The sampling theory model selection criteria we choose are weighted sum of squared error (SSE), unweighted SSE and Eforn's R^2. The unweighted SSE or the normal SSE is given by,

$$\sum_{i=1}^{n} \left[y_i - \hat{F}(x_i\hat{\theta}) \right]^2 \tag{101}$$

where $\hat{\theta}$ is the MLE of θ. Theoretically, $\hat{\theta}$ can be any estimator. Amemiya (1981) points out this unweighted SSE corresponds to the SSE in standard regression model. However, it does not have the same strong performance as it is in standard regression model because the probit or logit model is heteroscedastic regression model. Thus, the weighted SSE is recommended as a more reasonable criterion. The weighted SSE is given as,

$$\sum_{i=1}^{n} \frac{\left[y_i - \hat{F}(x_i\hat{\theta}) \right]^2}{\hat{F}(x_i\hat{\theta})[1 - \hat{F}(x_i\hat{\theta})]} \tag{102}$$

It is weighted by the estimated probability. Amemiya presents two reasons to choose this criterion over the unweighted one. First, it attaches higher weight to the squared error with larger variance. Second, if use the true probability instead of the estimated probability in the denominator, we can obtain a more efficient estimator of θ by minimizing the above weighted SSE than minimizing the unweighted SSE with respect to θ. I calculate both unweighted SSE and weighted SSE, and the results indeed show they do not always produce the same results. In fact, most of the cases, they yield opposite results. I also use Efron's R^2 in the experiments and it is defined as follows,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left[y_i - \hat{F}(x_i\hat{\theta})\right]^2}{\sum_{i=1}^{n} \left[y_i - \bar{y}\right]^2} \tag{103}$$

This R^2 is an analogue of R^2 in standard regression model.

The R^2 and unweighted SSE are related algebraically as follows. Given the unweighted SSE

$$\mathrm{USSE} = \sum_{i=1}^{n} \left[y_i - \widehat{F}(x_i\hat{\theta})\right]^2 \tag{104}$$

and the R^2 is,

$$\mathrm{R}^2 = 1 - \frac{\sum \left[y_i - \widehat{F}(x_i\hat{\theta})\right]^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{USSE}{\sum (y_i - \bar{y})^2} \tag{105}$$

we obtain

$$\Delta\mathrm{R}^2 = \mathrm{R}^2_{\log it} - \mathrm{R}^2_{probit} = \frac{\mathrm{USSE}_1 - \mathrm{USSE}_2}{\sum (y_i - \bar{y})^2} = -\frac{\Delta\mathrm{USSE}}{\sum (y_i - \bar{y})^2}. \tag{106}$$

Hence, ΔUSSE and $\Delta\mathrm{R}^2$ are the same except for the signs. We also test the other type of criteria focus on the prediction, both Bayesian and sampling theory, like the prediction-realization table and Cross Validation ratio based on the predictive probability[5]. They both turn out to be poor at discriminating probit and logit models.

[5] The prediction-realization table suggested by Franses and Paap (2001) . It calculate the prediction probability and yield a (2×2) table, with
p_{11} = the probability that the predicted $\tilde{y} = 1$ and realized $y = 1$;
p_{10} = the probability that the predicted $\tilde{y} = 1$ and realized $y = 0$;
p_{00} = the probability that the predicted $\tilde{y} = 0$ and realized $y = 0$;
p_{01} = the probability that the predicted $\tilde{y} = 0$ and realized $y = 1$;
where predicted $\tilde{y}$ is calculated by transforming $\tilde{X}\hat{\beta}$ using either normal or logistic functions. For example, if $\Phi(\tilde{X}\hat{\beta}) \geqq 0$, then $\tilde{y} = 1$; otherwise, $\tilde{y} = 0$. The correct prediction is represented by the sum of p_{11} and p_{00}. We would expect the true model yield higher $(p_{11} + p_{00})$. The CV is based on the predictive density. For logit and probit model, it is impossible to get the predictive density as

4.4 Sampling Experiments

4.4.1 Model Specification

In our sampling experiment, We replicate the Bayesian and sample theory model selection criteria that are discussed in the previous section r times. The data are generated first by the logit model and then by the probit model. We call them logit and probit data, respectively. Using the data so generated we obtain the estimates of the logit and probit model and compute the model selection criteria. In this set up one of the models is the true model.

The binary data of $y = 1$ or $y = 0$ is generated by

$$y = \begin{cases} 1.0 & \text{if } \ u \le F(x\beta) \\ 0 & \text{otherwise,} \end{cases} \tag{107}$$

where $F(x\beta)$ is either the logit or probit cumulative density function; u is drawn from the uniform distribution on (0, 1), and x is the covariates. This follows the Bernoulli

the predictive $\tilde{y}$ is either equal to 1 or 0. To construct the counterpart of the integration used in continous case, we use average of probability evaluated at the posterior parameters, that is,

$$\Phi(\tilde{y}_j) = \frac{1}{i}\sum_i \left\{F(\tilde{x}_j\beta_j^{(i)})P(\beta_j^{(i)}|data) \times y_j + \left[1 - F(\tilde{x}_j\beta_j^{(i)})\right] P(\beta_j^{(i)}|data) \times (1-y_j)\right\},\ j = 1,\ldots,n.$$

Here, $F(.)$ is either normal or logistic function; $P(\beta_j^{(i)}|data)$ is the posterior probability of parameters β_j at ith MCMC draw and y_j is the realized jth y. Normally, the CV ratio is obtained from the Kernel density. The procedure is as follows: first draw the kernel density of the predicted $\tilde{y}_r$ based on the MCMC sequence, where r is the rth data point. Then, evaluate the posterior density at the realized y_r. That is to locate the corresponding Kernel density $f(\tilde{y}_r^m)$ where $\tilde{y}_r^m = y_r, m$ is the mth MCMC replication. Next, obtain the posterior probability $P(y_r)$ at the realized y_r by $P(y_r) = \frac{1}{2}\left[f(\tilde{y}_r^m) + f(\tilde{y}_r^{m-1})\right]$. Finally, multiply those posterior probabilities and compare the ratio of the product from model one to that of the model two. If Model one is the true model or better model, then this ratio is expected to be greater than one. We use the predictive probability defined in the above equation as the proxy for kernel density of predictive $\tilde{y}$. Thus the ratio is formed as,

$$\frac{\prod_j P_{\text{probit}}(y_j)}{\prod_j P_{\text{logit}}(y_j)},\ j = 1,\ldots,n$$

where $P_{\text{probit}}(\cdot)$ stands for the posterior density under probit model, and $P_{\text{logit}}(\cdot)$ refer to the posterior density under logit model.

random number generation in Ross (2007).

First we set $x = 1.0$ (and thus $x\beta = \beta$). We call this the constant term model. Then we set $x = (1.0,\ x_2)$ and $x\beta = \beta_1 + \beta_2 x_2$ where x_2 is drawn from the uniform distribution on $(0,\ a)$. We call this two variable model. The constant term model is rarely used in applications of the logit and probit model. However, this model, as compared to the two variable model, demonstrates that the values of the model selection criteria depend on the values of the covariates. The first setting below yield a unbalanced data, with the mean of the percentage of $y = 1$ to be $12.5\% - 24.0\%$.

$$Y^* = X\beta,\ X = 1.0, \beta = -1.2 \tag{108}$$

The second specification obtains the balanced data by change the parameter β to be 1. The mean of the percentage of $y = 1$ is $44\% - 52\%$. It is defined as,

$$Y^* = X\beta,\ X = 1.0, \beta = 1 \tag{109}$$

The two variable model is sometimes used in biometrics. For example, we want to compare the effect of certain drug on the patients. We can easily extend to multiple independent variables based on the specific research goal. The multiple covariate experiment is discussed in later section. Again, there are two settings for this model. The first setting below yields a unbalanced data, with percentage of $y = 1$ to be $9.9\% - 14.9\%$.

$$Y^* = X\beta,\ X = 1 \sim a * U(0,1), \beta = \begin{pmatrix} 1.6 \\ -.2 \end{pmatrix}, a = 70 \tag{110}$$

Changing the parameter a which multiplies $U(0,1)$, the second specification is defined as,

$$Y^* = X\beta,\ X = 1 \sim ad * U(0,1), \beta = \begin{pmatrix} 1.6 \\ -.2 \end{pmatrix}, a = 15 \tag{111}$$

Then, the data become balanced with percentage of $y = 1$ to be $48.6\% - 56.1\%$. Here X is a (1002×2) matrix, with X_1 is constant and X_2 is uniformly distributed. We use the first 1000 data points to do the estimation and predict the last 2 data points.

4.4.2 MCMC Procedure

To draw the parameters β, we choose flat prior and set them equal to one. The log likelihood function for both models is given by,

$$l(\beta, \sigma|y, x) = \sum_i [y_i \times \ln(P_i) + (1 - y_i) \times \ln(1 - P_i)] \tag{112}$$

Then, the posterior density of β is defined as

$$f(\beta|y, x, \sigma) = \ln(prior) + \sum_i [y_i \times \ln(P_i) + (1 - y_i) \times \ln(1 - P_i)] \tag{113}$$

To draw the coefficient β, we use Metropolis-Hastings algorithm with random walk. The proposal density is given by,

$$\begin{pmatrix} \beta_1^{(i)} \\ \beta_2^{(i)} \end{pmatrix} \sim N\left\{ \begin{bmatrix} \beta_1^{(i-1)} \\ \beta_2^{(i-1)} \end{bmatrix},\ c \times \sigma^{2(i-1)}(X'X)^{-1} \right\}, \tag{114}$$

$$\text{where } \sigma^{2(i-1)} = \frac{(y^* - X\beta^{(i-1)})' \times (y^* - X\beta^{(i-1)})}{n}$$

We set

$$\beta^{(i)} = \begin{cases} \beta^{(i)} \text{ with probability } \alpha(\beta^{(i-1)}, \beta^{(i)}) \\ \beta^{(i-1)} \text{ with probability } 1 - \alpha(\beta^{(i-1)}, \beta^{(i)}) \end{cases}$$

where $\alpha = \min\left\{\frac{f(\beta^i)}{f(\beta^{i-1})},\ 1\right\}$, and $f(\cdot)$ is the posterior density defined by equation (113). Thus, β are drawn from normal distribution with previous draw $\beta^{(i-1)}$ as its mean. The initial value of β is the OLS estimate obtained by $\beta = Y/X$. Constant c is the scale parameter. We set $c = 2$ for probit model under both cases. For logit model, c is set to be 1.2 when data is generated by logit model; while $c = .3$ if data generated by probit model. We use the posterior mean of MCMC draws to evaluate the DIC and PDIC. However, the AIC, weighted SSE, unweighted SSE and R^2 are evaluated at MLE of β. The other Bayesian procedure, such as Gibbs sampler with data augmentation is also applied and yields similar coefficient estimates. However, it converges to the steady state slower than random walk does.

4.5 Results

We calculate the mean, median, maximum, minimum and standard deviation of ΔDIC, ΔPDIC, ΔAIC, ΔSSE, ΔUSSE, $\Delta\mathrm{R}^2$. We also present percentage of choosing the correct model for all Δ.The mean of MSE of β_2, mean of MSE of the marginal effect of β_2 is reported for two variable case. We set iteration r is set to be 50 to reduce the computing time. However, our experiments with iteration from 50, to 100, even to 1000 yield similar results.

4.5.1 Constant Term Model

Table 13 presents the results of the sampling experiments for the constant term model for the balanced data (*i.e.* the percentage of $y = 1$ being $44.6 - 51.9\%$). The summary statistics show that the distributions of all the model selection criteria center tightly around zero. Especially the sampling theory criteria of Δ SSE, Δ USSE and Δ AIC range around 10^{-10} to 10^{-6}. Amemiya (1981) and Long (1997) point out, there is no difference between the logit and probit model because the regression coefficients, β, can be adjusted to make the estimated logit and probit cumulative densities almost identical. This argument works well for the constant term model since there is no variability in the regressor. The percentages of choosing the right model, on the other hand, range from 22% to 74%.

We find very similar results for unbalanced data case in table 14. ΔDIC and ΔPDIC are close or equal to zero. ΔAIC, ΔSSE and ΔUSSE range from 10^{-10} to 10^{-6}. The percentages of choosing the right model, on the other hand, range from 30% to 70%. Therefore, no model selection critieria works for constant model no matter the data is balanced or unbalanced.

I also compare the AIC calculated based on MLE as in equation (95) and AIC based on MCMC draws of the parameters β. The latter one is calculated as follows,

$$AIC = \frac{1}{m} AIC \left(\theta^{(i)}\right) \tag{115}$$

Where m is the number of the iterations of MCMC, $\theta^{(i)}$ is the i-th draw of θ. I call this AIC "MCMC treated AIC". The comparison results are shown in tables 15 and 16. The mean of the ΔAIC_{MCMC} in all tables are less than zero. In additional, the percentage of $\Delta\text{AIC}_{MCMC} < 0$ is greater than 80% in table 16. These results indicate that MCMC treated AIC works slightly better for probit data than for logit data. If we compare the ΔAIC_{MCMC} and ΔAIC_{MLE} within each table, we may conclude that MCMC treated AIC works better than AIC based on MLE because all the ΔAIC_{MLE} are virtually zero. This may due to the fact that MLE is a point estimator.

To further verify the DIC, I compare the DIC from MCMC draw and DIC from exact posterior pdf. The difference is around 2 as shown in table 17, which is less than .2% even for the smallest DIC. Thus, the DIC is reliable. The DIC from MCMC is greater than DIC from the exact posterior for both logit and probit data. When probit model is true, the DIC difference for probit model is smaller than for logit model. However, such pattern is not found when logit model is true.

4.5.2 Regression Model with Two Variable

Logit Data In table 18, the true model is logit model. The percentage of $y = 1$ is $11.1\% - 14.9\%$ for unbalanced case and is $48.6\% - 55.2\%$ for balanced case. We find that the distribution of ΔDIC and ΔAIC shift to the right, indicating that as percentage of $y = 1$ becomes small, the ΔDIC and ΔAIC become large. This shows that ΔDIC and ΔAIC work well as the model selection criteria for unbalanced data. The percentage of ΔDIC>0 and the percentage of ΔDIC>0 are 64%, both greater than 50%. When $y = 1$ is $48.6\% - 55.2\%$, the sizes of ΔDIC and ΔAIC are close to zero. Thus, they cannot discriminate probit and logit model under balanced data case. We also find that ΔPDIC are close to zero for both unbalanced and balanced data, which is inconsistent with our expectation. It first appears to us that ΔPDIC does not work as a model selection criteria under either cases. However, we further study the property of PDIC and the discussion is provided in following section. For unbalanced data, the mean of MSE of β_2 for logit model is .0005, which is much smaller than for probit model of .0080. When data change to be balanced, the mean of MSE of β_2 for logit model is still much smaller than for probit model, indicating that mean of MSE of β_2 is a good model selection criterion.

We turn to the sampling theory criteria. For unbalanced data, both the unweighted SSE and the R^2 have certain discriminating ability. The ΔUSSE is greater than zero and the R^2 is less than zero, indicating the true model is the logit model. However, the difference between weighted SSE, ΔSSE, chooses probit model as it is less than zero, which contradict to the data generating process. We address this issue in the following section. In addition, the size of ΔR^2 and ΔUSSE are very small compared to that of ΔDIC and ΔAIC. For balanced data in table 18, all these criteria are centered at zero, showing that none of them works well for discriminating probit and logit model for balanced data.

Probit Data If data generated by probit model, we have results in table 19 much similar to those found for logit data. For unbalanced data where percentage of $y = 1$ is $9.9\% - 13.1\%$, ΔDIC and ΔAIC are both less than zero and shift to the left. The percentage of ΔDIC< 0 is 82%, and the percentage of ΔAIC< 0 is 76%. It indicates that ΔDIC and ΔAIC choose the probit model as the true model. ΔSSE and ΔUSSE are both less than zero, with % of ΔSSE< 0 is 100%, showing that ΔSSE choose the true model, probit model. Mean of MSE of β_2 works well as the model selection criteria for both unbalanced and balanced data, while ΔPDIC still close to zero for both data. The only difference between probit data and logit data is ΔSSE. In table 18, ΔSSE are less that zero, indicating it always choose the wrong model; while in table 19, ΔSSE always choose the true model as they are less than zero. For balanced case, this mean of MSE of probit is .2012, which is still bigger than the logit model of .0011.

More Experiments for Two Variables Regression To further verify our results, I change the parameters β and the parameter multiplying the uniform distribution $U(0, 1)$ to produce another set of data. The results are presented in tables 20 - 21. The data in Table 20 is generated by logit model, with $\beta_1 = .9$, $\beta_2 = -.1, ad = 100$, and case, the percentage of $y = 1$ is $9.8\% - 13.6\%$. For balanced data, $\beta_1 = 1.6$, $\beta_2 = -.1$, $ad = 20$ and % of $y = 1$ is $48\% - 52\%$. For table 21 the parameters are the same as those in table 20 except the data is generated by probit model. The results keep the same as those found in table 18-19. ΔDIC and ΔAIC work well and can be a model selection criteria for unbalanced data. For balanced data, the sizes of these two criteria are close to zero, and do not work as the model selection criteria. In addition, these two criteria are highly correlated. The mean of MSE of β_2 is a good criterion for both unbalanced and balanced case. ΔPDIC does not work compared to ΔDIC and ΔAIC. ΔDIC, ΔPDIC and ΔAIC work better when the true model is probit than when the true model is logit. If it is logit data, SSE never selects the correct model, i.e., logit model. If it is probit data, SSE always select the probit model, which is the correct model. The mean of the $\Delta\mathrm{R}^2$ always choose the correct model for unbalanced data, the percentage of $\Delta\mathrm{R}^2$ also indicate the choice of correct model except for logit data in table 20, where percentage of $\Delta\mathrm{R}^2 > 0$ is 48%, equal to the percentage of ΔUSSE> 0. We also observe the % of ΔSSE> 0 is 0 for logit data and % of ΔSSE< 0 is 100 for probit data. If we increase the out-of-sample period from 2 to 10, % of ΔPDIC> 0 rise from 26% to 56% for logit data under unbalanced case.

The inconsistent PDIC is related to the small number of out-of-sample points. Since PDIC is calculated based on MCMC draws, we calculate the $\hat{F}(x_i\hat{\theta}_{mcmc})$ and $\hat{F}(x_i\hat{\theta}_{mcmc})\times\left[1-\hat{F}(x_i\hat{\theta}_{mcmc})\right]$, it has large variance when using them to calculate cdf. If we change $p = 10$ to make the out-of-sample data increased, the % of ΔPDIC> 0 increased to 50%. for unbalanced data case. We further increase p to be 100, then the probability of choosing the correct model for ΔPDIC is over 50%. Thus, % of ΔPDIC> 0 tends to work by choosing the correct model if we increase the out-of-sample period. This can be verified in another way: we calculate DIC the same way as PDIC except using in-sample data in DIC vs.out-of-sample data in PDIC. In-sample-data is 1000, which is much bigger than the out-of-sample data of 2, so DIC works, but PDIC appears does not work.

The all-or-nothing results for Δ SSE may be explained by the formula for SSE:

$$\mathrm{SSE} = \sum_{i=1}^{n} \frac{\left[y_i - \widehat{F}(x_i\widehat{\beta})\right]^2}{\widehat{F}(x_i\widehat{\beta})[1-\widehat{F}(x_i\widehat{\beta})]}.$$

where $\widehat{\beta}$ is the MLE of β. When % of y=1 is 9–15%, more than 85% of the numerator,

$[y_i - \widehat{F}(x_i\widehat{\beta})]^2$, becomes $[\widehat{F}(x_i\widehat{\beta})]^2$ and the fraction under the summation sign is given by $1/(1/\widehat{F} - 1)$. Since the logit has a fatter tail than the probit, we have $\widehat{F}(\cdot)_{logit} > \widehat{F}(\cdot)_{probit}$, and this leads to

$$\left(\frac{1}{\dfrac{1}{\widehat{F}_{probit}} - 1}\right) < \left(\frac{1}{\dfrac{1}{\widehat{F}_{logit}} - 1}\right)$$

making SSE$_{logit}$ $>$SSE$_{probit}$. The results in Tables 18 and 19 show that even for the balanced data cases when roughly 50% of the time y equals 1, the over-estimation of $\widehat{F}_{logit}$ tends to influence the SSE.

4.5.3 Regression Model with Six Variable

We extend the two variable model to 6 variable (including constant) model and the results are displayed in table 22 and 23. Instead of assuming the distribution of the covariates $x's$, we use the data from the labor participation model in Greene's (2003, P682). The sample size in Greene's example is 753, we increase the sample size to 1000 by randomly resampling from the original data with replacement. The number of out-of-sample observations p is set to be 10 as previously. The latent variable Y^* is formed by $Y^* = x\beta + error$, where the coefficients β are also taken from Greene's estimated parameters. The logit data is obtained by generating the logistic error term, whereas probit data is produced by setting the error to be normally distributed. Then we change the latent variable Y^* to binary by

$$Y = \begin{cases} 0 & \text{if } Y^* \leq a \\ 1 & \text{otherwise} \end{cases} \tag{116}$$

By controlling a, we obtain the unbalanced and balanced data.

Table 22 is the results for logit data and table 23 is the results for probit data.

The results are similar to those in tables 18 -21. For unbalanced data, DIC, AIC and USSE perform well in choosing the correct model while SSE is always choose probit model. For balanced data, generally no model selection criteria works.

4.6 EPD Model

4.6.1 EPD and its Random Sampling Algorithm

In the previous section the experiments are conducted with logit and probit data. Hence, in all the experiments one of the logit and probit models is the true model. Suppose that if data do not come from the logit or probit distributions. Which model, the logit or probit model, will explain the data better? The variances of the logit and probit distributions can be made close to each other, but the kurtosis cannot be made close to each other. The kurtosis of the standard logit distribution is 4.2 while that of the probit is 3. We may expect if data come from a leptokurtic distribution the logit model may explain the data better than the probit model, whereas the probit model may better explain the data if they come from a platykurtic distribution.

To examine our conjecture, we generate data from an exponential power distribution (EPD):

$$f(x) = \exp\left[-\left|x\right|^{\alpha}\right] \bigg/ 2\Gamma\left(1+\frac{1}{\alpha}\right), \quad \text{where } -\infty < x < +\infty, \quad \alpha \geq 1 \tag{117}$$

This family of distribution is symmetric at zero; the variance σ^2 and kurtosis k are given by,

$$\sigma^2 = \Gamma(\frac{3}{\alpha})\Gamma(\frac{1}{\alpha}) \tag{118}$$

$$k = \Gamma(\frac{5}{\alpha})\Gamma(\frac{1}{\alpha}) \bigg/ \left[\Gamma(\frac{3}{\alpha})\right]^2 \tag{119}$$

Changing the shape parameter α yields the different distributions. For example, when $\alpha = 2$, $k = 1$, we have normal distribution; while $\alpha = 1, k = 6$ results in double exponential distribution; and $\alpha = \infty, k = 1.8$ produces uniform distribution. Therefore, the EPD can attain a broad range of kurtosis values and yield the special distributions. The random sampling of EPD adopted here is based on the proposed algorithm ED by Tadikamalla (1980). The algorithm is developed from Von Neumann's (1952) rejection method. The proposal density is double-exponential distribution with scale parameter β. It is defined as,

$$q(x, \beta) = \frac{\exp(-|x|/\beta)}{2\beta}, \ \beta > 0 \tag{120}$$

We draw the random number from the proposal density and use the acceptance-rejection rule to obtain the target random number. The algorithm details as follows,

step 1: set $A = \alpha$, $B = A^{A.}$

step 2: Generate the variate x by following rule: First, generate a uniform random number u. If $u > .5$, set $x = B(-\ln(2(1-u))$; Otherwise, set $x = B\ln(2u)$.

step 3: Obtain another uniform random number s.

step 4: If $\ln s \leq (-|x|^{\alpha} + |x|/B - 1 + A)$,return x as the desired variate. Otherwise, go back to step 2.

In our experiments, we test $\alpha = 1$ and $\alpha = 4$,which represent two different distributions. First, we generate the error term ϵ by above algorithm. The density in figure 32 demonstrate that when $\alpha = 1$,we have leptokurtic distribution. If $\alpha = 2$, the distribution is roughly normal. The distribution become very flatter if we set $\alpha = 4$.

The EPD binary data are generated by drawing the regression error terms from the EPD distribution given x and β:

$$y_i^* = \beta_1 + \beta_2 x_i + \epsilon_i, \quad i = 1, \cdots, n \tag{121}$$

where $x \sim U(0,\, a)$ and the regression coefficients and a will be determined to obtain the unbalanced and balanced binary y's:

$$y_i = \begin{cases} 1 & \text{if } \; y_i^* \leq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{122}$$

We generate data using equations (121) and (122) and call them EPD data. If we set $\alpha = 1$, and the kurtosis, γ_4, of the EPD distribution is

$$\gamma_4 = \frac{\Gamma\left(\frac{5}{\alpha}\right)\Gamma\left(\frac{1}{\alpha}\right)}{\Gamma\left(\frac{3}{\alpha}\right)^2}. \tag{123}$$

Hence, $\gamma_4 = 6.0$, which is leptokurtic.

4.6.2 Results for EPD Model

The sampling experiment is performed by generating 50 of DIC, PDIC, AIC, MSE of β_2, weighted SSE, unweighted SSE and R^2. When $\alpha = 1$ (leptokurtic distribution) and % of $y = 1$ is $15.0\% - 19.4\%$ in table 24, ΔDIC and ΔAIC indicate that logit model is preferred to probit model. The sampling theory criteria, ΔUSSE and ΔR^2 also choose logit model. In addition to compare the MSE of the coefficient, we also calculate the mean of MSE of marginal effect of β_2. In empirical study, marginal effect is a more useful measure than coefficients because the coefficients are difficult to interpret for probit, whereas marginal effect offer an intuitive interpretation. In the simulation, we can calculate the true marginal effect of β_2 because the true value of β_2 is given. The difference between the estimated marginal effect and the true marginal effect is calculated for both models. The MSE of the marginal effect is given by

$$\frac{1}{m}\sum_{i=1}^{m}\left[f(\bar{x}\beta^{(i)})\beta_2^{(i)} - f(\bar{x}\beta)\beta_2\right]^2$$

where m is the MCMC iteration. $f(\bar{x}\beta)\beta_2$ is the true marginal effect, and $f(\bar{x}\beta^{(i)})\beta_2^{(i)}$ is the estimated marginal effect of β_2.We expect the estimated marginal effect for the true model is closer to the true marginal effect than for the wrong model. In this case, we expect a larger MSE of marginal effect for probit model than for logit model. Mean of MSE of the marginal effect of β_2 is .524 for probit model, which is larger than the logit model of .015. However, mean of the weighted SSE and mean of MSE of β_2 indicate probit model is preferred to logit model. For balanced data case in table 24, the percentage of $y = 1$ is $56\% - 62\%$. The distribution of ΔDIC, ΔPDIC and ΔAIC are around zero. The percentage of ΔDIC indicate logit model is preferred to probit model. Mean of MSE of β_2 also choose logit model over probit model. However, mean of MSE of β_2 for probit model is smaller than for logit model. ΔSSE, ΔUSSE and ΔR^2do not work here as they are essentially zero. When $\alpha = 4$ (platykurtic distribution) and % of $y = 1$ is $7.4\% - 11.1\%$ in table 25, all model selection criteria here prefer probit model to logit model. All $\Delta < 0$. Mean of MSE of β_2 is much smaller for probit model than for logit model. Mean of MSE of β_2 for probit model is .003, only one seventh of .021 for logit model. ΔSSE and ΔUSSE are less than zero, and $\Delta\mathrm{R}^2$ is greater than zero. The percentage of $\Delta < 0$ in table 25 are 96%-100%, indicating the probit model is mostly picked by the above criteria. When data become balanced as shown in table 25, mean of MSE of β_2 for probit model is still smaller than for logit model. All the other criteria does not work as the model selection criteria.

4.7 Applications

4.7.1 Stock-Oil Model

In the previous section we found that if we know the kurtosis of a distribution from which binary data come, we may distinguish the logit and probit model if the binary data is unbalanced. In general we do not know the distribution the binary data originates, but sometimes such an information is available. For example, it is known that financial return data are leptokurtic. Often we are interested in learning whether the stock prices go up or down. Also, recently the relationship between the stock price and oil price has attracted attention.

When the price of crude oil decreases, the stock price generally rises (see Sadorsky (1999) and Cinder (2001), Nandha and Fall (2008)). Some connect the oil price shock to the recession of the U.S. economy (Hamilton (1996, 2003) and Gronwald (2008), among others.) Others argue that there is a stable negative relationship between stock and oil prices, but this relationship collapsed after 1998 (Miller and Ratti) (2008)). Is a high crude oil price a sign of future inflation and detriment to the stock return, or is it a proxy for general economic strength? These questions need further study perhaps using more complex models. Here we wish to capture the ups and downs of the stock and oil prices by a simple binary choice model. We apply the model selection criteria we discussed in the previous section. In the past few years both negative and positive relationships have been observed between the stock and oil prices. For example, NYMEX crude oil price story [6] show that between March 2007 and October 2007 both NYMEX crude oil and S&P500 price were up. From November 2007 to March 2008 the stock price tended to go down when the oil price went up. According to some news reports (Simons (2008)) the relationship between the return on crude oil and the return on S&P500 is unstable. Between May 2003 and August 2007 there was a positive relationship due to strong global growth and

[6] The article is accessible on internet: http://www.post1.org/wiki/NYMEX_crude_oil_price_records.

stimulative policies. In the most recent period of August 2007 and August 2008, there is a negative relationship.

Let us focus on the period between January 2007 and March 2008. We use daily Europe Brent Spot Price FOB (Dollars per Barrel) and daily S&P500 index. The Europe Brent Spot Price is slightly different from NYMEX. Both can be obtained from Energy Information Administration and they are produced by two major sources. WIT-Cushing Oklahoma and Brent in Europe. Cushing is a price settlement point for West Texas Intermediate on the New York Mercantile Exchange (NYMEX) and has been quoted as the most significant trading hub for crude oil in North America. However, it has lost the leading price indicator status since April 13, 2007 because a huge stockpile at the facility has caused prices to be artificially low. Accordingly, we use Brent in Europe data.

The logarithms of S&P500 and crude oil price are shown in Figure 33. There are presences of negative relationship in some subperiods. From January 2007 to October 2007, the crude oil price rose from below \$60 per barrel to nearly \$80. S&P500 also soared to an all-time high of 1565.15. From November 2007 to March 2008 the crude oil price continue surging while S&P500 plunged a 19-months low on March 17. Since then, both oil and stock prices have been rising.

We are also interested in the kurtosis of the returns on S&P500. In our simulation we found that the logit model better explains the data than the probit model when the distribution from which the binary data originate is leptokurtic. Using MCMC algorithms the kurtosis, γ_4, of the S&P500 return data is estimated in the regression model

$$Y_{SP\&500,t} = \beta_1 + \beta_2 \Delta\, X_{oil,t-1} + \epsilon_t \tag{124}$$

where ϵ_t follows the EPD distribution, $Y^*_{SP500,t} = \Delta\, y_t = \ln\, y_t - \ln\, y_{t-1}$; y_t is the daily S&P index; $\Delta\, X_{oil,t-1}$ is the change in log of the daily crude oil price. The detailed

MCMC algorithm to draw the coefficients (α, β, σ) is listed in Appendix V.A.

The posterior mean and standard deviation of γ_4 are 5.284 and 0.771, respectively. The posterior probability density of shape parameter α and Kurtosis γ_4 are presented in Figures 34 and 35. It is slightly skewed to the right, and it clearly lies in the range greater than 3. The stylized fact that financial return data are leptokurtic holds true for the S&P500 returns.

We change $Y^*_{SP500,t}$ to the binary data by

$$Y_t = \begin{cases} 0 & \text{if } Y^*_{SP500,t} \leq a \\ 1 & \text{otherwise} \end{cases} \tag{125}$$

To obtain balanced data we set $a = 0$ and we set $a = 0.01$ to obtain unbalanced data. The percentages of $Y = 1$ are 53% for the balanced data and 13.6% for the unbalanced data.

The posterior means and standard deviations of β_1 and β_2 for the logit and probit models are given in Tables 26 and 27 for the unbalanced data and for the balanced data, respectively. The model selection criteria are also presented.

The negative β_2 shown in Tables 26 and 27 indicates that the stock price and crude oil price have a negative relationship in the period of 2007 to 2008. The estimates of β_2 indicate that the relationship between the stock return and oil return is not strong. Among the model selection criteria Δ DIC seems to be the reliable measure since all other criteria are virtually zero. When the data is balanced Δ DIC is 0.248 but when the data is unbalanced Δ DIC is 0.517 indicating that the logit model is preferred over the probit model. Since the stock return data is leptokurtic we expect that the logit model performs better than the probit model when data are unbalanced.

4.7.2 Labor Participation Model

We also apply the labor force participation model presented in Greene's (2003). The model is specified as,

$$Y = F(cons\tan t, age, age^2, income, eduction, kids) \tag{126}$$

where Y is a dummy of women work in 1975. We scale the family income by 10000. Greene fits a probit model to above regression. Here, we apply both probit and logit model as we want to compare them. The results from our MCMC estimates and the probit results from Greene's are presented in table 28. Our coefficient estimates for probit model and the significance level are similar to those found in Greene's probit model estimates. For these data, $Prob(Y = 1)$ is 0.57, indicating a balanced data. From our sampling experiments in section 3, we expect none of the model selection criteria works. The results in table 28 show that ΔDIC, ΔPDIC, ΔAIC,ΔSSE and ΔUSSE are all close to zero, and the sign are mixed. ΔDIC is positive, ΔPDIC, ΔAIC, and ΔSSE are slightly negative, ΔUSSE is equal to zero. Thus, there is virtually no different for logit and probit model. The estimates from logit model looks radically different, however, if we scale the coefficients by 1.6-1.8, it produces the coefficients of probit (except age^2), within the range of the scaling parameters given by Amemiya (1981) and Long (1997). In addition, .we calculate the marginal effect of both probit and logit models given by

$$\frac{1}{m}\sum_{i=1}^{m} f(\bar{x}\beta^{(i)})\beta^{(i)} \tag{127}$$

where m is the MCMC iteration, $\beta^{(i)}$ is the i-th MCMC draw of β.Comparison of the marginal effects and the marginal effects from Greene's estimates in table 29 shows that the results from these two models are nearly identical, and they are the same as those from Greene's estimates.

4.8 Conclusion Remarks

There are several conclusions we can draw. If there is only the constant term in the regression model, all the model selection criteria considered in this essay are distributed tightly around zero. This is especially true for the sampling theory criteria of Δ AIC, Δ SSE, and Δ USSE. With two variables in the regression we find that Δ DIC and Δ AIC tend to choose the correct model. Δ SSE always chooses the probit model when the data are unbalanced. If data are balanced no criterion works.

We have shown that if we have knowledge on the kurtosis of the distribution from which binary data are generated, we can discriminate the logit and probit model if the data are unbalanced.

The critical values for $\Delta > 0$ vary with the regression model and with covariates in the regression model. There are some attempt to find critical values of the model selection criteria that yield 90% or 95% acceptance or rejection of a model. For example Burnham and Anderson (1998) come up with such critical values for the AIC. Hahn and Soyer (2005) follow Burnham and Anderson to choose critical values.

We changed the sample size from 1,000 to 200. The results show that the model selection criteria do not work well even for unbalanced data. This may be due to the fact that we need a large sample to have enough observations in the tails. Arguing that we have knowledge on the kurtosis of financial return data, we showed that the ups and downs of S&P500 returns are better explained by the logit model if the data are unbalanced.

Appendix IV.A: MCMC procedure to draw (α, β, σ)

We choose flat prior and set them equal to one. The log likelihood function for EPD models is given by,

$$
\begin{aligned}
l(\alpha, \beta, \sigma|y, x) &= -(n+1) \times \ln(\sigma) - n \times \ln gamma(\frac{1}{\alpha}) && (128)\\
& -n \times \frac{(\alpha+1)}{\alpha} \times \ln(2) - \frac{.5}{\sigma^{\alpha}} \times \sum_i |y_i - x_i\beta|^{\alpha} && (129)
\end{aligned}
$$

We draw the coefficients in blocks. To draw the coefficient α,we use Modified Efficient Jump. The proposal density inverted Gaussian given by,

$$f(x) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right], x > 0, \mu > 0, \lambda > 0$$

We use Devroye (1986)'s random number generation algorithm as follows,

1. Generate a $N(0,1)$.
2. Set $Y = N^2$.
3. Set $X_1 = \mu + \frac{\mu^2 Y}{2\lambda} - \frac{\mu}{2\lambda}\sqrt{4\mu\lambda Y + \mu^2 Y^2}$
4. Generate a uniform random variate, $U(0,1)$.If $U \leq \frac{\mu}{\mu + X_1}$,set $X = X_1$,else set $X = \frac{\mu^2}{X_1}$.

We obtain λ and μ from Simpson's rule.

We set

$$A^{(i)} = \begin{cases} \alpha^{(i)} \text{ with probability } A(\alpha^{(i-1)}, \alpha^{(i)}) \\ \alpha^{(i-1)} \text{ with probability } 1 - A(\alpha^{(i-1)}, \alpha^{(i)}) \end{cases}$$

where $A = \min\left\{\frac{p(\alpha^i, \beta^{i-1}, \sigma^{i-1}) q(\alpha^i, \beta^{i-1}, \sigma^{i-1})}{p(\alpha^{i-1}, \beta^{i-1}, \sigma^{i-1}) q(\alpha^{i-1}, \beta^{i-1}, \sigma^{i-1})},\ 1\right\}$, and $p(\cdot)$ is the posterior density defined by equation (128). $q(\cdot)$ is the proposal density for α.

To draw the coefficient β, we use Metropolis-Hastings algorithm. The proposal density is given by,

$$\begin{pmatrix} \beta_1^{(i)} \\ \beta_2^{(i)} \end{pmatrix} \sim N\left\{\left[(X'X)^{-1}X'Y\right],\ \sigma^{2(i-1)}(X'X)^{-1}\right\}$$

We set

$$\beta^{(i)} = \begin{cases} \beta^{(i)} \text{ with probability } A(\beta^{(i-1)}, \beta^{(i)}) \\ \beta^{(i-1)} \text{ with probability } 1 - A(\beta^{(i-1)}, \beta^{(i)}) \end{cases}$$

where $A = \min\left\{\frac{p(\alpha^i,\beta^i,\sigma^{i-1})q(\alpha^i,\beta^i,\sigma^{i-1})}{p(\alpha^i,\beta^{i-1},\sigma^{i-1})q(\alpha^i,\beta^{i-1},\sigma^{i-1})},\ 1\right\}$, and $p(\cdot)$ is the posterior density defined by equation (128). $q(\cdot)$ is the proposal density for β.Thus, β are drawn from normal distribution with OLS estimator as its mean.

The σ is draw from inverted gamma and is accepted or rejected by,

$$\sigma^{(i)} = \begin{cases} \sigma^{(i)} \text{ with probability } A(\sigma^{(i-1)}, \sigma^{(i)}) \\ \sigma^{(i-1)} \text{ with probability } 1 - A(\sigma^{(i-1)}, \sigma^{(i)}) \end{cases}$$

where $A = \min\left\{\frac{p(\alpha^i,\beta^i,\sigma^i)q(\alpha^i,\beta^i,\sigma^i)}{p(\alpha^i,\beta^i,\sigma^{i-1})q(\alpha^i,\beta^i,\sigma^{i-1})},\ 1\right\}$, and $p(\cdot)$ is the posterior density defined by equation (128). $q(\cdot)$ is the proposal density for σ.

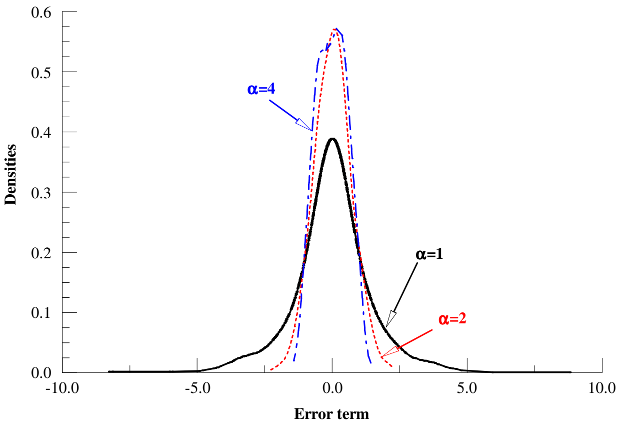


Figure 32. Kernel Densities of Error Term from EPD

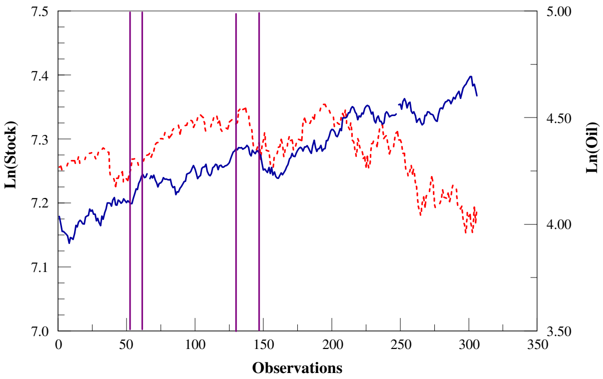


Figure 33. Stock Market Price and Crude Oil Price (Jan. 2007-Mar. 2008)

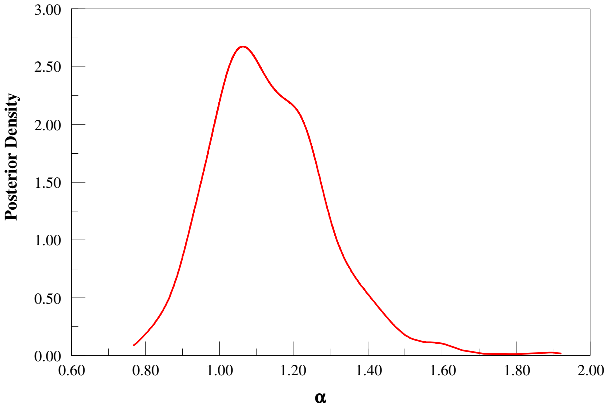


Figure 34. Posterior Density of α: Change in Ln(S&P500)

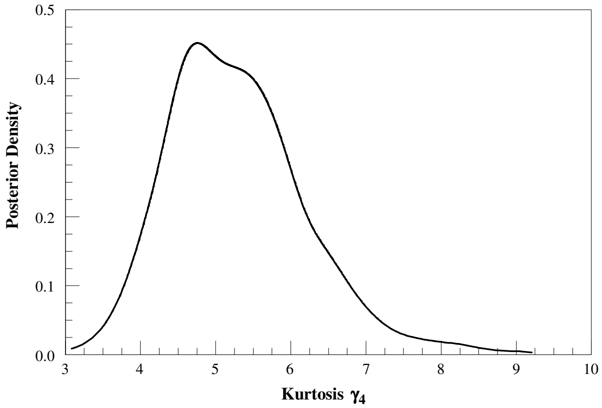


Figure 35. Posterior Density of Kurtosis: Change in Ln(S&P500)

Table 13: Constant Term Regression Model: Balanced Data

	Logistic Model is True			
	$\beta_1 = -0.1, \%$ of $y = 1$ is $44.6 - 51.9\%$			
	mean	max	min	std
% of $y = 1$	48.60	51.90	44.60	1.397
Δ DIC	.016	.320	$-.332$	.158
Δ PDIC	.001	.021	$-.021$	.009
Δ AIC	-1.85×10^{-8}	3.14×10^{-7}	-3.563×10^{-7}	7.18×10^{-8}
Δ SSE	1.193×10^{-7}	2.391×10^{-6}	-2.429×10^{-6}	6.454×10^{-7}
Δ USSE	-4.6×10^{-9}	7.8×10^{-9}	-8.88×10^{-8}	1.79×10^{-8}
% of choosing the correct (Logit) model				
Δ DIC	54%			
Δ PDIC	66%			
Δ AIC	74%			
Δ SSE	72%			
Δ USSE	74%			
	Probit Model is True			
	$\beta_1 = -0.1, \%$ of $y = 1$ is $44.2 - 51.5\%$			
	mean	max	min	std
% of $y = 1$	47.77	51.50	44.20	1.529
Δ DIC	.100	.898	$-.984$	.408
Δ PDIC	.001	.055	$-.052$	.017
AIC	-3.63×10^{-8}	4×10^{-10}	0×10^{-10}	1×10^{-10}
Δ SSE	1.336×10^{-7}	2.387×10^{-6}	-2.996×10^{-6}	9.468×10^{-7}
Δ USSE	-9×10^{-9}	6.4×10^{-9}	8.77×10^{-8}	2.38×10^{-8}
% of choosing the correct (Probit) model				
Δ DIC	40%			
Δ PDIC	42%			
Δ AIC	26%			
Δ SSE	22%			
Δ USSE	26%			

Notes: Δ DIC = DIC of probit − DIC of logit
Δ PDIC = PDIC of probit − PDIC of logit (ten period ahead prediction)
Δ AIC = AIC of probit − AIC of logit
Δ SSE = SSE of probit − SSE of logit
Δ USSE = USSE of probit − USSE of logit
$\Delta > 0$ indicates the choice of the logit model; otherwise choose Probit.
$r = 50$ (r is the number of replications.)

Table 14: Constant Term Regression Model: Unbalanced Data

	Logistic Model is True			
	$\beta_1 = -1.2$,% of $y = 1$ is $20.5 - 27.9$%			
	mean	max	min	std
% of $y = 1$	24.026	27.900	20.500	1.531
Δ DIC	.020	.269	$-.250$	.126
Δ PDIC	$-.001$	.013	$-.009$	.005
Δ AIC	5.70×10^{-8}	1.833×10^{-7}	-7.17×10^{-8}	6.66×10^{-8}
Δ SSE	-5.189×10^{-6}	1.232×10^{-5}	-1.513×10^{-6}	7.574×10^{-6}
Δ USSE	1.060×10^{-8}	3.30×10^{-8}	-1.17×10^{-8}	1.12×10^{-8}
% of choosing the correct (Logit) model				
Δ DIC 56%				
Δ PDIC 44%				
Δ AIC 66%				
Δ SSE 34%				
Δ USSE 66%				

	Probit Model is True			
	$\beta_1 = -1.2$,% of $y = 1$ is $10.4 - 15.3$%			
	mean	max	min	std
% of $y = 1$	12.55	15.30	10.40	.983
Δ DIC	.038	.498	$-.748$	.295
Δ PDIC	.000	.028	$-.010$	.007
AIC	-7.5×10^{-9}	1.551×10^{-7}	-7.45×10^{-8}	3.06×10^{-8}
Δ SSE	1.01×10^{-6}	8.952×10^{-6}	-3.223×10^{-5}	5.436×10^{-6}
Δ USSE	-8×10^{-10}	1.45×10^{-8}	-7.3×10^{-9}	2.9×10^{-9}
% of choosing the correct (Probit) model				
Δ DIC 42%				
Δ PDIC 60%				
Δ AIC 70%				
Δ SSE 30%				
Δ USSE 70%				

Notes: Δ DIC = DIC of probit − DIC of logit
Δ PDIC = PDIC of probit − PDIC of logit (ten period ahead prediction)
Δ AIC = AIC of probit − AIC of logit
Δ SSE = SSE of probit − SSE of logit
Δ USSE = USSE of probit − USSE of logit
Δ > 0 indicates the choice of the logit model; otherwise choose Probit.
$r = 50$ (r is the number of replications.)

Table 15: Comparison of AIC_{MLE} and $\text{AIC}_{\text{MLCMC}}$ for Logit Data

	% of $y = 1$ is 20.6 – 27.9%			% of $y = 1$ is 44.6 – 51.9.%		
	$\beta_1 = -1.2$			$\beta_1 = -.1$		
	mean	median	std	mean	median	std
$y = 1$	24.03%	24.05%	1.53	48.60%	48.75%	1.40
$\Delta\,\text{AIC}_{MLE}$ $\Delta\,\text{AIC}_{MCMC}$	5.7×10^{-8} $-.003$	5.0×10^{-8} $-.001$	.000 .004	-1.9×10^{-8} $-.012$	6×10^{-10} $-.008$	7.2×10^{-8} .011
% of $\{\Delta\,\text{AIC}_{MLE} > 0\}$ = 64% % of $\{\Delta\,\text{AIC}_{MCMC} > 0\}$ = 32%				% of $\{\Delta\,\text{AIC}_{MLE} > 0\}$ = 74% % of $\{\Delta\,\text{AIC}_{MCMC} > 0\}$ = 8%		

Notes: $\Delta\,\text{AIC}$ = AIC of probit − AIC of Logit
$\Delta > 0$ indicates the choice of the logit model.

Table 16: Comparison of AIC_{MLE} and $\text{AIC}_{\text{MLCMC}}$ for Probit Data

	% of $y = 1$ is 10.4 – 15.3%			% of $y = 1$ is 44.2 – 51.5.%		
	$\beta_1 = -1.2$			$\beta_1 = -.1$		
	mean	median	std	mean	median	std
$y = 1$	12.55%	12.60%	.98	47.77%	47.75%	1.53
$\Delta\,\text{AIC}_{MLE}$ $\Delta\,\text{AIC}_{MCMC}$	7.5×10^{-9} $-.011$	1.9×10^{-9} $-.003$	3.1×10^{-8} .016	-3.6×10^{-8} $-.041$	8.0×10^{-10} $-.020$	1×10^{-10} .054
% of $\{\Delta\,\text{AIC}_{MLE} < 0\}$ = 70% % of $\{\Delta\,\text{AIC}_{MCMC} < 0\}$ = 84%				% of $\{\Delta\,\text{AIC}_{MLE} < 0\}$ = 26% % of $\{\Delta\,\text{AIC}_{MCMC} < 0\}$ = 88%		

Notes: $\Delta\,\text{AIC}$ = AIC of probit − AIC of Logit
$\Delta < 0$ indicates the choice of the probit model.

Table 17: Difference of DIC by MCMC and DIC by Exact Posterior Pdf

	Logit data		Probit data	
	$\beta = -2.0$	$\beta = -.1$	$\beta = -2.0$	$\beta = -.1$
$\Delta DIC1$	2.0028	2.0588	1.9802	2.0639
$\Delta DIC2$	1.9883	2.2822	2.4717	2.1841
$\bar{y}$	.006	.485	.027	.47

Notes: $\Delta DIC1 = DIC_{MCMC} - DIC_{EXACT}$ for probit model
$\Delta DIC2 = DIC_{MCMC} - DIC_{EXACT}$ for logit mode

Table 18: Two variables Regression Model: Logit Data

	% of $y = 1$ is 11.1–14.9%				% of $y = 1$ is 48.6–55.2%			
	$\beta_1 = 1.6,\ \beta_2 = -.2\ ad = 70$				$\beta_1 = 1.6,\ \beta_2 = -.2\ ad = 15$			
	mean	max	min	std	mean	max	min	std
% of $y = 1$	12.36	14.90	11.10	.72	52.47	55.20	48.60	1.478
Δ DIC	1.756	21.225	-3.337	4.018	$-.033$	.531	-1.052	.335
Δ PDIC	$-.008$	.024	$-.018$	.006	.000	.031	$-.031$	.013
Δ AIC	1.790	16.608	-3.460	3.579	.000	.633	$-.796$	.279
Δ SSE	$-.0367$	$-.0212$	$-.0616$	.0103	$-.0006$	.0014	$-.0020$	.0007
Δ USSE	.1408	1.4270	$-.2964$	.3269	.0053	.1376	$-.1281$	.0517
Δ R^2	$-.0013$	.0030	$-.0125$	.0030	.0000	.0005	$-.0006$	.0002
% of choosing the correct (logit) model					% of choosing the correct (logit) model			
Δ DIC 64% Δ PDIC 2% Δ AIC 64%					Δ DIC 46% Δ PDIC 50% Δ AIC 56%			
Δ SSE 0% Δ USSE 60%					Δ SSE 18% Δ USSE 54%			
Mean of MSE of β_2								
Probit: .0080 Logit: .0005					Probit: .0061 Logit: .0003			

Notes: Δ DIC = DIC of probit $-$ DIC of logit
Δ PDIC = PDIC of probit $-$ PDIC of logit
Δ AIC = AIC of probit $-$ AIC of logit
Δ SSE = SSE of probit $-$ SSE of logit
Δ USSE = USSE of probit $-$ USSE of logit
$\Delta > 0$ indicates the choice of the logit model.

Table 19: Two Variables Regression Model: Probit Data

	% of $y = 1$ is 9.9–13.1%				% of $y = 1$ is 49.8–56.1.%			
	$\beta_1 = 1.6,\ \beta_2 = -.2\ ad = 70$				$\beta_1 = 1.6,\ \beta_2 = -.2\ ad = 15$			
	mean	max	min	std	mean	max	min	std
% of $y = 1$	11.40	13.10	9.90	.703	53.43	56.10	49.80	1.378
Δ DIC	$-.990$	3.147	-3.588	1.269	$-.208$	1.584	-1.922	.822
Δ PDIC	$-.002$	$-.001$	$-.004$	.001	$-.002$	.039	$-.086$	.035
Δ AIC	$-.826$	3.158	-3.195	1.229	$-.184$	1.656	-1.690	.825
Δ SSE	$-.0269$	$-.0122$	$-.0467$	.0071	$-.0053$	.0000	$-.0107$	.0021
Δ USSE	$-.0146$	.2201	$-.2960$	.1204	$-.0210$	.2701	$-.2406$	.1225
Δ R^2	.0001	.0029	$-..0022$	.0012	.0001	.0010	$-.0011$	.0005
% of choosing correct (probit) model					% of choosing correct (probit) model			
Δ DIC	82%				Δ DIC	60%		
Δ PDIC	100%				Δ PDIC	42%		
Δ AIC	76%				Δ AIC	58%		
Δ SSE	100%				Δ SSE	100%		
Δ USSE	50%				Δ USSE	64%		
Mean of MSE of β_2								
Probit: .0007	Logit: .0311				Probit: .0061	Logit: .0003		

Notes: Δ DIC = DIC of probit − DIC of logit
Δ PDIC = PDIC of probit − PDIC of logit
Δ AIC = AIC of probit − AIC of logit
Δ SSE = SSE of probit − SSE of logit
Δ USSE = USSE of probit − USSE of logit
$\Delta < 0$ indicates the choice of the probit model.

Table 20: Two variables Regression Model: Logit Data (II)

	% of $y = 1$ is 9.8–13.6%				% of $y = 1$ is 43.7-51.5%			
	$\beta_1 = .9,\ \beta_2 = -.1\ ad = 100$				$\beta_1 = .9,\ \beta_2 = -.1\ ad = 20$			
	mean	max	min	std	mean	max	min	std
% of $y = 1$	12.12	13.60	9.80	.782	48.19	51.50	43.70	1.64
Δ DIC	1.187	6.918	-3.956	2.732	$-.053$	.370	$-.518$	.214
Δ PDIC	.006	.404	$-.101$	.065	.001	.027	$-.027$	.008
Δ AIC	1.244	7.207	-3.979	2.809	$-.019$	.222	$-.291$	.112
Δ SSE	$-.0302$	$-.0210$	$-.0393$	.0046	.0000	.0006	$-.0005$	.0002
Δ USSE	.1581	.9557	$-.3038$	.2866	$-.0028$	.0532	$-.0544$	.0240
Δ R^2	$-.0015$	.0034	$-.0099$	.0028	.0000	.0002	$-.0002$	.0001
% of choosing correct (logit) model					% of choosing correct (logit) model			
	Δ DIC 62% Δ PDIC 26% Δ AIC 62%				Δ DIC 34% Δ PDIC 64% Δ AIC 44%			
	Δ SSE 0% Δ USSE 70%				Δ SSE 40% Δ USSE 48%			
Mean of MSE of β_2								
Probit: .0021 Logit: .0001					Probit: .0015 Logit: .0001			

Notes: Δ DIC = DIC of probit − DIC of logit
Δ PDIC = PDIC of probit − PDIC of logit
Δ AIC = AIC of probit − AIC of logit
Δ SSE = SSE of probit − SSE of logit
Δ USSE = USSE of probit − USSE of logit
$\Delta > 0$ indicates the choice of the logit model.

Table 21: Two Variables Regression Model: Probit Data (II)

	% of $y = 1$ is 9.8–11.5%				% of $y = 1$ is 42.7–50.3.%			
	$\beta_1 = .9,\ \beta_2 = -.1,\ ad = 100$				$\beta_1 = .9,\ \beta_2 = -.1,\ ad = 20$			
	mean	max	min	std	mean	max	min	std
% of $y = 1$	11.40	11.50	9.85	8.60	47.45	50.30	42.70	1.734
Δ DIC	−.990	−1.488	1.077	−3.688	−.148	.681	−1.356	.427
Δ PDIC	−.002	−.005	−.002	−.011	−.002	.020	−.049	.013
Δ AIC	−.826	−1.530	1.095	−3.769	−.082	.698	−1.094	.387
Δ SSE	−.0269	−.0284	−.0193	−.0423	−.0007	.0016	−.0033	.0010
Δ USSE	−.0146	−.0579	.3244	−.2800	−.0097	.1447	−.1671	.0701
Δ R^2	.0001	.0006	.0032	−..0034	.0000	.0007	−.0006	.0003
% of choosing correct (probit) model					% of choosing correct (probit) model			
Δ DIC 86%					Δ DIC 64%			
Δ PDIC 100%					Δ PDIC 54%			
Δ AIC 82%					Δ AIC 54%			
Δ SSE 100%					Δ SSE 80%			
Δ USSE 64%					Δ USSE 56%			
Mean of MSE of β_2								
Probit: .0001 Logit: .0066					Probit: .0001 Logit: .0042			

Notes: Δ DIC = DIC of probit − DIC of logit
Δ PDIC = PDIC of probit − PDIC of logit
Δ AIC = AIC of probit − AIC of probit
Δ SSE = SSE of probit − SSE of logit
Δ USSE = USSE of probit − USSE of logit
$\Delta < 0$ indicates the choice of the probit model.

Table 22: Six Variables Regression Model: Logit Data

	% of $y = 1$ is 14.8-15.2%				% of $y = 1$ is 50.0-50.3%			
	mean	max	min	std	mean	max	min	std
% of $y = 1$	15.00	15.16	14.81	0.043	50.00	50.30	49.90	.055
Δ DIC	2.294	11.129	−3.109	2.934	.577	3.842	−2.457	1.266
Δ PDIC	0.018	0.350	−0.147	0.090	0.002	0.075	−0.072	0.034
Δ AIC	1.915	10.645	−3.236	2.840	0.055	3.586	−2.940	1.305
Δ SSE	−0.007	−0.003	−0.009	0.001	−0.001	0.001	−0.003	0.001
Δ USSE	0.318	1.353	−0.265	0.358	0.032	0.660	−0.366	0.180
% of choosing the correct (logit) model					% of choosing the correct (logit) model			
	Δ DIC	80%			Δ DIC	72%		
	Δ PDIC	52%			Δ PDIC	48%		
	Δ AIC	80%			Δ AIC	54%		
	Δ SSE	0%			Δ SSE	2%		
	Δ USSE	82%			Δ USSE	54%		

Notes: Δ DIC = DIC of probit − DIC of logit
Δ PDIC = PDIC of probit − PDIC of logit (ten period ahead prediction)
Δ AIC = AIC of probit − AIC of probit
Δ SSE = SSE of probit − SSE of logit
Δ USSE = USSE of probit − USSE of logit
$\Delta > 0$ indicates the choice of the logit model.
$r = 50$ (r is the number of replications.)

Table 23: Six Variables Regression Model: Probit Data

	% of $y = 1$ is 14.9-15.1%				% of $y = 1$ is 49.8-50.1%			
	mean	max	min	std	mean	max	min	std
% of $y = 1$	15.00	15.06	14.85	0.036	50.00	50.07	49.80	0.045
Δ DIC	-0.687	3.750	-7.692	1.974	0.373	1.187	-1.043	0.556
Δ PDIC	-0.003	0.666	-0.109	0.108	-0.001	0.038	-0.058	0.020
Δ AIC	$-.1.027$	3.251	-8.070	1.934	-0.068	0.717	-1.306	0.444
Δ SSE	-0.003	-0.002	-0.006	0.001	0.000	0.000	-0.001	0.000
Δ USSE	-0.108	0.401	-0.902	0.256	. $-$ 0.005	0.130	-0.181	0.076
% of choosing the correct (probit) model					% of choosing the correct (probit) model			
Δ DIC 56%					Δ DIC 32%			
Δ PDIC 58%					Δ PDIC 64%			
Δ AIC 66%					Δ AIC 46%			
Δ SSE 100%					Δ SSE 94%			
Δ USSE 62%					Δ USSE 50%			

Notes: Δ DIC = DIC of probit $-$ DIC of logit
Δ PDIC = PDIC of probit $-$ PDIC of logit (ten period ahead prediction)
Δ AIC = AIC of probit $-$ AIC of probit
Δ SSE = SSE of probit $-$ SSE of logit
Δ USSE = USSE of probit $-$ USSE of logit
$\Delta < 0$ indicates the choice of the probit model.
$r = 50$ (r is the number of replications.)

Table 24: EPD Model with $\alpha = 1$

	% of $y = 1$ is 15.0–19.4%					% of $y = 1$ is 56.0–62.8%				
	$\beta_1 = .1,\ \beta_2 = -6,\ ad = .5$					$\beta_1 = .1,\ \beta_2 = .4,\ ad = .5$				
	mean	median	max	min	std	mean	median	max	min	std
% of $y = 1$	16.90	16.80	19.40	15.00	.924	58.98	58.75	62.80	56.00	1.594
Δ DIC	1.923	1.790	5.406	-1.035	1.428	.194	.199	.511	$-.036$	.119
Δ PDIC	.008	.033	.051	$-.135$	.058	.000	.000	.027	$-.016$	.008
Δ AIC	1.867	1.800	5.357	-1.126	1.446	$-.002$	.000	.217	$-.031$	.008
Δ SSE	$-.0075$	$-.0076$	$-.0013$	$-.0128$	.0026	.0000	.0000	.0001	.0000	.0000
Δ USSE	.3134	.3195	.8886	$-.2683$	.2287	$-.0004$	.0000	.0041	$-.0067$	.0018
$\Delta\,\mathrm{R}^2$	$-.0022$	$-.0022$	.0020	$-.0064$	.0016	.0000	.0000	.0000	.0000	.0000
	% of $\{\Delta\,\mathrm{DIC} > 0\} = 80\%$ % of $\{\Delta\,\mathrm{PDIC} > 0\} = 80\%$ % of $\{\Delta\,\mathrm{AIC} > 0\} = 88\%$					% of $\{\Delta\,\mathrm{DIC} > 0\} = 54\%$ % of $\{\Delta\,\mathrm{PDIC} > 0\} = 52\%$ % of $\{\Delta\,\mathrm{AIC} > 0\} = 44\%$				
	% of $\{\Delta\,\mathrm{SSE} > 0\} = 0\%$ % of $\{\Delta\,\mathrm{USSE} > 0\} = 90\%$					% of $\{\Delta\,\mathrm{SSE} > 0\} = 56\%$ % of $\{\Delta\,\mathrm{USSE} > 0\} = 44\%$				
Mean of MSE of β_2										
Probit: 3.2681 Logit: 4.2840						Probit: .0652 Logit: .1918				
Mean of MSE of Marginal Effect of β_2										
Probit: .524 Logit: .015						Probit: .054 Logit: .010				

Notes: Δ DIC = DIC of probit $-$ DIC of logit
Δ PDIC = PDIC of probit $-$ PDIC of logit
Δ AIC = AIC of probit $-$ AIC of probit
Δ SSE = SSE of probit $-$ SSE of logit
Δ USSE = USSE of probit $-$ USSE of logit
$\Delta > 0$ indicates the choice of the logit model.

Table 25: EPD Model with $\alpha = 4$

	% of $y = 1$ is 7.4–11.0%					% of $y = 1$ is 57.8–64.2%				
	$\beta_1 = .1,\ \beta_2 = -6,\ ad = .5$					$\beta_1 = .1,\ \beta_2 = .4,\ ad = .5$				
	mean	median	max	min	std	mean	median	max	min	std
% of $y = 1$	9.38	9.40	11.00	7.40	.768	61.11	61.00	64.20	57.80	1.312
Δ DIC	−3.419	−3.517	.327	−5.083	1.196	.142	.136	.440	−.187	.156
Δ PDIC	−.013	−.012	−.008	−.017	.002	−.001	.000	.032	−.020	.010
Δ AIC	−3.517	−3.652	−.350	−5.241	1.166	.003	.000	.082	−.026	.015
Δ SSE	−.0227	−.0221	−.0162	−.0350	.0039	.0000	.0000	.0000	−.0002	.0000
Δ USSE	−.2682	−.2853	.0931	−.5898	.1413	.0007	.0000	.0190	−.0060	.0035
Δ R^2	.0031	.0033	.0071	−..0011	.0016	.0000	.0000	.0000	−.0001	.0000
	% of $\{\Delta\,\text{DIC} < 0\} = 98\%$					% of $\{\Delta\,\text{DIC} < 0\} = 24\%$				
	% of $\{\Delta\,\text{PDIC} < 0\} = 100\%$					% of $\{\Delta\,\text{PDIC} < 0\} = 48\%$				
	% of $\{\Delta\,\text{AIC} < 0\} = 100\%$					% of $\{\Delta\,\text{AIC} < 0\} = 48\%$				
	% of $\{\Delta\,\text{SSE} < 0\} = 100\%$					% of $\{\Delta\,\text{SSE} < 0\} = 60\%$				
	% of $\{\Delta\,\text{USSE} < 0\} = 96\%$					% of $\{\Delta\,\text{USSE} < 0\} = 46\%$				
Mean of MSE of β_2										
Probit: 29.4254 Logit: 221.9853						Probit: .0780 Logit: .3219				
Mean of MSE of Marginal Effect of β_2										
Probit: .003 Logit: .021						Probit: .030 Logit: .010				

Notes: Δ DIC = DIC of probit − DIC of logit
Δ PDIC = PDIC of probit − PDIC of logit
Δ AIC = AIC of probit − AIC of probit
Δ SSE = SSE of probit − SSE of logit
Δ USSE = USSE of probit − USSE of logit
$\Delta\ <\ 0$ indicates the choice of the probit model.

Table 26: Posterior Summary Statsistics and Model Selection Criteria: Stock-Oil Data Unbalanced Data

		mean	std
logit	β_1	−1.880	0.163
	β_2	−0.004	0.017
probit	β_1	−1.106	0.094
	β_2	−0.002	0.010
Δ DIC 0.517			
Δ PDIC −0.015			
Δ AIC 0.010			
Δ SSE 0.000			
Δ USSE 0.001			

Table 27: Posterior Summary Statsistics and Model Selection Criteria: Stock-Oil model Balanced Data

		mean	std
logit	β_1	0.190	0.171
	β_2	−0.106	0.232
probit	β_1	0.114	0.111
	β_2	−0.064	0.150
Δ DIC 0.248			
Δ PDIC −0.002			
Δ AIC 0.000			
Δ SSE 0.000			
Δ USSE 0.000			

Table 28: Posterior Summary Statsistics and Model Selection Criteria: Labor Participation Model

		const	age	age^2	Income	Edu.	kids
logit	mean	−6.401	0.292	−0.004	0.066	0.154	−0.772
	std	2.387	0.111	0.001	0.069	0.037	0.204
probit	mean	−4.105	0.187	−0.002	0.037	0.094	−0.448
	std	1.431	0.067	0.001	0.042	0.023	0.138
Greene's	mean	−4.157	0.185	−0.002	0.046	0.098	−0.450
	std	1.402	0.066	0.001	0.042	0.023	0.130
Δ DIC 0.449							
Δ PDIC −0.047							
Δ AIC −0.296							
Δ SSE 0.001							
Δ USSE −0.027							

Table 29: Marginal Effects: Labor Participation Model

	const	age	Income	Edu.	kids
logit	−1.555	0.071	0.016	0.037	−0.175
probit	−1.603	0.073	0.015	0.037	−0.175
Greene's	−1.620	0.072	0.018	0.038	−0.175

[1] Ait-Sahalia, Y (1996). Testing continuous time models of the spot interest rate. Review of Financial Studies 9, 385-426.

[2] Albert, J.H. and S. Chib.(1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. Journal of Business and Economic Statistics 11, 1-15

[3] Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data., Journal of the American Statistical Association 88, 669-679

[4] Alqallaf F., and P. Gustafson (2001). On cross-validation of Bayesian models. The Canadian Journal of Statistics, Vol. 29, No.2, 333-340

[5] Alvarez, R.M and J. Nagler (2001). Correlated disturbances in discrete choice models: a comparison of multinomial probit and logit models. Political Science 42, 55-96

[6] Amemiya, T. (1981). Qualitative response models A survey. Journal of Economic Literature, 19, 1483-1536.

[7] Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. Biometrika 2007 94(2):443-458

[8] Andersen, T.G. , T. Bollerslev, F.X. Diebold and P.Labys (2003). Modeling and forecasting realized volatility. Econometrica 71, 579-625

[9] Ang, A., and G. Bekaert, G. (2002a). International asset allocation with regime shifts. Review of Financial Studies, 15, 4, 1137-187.

[10] Ashfold, J.R. and R.R. Sowden (1970), Multi-variate probit and analysis. Biometrics 26, 535-546

[11] Bekaert G., R.J Hodrick and D. Marshall (2001), Peso problem explanations for term structure anomalies. Journal of Monetary Economics, 48, 2,(October), 241-70.

[12] Bekara, M., and Fleury, G. (2003). Model Selection Using Cross Validation Bayesian Predictive Densities. Conference paper, WISP 2003, Budapest, Hungary

[13] Berger, J.O. and L.R Pericchi.(1992). The intrinsic bayes factor. technical report. Department of Statistics, Purdue University, West Lafayette.

[14] Bliss, C. I. (1934a). The method of probits. Science 79, 38-39

[15] Bliss, C. I. (1934b). The method of probits. Science 79, 400-401.

[16] Box, G.E.P. and G.C. Tiao (1973). Bayesian Inference in Statistical Analysis. P157-158, Addison-Wesley

[17] Burnham, K.P. and D.R. Anderson (1998). Model Selection and Inference: A Practical Information-Theoretic Approach. New York,: Springer.

[18] Cai, J. (1994). A Markov model of switching-regime ARCH. Journal of Business & Economic Statistics, 12(3), 309-316

[19] Carlin, N.G., B.P. Polson, and D.S Stoffer (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. Journal of the American Statistical Association, 87

[20] Chakrabarti, A., and Ghosh, J. K. (2006). Some Aspects of Bayesian Model Selection for Prediction. Conference paper, ISBA 8th World Meeting on Bayesian Statistics.

[21] Chambers, E.A. and D.R. Cox (1967). Discrimination between alternative binary response models. Biometrika 54, 573-578.

[22] Chan, K.S.(1986). On estimating thresholds in autoregressive models. Journal of Times Series Anal. 7, 179-190.

[23] Chen, C.W.S and J. C. Lee (1995). Bayesian Inference of Threshold Autoregressive Models. Journal of Time Series Analysis 16, 383-392

[24] Chib, S. and E. Greengerg (1994). Bayes Inference of Threshold Autoregressive models. Journal of Econometrics 64, 183-206.

[25] Chib, S. and E. Greenberg (1998). Analysis of multivariate probit models. Biometrika 85, 2, 347-361.

[26] Ciner, C. (2001). Energy Shocks and Financial Markets: Nonlinear Linkages. Studies in Non-Linear Dynamics and Econometrics, 5, 203-12

[27] Cramer, J.S. (2003). *Logit Models From Economics and Other Fields.* Chapter 9, Cambridge University Press

[28] Davidson, R. and J.G. MacKinnon (1993). *Estimation and Inference in Econometrics.* P515, New York: Oxford.

[29] Derker, M.J. (1997), Markov switching in GARCH processes and mean-reverting stock-market volatility. Journal of Business & Economic Statistics, 15,1, 26-34

[30] Devroye, L. (1986), *Non-Uniform Random Variate Generation*, Springer-Verlag, New York.

[31] Dow, J.K, and J.W. Endersby (2004). Multinomial probit and multinomial logit: a comparison of choice models for voting research, Electoral studies 23, 107-122.

[32] Durland J. and T. MAcCurdy, T. (1994). Duration dependent transitions in a Markov model of U.S. GNP growth. Journal of Business and Economic Statistics, 12, 279-288

[33] Engel C. and J. Hamilton (1990). Long swings in the dollar: are they in the data and do markets know It? The American Economic Review, 80, 4, 689-713

[34] Efron, B (1983)., Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. Journal of the American Statistical Association, 78: 316-331

[35] Evans, M.D. and k.K. Lewis (1994). Do expected shifts in inflation affect estimates of the long-run Fisher relation. Journal of Finance, 50, 1, 225-53

[36] Evans, M.D. and P. Wachtel (1993). Were price changes during the Great Depression anticipated? Evidence from nominal interest rates. Journal of Monetary Economics, 32(1), 3-34

[37] Fahrmeir, L. and G. Tutz. (2001). *Multivariate Statistical Modeling Based on Generalized Linear Models* (2nd ed.). New York: Springer

[38] Forbes, C.S. G. RJ. Karlb and P. Kofman.(1999). Bayesian arbitrage threshold analysis. Journal of Business & Economic Statistics, American Statistical Association, 17(3), 364-72.

[39] Francq, C. and J.M. Zakoian (2001). Stationary of multivariate Markov-switching ARMA modes. Journal of Econometrics, (102), 339-64

[40] Franses P.H. and R. Paap (2001). *Quantitative Models in Marketing Research.* Cambridge University Press.

[41] Garcia, R., and P. Perron (1996). An analysis of the real interest rate under regime shifts. Review of Economics and Statistics 78, 111-25.

[42] Gaddum, J. H. (1933), Reports on biological standard III. methods of biological assay depending on a auantal response. London: Medical Research Council. Special Report Series of the Medical Research Council, no. 183.

[43] Geisser, S. (1975). The predictive sample reuse method with Application, J. Am. Statist. Ass., 70: 320-328

[44] Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. J. Am. Statist. Ass.,74: 153-160.

[45] Gelfand, A.E. (1996). Model determination using sampling based methods. In Markov chain Monte Carlo in practice. (W.R.Gilks, S.Richardson and D.J. Spiegelhalter, Eds). London: Chapman and Hall, 145-162

[46] Gelfand, A. E. and D.K. Dey (1994). Bayesian model choice: asymptotes and exact calculations. Journal of the Royal Statistical Society B, 56(3): 501–514.

[47] Gelfand, A.E., D.K. Dey and H. Chang (1992). Model determination using predictive distributions with implementation via sampling-based methods. Bayesian Statistics 4 (eds J. Bernardo, J.O. Berger, A.P. Dawid and A.F. M. Smith), 147-167, Oxford: University Press

[48] Geisser, S. (1988). The future of statistics in retrospect. Bayesian Statistics 3 (eds J. Bernardo, M.H. DeGroot, D.V. Lindley and A.F. M. Smith), 147-158. Oxford: Oxford University Press.

[49] Geweke, J. and N. Terui (1993). Bayesian threshold autoregressive models for nonlinear time series. Journal of Time Series Analysis, 14(5), 441-454.

[50] Gill, J. (2001). *Generalized Linear Models: A Unified Approach.* 30-34, Thousand Oaks, CA: Sage

[51] Goldman, E. and T. Agheyegbe (2005). Estimation of threshold time series models using efficient jump MCMC. Working paper, Pace University

[52] Goldman, E. and H. Tsurumi (2005). Bayesian analysis of a doubly-truncated regression model with ARMA-GARCH error. Studies in Nonlinear Dynamics & Econometrics*,* forthcoming

[53] Gospodinov N., (2005), Testing for threshold nonlinearity in short-term interest rates". Journal of Financial Econometrics *3*, 344-71

[54] Greene, W.H. (2003). Econometric Analysis (5rd ed.). P675, Upper Saddle River, NJ: Prentice-Hall.

[55] Gronwald, M. (2008). Large oil shocks and the US economy: Infrequent incidents with large effects. Energy Journal, 29, 151-71

[56] Hass, M., S. Mittnik and M.S. Paolella (2004). "A new approach to Markov switching GARCH models. *Journal of* Financial Econometrics, 2(4), 493-530

[57] Hahn, E.D. and R. Soyer (2005). Probit and logit models: differences in the multivariate realm. Working paper, http://home.gwu.edu/~soyer/mv1h.pdf

[58] Hamilton, J.D. (1988). Rational expectation econometric analysis of changes in regime: An investigation of the term structure of interest rates. Journal of Economic Dynamics and Control, June/September, 12, 385-423

[59] Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycles. Econometrica, 57(2), 357-84

[60] Hamilton, J.D. and R.Susmel (1994). Autoregressive conditional heteroskedasticity and changes in regime. Journal of Econometrics, 64(1-2), 307-333

[61] Hamilton, J.D. (1996). This is what happened to the Oil price-macroeconomy relationship. Journal of Monetary Economics, 38, 215-20.

[62] Hamilton, J. D. (2003). What is an oil shock?. Journal of Econometrics, 113, 363-98.

[63] Hansen, B.E.(1997). Inference in TAR models. Studies in Nonlinear Dynamics and Econometrics 2, 119-31

[64] Hardin, J. and J. Hilbe (2001). Generalized Linear Models and Extensions. 103, College Station, TX: Stata Press

[65] Hark Yoo, B (2006), A Bayesian analysis of Markov switching models with ARMA-GARCH error. Working paper, Bank of Korea.

[66] Hastie T., R. Tibshirani and J. Friedman (2001). The Elements of Statistical Learning. New York: Springer.

[67] Huang, M.T. (2004). Bayesian analysis of multivariate threshold error correction model for futures index. Working Paper, National Tsing-Hua University, Taiwan.

[68] Karny M., P. Nedoma and V. Šmí¿dl (2005). On cross-validation of controlled dynamic models: Bayesian approach. Preprints of the 16th World Congress of the International Federation of Automatic Control, IFAC, Prague, 1-6

[69] Kass, R.E. and A.E. Raftery (1995). Bayes factors. Journal of the American Statistical Association 90, 773-794

[70] Koop, G. and S.M. Pooter (1999). Are apparent findings of nonlinearity due to structural instability in economic time series? Staff Reports 59, Federal Reserve Bank of New York.

[71] Liu, X. and M. J. Daniels (2006). A new algorithm for simulating a correlation matrix based on parameter expansion and re-parameterization. Journal of computational and graphical statistics, Vol.15, 4, 897-914

[72] Liu J.S. and Y.N.Wu, (1999). Parameter expansion for data augmentation. Journal of the American Statistical Association, 94, 448, 1264-1274

[73] Long, J.S. (1997). Regression Models for Categorical and Limited Dependent Variables. 42-43. Thousand Oaks, CA: Sage.

[74] Lunde, A. and A. Timmermann (2000), Duration dependence in stock prices: An analysis of bull and bear markets. Econometric Society World Congress 2000 Contributed Papers 1216, Econometric Society.

[75] McCulloch, D.G., D.R. McKenzie and C.M. Goringe (2000). Ab initio simulations of the structure of amorphous carbon. Phys. Rev. B 61, 2349-2355

[76] McCulloch, R. and P.E. Rossi (1994). An exact likelihood analysis of the multinomial probit model. Journal of Econometrics, vol. 64(1-2), 207-240.

[77] McFadden, D. (2001). Economic choice, American Economic Review 91, 352-370. Nobel prize acceptance speech.

[78] Miller, J.I and R.A Rattie (2008). Crude oil and stock market: stability, instability and bubbles. Working paper, Department of Economics, University of Missouri , http://economics.missouri.edu/working-papers/2008/WP0810_millerz_ratti.pdf

[79]Nakatsuma, T. (2000). Bayesian analysis of ARMA-GARCH models: A Markov chain sampling approach. Journal of Econometrics 95, 57-69

[80] Nandha, M. and R. Faff (2008). Does oil move equity prices? A global view. Energy Economics, 30, 986-97.

[81] Pearl, R. and L. J. Reed (1920). On the rate of growth of the population of the United States since 1870 and its mathematical representation. Proceedings of the National Academy of Sciences 6, 275-288.

[82] Pearl, R. and L. J. Reed (1922). A further note on the mathematical theory of population growth. Proceedings of the National Academy of Sciences 8, 365-368.

[83] Pearl, R. and L. J. Reed (1923). On the mathematical theory of population growth. Metron 5, 6-19.

[84] Perron, P. (1989). The great crash, the oil price shock, and the unit root hypothesis. Econometrica, 57, 1361-401

[85] Pertuccelli, J.D. and N. Davis (1986). A Portmanteau test for self exciting threshold autoregressive-type nonlinearity in time series. Biometrika 73, 687-94

[86] Pettit, L.I and K.D.S. Young (1990), Measuring the effect of observation on Bayes factors. Biometrika*,* 77, 455-466.

[87] Phann, G.A., P.C. Schotman and R. Tschering (1996). Non-linear interest rate dynamics and implications for the term structure. Journal of Econometrics*,* 74, 149-76.

[88] Power, D.A. and Y. Xie (2000). Statistical methods for Categorical Data Analysis. San Diego: Academic Press.

[89] Quinne, K.M., A.D. Matin and A.B. Whitford (1999). Voter choice in multi-party democracies: a test of competing theories and models. American Journal of Political Science *43*, 1231-1247

[90] Rapach, D.E. and J.K. Strauss (2005). Structural breaks and GARCH models of exchange rate volatility. Technical Report, Saint Louis University.

[91] Ricketts, N. and D. Rose (1995). Inflation, learning and monetary policy regimes in the G-7 economies. Working Paper, Bank of Canada, 95-6

[92] Ritter, C. and M. A. Tanner (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. Journal of the American Statistical Association 87 (419), 861-868.

[93] Robert, C.P. and D.M.Titterington (2002). Discussion of a paper by D.J.Spiegelhalter, et. al. J.R. Statist*, Soc.* B64, 621 2.

[94] Rossi, P.E. and G. M. Allenby (2003). Bayesian statistics and marketing. Marketing Science *22*, 304-328.

[95] Sadorsky, P. (1999). Oil price shocks and stock market activity. Energy Economics, 21, 449-69.

[96] Shao, J. (1993). Linear model selection by Cross-validation. J. Am. Statist. Assoc. 88: 486-494.

[97] Simons, H. (08/12/2008), Will crude grease the market.
http://www.thestreet.com/print/story/10432944.html.

[98] Smith, D.R. (2000). Markov-switching and stochastic volatility diffusion models of short-term interest rates. Finance Division, Faculty of Commerce University of British Columbia.

[99] Spector, L. and M. Mazzeo (1980). Probit analysis and economic education. Journal of Economic Education, 11, 37-44.

[100] Spiegelhalter, D.J., N.G. Best, B.P.Carlin, and A.van der Linde (2003). Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society: Series B 64, 583-639

[101] Stanton, R. (1997). A nonparametric model of term structure dynamics and the market price of interest rate risk. Journal of Finance 52, 1973-2002

[102] Stephen, F.G. .(1996), Modeling the conditional distribution of interest rates as a regime-switching process. Journal of Financial Economics, (42), 27-62

[103] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). J.R. Statist.Soc. B42; 213-220

[104] Stone, M. (1977a). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society, B39: 44-47

[105] Tadikamalla, P.R. (1980). Random sampling from the exponential power distribution. Journal of the American Statistical Association, Vol.75, No.371, 683-686.

[106] Tong, H. (1990). Non-linear Times Series: A Dynamical Approach. Oxford University Press.

[107] Tsay, R.S. (1989). Testing and modeling threshold autoregressive processes. Journal of the American Statistical Association, 84, 231-40.

[108] Tsay, R.S. (1998). Testing and modeling multivariate threshold models. Journal of the American Statistical Association 93, 1188-202.

[109] Vehtari, A., and Lampinen J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. Neural Computation 14 (10): 2339-2468, The MIT Press

[110] Verhulst, P.F. (1845). Recherches mathematiques sur la loi d'accoissement de la population. Nouveaux Memoire de l'académie Royale des Sciences ,des Lettres et des Beaux-Arts de Belgique *18*, 1-32

[111] Von Neumann, J. (1951). Various techniques used in connection with random digits. in *M*onte Carlo Methods, National Bureau of Standards Applied Mathematics Series, No.12

[113] Waston, M.W. (1999). Explaining the increased variability in long term interest rates. Federal Reserve Bank of Richmond Economic Quarterly 85, 71-96

[114] Wilson, E. B. (1925). The logistic or autocatalytic grid. *P*roceedings of the National Academy of Science 11, 451-456.

[115] Winsor, C.P. (1932). A comparison of certain symmetrical growth curves., Journal of the Washington Academy of Science 22, 73-84.

[116] Yu, K. and J. Zhang (2005). A three-parameter asymmetric Laplace distribution and its extension. Communications in Statistics-Theory and Methods, 34, 1867-1879

[117] Yule, G.U. (1995). The growth of population and the factors which control it. Journal of the Royal Statistical Society 138, 1-59

Curriculum Vita

Guo Chen

2004-2009	Ph.D. in Economics, Rutgers University
2002-2004	M.A. in Economics, Northeastern University
1993-1997	B.A. in Science and English, University of Science and Technology Beijing, China

2007-2009	Research Assistant, Institute for Health, Rutgers University
2005-2007	Teaching Assistant, Department of Economics, Rutgers University
2003-2004	Teaching Assistant, Department of Economics, Northeastern University
1998-2001	Account Executive, Huson Bay Consulting, Beijing, China
1997-1998	Project Administrator, Science & Technology Development Bureau, Fuzhou, China