THREE ESSAYS ON BAYESIAN ECONOMETRICS

XIAOBIN LIU

SINGAPORE MANAGEMENT UNIVERSITY

2018

Three Essays on Bayesian Econometrics

Xiaobin Liu

A DISSERTATION

In

ECONOMICS

Presented to the Singapore Management University in Partial Fulfilment

of the Requirements for the Degree of PhD in Economics

2018

Supervisor of Dissertation

PhD in Economics, Programme Director

Three Essays on Bayesian Econometrics

by
Xiaobin Liu

Submitted to School of Economics in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Economics

Dissertation Committee:

Jun Yu (Supervisor/Chair)
Lee Kong Chian Professor of Economics and Finance
Singapore Management University

Peter C.B. Phillips
Sterling Professor of Economics and Statistics
Yale University
Distinguished Term Professor of Economics
Singapore Management University

Thomas J. Sargent
Professor of Economics
New York University
Distinguished Term Professor of Economics
Singapore Management University

Jun Tu
Associate professor of Finance
Singapore Management University

Singapore Management University
2018

Abstract

Three Essays on Bayesian Econometrics

Xiaobin Liu

My dissertation consists of three essays which contribute new theoretical results to Bayesian econometrics.

Chapter 2 proposes a new Bayesian test statistic to test a point null hypothesis based on a quadratic loss. The proposed test statistic may be regarded as the Bayesian version of the Lagrange multiplier test. Its asymptotic distribution is obtained based on a set of regular conditions and follows a chi-squared distribution when the null hypothesis is correct. The new statistic has several important advantages that make it appealing in practical applications. First, it is well-defined under improper prior distributions. Second, it avoids Jeffrey-Lindley's paradox. Third, it always takes a non-negative value and is relatively easy to compute, even for models with latent variables. Fourth, its numerical standard error is relatively easy to obtain. Finally, it is asymptotically pivotal and its threshold values can be obtained from the chi-squared distribution.

Chapter 3 proposes a new Wald-type statistic for hypothesis testing based on Bayesian posterior distributions. The new statistic can be explained as a posterior version of Wald test and have several nice properties. First, it is well-defined under improper prior distributions. Second, it avoids Jeffreys-Lindley's paradox. Third, under the null hypothesis and repeated sampling, it follows a χ^2 distribution asymptotically, offering an asymptotically pivotal test. Fourth, it only requires inverting the posterior covariance for the parameters of interest. Fifth and perhaps most importantly, when a random sample from the posterior distribution (such as an MCMC

output) is available, the proposed statistic can be easily obtained as a by-product of posterior simulation. In addition, the numerical standard error of the estimated proposed statistic can be computed based on the random sample. The finite-sample performance of the statistic is examined in Monte Carlo studies.

Chapter 4 proposes a quasi-Bayesian approach for structural parameters in finite-horizon life-cycle models. This approach circumvents the numerical evaluation of the gradient of the objective function and alleviates the local optimum problem. The asymptotic normality of the estimators with and without approximation errors is derived. The proposed estimators reach the efficiency bound in the general methods of moment (GMM) framework. Both the estimators and the corresponding asymptotic covariance are readily computable. The estimation procedure is easy to parallel so that the graphic processing unit (GPU) can be used to enhance the computational speed. The estimation procedure is illustrated using a variant of the model in Gourinchas and Parker (2002)

Table of Contents

Acknowledgements

I would like to express my deepest gratitude to my advisor, Professor Jun Yu, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. I appreciate all his contributions of time and ideas to make my Ph.D. experience productive and stimulating. The joy and enthusiasm he has for his research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. Professor Yu led me into the door of academic research and set a role model for me.

I own a special debt to Professor Peter C.B. Phillips for his wonderful course on Econometrics, valuable feedback on my research. I hope that I could be as lively, enthusiastic, and energetic as him and to someday be able to command an audience as well as he can. I would like to thank Professor Thomas Sargent for his encouragement, insightful comments, and hard questions. I thank Professor Jun Tu for reviewing my thesis and give me very valuable feedback. My sincere thanks also go to Professor Liangjun Su, Professor Sainan Jin, Professor Yichong Zhang and Professor Anthony Tay for their kind helps during my stay at SMU.

I would like to give greateest thanks to my friend Professor Yong Li for the stimulating discussions and instruction during the whole process. I have learnt a lot from him about the Bayesian econometrics.

Also I thank my dear friends Wenxin Huang, Han Chen, Sijia Xu, Peng Liu, Liang Jiang and Lu Peng.

Last but not least, I wish to thank my family, especially my wife, Lingzi, for all their love and encouragement. They were always there cheering me up and stood by me through the good time and bad.

Chapter 1 Introduction

In many practical appliations, the maximum likelihood estimator (MLE) or the classical extremum is adopted to obtained the estimator of interest and afterwards we do the inference based on the estimated results. However, in many cases, the MLE or the classical extremum estimators may be too difficult to obtain computationally. One typical example is the entire class of non-linear and non-Gaussian state space models because of their likelihood function is very hard to calculate numerically, making the MLE nearly impossible to obtain. Not surprisingly, Bayesian Markov chain Monte Carlo (MCMC) methods have emerged as the leading estimation tool to deal with non-linear and non-Gaussian state space models. Besides, there are many other examples in economics where the classical extremum estimators are subject to the curse of dimensionality in computation and some numerical problems. To deal with these problem, the Bayesian MCMC methods are also widely used.

Chapter 2 and Chapter 3 develop approachs to test a point null hypothesis based on the Bayesian posterior distribution. Hypothesis testing plays a fundamental role in making statistical inference in empirical applications. Testing a point null hypothesis is important for checking statistical evidence from data to support or to be against a particular theory because theory often can be reduced to a testable hypothesis. In many cases, the posterior distribution of parameters is available in the form of a random sample (such as MCMC sample). So that Chapter 2 and Chapter 3 propose two different statistics for the point-null hypothesis problem when the posterior distribution of parameters is available.

The statistic in Chapter 2 is equivalent to the LM statistic asymptotically. But the one in Chapter 3 can be understood as the posterior version of the well-known Wald statistic that has been used widely in practical applications. These two statistics shares some desirable properties. First, they are well-defined under improper prior distributions. Second, they avoids Jeffreys-Lindley's paradox. Third, their asymptotic distribution is a χ^2 distribution under the null hypothesis and repeated sampling, so that the threshold values are easy to obtain. Fourth, their NSE can be easily obtained. Fifth, compare with the classic Bayes factors, they are relatively easy to compute, they are all by-products of the posterior sampling.

In particular, the statistic in Chapter 3 is much easier to compute as it is only based on the posterior mean and posterior variance of the parameters of interest, which implies only posterior sampling for the alternative hypothesis is needed. At last, the statistic in Chapter 3 can be used to test hypotheses that imposes nonlinear relationships among the parameters of interest, for which the BF is difficult to use.

In Chapter 4, I propose a quasi-Bayesian estimator is introduced for structural parameters in finite-horizon life-cycle models. The asymptotic normality of the estimator is derived when an analytical solution for the model exists. When the policy functions are not analytically available, it is shown that if the approximation errors caused by numerical solving vanish fast enough, the estimator remains to be asymptotically normal. Further, it is shown that the estimator reaches the efficiency bound in the GMM framework. In the proposed method, the usual optimization procedure is converted into a sampling procedure, thereby avoiding the numerical evaluation for the gradient of objective function and alleviating the local optimum problem. The estimator and associated asymptotic covariance can be computed simultaneously. The estimation procedure is also easy to parallelize, facilitating a GPU-based and adaptive algorithm to enhance computational efficiency.

Chapter 2 A Bayesian Chi-Squared Test for Hypothesis Testing

2.1 Introduction

This paper is concerned with statistical testing of a point null hypothesis after a Bayesian Markov chain Monte Carlo (MCMC) method has been used to estimate the models. Testing for a point null hypothesis is prevalent in economics although its importance is debatable. In the meantime, Bayesian MCMC methods have found more and more applications in economics because they make it possible to fit increasingly complex models, including latent variable models (Shephard, 2005), dynamic discrete choice models (Imai, Jain and Ching, 2009) and dynamic general equilibrium models (DSGE) (An and Schorfheide, 2007).

In the Bayesian paradigm, the Bayes factor (BF) is the gold standard for Bayesian model comparison and Bayesian hypothesis testing (Kass and Raftery 1995; Geweke, 2007). Unfortunately, the BF is not problem-free. First, the BF is sensitive to the prior and subject to Jeffreys-Lindley's paradox; see for example, Kass and Raftery (1995), Poirier (1995), Robert (1993, 2001). Second, the calculation of the BF for hypothesis testing generally requires the evaluation of marginal likelihood which is a marginalization over the unknown quantities. In many cases, the evaluation of marginal likelihood is difficult. Not surprisingly, alternative strategies have been proposed to test a point null hypothesis in the Bayesian literature. These methods can be classified into two classes.

In the first class, refinements are made to the BF to overcome the theoretical and computational difficulties. For example, to reduce the influence of the prior on the BF, one may split the data into two parts, a training sample and a sample for statistical analysis. The training sample is used to update the non-informative prior and to obtain a new proper informative prior, as in the fractional BF (O'Hagan, 1995). In practice, however, this strategy is not always satisfactory because it relies on an arbitrary division of the data. To alleviate this difficulty, Berger and Perrichi (1996) proposed the so-called intrinsic BF which is based on the minimal training sample that results in proper posteriors. In general, the minimal training sample is not unique. Hence, the intrinsic BF is obtained by averaging the partial BFs calculated from all possible minimal training samples. Unfortunately, the intrinsic BF is computationally demanding, especially for latent variable models. O'Hagan (1995) discussed properties of the fractional and the intrinsic BFs.

In the second class, instead of refining the BF methodology, several interesting Bayesian approaches have been proposed for hypothesis testing based on the decision theory. For example, Bernardo and Rueda (2002, BR hereafter) showed that the BF for the Bayesian hypothesis testing can be regarded as a decision problem with a simple zero-one discrete loss function. However, the zero-one discrete function requires the use of non-regular (not absolutely continuous) prior and this is why the BF leads to Jeffreys-Lindley's paradox. BR further suggested using a continuous loss function, based on the well-known continuous Kullback-Leibler (KL) divergence function. As a result, it was shown in BR that their Bayesian test statistic does not depend on any arbitrary constant in the prior. However, BR's approach has some disadvantages. First, the analytical expression of the KL loss function required by BR is not always available, especially for latent variable models. Second, the test statistic is not a pivotal quantity. Consequently, BR had to use subjective threshold values to test the hypothesis.

To deal with the computational problem in BR in latent variable models, Li and Yu (2012, LY hereafter) proposed a new test statistic based on the $\mathcal{Q}$ function in the Expectation-Maximization (EM) algorithm of Dempster, et al. (1977). LY showed that the new statistic is well-defined under improper priors and easy to compute for latent variable models. Following the idea of McCulloch (1989), LY proposed to choose the threshold values based on the Bernoulli distribution. However, like the test statistic proposed by BR, the test statistic proposed by LY is not pivotal. Moreover, it is not clear if the test statistic of LY can resolve Jeffreys-Lindley's paradox.

Based on the difference between the deviances, Li, Zeng and Yu (2014, LZY hereafter) developed another Bayesian test statistic for hypothesis testing. This test statistic is well-defined under improper priors, free of Jeffreys-Lindley's paradox, and not difficult to compute. Moreover, its asymptotic distribution can be derived and one may obtain the threshold values from the asymptotic distribution. Unfortunately, in general the asymptotic distribution depends on some unknown population parameters and hence the test is not pivotal.

In the present paper, we propose an asymptotically pivotal Bayesian test statistic, based on a quadratic loss function, to test a point null hypothesis within the decision-theoretic framework. The new statistic has the several nice properties that makes it appeal in practice after the models are estimated by Bayesian MCMC methods. First, it is well-defined under improper prior distributions. Second, it is immune to Jeffreys-Lindley's paradox. Third, it is easy to compute. The main computational effort is to get the first derivative of the likelihood function with respect to the parameters. For latent variable models, the first derivative can be easily evaluated from the MCMC output with the help of the EM algorithm. Fourth, its numerical standard error (NSE) can be relatively easy to obtain. Finally, the asymptotic distribution of the test statistic follows the chi-squared distribution and hence the test is asymptotically pivotal.

Under a set of regularity conditions, we show that if the null hypothesis is correct our test statistic is asymptotically equivalent to the Lagrange multiplier (LM) statistic, a very popular test statistic in the frequentist's paradigm for testing a point null hypothesis. However, our proposed test has several important advantages over the LM test. First, it can incorporate the prior information to improve statistical inference. Second, the implementation of the LM test requires maximum likelihood (ML) estimation of the model under the null hypothesis. For some models, such as latent variable models and DSGE models, it is generally hard to do ML and, hence, to compute the LM statistic. Bayesian MCMC has been used to fit models with increasing complexity. The proposed test is the by-product of the Bayesian posterior output and hence easier to implement than the LM test. Third, unlike the LM test that can take a negative value in finite sample, our test always takes a nonnegative value. Finally, unlike the LM test, the new test does not need to invert any matrix. This advantage is useful when the dimension of the parameter space is high .

The paper is organized as follows. Section 2 reviews the Bayesian literature on testing a point null hypothesis from the viewpoint of the decision theory. Section 3 develops the new Bayesian test statistic, establishes its asymptotic properties, discusses how to compute it and its NSE from the MCMC outputs. Section 4 illustrates the new method by using three real examples in economics and finance. Section 5 concludes the paper. Appendix collects the proof of all the theoretical results and the derivation of the test statistic in the examples.

2.2 Bayesian Hypothesis Testing under Decision Theory

2.2.1 Testing a point null hypothesis

Let the observable data, $\mathbf{y} = (y_1, y_2, \cdots, y_n)' \in \mathbf{Y}$. A probability model $M \equiv \{p(\mathbf{y}|\theta, \psi)\}$ is used to fit the data. We are concerned with a point null hypothesis testing problem which may arise from the prediction of a particular theory. Let $\theta \in \blacksquare$ denote a vector of p-dimensional parameters of interest and $\psi \in \blacksquare$ a vector of q-dimensional nuisance parameters. The problem of testing a point null hypothesis is given by

$$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta \neq \theta_0 \end{cases}. \tag{2.2.1}$$

The hypothesis testing may be formulated as a decision problem. It is obvious that the decision space has two statistical decisions, to accept H_0 (name it d_0) or to reject H_0 (name it d_1). Let $\{\mathscr{L}[d_i,(\theta,\psi)], i=0,1\}$ be the loss function of statistical decision. Hence, a natural statistical decision to reject H_0 can be made when the expected posterior loss of accepting H_0 is sufficiently larger than the expected posterior loss of rejecting H_0, i.e., when

$$\mathbf{T}(\mathbf{y},\theta_0) = \int_\Theta \int_\Psi \{\mathscr{L}[d_0,(\theta,\psi)] - \mathscr{L}[d_1,(\theta,\psi)]\}\, p(\theta,\psi|\mathbf{y})\mathrm{d}\theta\mathrm{d}\psi > c \geq 0,$$

where $\mathbf{T}(\mathbf{y},\theta_0)$ is a Bayesian test statistic; $p(\theta,\psi|\mathbf{y})$ is the posterior distribution with some given prior $p(\theta,\psi)$; c is a threshold value. Let $\triangle\mathscr{L}[H_0,(\theta,\psi)] = \mathscr{L}[d_0,(\theta,\psi)] - \mathscr{L}[d_1,(\theta,\psi)]$ be the net loss difference function which can generally be used to measure the evidence against H_0 as a function of (θ,ψ). Hence, the Bayesian test statistic can be rewritten as

$$\mathbf{T}(\mathbf{y},\theta_0) = E_{\vartheta|\mathbf{y}}\left(\triangle\mathscr{L}[H_0,(\theta,\psi)]\right).$$

2.2.2 A literature review

The BF is defined as the ratio of the two marginal likelihood functions, namely,

$$BF_{01} = \frac{p(\mathbf{y}|M_0)}{p(\mathbf{y}|M_1)},$$

where $M_0 := \{p(\mathbf{y}|\theta_0,\psi), \psi \in \Psi\}$ is the model under the null; $M_1 := M$ is the model under the alternative. The two marginal likelihood functions are defined as

$$p(\mathbf{y}|M_0) = \int_{\Psi} p(\mathbf{y}|\theta_0,\psi)p(\psi|\theta_0)\mathrm{d}\psi,$$
$$p(\mathbf{y}|M_1) = \int_{\Theta}\int_{\Psi} p(\mathbf{y}|\theta,\psi)p(\psi|\theta)p(\theta)\mathrm{d}\theta\mathrm{d}\psi.$$

The BF corresponds to the use of the zero-one discrete loss function, namely,

$$\triangle\mathscr{L}[H_0,(\theta,\psi)] = \begin{cases} -1 & \text{if } \theta = \theta_0 \\ 1 & \text{if } \theta \neq \theta_0 \end{cases},$$

and in this case, with $c = 0$, we

$$\text{Reject } H_0 \text{ iff } BF_{01} = \frac{\int_{\Psi} p(\mathbf{y}|\theta_0,\psi)p(\psi|\theta_0)\mathrm{d}\psi}{\int_{\Theta}\int_{\Psi} p(\mathbf{y}|\theta,\psi)p(\psi|\theta)p(\theta)\mathrm{d}\theta\mathrm{d}\psi} < 1.$$

Remark 2.2.1. *The BF has several disadvantages. If the Jeffreys or the reference prior (Jeffreys, 1961) is used to reflect the objectiveness, the BF is not well-defined since it depends on an arbitrary constant (BR, 2002). In addition, if a proper prior with a large spread is used to represent the prior ignorance, the BF has a tendency to favor the null hypothesis, giving rise to Jeffreys-Lindley's paradox; see Poirier (1995), Robert (1993, 2001). Moreover, for many models in economics, such as latent variable models and the DSGE models, the marginal likelihood and, hence, the BF are very difficult to evaluate; see Han and Carlin (2001) for a good review of methods for calculating the BF from the MCMC output.*

BR (2002) suggested using a continuous loss function based on the KL divergence,

$$KL[p(x),q(x)] = \int p(x)\log\frac{p(x)}{q(x)}dx, \tag{2.2.2}$$

where $p(x)$ and $q(x)$ are any two regular probability density functions (pdf). The corresponding Bayesian test statistic is:

$$\mathbf{T}_{BR}(\mathbf{y},\theta_0)=E_{\vartheta|\mathbf{y}}\left(\min\left\{KL\left[p(\mathbf{y}|\theta,\psi),p(\mathbf{y}|\theta_0,\psi)\right],KL\left[p(\mathbf{y}|\theta_0,\psi),p(\mathbf{y}|\theta,\psi)\right]\right\}\right). \tag{2.2.3}$$

Remark 2.2.2. *It is shown in BR (2002) that $\mathbf{T}_{BR}(\mathbf{y},\theta_0)$ is well-defined under improper distributions. This is an important advantage over the BF. However, the BR test is not without its problems. First, the KL divergence function often does not have a closed-form expression. Consequently, $\mathbf{T}_{BR}(\mathbf{y},\theta_0)$ may be difficult to compute. Second, BR suggested choosing threshold values based on the normal distribution to implement the test. Unfortunately, the choice of the normal distribution and, hence, the threshold values is subjective and lacks of rigorous statistical justifications. A different distribution will lead to different threshold values.*

To alleviate the computational problems of $\mathbf{T}_{BR}(\mathbf{y},\theta_0)$ in the context of latent variable models, LY (2012) proposed a new loss difference function, based on the $\mathscr{Q}$ function used in the EM algorithm (Dempster, Laird and Rubin, 1977). Let $\mathbf{z}=(z_1,z_2,\cdots,z_n)'$ denote the latent variables and $\mathbf{x}=(\mathbf{y}',\mathbf{z}')'$. Let $p(\mathbf{y}|\vartheta)$ and $p(\mathbf{x}|\vartheta)$ $(:=p(\mathbf{y},\mathbf{z}|\vartheta))$ be the observed data likelihood function and the complete data likelihood function, respectively. The relationship between these two likelihood functions is

$$p(\mathbf{y}|\vartheta)=\int p(\mathbf{y},\mathbf{z}|\vartheta)d\mathbf{z}.$$

For any ϑ_1 and ϑ_2, the $\mathscr{Q}$ function is:

$$\mathscr{Q}(\vartheta_1|\vartheta_2)=E_{z|\mathbf{y},\vartheta_2}\left[\log p(\mathbf{y},\mathbf{z}|\vartheta_1)\right].$$

Compared with the observed data likelihood function $p(\mathbf{y}|\vartheta)$, the $\mathscr{Q}$ function is easier to evaluate in latent variable models. In particular, when the analytical expression of $p(\mathbf{y}|\vartheta)$ is not available, the $\mathscr{Q}$ function can be easily approximated from the MCMC output via,

$$\mathscr{Q}(\vartheta_1|\vartheta_2) \approx \frac{1}{G}\sum_{g=1}^{G}\log p\left(\mathbf{y}, \mathbf{z}^{(g)}|\vartheta_1\right),$$

where $\{\mathbf{z}^{(g)}, g = 1, 2, \cdots, G\}$ are the effective MCMC draws from the posterior distribution $p(\mathbf{z}|\mathbf{y}, \vartheta_2)$. Let $\vartheta_0 = (\theta_0, \psi)$. LY (2012) defined a new continuous net loss difference function as:

$$\triangle\mathscr{L}(\vartheta, \vartheta_0) = \{\mathscr{Q}(\vartheta, \vartheta) - \mathscr{Q}(\vartheta_0, \vartheta)\} + \{\mathscr{Q}(\vartheta_0, \vartheta_0) - \mathscr{Q}(\vartheta, \vartheta_0)\},$$

and proposed a Bayesian test statistic as:

$$\mathbf{T}_{LY}(\mathbf{y}, \theta_0) = E_{\vartheta|\mathbf{y}}\left[\triangle\mathscr{L}(\vartheta, \vartheta_0)\right].$$

Remark 2.2.3. *It is shown in LY (2012) that the test statistic, $T_{LY}(\mathbf{y}, \theta_0)$, is well-defined under improper priors and also easy to compute. However, this test statistic has some practical disadvantages. First, like the test statistic of BR, some threshold values have to be specified. Following the idea of McCulloch (1989), LY (2012) proposed to choose threshold values based on the Bernoulli distribution. Unfortunately, the choice of the Bernoulli distribution is arbitrary. If another distribution is used, the threshold values will be different. Second, it is not clear whether this test statistic is immune to Jeffreys-Lindley's paradox.*

Aiming to alleviate Jeffreys-Lindley's paradox, LZY (2014) developed an alternative Bayesian test statistic based on the Bayesian deviance. The net loss function and the test statistic are given, respectively, by

$$\triangle\mathscr{L}[H_0, (\theta, \psi)] = 2\log p(\mathbf{y}|\theta, \psi) - 2\log p(\mathbf{y}|\theta_0, \psi),$$

$$\mathbf{T}_{LZY}(\mathbf{y}, \theta_0) = 2\int\left[\log p(\mathbf{y}|\theta, \psi) - \log p(\mathbf{y}|\theta_0, \psi)\right]p(\theta, \psi|\mathbf{y})\mathrm{d}\theta\mathrm{d}\psi. \qquad (2.2.4)$$

$\mathbf{T}_{LZY}$ can be understood as the Bayesian version of the likelihood ratio test.

However, for latent variable models, the likelihood function $p(\mathbf{y}|\theta,\psi)$ generally is not available in closed-form. To achieve computational tractability, under some regularity conditions, LZY (2014) gave an asymptotically equivalent form for $\mathbf{T}_{LZY}(\mathbf{y},\theta_0)$, i.e.,

$$\begin{aligned}\mathbf{T}_{LZY}^{*}(\mathbf{y},\theta_0) &= 2D+2\left[\log p(\bar{\theta},\bar{\psi})-\log p(\bar{\psi}|\theta_0)\right]-2\left[\int \log p(\theta|\psi)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\right] \\ &\quad -\left[p+q-\mathbf{tr}[-L_{0n}^{(2)}(\bar{\psi})V_{22}(\bar{\vartheta})]\right],\end{aligned}$$

where $\bar{\vartheta}=(\bar{\theta},\bar{\psi})'$ is the posterior mean of ϑ under H_1, $\bar{\vartheta}_*=(\theta_0,\bar{\psi})'$, $\bar{\vartheta}_b=(1-b)\bar{\vartheta}_*+b\bar{\vartheta}$, for $b\in[0,1]$, $S(\mathbf{x}|\vartheta)=\partial\log p(\mathbf{x}|\vartheta)/\partial\vartheta$, $D=\int_0^1\left\{(\bar{\theta}-\theta_0)'\left[E_{\mathbf{z}|\mathbf{y},\bar{\vartheta}_b}\left(S_1(\mathbf{x}|\bar{\vartheta}_b)\right)\right]\right\}\mathrm{d}b$ the subvector of $S(\mathbf{x}|\vartheta)$ corresponding to θ, $V_{22}(\bar{\vartheta})=E[(\psi-\bar{\psi})(\psi-\bar{\psi})'|\mathbf{y},H_1]$, the submatrix of $V(\bar{\vartheta})$ corresponding to ψ, and $L_{0n}^{(2)}(\psi)=\partial^2\log p(\mathbf{y},\psi|\theta_0)/\partial\psi\partial\psi'$.

Remark 2.2.4. *As shown in LZY (2014),* $\mathbf{T}_{LZY}^{*}(\mathbf{y},\theta_0)$ *appeals in four aspects. First, it is well-defined under improper priors. Second, it does not suffer from Jeffreys-Lindley's paradox and, hence, can be used under non-informative vague priors. Third, it is easy to compute. Furthermore, for latent variable models,* $\mathbf{T}_{LZY}^{*}(\mathbf{y},\theta_0)$ *only involves the first and the second derivatives which is easy to evaluate from the MCMC output with the help of the EM algorithm. Finally, LZY (2014) derived the asymptotic distribution of* $\mathbf{T}_{LZY}^{*}(\mathbf{y},\theta_0)$*. When* θ *and* $\bar{\psi}$ *are orthogonal, the asymptotic distribution is determined by the chi-squared distribution. In this case the test is asymptotically pivotal and the thresholds can be obtained form the asymptotic distribution. Unfortunately, in general the test is not asymptotically pivotal because the asymptotic distribution depends on some unknown population parameters.*

2.3 Bayesian Hypothesis Testing Based on a Quadratic Loss

2.3.1 The test statistic

To deal with the non-pivotal problem, in this section, we develop a new Bayesian test statistic for hypothesis testing. The new statistic shares all the nice features of the LZY statistic. First, it is motivated from the decision-theoretic perspective. Second, it is well-defined under improper prior distributions. Third, it is immune to Jeffreys-Lindley's paradox. Fourth, it is easy to compute. However, unlike the LZY statistic, the new statistic is asymptotically pivotal and the threshold can be easily obtained from its asymptotic distribution.

To fix the idea, let

$$s(\vartheta) = \frac{\partial \log p(\mathbf{y}|\vartheta)}{\partial \vartheta}, C(\vartheta) = s(\vartheta)s(\vartheta)',$$

where $s(\vartheta)$ is the score function and $\vartheta = (\theta, \psi)$. We define a quadratic loss function as:

$$\triangle\mathscr{L}[H_0, \vartheta] = (\theta - \bar{\theta})' C_{\theta\theta}(\bar{\vartheta}_0)(\theta - \bar{\theta}), \tag{2.3.1}$$

where $C_{\theta\theta}(\vartheta)$ is the submatrix of $C(\vartheta)$ corresponding to θ and is semi-positive definite, $\bar{\vartheta}_0 = (\theta_0, \bar{\psi}_0)$ is the Bayesian estimator of ϑ under H_0, $\bar{\theta}$ is the Bayesian estimator of θ under H_1. Based on this quadratic loss, we propose the following Bayesian test statistic:

$$\mathbf{T}(\mathbf{y}, \theta_0) = \int \triangle\mathscr{L}[H_0, \vartheta] p(\vartheta|\mathbf{y}) \mathrm{d}\vartheta = \int (\theta - \bar{\theta})' C_{\theta\theta}(\bar{\vartheta}_0)(\theta - \bar{\theta}) p(\vartheta|\mathbf{y}) \mathrm{d}\vartheta, \tag{2.3.2}$$

where $p(\vartheta|\mathbf{y})$ is the posterior distribution of ϑ under H_1.

Remark 2.3.1. *Clearly $\mathbf{T}(\mathbf{y}, \theta_0)$ depends on the posterior distribution directly. The prior information only influences the test statistic via the posterior distribution.*

Remark 2.3.2. *Since the posterior distribution $p(\vartheta|\mathbf{y})$ is independent of an arbitrary constant in the prior distributions, both $s(\vartheta)$ and $C_{\theta\theta}(\bar{\vartheta}_0)$ are independent of the arbitrary constant. As a result, $\mathbf{T}(\mathbf{y},\theta_0)$ is well-defined under improper priors.*

Remark 2.3.3. *Under some regular condition, we will show in Theorem 2.3.1 below that the proposed test converges to the χ^2 distribution and hence it is not subject to Jeffreys-Lindley's paradox, at least when the sample size is large. To see how it can avoid Jeffreys-Lindley's paradox, consider the example discussed in LZY (2014). Let $y \sim N(\theta,\sigma^2)$ with a known σ^2 and we test the null hypothesis $H_0:\theta=0$. Let the prior distribution of θ be $N(\mu,\tau^2)$ with $\mu=0$. LZY showed that the posterior distribution of θ is $N(\mu(y),\omega^2)$ with*

$$\mu(y)=\frac{\sigma^2\mu+\tau^2 y}{\sigma^2+\tau^2},\omega^2=\frac{\sigma^2\tau^2}{\sigma^2+\tau^2},$$

and BF is

$$BF_{10}=\frac{1}{BF_{01}}=\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left[\frac{\tau^2y^2}{2\sigma^2(\sigma^2+\tau^2)}\right].$$

As $\tau^2\to+\infty$, $BF_{10}\to 0$, suggesting the test always supports H_0, whether or not H_0 holds true, giving rise to Jeffreys-Lindley's paradox. On the other hand, it is easy to show that

$$C_{\theta\theta}(\bar{\vartheta}_0)=\frac{y^2}{\sigma^4},\ \textit{and}\ \mathbf{T}(y,0)=\frac{y^2}{\sigma^4}\int(\theta-\bar{\theta})^2p(\theta|y)d\theta=\frac{\omega^2y^2}{\sigma^4}.$$

As $\tau^2\to+\infty$, $\mu(y)\to y$, $\omega^2\to\sigma^2$, and, hence, $\mathbf{T}(y,0)\to y^2/\sigma^2$ which is distributed as $\chi^2(1)$ when H_0 is true. Consequently, our proposed test statistic is immune to Jeffreys-Lindley's paradox.

Remark 2.3.4. *To calculate $\mathbf{T}(\mathbf{y},\theta_0)$, the first derivatives of the observed-data likelihood function must be evaluated. For most latent variable models, the first derivatives are difficult to evaluate directly because the observed-data likelihood function is not available in closed-form. There are several approaches to calculate the first derivatives from the MCMC output.*

First, the first derivatives can be approximated using the EM algorithm in connection with the data augmentation technique. For any ϑ and ϑ^* in the support space of ϑ, it was shown in Dempster et al. (1977) that

$$s(\vartheta) = \frac{\partial \log p(\mathbf{y}|\vartheta)}{\partial \vartheta} = \frac{\partial \mathscr{Q}(\vartheta|\tilde{\vartheta})}{\partial \vartheta}|_{\tilde{\vartheta}=\vartheta} = \int \frac{\partial \log p(\mathbf{y}, \mathbf{z}|\vartheta)}{\partial \vartheta} p(z|\mathbf{y}, \vartheta) dz.$$

Hence, based on the MCMC output, the first derivative can be approximated by:

$$s(\vartheta) \approx \frac{1}{G} \sum_{g=1}^{G} \left\{ \frac{\partial \log p(\mathbf{y}, z^{(g)}|\vartheta)}{\partial \vartheta} \right\},$$

where $\{z^{(g)}, g = 1, 2, \cdots, G\}$ are effective MCMC draws from the posterior distribution $p(z|\mathbf{y}, \vartheta)$ due to the use of data augmentation.

Second, for the dynamic state space models, more efficient approaches are available to compute the first derivatives. For example, for Gaussian linear state space models the Kalman filter is computationally very efficient for computing the first derivatives (Harvey, 1989). For non-Gassuian nonlinear state space models, the particle filter is an efficient approach for computing the first derivatives. See, for example, Poyiadjis, et al (2011) and Doucet and Shephard (2012) for recent contributions in using the particle filter to approximate the score functions. Doucet and Johansen (2011) gives an excellent review of the literature on the particle filter.

Remark 2.3.5. *It is known that the BF is the ratio of two marginal likelihoods. For model M (corresponding to either the null hypothesis or the alternative hypothesis), as shown in Chib (1995) based on Bayes' theorem, the log-marginal likelihood may be calculated by*

$$\log p(\mathbf{y}|\vartheta, M) + \log p(\vartheta|M) - \log p(\vartheta|\mathbf{y}, M), \tag{2.3.3}$$

where $p(\mathbf{y}|\vartheta, M)$ is the observed likelihood function, $p(\vartheta|M)$ is the prior distribution, and $p(\vartheta|\mathbf{y}, M)$ is the posterior distribution, ϑ is an appropriately selected high density point in the estimated model. Chib (1995) suggested using the posterior *mean,* $\bar{\vartheta}$.

The second term is the log prior density which is easy to calculate. The third quantity, $p(\vartheta|\mathbf{y},M)$*, is the posterior density and only known up to a constant. Based on the Gibbs sampler and the Metropolis-Hastings algorithm, Chib (1995) and Chib and Jeliazkov (2001) proposed methods to approximate* $p(\vartheta|\mathbf{y},M)$*. These methods are generally applicable to a wide class of models. When the parameter* ϑ *is high-dimensional, however, estimating* $p(\vartheta|\mathbf{y},M)$ *is computationally demanding. The first term,* $p(\mathbf{y}|\vartheta,M)$*, is easy to evaluate when it has an analytical expression. For many models, including the dynamic latent variable models, however, the first term,* $p(\mathbf{y}|\vartheta,M)$*, is marginalized over the latent variables such as* $\mathbf{z}$*, that is,*

$$p(\mathbf{y}|\vartheta,M) = \int p(\mathbf{y},\mathbf{z}|\vartheta,M)d\mathbf{z} = \int p(\mathbf{y}|\mathbf{z},\vartheta,M)p(\mathbf{z}|\vartheta,M)d\mathbf{z}$$

Often integration is of high-dimension and has to be evaluated numerically. Unfortunately, mimicking the strategy in Remark 3.4 by averaging $p(y,z^{(m)}|\theta,M)$ *over the effective draws* $\{z^{(g)}, g = 1,2,\cdots G\}$ *from* $p(z|\theta,M)$ *is numerically unstable because the expectation is taken with respect to the prior distribution. Whereas, computing* $s(\vartheta)$ *in Remark 3.4 is taken with respect to the posterior distribution. All these problems make it difficult to evaluate the marginal likelihood* $\log p(y|M)$ *and BF. To calculate* $T(\mathbf{y},\theta_0)$*, the main computational effort is to evaluate the first derivatives of* $\log p(y|\vartheta,M)$*, which can be achieved by the EM algorithm, the Kalman filter or the particle filter, as remarked earlier. Thus, there is a computational advantage in the proposed test over the BF.*

Since $\mathbf{T}(\mathbf{y},\theta_0)$ is calculated from the MCMC output, it is important to assess the NSE for measuring the magnitude of simulation error. When the observed likelihood function $p(\mathbf{y}|\vartheta)$ has a closed-form expression, the first derivative and $C_{\theta\theta}(\bar{\vartheta}_0)$ are also available analytically. Let

$$f(\theta) = (\theta - \bar{\theta})' C_{\theta\theta}(\bar{\vartheta}_0)(\theta - \bar{\theta}).$$

Then, we have

$$\mathbf{T}(\mathbf{y},\theta_0) = E_{\vartheta|\mathbf{y}}[f(\theta)|\mathbf{y}], \widehat{\mathbf{T}}(\mathbf{y},\theta_0) = \frac{1}{G}\sum_{g=1}^{G} f\left(\theta^{(g)}\right),$$

where $\theta^{(g)}, g = 1,2,\cdots,G$ are random draws from the posterior distribution $p(\vartheta|\mathbf{y})$.

If $\theta^{(g)}, g = 1,2,\cdots,G$ are independent random samples, it can be shown that

$$Var\left(\widehat{\mathbf{T}}(\mathbf{y},\theta_0)\right) = Var\left(G^{-1}\sum_{g=1}^{G} f\left(\theta^{(g)}\right)\right) = \frac{1}{G}Var\left(f\left(\theta^{(g)}\right)\right).$$

A consistent estimator of $Var\left(f\left(\theta^{(g)}\right)\right)$ is given by

$$G^{-1}\sum_{g=1}^{G}\left(f\left(\theta^{(g)}\right) - \widehat{\mathbf{T}}(\mathbf{y},\theta_0)\right)\left(f\left(\theta^{(g)}\right) - \widehat{\mathbf{T}}(\mathbf{y},\theta_0)\right)'.$$

If $\theta^{(g)}, g = 1,2,\cdots,G$ are dependent random samples, following Newey and West (1987), a consistent estimator of $Var\left(\widehat{\mathbf{T}}(\mathbf{y},\theta_0)\right)$ is

$$\frac{1}{G}\left[\Omega_0 + \sum_{k=1}^{q}\left(1 - \frac{k}{q+1}\right)\left(\Omega_k + \Omega_k'\right)\right], \tag{2.3.4}$$

where

$$\Omega_k = G^{-1}\sum_{g=k+1}^{G}\left(f\left(\theta^{(g)}\right) - \widehat{\mathbf{T}}(\mathbf{y},\theta_0)\right)\left(f\left(\theta^{(g)}\right) - \widehat{\mathbf{T}}(\mathbf{y},\theta_0)\right)',$$

and q is a positive integer at which the autocorrelation tapers off. In the applications, we set $q = 10$.

When the observed likelihood function $p(\mathbf{y}|\vartheta)$ does not have an analytical expression, another approach for assessing the NSE is given below. Note that

$$\begin{aligned}
\mathbf{T}(\mathbf{y},\theta_0) &= \int (\theta-\bar{\theta})' C_{\theta\theta}(\bar{\vartheta}_0)(\theta-\bar{\theta})\, p(\vartheta|\mathbf{y})\, d\vartheta \\
&= \int \mathbf{tr}\left[(\theta-\bar{\theta})' C_{\theta\theta}(\bar{\vartheta}_0)(\theta-\bar{\theta})\right] p(\vartheta|\mathbf{y})\, d\vartheta \\
&= \mathbf{tr}\left[C_{\theta\theta}(\bar{\vartheta}_0)\int (\theta-\bar{\theta})(\theta-\bar{\theta})' p(\vartheta|\mathbf{y})\, d\vartheta\right],
\end{aligned}$$

and that

$$s_\theta(\vartheta) = \int \frac{\partial \log p(\mathbf{y},\mathbf{z}|\vartheta)}{\partial\theta} p(\mathbf{z}|\mathbf{y},\vartheta)\,\mathrm{d}\mathbf{z}.$$

We can estimate $s_\theta\left(\bar{\vartheta}_0\right)$ by

$$\widehat{h}_1 = \frac{1}{G}\sum_{g=1}^{G}\frac{\partial \log p\left(\mathbf{y},\mathbf{z}^{(g)}|\bar{\vartheta}_0\right)}{\partial\theta} = \frac{1}{G}\sum_{g=1}^{G} h_1^{(g)}$$

where $\{\mathbf{z}^{(g)}, g=1,2,\cdots,G\}$ are efficient random draws from $p\left(\mathbf{z}|\bar{\vartheta}_0,\mathbf{y}\right)$. Furthermore, we get

$$\int (\theta-\bar{\theta})(\theta-\bar{\theta})' p(\vartheta|\mathbf{y})\,\mathrm{d}\vartheta \approx \widehat{H}_2 = \frac{1}{G}\sum_{g=1}^{G}\left(\theta^{(g)}-\bar{\theta}\right)\left(\theta^{(g)}-\bar{\theta}\right)' = \frac{1}{G}\sum_{g=1}^{G} H_2^{(g)}.$$

Then, we have

$$\widehat{\mathbf{T}}(\mathbf{y},\theta_0) = \mathbf{tr}\left(\widehat{h}_1\widehat{h}_1'\widehat{H}_2\right).$$

Following the notations of Magus and Neudecker (2002) about matrix derivatives, let

$$\widehat{h}_2 = vech\left(\widehat{H}_2\right),\ h_2^{(g)} = vech\left(H_2^{(g)}\right),\ \widehat{\mathbf{h}} = \left(\widehat{h}_1', \widehat{h}_2'\right)'.$$

Note that the dimension of $\widehat{h}_1$ is $p\times 1$ and the dimension of $\widehat{h}_2$ is $p^*\times 1, p^* = p(p+1)/2$. Hence, we have

$$\begin{aligned}
\frac{\partial\widehat{\mathbf{T}}(\mathbf{y},\theta_0)}{\partial\widehat{h}} &= vec(I_p)'\left[\left(\left(\widehat{h}_1'\widehat{H}_2\right)'\otimes I_p\right)\frac{\partial\widehat{h}_1}{\partial\widehat{h}} + \left(\widehat{H}_2'\otimes\widehat{h}_1\right)\frac{\partial\widehat{h}_1'}{\partial\widehat{h}} + \left(I_p\otimes\widehat{h}_1\widehat{h}_1'\right)\frac{\partial\widehat{H}_2}{\partial\widehat{h}}\right] \\
&= vec(I_p)'\left[\left(\widehat{H}_2'\widehat{h}_1\otimes I_p + \widehat{H}_2'\otimes\widehat{h}_1\right)\frac{\partial\widehat{h}_1}{\partial\widehat{h}} + \left(I_p\otimes\widehat{h}_1\widehat{h}_1'\right)\frac{\partial\widehat{H}_2}{\partial\widehat{\mathbf{h}}}\right].
\end{aligned}$$

where I_p is the p-dimensional identity matrix and

$$\frac{\partial \widehat{h}_1}{\partial \widehat{h}} = \frac{\partial \left(\widehat{h}_1'\right)}{\partial \widehat{h}} = \left[I_p, 0_{p\times p^*}\right], \frac{\partial \widehat{H}_2}{\partial \widehat{h}} = \left[0_{p^2\times p}, \left(\frac{\partial \widehat{H}_2}{\partial \widehat{h}_2}\right)_{p^2\times p^*}\right] = \left[0_{p^2\times p}, \frac{\partial vec(\widehat{H}_2)}{\partial \widehat{h}_2}_{p^2\times p^*}\right].$$

By the Delta method,

$$Var\left(\widehat{\mathbf{T}}(\mathbf{y},\theta_0)\right) = \frac{\partial \widehat{\mathbf{T}}(\mathbf{y},\theta_0)}{\partial \widehat{\mathbf{h}}} Var\left(\widehat{h}\right)\left(\frac{\partial \widehat{\mathbf{T}}(\mathbf{y},\theta_0)}{\partial \widehat{h}}\right)'.$$

Again, following Newey and West (1987), a consistent estimator can be given by

$$Var(\widehat{\mathbf{h}}) = \frac{1}{G}\left[\Omega_0 + \sum_{k=1}^{q}\left(1 - \frac{k}{q+1}\right)\left(\Omega_k + \Omega_k'\right)\right],$$

where

$$\Omega_k = G^{-1}\sum_{g=k+1}^{G}\left(\mathbf{h}^{(g)} - \widehat{\mathbf{h}}\right)\left(\mathbf{h}^{(g)} - \widehat{\mathbf{h}}\right)'.$$

Remark 2.3.6. *Based on (2.3.3), Chib (1995) provided a method to calculate the NSE for estimating* $\log p(\vartheta|\mathbf{y},M)$*. When* $\log p(\mathbf{y}|\vartheta,M)$ *is available in closed-form, the NSE of the estimate of* $\log p(\mathbf{y}|M)$ *is the same as that of* $\log p(\vartheta|\mathbf{y},M)$ *because both* $p(\mathbf{y}|\vartheta,M)$ *and* $p(\vartheta|M)$ *can be computed without incurring simulation errors. However, when* $p(\mathbf{y}|\vartheta,M)$ *does not have a closed-form expression, it has to be calculated by a simulation-based method (such as the EM algorithm or the particle filters) and there will be the NSE for estimating it. In this case, it will be difficult to obtain the NSE of* $\log p(\mathbf{y}|\vartheta,M)$*. Relative to* $\log p(\vartheta|\mathbf{y},M)$ *whose order of magnitude is often* $O_p(1)$*,* $\log p(\mathbf{y}|\vartheta,M)$ *is typically* $O_p(n)$ *so that* $\log p(\mathbf{y}|\vartheta,M)$ *is dominant in* $\log p(\mathbf{y}|M)$*. Consequently, one cannot ignore the NSE of* $\log p(\mathbf{y}|\vartheta,M)$ *when calculating the NSE of* $\log p(\mathbf{y}|M)$*. As a result, it will be very difficult to obtain the NSE of the estimate of* $\log p(\mathbf{y}|M)$ *and hence that of the BF. The ease with which one can calculate the NSE of the estimate of* $\mathbf{T}(\mathbf{y},\theta_0)$ *is another important advantage of the proposed test over the BF.*

2.3.2 The threshold value

To implement the proposed test, a threshold value, c, has to be specified, i.e.,

$$\text{Accept } H_0 \text{ if } \mathbf{T}(\mathbf{y},\theta_0) \leq c; \text{ Reject } \; H_0 \; \text{ if } \mathbf{T}(\mathbf{y},\theta_0) > c.$$

This section obtains the asymptotic distribution of the test statistic under H_0 and establishes the link between the test statistic and the LM test. To do so, following LZY (2014), we first impose a set of regularity conditions.

Assumption 1: There exists a finite sample size n^*, so that, for $n > n^*$, there is a local maximum at $\widehat{\vartheta}$ (i.e., posterior mode) such that $L_n^{(1)}(\widehat{\vartheta}) = 0$ and $L_n^{(2)}(\widehat{\vartheta})$ is negative definite, where $L_n(\vartheta) = \log p(\vartheta|\mathbf{y})$, $L_n^{(1)}(\vartheta) = \partial \log p(\vartheta|\mathbf{y})/\partial\vartheta$, $L_n^{(2)}(\vartheta) = \partial^2 \log p(\vartheta|\mathbf{y})/\partial\vartheta\partial\vartheta'$.

Assumption 2: The largest eigenvalue λ_n of $-L_n^{-(2)}(\widehat{\vartheta})$ goes to zero when $n \to \infty$.

Assumption 3: For any $\varepsilon > 0$, there exists an integer N and some $\delta > 0$ such that for any $n > \max\{N, n^*\}$ and $\vartheta \in H(\widehat{\vartheta},\delta) = \{\vartheta : ||\vartheta - \widehat{\vartheta}|| \leq \delta\}$, $L_n^{(2)}(\vartheta)$ exists and satisfies

$$-A(\varepsilon) \leq L_n^{(2)}(\vartheta) L_n^{-(2)}(\widehat{\vartheta}) - \mathbf{E}_{p+q} \leq A(\varepsilon),$$

where $\mathbf{E}_{p+q}$ is an identity matrix and $A(\varepsilon)$ is a positive semi-definite symmetric matrix whose largest eigenvalue goes to zero as $\varepsilon \to 0$.

Assumption 4: For any $\delta > 0$, as $n \to \infty$,

$$\int_{\Omega - H(\widehat{\vartheta},\delta)} p(\vartheta|\mathbf{y}) d\vartheta \to 0,$$

where Ω is the support space of ϑ.

Assumption 5: The likelihood function under both the null hypothesis and the alternative hypothesis is regular so that the standard ML theory can be applied. Furthermore, if the null hypothesis is true, let $\vartheta_0 = (\theta_0, \psi_0)$ be true value of ϑ, as

$n \to \infty$, for any null sequence $k_n \to 0$, so that,

$$\sup_{||\vartheta - \vartheta_0|| < k_n} n^{-1} ||\mathbf{I}(\vartheta) - \mathbf{I}(\vartheta_0)|| \xrightarrow{p} 0,$$

where $\mathbf{I}(\vartheta) = \partial^2 \log p(\mathbf{y}|\vartheta)/\partial\vartheta\partial\vartheta'$.

Remark 2.3.7. *In the literature, Assumptions 1-4 have been used to develop the Bayesian large sample theory; see, for example, Chen (1985). Assumption 5 is a fundamental regularity condition for developing the standard ML theory. Based on these regularity conditions, LZY (2014) showed that*

$$\begin{aligned}
\bar{\vartheta} &= E\left[\vartheta|\mathbf{y}, H_1\right] = \int \vartheta p(\vartheta|\mathbf{y}) d\vartheta = \widehat{\vartheta} + o_p(n^{-1/2}), \\
V(\widehat{\vartheta}) &= E\left[(\vartheta - \widehat{\vartheta})(\vartheta - \widehat{\vartheta})'|\mathbf{y}, H_1\right] = -L_n^{-(2)}(\widehat{\vartheta}) + o_p(n^{-1}).
\end{aligned}$$

When the null hypothesis holds, we also have

$$\begin{aligned}
\bar{\psi}_0 &= E\left[\psi|\mathbf{y}, H_0\right] = \int \psi p(\psi|\mathbf{y}, \theta_0) d\psi = \widehat{\psi}_0 + o_p(n^{-1/2}), \\
V_0(\widehat{\psi}_0) &= E\left[(\psi - \widehat{\psi}_0)(\psi - \widehat{\psi}_0)'|\mathbf{y}, H_0\right] = -L_{0n}^{-(2)}(\widehat{\psi}_0) + o_p(n^{-1}),
\end{aligned}$$

where $L_{0n}^{(2)}(\psi_0) = \log p(\psi|\theta_0, \mathbf{y})/\partial\psi\partial\psi'|_{\psi=\psi_0}$ *and* $\widehat{\psi}_0$ *is the local maximum of* $\log p(\psi|\mathbf{y}, \theta_0)$ *under* H_0.

Lemma 2.3.1. *Let*

$$\mathbf{J}(\vartheta) = \mathbf{I}^{-1}(\vartheta).$$

When the null hypothesis is true, and $\vartheta_0 = (\theta_0, \psi_0)$ *is the true value of* ϑ*, for any consistent estimator* $\tilde{\vartheta}$ *of* ϑ*, we have*

$$\begin{aligned}
&\mathbf{I}(\vartheta_0) = O_p(n), \mathbf{I}(\tilde{\vartheta}) = \mathbf{I}(\vartheta_0) + o_p(n) = O_p(n), \\
&\mathbf{J}(\vartheta_0) = O_p(n^{-1}), \mathbf{J}(\tilde{\vartheta}) = \mathbf{J}(\vartheta_0) + o_p(n^{-1}) = O_p(n^{-1}).
\end{aligned}$$

Lemma 2.3.2. *Let* $\widehat{\vartheta}_0 = (\theta_0, \widehat{\psi}_0)$ *be the posterior mode of* ϑ *under the null hypoth-*

esis. Under Assumptions 1-5 and when the null hypothesis is true, we have

$$s(\widehat{\vartheta}_0) = O_p(n^{1/2}), s(\bar{\vartheta}_0) = O_p(n^{1/2}), C(\widehat{\vartheta}_0) = O_p(n),$$
$$C(\bar{\vartheta}_0) = C(\widehat{\vartheta}_0) + o_p(n) = O_p(n).$$

Let the LM statistic (Breusch and Pagan, 1980) be

$$LM = s_\theta(\widehat{\vartheta}_{m0}) \left[-\mathbf{J}_{\theta\theta}(\widehat{\vartheta}_{m0}) \right] s_\theta(\widehat{\vartheta}_{m0}),$$

where $\widehat{\vartheta}_{m0} = (\theta_0, \widehat{\psi}_{m0})$ is the ML estimator of ϑ under the null hypothesis, $s_\theta(\vartheta)$ is the score function corresponding to θ, $\mathbf{J}_{\theta\theta}(\vartheta)$ is the submatrix of $\mathbf{J}(\vartheta)$ corresponding to θ.

Theorem 2.3.1. *Under Assumptions 1-5, we can show that*

$$\mathbf{T}(\mathbf{y}, \theta_0) = s_\theta(\widehat{\vartheta}_0) \left[-L_{n,\theta\theta}^{-(2)}(\widehat{\vartheta}) \right] s_\theta(\widehat{\vartheta}_0) + o_p(1), \tag{2.3.5}$$

where $L_{n,\theta\theta}^{-(2)}$ is the submatrix of $L_n^{-(2)}(\vartheta)$ corresponding to θ. Furthermore, when the null hypothesis is true and the likelihood dominates the prior, we have

$$\mathbf{T}(\mathbf{y}, \theta_0) = LM + o_p(1) \xrightarrow{d} \chi^2(p). \tag{2.3.6}$$

Remark 2.3.8. *From Equation (2.3.6), $\mathbf{T}(\mathbf{y}, \theta_0)$ may be regarded as the Bayesian version of the LM statistic. However, the LM test is a frequentist test which is based on ML estimation of the model in the null hypothesis whereas our test is a Bayesian test which is based on the posterior quantities of the models under both the null hypothesis as well as the alternative hypothesis.*

Remark 2.3.9. *In Theorem 2.3.1, we can see that under the null hypothesis, the asymptotic distribution of $\mathbf{T}(\mathbf{y}, \theta_0)$ always follows the χ^2 distribution and, hence, is independent of the nuisance parameters. This suggests that the new test is asymptotically pivotal, a property that compares favorably with the use of the subjective* *threshold values as in BR (2002) and LY (2012).*

Remark 2.3.10. *When the likelihood dominates the prior, the posterior mode, $\widehat{\vartheta}$, reduces to the ML estimator of ϑ under the alternative hypothesis, and the posterior mode, $\widehat{\vartheta}_0 = (\theta_0, \widehat{\psi}_0)$, reduces to the ML estimator of ϑ under the null hypothesis. From Equation (2.3.5), we can see that*

$$\mathbf{T}(\mathbf{y}, \theta_0) = s_\theta(\widehat{\vartheta}_0)\left[-L_{n,\theta\theta}^{-(2)}(\widehat{\vartheta})\right] s_\theta(\widehat{\vartheta}_0) + o_p(1) = -s_\theta(\widehat{\vartheta}_0)\left[\mathbf{J}_{\theta\theta}(\widehat{\vartheta})\right] s_\theta(\widehat{\vartheta}_0) + o_p(1).$$

If the null hypothesis is false, according to the standard ML theory, we get

$$\mathbf{J}(\vartheta_0) = \mathbf{J}(\widehat{\vartheta}) + o_p(n^{-1}) \neq \mathbf{J}(\widehat{\vartheta}_0) + o_p(n^{-1}).$$

except that $\mathbf{J}(\vartheta)$ is independent on ϑ. This is because, under the alternative, $\widehat{\vartheta}$ is a consistent estimator of ϑ whereas $\widehat{\vartheta}_0$ is not.

$$\begin{aligned}
\mathbf{T}(\mathbf{y}, \theta_0) &= -s_\theta(\widehat{\vartheta}_0)'\mathbf{J}_{\theta\theta}(\widehat{\vartheta}) s_\theta(\widehat{\vartheta}_0) + o_p(1) \\
&\neq -s_\theta(\widehat{\vartheta}_0)'\mathbf{J}_{\theta\theta}(\widehat{\vartheta}_0) s_\theta(\widehat{\vartheta}_0) + o_p(1) \\
&= \mathbf{LM} + o_p(1).
\end{aligned}$$

Remark 2.3.11. $\mathbf{T}(\mathbf{y}, \theta_0)$ *can incorporate the prior information to improve statistical inference when the sample size is small. This property is shared by the BF but not by the LM test. To illustrate the idea, consider a simple example, where $y_1, ..., y_n \sim N(\theta, \sigma^2)$ with a known variance $\sigma^2 = 1$. The true value of θ is set at $\theta_0 = 0.25$. The prior distribution of θ is set as $N(\mu_0, \tau^2)$. The simple point null hypothesis is $H_0 : \theta = 0$. It can be shown that*

$$\begin{aligned}
2\log BF_{10} &= \frac{(n\bar{y}\tau^2 + \mu_0\sigma^2)^2}{(\sigma^2+\tau^2)(\sigma^2\tau^2)} + \log\frac{\sigma^2}{n\tau^2+\sigma^2}, \\
\mathbf{T}(\mathbf{y}, \theta_0) &= \frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}\left[\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau^2}\right]^2, LM = \frac{n\bar{y}^2}{\sigma^2},
\end{aligned}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$. When $n \longrightarrow \infty$, $\mathbf{T}(\mathbf{y}, \theta_0) \longrightarrow LM$ and the asymptotic distribution *for both* $\mathbf{T}(\mathbf{y},\theta_0)$ *and LM is* $\chi^2(1)$. *Let us consider the case that corresponds to an informative prior* $N(0.25,10^{-4})$ *and compare it to the case that corresponds to a non-informative prior* $N(0,10^4)$. *Table 1 reports* $2\log BF_{10}$, $T(\mathbf{y},\theta_0)$, *and LM when* $n = 10, 100, 1000, 10000$ *under these two priors. It can be seen that both the BF and the new test depend on the prior (although the BF tends to choose the wrong model under the vague prior even when the sample size is very large) while the LM test is independent of the prior. When* $n = 10$, $\mathbf{T}(\mathbf{y},\theta_0)$ *correctly rejects the null hypothesis when the prior is informative but fails to reject it when the prior is vague. In this case, the LM test fails to reject the null hypothesis under both priors.*[1]

Table 2.1: Comparison of $2\log BF_{10}$, $\mathbf{T}(\mathbf{y},\theta_0)$, and LM

Prior	$N(0.25,10^{-4})$				$N(0,10^4)$			
n	10	100	1000	10000	10	100	1000	10000
$2\log BF_{10}$	624.69	643.11	753.13	2601.01	-11.56	-13.45	-15.87	-18.16
$\mathbf{T}(\mathbf{y},\theta_0)$	624.13	636.81	684.81	1300.98	0.025	13.03	59.84	676.49
LM	0.025	13.03	59.84	676.49	0.025	13.03	59.84	676.49

Remark 2.3.12. *It is well known that the BF is conservative compared to the likelihood ratio test; see, for example, Edwards et al. (1963), Kass and Raftery (1995), Li, et al (2014). Our test is also less conservative than the BF since it is asymptotically pivotal. To illustrate this property, we consider the example in Remark 3.12 of Li, et al (2014). Let* $y_1,...,y_n \sim N(\theta,1)$. *The prior distribution of* θ *can be set as* $N(0,\tau^2)$. *We want to test the simple point null hypothesis* $H_0 : \theta = 0$. *Suppose* $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i = \sqrt{6.634897/n}$ *so that the critical level of the LM test is always kept at 99%. In this case, it can be shown that* $2\log BF_{10} = \frac{n\tau^2}{n\tau^2+1}(\sqrt{n}\bar{y})^2 - \log(n\tau^2+1)$, $\mathbf{T}(\mathbf{y},\theta_0) = \frac{n\tau^2}{n\tau^2+1}(\sqrt{n}\bar{y})^2$ *and LM=* $(\sqrt{n}\bar{y})^2$. *According to Fisher's scale, we have*

[1]To implement the LM test,we use the following Fisher's scale. Let α be the critical level and $P = 1-\alpha$. If P is between 95% and 97.5%, the evidence for the alternative is "moderate"; between 97.5% and 99%, "substantial"; between 99% and 99.5%, "strong"; between 99.5% and 99.9%, "very strong"; larger than 99.9%, "overwhelming". To implement the BF we use Jeffreys' scale instead. If $\log BF_{10}$ is less than 0, there is "negative" evidence for the alternative; between 0 and 1, "not worth more than a bare mention"; between 1 and 3, "positive"; between 3 and 5, "strong"; larger than 5, "very strong".

"strong" evidence for the alternative hypothesis based on the LM test. Table 2 reports $2\log BF_{10}$, $\mathbf{T}(\mathbf{y},\theta_0)$, *LM when* $\tau=1$. *It can be seen that the BF finds the evidence for the alternative hypothesis to be "positive" when* $n=10$. *The evidence turns to be "not worth more than a bare mention" when* $n=100$, *but to "negative" when* $n=1000,10000$. *This result is consistent with the conservative property of the BF relative to the LM test. In the meantime, our test statistic is slightly more conservative than the LM test although the difference is smaller and the two statistics converge to each other as the sample size grows. When the user is conservative and has a highly informative prior, we caution against the idea of basing the hypothesis testing solely on the proposed test.*

Table 2.2: Comparison of $2\log BF_{10}$, $\mathbf{T}(\mathbf{y},\theta_0)$, and LM when the prior distribution of θ is $N(0,1)$ and $\bar{y}=\sqrt{6.634897/n}$ so that the critical level of LM is always 99%.

n	10	100	1000	10000
$2\log BF_{10}$	3.63383	1.95408	-0.28049	-2.57621
Decision	positive	not worth mention	negative	negative
$\mathbf{T}(\mathbf{y},\theta_0)$	6.03170	6.56920	6.62830	6.63420
LM	6.63490	6.63490	6.63490	6.63490

The implementation of the LM test requires the ML estimation of the null model. When it is hard to do the ML estimation, it will be difficult to calculate the LM statistic. This is the case for many models that involve latent variables. However, as long as the Bayesian MCMC methods are applicable, our test can be implemented. Moreover, our method offers two additional advantages over the LM test, which we explain below.

Remark 2.3.13. *We have shown that when the alternative hypothesis is correct, our test statistic is not close to the LM test. In this case, our test continues to take a nonnegative value whereas the LM test can take a negative value. This is because, in our test, the weight matrix* $C_{\theta\theta}(\bar{\vartheta}_0)$ *remains at least semi-positive definite so that* $\mathbf{T}(\mathbf{y},\theta_0)$ *is not negative. When* θ_0 *is further away from the true value of* θ, $s_\theta(\bar{\vartheta}_0)$

and $C_{\theta\theta}(\bar{\vartheta}_0)$ will be further way from zero. Consequently, $\mathbf{T}(\mathbf{y},\theta_0)$ will be larger so that it can discriminate H_0 against H_1. Whereas, when θ_0 is further away from the true value of θ, the weight matrix $-\mathbf{J}(\widehat{\vartheta}_{m0})$ in the LM statistic may not be positive definite. This may cause some difficulties in using the LM test.

To illustrate the remark, consider the following example where $y_t \sim N(0,\sigma^2), t = 1,2\cdots,n$, and the true value of σ^2 is 0.1. We would like to test

$$H_0 : \sigma^2 = 1, H_1 : \sigma^2 \neq 1.$$

In this case, we have

$$\mathbf{I}(\vartheta) = \mathbf{I}(\sigma^2) = \frac{\partial^2 \log p(\mathbf{y}|\sigma^2)}{\partial\sigma^2\partial\sigma^2} = \frac{n}{2\sigma^4} - \frac{\sum_{t=1}^n y_t^2}{\sigma^6}.$$

When n is large enough, we know that $\sum_{t=1}^n y_t^2/n \approx 0.1$ and, hence,

$$\begin{aligned}
\mathbf{I}(\widehat{\vartheta}_{m0}) &= \mathbf{I}(\sigma^2 = 1) = \frac{n}{2} - \sum_{t=1}^n y_t^2 = \frac{n}{2}\left(1 - 2\frac{\sum_{t=1}^n y_t^2}{n}\right) \approx 0.4n > 0, \\
-\mathbf{J}(\widehat{\vartheta}_{m0}) &= \frac{1}{-\mathbf{I}(\widehat{\vartheta}_{m0})} = -\frac{1}{0.4n} < 0.
\end{aligned}$$

Consequently, the LM statistic is negative. Whereas, for our statistic, we have

$$\begin{aligned}
C_{\theta\theta}(\bar{\vartheta}_0) &= \frac{1}{4}\left(n - \sum_{t=1}^n y_t^2\right)^2, \bar{\sigma}^2 = \int \sigma^2 p(\sigma^2|\mathbf{y})\mathrm{d}\sigma^2, \\
\mathbf{T}(\mathbf{y},\sigma^2 = 1) &= \int (\sigma^2 - \bar{\sigma}^2)^2 C_{\theta\theta}(\bar{\vartheta}_0) p(\sigma^2|\mathbf{y})\mathrm{d}\sigma^2.
\end{aligned}$$

Hence, the proposed test does not suffer from the same problem as the LM test.

Remark 2.3.14. *The implementation of the LM test requires the inversion of $-\mathbf{I}(\vartheta_0)$. When the dimension of ϑ is high, such an inversion may be numerically challenging. Whereas, to calculate $\mathbf{T}(\mathbf{y},\theta_0)$, one does not need to invert any matrix.*

2.4 Empirical Illustrations

In this section, we illustrate the proposed test statistic using three popular examples in economics and finance. The first example is a simple linear regression model where the BF and our proposed test statistic both have analytical expressions. We hope to compare them and pay particular attention to their sensitivity with respect to the prior. The second example is a probit model. In this example, the observed data likelihood is available in closed-form, facilitating the comparison of the BF, the LM and our proposed test. We consider both the joint and individual point null hypothesis tests. The third example is a stochastic conditional duration (SCD) model, where the duration is latent. In this example, the analytical expression of the observed data likelihood does not exist so that the implementation of the LM test is very difficult. Hence, we only compare the BF and our proposed test. However, it is difficult to compute the NSE of the BF in this example.

2.4.1 Hypothesis testing in linear regression models

The first example is the simple linear regression model:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \;\; \varepsilon_i \sim i.i.d.\, N\left(0, \sigma^2\right), i = 1, ..., n. \tag{2.4.1}$$

We would like to test $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$. Assume that the prior distributions for (α, β) and σ^2 are normal and inverse gamma, respectively,

$$(\alpha, \beta)' \sim N\left(\tilde{\mu}, \sigma^2 \tilde{V}\right), \quad \sigma^2 \sim IG(a, b),$$

where $\tilde{\mu} = \left(\mu_\alpha, \mu_\beta\right)'$, $\tilde{V} = diag\left(V_\alpha, V_\beta\right)$.

The marginal likelihood for the model under H_0 is given by

$$p\left(\mathbf{y}|M_0\right) = \frac{b^a \Gamma\left(a + \frac{n}{2}\right)}{(2\pi)^{\frac{n}{2}} \Gamma(a)} \sqrt{\frac{1}{nV_\alpha + 1}} \left[b + \frac{1}{2}\left(\left(\mathbf{y} - \beta_0 \mathbf{x}\right)' \left(\mathbf{y} - \beta_0 \mathbf{x}\right) + \frac{\mu_\alpha^2}{V_\alpha} - \frac{\mu_\alpha^*}{V_\alpha^*}\right)\right]^{-\left(a + \frac{n}{2}\right)},$$

where $V_\alpha^* = \frac{V_\alpha}{nV_\alpha+1}$, $\mu_\alpha^* = V_\alpha^*\left(\sum_{i=1}^n (y_i - \beta_0 x_i) + \frac{\mu_\alpha}{V_\alpha}\right) = V_\alpha^*\left(\iota'(\mathbf{y} - \beta_0\mathbf{x}) + \frac{\mu_\alpha}{V_\alpha}\right)$ with $n \times 1$ vector $\iota = (1,...,1)'$. The marginal likelihood for the model under H_1 is given by

$$p(\mathbf{y}|M_1) = \frac{b^a\sqrt{|V^*|}\Gamma\left(a+\frac{n}{2}\right)}{(2\pi)^{\frac{n}{2}}\Gamma(a)\sqrt{|\tilde{V}|}}\left[b+\frac{1}{2}\left((\tilde{\mu})'\tilde{V}^{-1}\tilde{\mu}+\mathbf{y}'\mathbf{y}-(\mu^*)'V^{*-1}\mu^*\right)\right]^{-\left(a+\frac{n}{2}\right)},$$

where $V^* = \left(\tilde{V}^{-1} + X'X\right)^{-1}$, $\mu^* = V^*\left(\tilde{V}^{-1}\tilde{\mu} + X'\mathbf{y}\right)$, $X = (\iota, \mathbf{x})$. The derivation is given in Appendix .1.4. Hence, in this simple model, $\mathbf{BF}_{10} = p(\mathbf{y}|M_1)/p(\mathbf{y}|M_0)$ has an analytical expression. Furthermore, the analytical expression of the proposed statistic can be given by

$$\mathbf{T}(\mathbf{y}, \beta_0) = \frac{2sV_{22}^*}{\nu - 2}C_{\theta\theta}\left(\bar{\vartheta}_0\right),$$

where $C_{\theta\theta}\left(\bar{\vartheta}_0\right) = \frac{1}{\bar{\sigma}_0^4}\left[\mathbf{x}'(\mathbf{y} - \bar{\alpha}_0\iota - \beta_0\mathbf{x})\right]^2$, $\bar{\sigma}_0^4 = \left(\bar{\sigma}_0^2\right)^2$, $\bar{\alpha}_0$ and $\bar{\sigma}_0^2$ are the posterior means of α and σ^2 under H_0, $\nu = 2a+n$, $s = \frac{1}{\nu}\left[b + \frac{1}{2}\left((\tilde{\mu})'\tilde{V}^{-1}\tilde{\mu} + \mathbf{y}'\mathbf{y} - (\mu^*)'V^{*-1}\mu^*\right)\right]$ and V_{22}^* is the submatrix of V^* corresponding to β. The derivation is also given in the same Appendix.

We now analyze a model in Brooks (2008, Page 40) where the return on a spot price is linked to the return on a futures price, i.e.,

$$\Delta\log(s_t) = \alpha + \beta\Delta\log(f_t) + \varepsilon_t,\ \varepsilon_t \sim i.i.d.N(0, \sigma^2),$$

where $\Delta\log(s_t)$ is the log-difference of the spot S&P500 index and $\Delta\log(f_t)$ is the log-difference of the S&P500 futures price, and β captures the optimal hedge ratio. We would like to test if $\beta = \beta_0 = 1$.

The hyperparameters are set at

$$\mu_a = 0, V_a = 10^3, \mu_\beta = 0, a = 0.001, b = 0.001.$$

In addition, we allow the prior variance of β, V_β, to vary so that we can examine how the prior influences the BF and $\mathbf{T}(\mathbf{y},\beta_0)$. Since both the priors and the likelihood function are in the Normal-Gamma form, we can directly draw samples from their posterior joint distributions under H_0 and H_1. In particular, 35,000 random draws are sampled from the posterior distributions for Bayesian statistical inference.

Table 3 reports $\log BF_{10}$, $\mathbf{T}(\mathbf{y},\beta_0)$, the posterior means and the posterior standard errors of all the parameters under H_1 for different values of V_β. From Table 3, we observe that the posterior quantities of all three parameters are robust to V_β. However, $\log BF_{10}$ is very sensitive to V_β. In particular, $\log BF_{10}$ decreases as V_β increases. When the prior variance V_β is moderate, $\log BF_{10}$ is more than 0 and tends to reject the null hypothesis. When V_β is sufficiently large, $\log BF_{10}$ is less than 0 and does not reject the null hypothesis. This observation clearly demonstrates that the BF is subject to Jeffreys-Lindley's paradox. On the contrary, $\mathbf{T}(\mathbf{y},\beta_0)$ takes nearly identical values with different V_β. Therefore, $\mathbf{T}(\mathbf{y},\beta_0)$ is immune to Jeffreys-Lindley's paradox. The asymptotic distribution of $\mathbf{T}(\mathbf{y},\beta_0)$ under H_0 is $\chi^2(1)$, and the 99.9 percentile of $\chi^2(1)$ is 10.83. $\mathbf{T}(\mathbf{y},\beta_0)$ is much larger than 10.83 in all cases, suggesting that the null hypothesis is rejected under the 99.9% probability level.

To investigate the sensitivity of our proposed test statistic and BF, Table 4 reported $\log BF_{10}$, $T(\mathbf{y},\beta_0)$, the posterior mean and the posterior standard error of all the parameters under different values of (a,b) given the prior hyperparameters $(\mu_a = 0, V_a = 10^3, \mu_\beta = 0, V_\beta = 10^{12})$. The results clearly show the sensitivity of the BF to the prior because the BF values change the sign. In the contrast, our test statistic does not change a lot and always supports the alternative hypothesis.

2.4.2 Hypothesis testing in discrete choice models

The probit model is widely used to analyze binary choice data. In this section, we fit the probit model to a dataset originally used in Mroz (1987). Since the observed data likelihood in the probit model is available in closed-form, we can directly compute the proposed Bayesian test statistic $\mathbf{T}(\mathbf{y},\theta_0)$ based on the MCMC output. Also, the LM test can be easily obtained.

Table 2.3: $\log BF_{10}$, $\mathbf{T}(\mathbf{y}, \beta_0)$, the posterior means and standard errors of ϑ under H_1

	$V_\beta = 0.1$	$V_\beta = 100$	$V_\beta = 10^5$	$V_\beta = 10^{22}$	$V_\beta = 10^{25}$	$V_\beta = 10^{35}$
$\log BF_{10}$	14.7354	11.2948	7.8409	-11.7311	-15.1849	-26.6979
$\mathbf{T}(\mathbf{y}, \theta_0)$	14.9596	15.1693	15.1696	15.1696	15.1696	15.1696
$\bar{\beta}$	0.1220	0.1232	0.1248	0.1243	0.1236	0.1233
$SE(\bar{\beta})$	0.1331	0.1338	0.1334	0.1347	0.1343	0.1337
$\bar{\alpha}$	0.3603	0.3633	0.3587	0.3655	0.3634	0.3622
$SE(\bar{\alpha})$	0.4438	0.4445	0.4423	0.4449	0.4477	0.4435
$\bar{\sigma}^2$	12.5972	12.5792	12.5790	12.5621	12.5741	12.5817
$SE(\bar{\sigma}^2)$	2.2913	2.2768	2.2941	2.2693	2.2785	2.2936

Table 2.4: $\log BF_{10}$, $\mathbf{T}(\mathbf{y}, \beta_0)$, the posterior means and standard errors of ϑ under different hyperparameters pairs (a, b)

	$(0.001, 0.001)$	$(0.1, 0.1)$	$(0.1, 0.01)$	$(1, 0.1)$	$(2, 0.001)$
$\log BF_{10}$	-0.21813	-0.17002	-0.16703	0.29718	0.81978
$\mathbf{T}(\mathbf{y}, \theta_0)$	15.2158	15.2628	15.2635	15.8318	16.3386
$\bar{\beta}$	0.12378	0.12378	0.12378	0.12291	0.12295
$SE(\bar{\beta})$	0.13399	0.1338	0.13378	0.13249	0.13056
$\bar{\alpha}$	0.35948	0.35948	0.35948	0.36327	0.36223
$SE(\bar{\alpha})$	0.13399	0.1338	0.13378	0.13249	0.13056
$\bar{\sigma}^2$	12.5948	12.5584	12.5556	12.2152	11.84
$SE(\bar{\sigma}^2)$	2.2909	2.2805	2.28	2.1805	2.0787

Table 2.5: Bayesian and ML estimates and their standard errors

	Bayesian Method		ML Method	
	Posterior Mean	Posterior SE	Estimate	SE
ϑ_0	0.2576	0.5125	0.2701	0.5086
ϑ_1	-1.2146×10^{-2}	4.8169×10^{-3}	-1.2024×10^{-2}	4.8398×10^{-3}
ϑ_2	0.1323	2.5451×10^{-2}	0.1309	2.5254×10^{-2}
ϑ_3	0.1242	1.8706×10^{-2}	0.1233	1.8716×10^{-2}
ϑ_4	-1.9×10^{-3}	6.0366×10^{-4}	-1.8871×10^{-3}	6×10^{-4}
ϑ_5	-5.3083×10^{-2}	8.4437×10^{-3}	-5.2853×10^{-2}	8.4772×10^{-3}
ϑ_6	-0.8752	0.1187	-0.8683	0.1185
ϑ_7	3.7766×10^{-2}	4.2809×10^{-2}	3.6005×10^{-2}	4.3477×10^{-2}

In the probit model, we take the married women's labor force participation ($inlf$) as the binary dependent variable (y) and $nwifeinc, educ, exper, expersq, age, kedslt6$, and $kidsge6$ are taken as independent variables; see Wooldridge (2002) for detailed explanation of these variables. The latent variable representation of the model is given by

$$z = \vartheta_0 + \vartheta_1 nwifeinc + \vartheta_2 educ + \vartheta_3 exper + \vartheta_4 expersq + \vartheta_5 age + \vartheta_6 kedslt6 + \vartheta_7 kidsge6 + e,$$

where z is the latent variable, e follows a standard normal distribution, and $inlf$ takes value 1 if $z > 0$, and 0 otherwise.

Proper but vague priors are used for all the regression coefficients. Specifically, each element of ϑ is assumed to follow the normal distribution with mean 0 and variance 10^8. In this example, we test a joint point null hypothesis and an individual point null hypothesis. In particular, we test whether $exper$ and $expersq$ have the joint explanatory power for y and whether $kidsge6$ has the explanatory power for y. Hence, the null hypothesis is $\vartheta_7 = 0$ in the individual test and $\vartheta_3 = \vartheta_4 = 0$ in the joint test.

Following Koop (2003), 35,000 draws are obtained using the Gibbs sampler under H_0 and H_1 with the first 10,000 samples discarded as burning-in samples. The convergence of Markov chains is monitored using the statistic of Heidelberger and Welch (1983). The parameter estimates and their corresponding standard errors under H_1 for both the Bayesian method and the ML method are reported in Table 2.5. For the Bayesian method, we report the posterior means and the posterior standard errors. For the ML method, we report the ML estimates and the asymptotic standard errors. Clearly, the difference between the two sets of results is small.

Since $\mathbf{T}(\mathbf{y},\theta_0)$ does not have a closed-form expression, we can obtain its estimate, $\widehat{\mathbf{T}}(\mathbf{y},\theta_0)$, from the MCMC outputs. The estimate and the NSE (in the bracket) are reported in Table 2.6. Since the observed likelihood function has an analytical expression, the LM test can be easily obtained and is reported in Table 2.6. In addition, the estimator of $\log BF_{10}$ and its NSE are also reported in Table 2.6. The details about the derivation of these statistics are given in Appendix .1.5.

For the individual test, the asymptotic distribution of $\mathbf{T}(\mathbf{y},\theta_0)$ under H_0 is $\chi^2(1)$ whose 95 percentile is 3.8415. According to $\widehat{\mathbf{T}}(\mathbf{y},\theta_0)$ and the LM statistic, the hypothesis $\vartheta_7 = 0$ cannot be rejected, suggesting that *kidsge*6 does not have a significant explanatory power on y. Furthermore, these two values are very close to each other, consistent with the result in Theorem 3.1. What is more, the BF also strongly support the null hypothesis, reinforcing the conclusion drawn from the other two statistics. The NSEs of the new test and $\widehat{\log BF_{10}}$ are of smaller order of magnitude than the corresponding statistics.

For the joint test, the asymptotic distribution of $\mathbf{T}(\mathbf{y},\theta_0)$ under H_0 is $\chi^2(2)$ whose 99.99 percentile is 18.42. $\widehat{\mathbf{T}}(\mathbf{y},\theta_0)$ is much larger than 18.42, suggesting that the null hypothesis is rejected under the 99.99% probability level. Similarly,the LM statistic is much larger than the 99.99 percentile of $\chi^2(2)$ and rejects the null hypothesis. The BF also strongly supports the alternative hypothesis. The three statistics all provide the "strong" evidence that *exper* and *expersq* have the joint explanatory power on y. Furthermore, the difference between $\widehat{\mathbf{T}}(\mathbf{y},\theta_0)$ and the LM statistic is significant. It suggests that these two test statistics may differ significantly when the null hypothesis is not held, consistent with Remark 3.10. The NSEs of the new test and $\widehat{\log BF_{10}}$ are of smaller order of magnitude than the corre-

Table 2.6: The proposed test statistic, the LM test statistic, $\widehat{\log BF_{10}}$, and the numerical standard errors of the proposed test statistic and $\widehat{\log BF_{10}}$ (in the bracket)

H_0	$\widehat{\mathbf{T}}(\mathbf{y},\ \theta_0)$	LM	$\widehat{\log BF_{10}}$
$\vartheta_7 = 0$	0.6805 (0.0204)	0.6861	-12.1454 (0.0226)
$\vartheta_3 = \vartheta_4 = 0$	126.7931 (3.7603)	99.088	21.9721 (0.021)

sponding statistics.

2.4.3 Hypothesis testing in stochastic conditional duration models

The third example is a simple extension of the stochastic conditional duration (SCD) model of Bauwens and Veredas (2004) given by

$$\begin{cases} d_t = \exp(\varphi_t)\,\varepsilon_t & \varepsilon_t \sim Exp(1), \\ \varphi_t = \phi\varphi_{t-1} + \alpha + x_t'\beta + \sigma\epsilon_t & \epsilon_t \sim N(0,1), \\ \varphi_1 \sim N\left(\frac{\alpha + x_1'\beta}{1-\phi}, \frac{\sigma^2}{1-\phi^2}\right), \end{cases}$$

for $t = 1, ..., T$. In this model, d_t is the adjusted duration; φ_t is the latent variable which is potentially serially correlated and $|\phi|$ is assumed to be less than 1; $\beta = (\beta_1, \beta_2)'$, $x_t' = (P_{t-1}, VOL_{t-1})$, where P_{t-1} is the price of the underlying stock at time $t-1$ and VOL_{t-1} is the trading volume of the stock at time $t-1$; ε_t and ϵ_t are independent random errors.

The data, collected from the TAQ database, are the time intervals (durations) between transactions for IBM between September 3, 1996 and September 30, 1996. Following Bauwens and Veredas (2004), the transaction data before 9:30 and after 16:00 are excluded and the simultaneous trades are treated as one single transaction. As a result, we are left with 17,103 raw durations.

Following Engle and Russell (1998), we adjust the raw durations using the daily season factor $\Psi(t_i)$ which is assumed to be a cubic spline with each node being the

Table 2.7: The posterior means and posterior standard errors of all the parameters under the three null hypotheses and the alternative hypothesis for the SCD model

Parameter	α		ϕ		σ^2		β_1		β_2	
Hypothesis	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
H_1	.1147	.0364	.9473	.0061	.0209	.0028	-.1105	.0363	-.0099	.0015
$H_0 : \beta_1 = \beta_2 = 0$	-.0052	.0014	.9523	.0059	.0204	.0028	-	-	-	-
$H_0 : \beta_1 = 0$	.0039	.0018	.9498	.0049	.0204	.0023	-	-	-.0093	.0015
$H_0 : \beta_2 = 0$	.0849	.0354	.9504	.0055	.0208	.0025	-0.0904	.0356	-	-

average duration on each half hour from 9:30 to 16:00, i.e.,

$$d_{t_i} = \frac{D_{t_i}}{\Psi(t_i)},$$

where D_{t_i} is the raw durations. Similar adjustments are also made to the prices and the volumes. We first test whether or not the price and the traded volume at time $t-1$ have a joint impact on the duration at time t, i.e., $\beta_1 = \beta_2 = 0$. Furthermore, we also test whether the individual effect is significant or not, i.e., $\beta_1 = 0$ and $\beta_2 = 0$.

Because the observed-data likelihood function is not available in closed-form, it is very hard to calculate the LM statistic even for the model under the null hypothesis. However, since the complete-data likelihood function has an analytical expression, the data augmentation technique facilitates the Bayesian MCMC estimation of the models. As a result, the proposed statistic is easy to calculate and the detailed derivation of $\widehat{\mathbf{T}}(\mathbf{d}, \theta_0)$ is reported in Appendix .1.6. The prior distributions for parameters are given as follows,

$$\phi = 2\phi^* - 1,\ \phi^* \sim Beta(1,1),\ \sigma^2 \sim IG(0.01, 0.01),$$

$$(\alpha, \beta')' | \sigma^2 \sim N(\mathbf{0}, 100 \times \sigma^2 I_3).$$

where I_3 is 3×3 identity matrix. 55,000 MCMC draws are obtained with the first 15,000 being treated as the burn-in samples. Again, we use the statistic of Heidelberger and Welch (1983) to check the convergence of all the chains. The posterior means and posterior standard errors of all the parameters under the three null hypotheses and the alternative hypothesis are reported in Table 2.7.

Table 2.8: The proposed test statistic, $\widehat{\log BF_{10}}$, their computing time (in seconds), and the numerical standard errors of the proposed test statistic (in the bracket)

	$\beta_1 = \beta_2 = 0$	$\beta_1 = 0$	$\beta_2 = 0$
$\widehat{\mathbf{T}}(\mathbf{d}, \theta_0)$	17.8312 (0.6262)	2.3209 (0.2979)	14.8087 (0.4107)
Time for $\widehat{\mathbf{T}}(\mathbf{d}, \theta_0)$(s)	4116.1709	4634.9620	3840.8727
$\widehat{\log BF_{10}}$	17.9863	0.8196	18.9603
Time for $\log BF_{10}$(s)	6889.1324	6913.0687	6510.5200

Table 2.8 reports the values of the new statistic and the BF and the computing time (in seconds) of the new test in the three cases. For hypotheses $\beta_1 = \beta_2 = 0$ and $\beta_2 = 0$, $\widehat{\mathbf{T}}(\mathbf{d}, \theta_0)$ strongly reject the null hypothesis, even under the 99.9% probability level. This is consistent with the BFs, which also strongly support the model under H_1. For hypothesis $\beta_1 = 0$, the BF does not find strong evidence for the alternative hypothesis with "not worth more than a bare mention" evidence. Our proposed statistic also fails to reject the null hypothesis at the 95% probability level.

Finally, from Table 2.8, we can show that the new statistic takes less time to compute than the BF. Moreover, the NSEs of the new test are of a smaller order of magnitude than the corresponding statistics. However, the NSEs of the BFs are difficult to obtain because the log-likelihood is not available in closed-form for the SCD model.

2.5 Conclusion

In this paper, we have proposed a new Bayesian test statistic to test a point null hypothesis based on a quadratic loss function. Under the null hypothesis and a set of regularity conditions, we show that our test is asymptotically equivalent to frequentist's LM test and follows a chi-squared distribution asymptotically. The proposed method is illustrated using a simple linear regression model, a discrete choice model and a stochastic conditional duration model.

The main advantages of the proposed test statistic are as follows. Relative to the BF, (i) it is well-defined under improper prior distributions; (ii) it is immune to Jeffreys-Lindley's paradox; (iii) it is easy to compute, even for the latent variable models; (iv) its asymptotic distribution is pivotal so that the threshold values are easy to obtain; (v) its NSE can be easily obtained. Relative to the LM test, (i) it can incorporate the prior information to improve hypothesis testing when the sample size is small; (ii) it does not suffer from the problem of taking negative values; (iii) it does not need to invert any matrix.

Chapter 3 A Posterior-Based Wald-Type Statistic for Hypothesis Testing

3.1 Introduction

This paper develops an approach to test a point null hypothesis based on the Bayesian posterior distribution. The statistic can be understood as the posterior version of the well-known Wald statistic that has been used widely in practical applications. The Wald statistic is often based on the maximum likelihood estimator (MLE) or the classical extremum estimators (denoted by $\widehat{\theta}$) of the parameter(s) of interest (denoted θ). Typically one kind of squared difference between $\widehat{\theta}$ and θ is shown to follow a χ^2 distribution asymptotically under the null hypothesis, producing an asymptotically pivotal test.

However, in many practical applications, the MLE or the classical extremum estimators may be too difficult to obtain computationally. For example, for the entire class of non-linear and non-Gaussian state space models, the likelihood function is very hard to calculate numerically, making the MLE nearly impossible to obtain. Not surprisingly, Bayesian MCMC methods have emerged as the leading estimation tool to deal with non-linear and non-Gaussian state space models. There are many other examples in economics where the classical extremum estimators are subject to the curse of dimensionality in computation and some numerical problems. To circumvents this problem, Chernozhukov and Hong (2003) introduced a class of quasi-Bayesian methods that allow users to employ MCMC to simulate a random sequence of draws such that the marginal distribution of the sequence is the same as the quasi-posterior distribution of parameters.

The central question we ask in this paper is how to test a point null hypothesis with the posterior distribution of parameters being available. Testing a point null hypothesis is important for checking statistical evidence from data to support or to be against a particular theory because theory often can be reduced to a testable hypothesis. In many cases, the posterior distribution of parameters is available in the form of a random sample (such as MCMC sample).

Broadly speaking, there are three posterior-based methods available in the literature for hypothesis testing. The first one is the Bayes factor (BF) which compares the posterior odds of the two competing theories corresponding to the null and alternative hypotheses (Kass and Raftery, 1995). Unfortunately, BFs are subject to a few theoretical and practical problems. First, BFs are not well-defined under improper priors. Second, BFs are subject to Jeffreys-Lindley's paradox. That is, they tend to choose the null hypothesis when a very vague prior is used for parameters in the null hypothesis; see Kass and Raftery (1995), Poirier (1995). Third, the calculation of BFs generally involves evaluation of marginal likelihood. In many cases, evaluation of marginal likelihood is difficult. Several strategies have been proposed in the literature to address some of these difficulties. For example, to deal with the first two problems, when calculating BFs one may use a highly informative prior which is data-dependent. To make it data-dependent, one may split the data into two parts, one as a training set, the other for statistical analysis. The training data can be used to update a prior (whether it is improper or vague) to generate a proper informative prior which is subsequently used to analyze the remaining data. See the fractional BF of O'Hagan (1995), and the intrinsic BF of Berger (1985). To address the computational problem, one can use the methods of Chib (1995) and Chib and Jeliazkov (2001) to compute BFs.

The second posterior-based method is to use credible intervals for point identified parameters and credible sets for partially identified parameters. This line of approaches has drawn a great deal of attentions among econometricians and statisticians in recent years; see Chernozhukov and Hong (2003), Moon and Schorfheide (2012), Norets and Tang (2013), Kline and Tamer (2016), Liao and Simoni (2015), Chen, et al (2016). Except Chernozhukov and Hong (2003), all the other studies focus in developing credible sets in partially identified models. Most of these studies justify credible sets using large-sample theory under repeated sampling.

The third method is based on the statistical decision theory. The idea begins with Bernardo and Rueda (2002, BR hereafter) where they demonstrated that the BF can be regarded as a decision problem with a simple zero-one loss function when it is used for point hypothesis testing. It is this zero-one loss that leads to Jeffreys-Lindley's paradox. BR further suggested using the continuous Kullback-Leibler (KL) divergence function as the loss functions to replace the zero-one loss. Subsequent extensions include Li and Yu (2012), Li, Zeng and Yu (2014) and Li, Liu and Yu (2015, LLY hereafter) where different continuous loss functions or net loss functions were used. The justification of these extensions is made by large-sample theory under repeated sampling.

In this paper, following the third line of approach, we propose a Wald-type statistic for hypothesis testing based on posterior distributions. The new statistic is well-defined under improper prior distributions and avoids Jeffreys-Lindley's paradox. It is asymptotically equivalent to the Wald statistic under the null hypothesis, and hence, follows a χ^2 distribution asymptotically. It is a by-product of posterior simulation, requiring almost no coding effort and little computational cost.

The paper is organized as follows. Section 2 reviews existing posterior-based statistics for hypothesis testing in the statistical decision framework. Section 3 develops the new statistic and establishes its large-sample theory. Section 4 explains how to implement the proposed test for an important class of models – latent variable models – where posterior analysis is routinely used. Section 5 investigates finite-sample properties of the proposed statistic using simulated data. Section 6 gives two real-data applications of the proposed method. Section 7 concludes the paper. Appendix collects the proof of theoretical results.

3.2 Hypothesis Testing based on Statistical Decision

It is assumed that a probability model $M \equiv \{p(y|\vartheta)\}$ is used to fit data $y := (y_1, \ldots y_n)'$ where $\vartheta := (\theta', \psi')' \in \Theta$. We are concerned with testing a point null hypothesis which may arise from the prediction of a particular theory. Let $\theta \in \Theta_\theta$ denote a vector of q_θ-dimensional parameters of interest and $\psi \in \Theta_\psi$ a vector of q_ψ-dimensional nuisance parameters, where $\Theta = \Theta_\theta \times \Theta_\psi$. The testing problem is given by

$$\begin{cases} H_0: & \theta = \theta_0, \\ H_1: & \theta \neq \theta_0. \end{cases} \tag{3.2.1}$$

In the statistical decision framework, hypothesis testing may be understood as follows. There are two statistical decisions in the decision space, accepting H_0 (name it d_0) or rejecting H_0 (name it d_1). Let $\{\mathscr{L}[d_i, \theta, \psi], i = 0, 1\}$ be the loss function of the statistical decision associated with d_i. When the expected posterior loss of accepting H_0 is sufficiently larger than the expected posterior loss of rejecting H_0, we reject H_0. That is, H_0 is rejected if

$$\begin{aligned} T(y, \theta_0) &= \int_{\boldsymbol{\Theta}} \{\mathscr{L}(d_0, \theta, \psi) - \mathscr{L}(d_1, \theta, \psi)\} p(\theta, \psi|y) \mathrm{d}\theta \mathrm{d}\psi \\ &= \int_{\boldsymbol{\Theta}} \triangle\mathscr{L}(H_0, \theta, \psi) p(\theta, \psi|y) \mathrm{d}\theta \mathrm{d}\psi \\ &= E_{\vartheta|y}(\triangle\mathscr{L}(H_0, \theta, \psi)) > c \geq 0, \end{aligned}$$

where $T(y, \theta_0)$ is a posterior-based statistic, $p(\theta, \psi|y)$ is the posterior distribution, c is a threshold value, $\triangle\mathscr{L}(H_0, \theta, \psi) := \mathscr{L}(d_0, \theta, \psi) - \mathscr{L}(d_1, \theta, \psi)$ is the net loss function.

BR showed that when the equal prior $p(\theta = \theta_0) = p(\theta \neq \theta_0) = \frac{1}{2}$ is used, $c = 0$, and the net loss function is taken as

$$\Delta\mathscr{L}(H_0, \theta, \psi) = \begin{cases} -1, & \text{if } \theta = \theta_0 \\ 1, & \text{if } \theta \neq \theta_0 \end{cases},$$

then $T(y, \theta_0) > 0$ is equivalent to the following decision rule based on the BF: reject H_0 if

$$\mathrm{BF}_{10} = \frac{p(y|H_1)}{p(y|H_0)} = \frac{\int p(y, \vartheta) \mathrm{d}\vartheta}{\int p(y, \psi|\theta_0) \mathrm{d}\psi} > 1.$$

While the BF serves as the gold standard for model comparison after posterior distributions are obtained for candidate models, it suffers from several theoretical and computational difficulties when it is used to test a point null hypothesis. First, it is not well-defined under improper priors. Second, it leads to Jeffreys-Lindley's paradox when a very vague prior is used. Third, BF_{10} requires evaluating the two marginal likelihood functions, $p(y|H_i), i = 0, 1$. Clearly, this involves marginalizations over ψ and over ϑ. Fourth, if ϑ is high-dimensional so that the integration is a high-dimensional problem, calculating $p(y|H_i), i = 0, 1$ will be difficult numerically although there have been several interesting methods proposed in the literature to compute the BF from MCMC output; see, for example, Chib (1995), and Chib and Jeliazkov (2001).

In the statistical decision framework, several statistics have been proposed for testing a point null hypothesis. Poirier (1997) developed a loss function approach for hypothesis testing for models without latent variables. BR (2002) suggested choosing the loss function to be the KL divergence function. The large-sample theory of the test statistics of BR has not been developed although it is well-defined under improper priors and can solve Jeffreys-Lindley's paradox.[1]

In a recent paper, LLY (2015) proposed the following quadratic net loss function

$$\Delta\mathscr{L}(H_0, \theta, \psi) = \left(\theta - \bar{\theta}\right)' C_{\theta\theta}\left(\bar{\vartheta}_0\right)\left(\theta - \bar{\theta}\right), C(\vartheta) = \left\{\frac{\partial \log p(y, \vartheta)}{\partial \theta}\right\}\left\{\frac{\partial \log p(y, \vartheta)}{\partial \theta}\right\}',$$

where $\bar{\vartheta} = \left(\bar{\theta}', \bar{\psi}'\right)'$ and $\bar{\vartheta}_0 = \left(\theta_0', \bar{\psi}_0'\right)'$ are the posterior mean under H_0 and H_1, respectively, $C_{\theta\theta}$ is the submatrix of C corresponding to θ. The statistic correspond-

[1]Given that the KL function is not analytically available for most latent variable models, Li and Yu (2012) suggested basing the loss function on the Q-function used in the EM algorithm. However, its large-sample theory has not been developed. On the other hand, Li, Zeng and Yu (2014) suggested using the deviance function to be the loss function. large-sample theory of the test statistic is derived. Unfortunately, in general the asymptotic distribution depends on some unknown population parameters and hence the test is not pivotal asymptotically.

ing to this net loss function is given by

$$T_{LLY}(y,\theta_0) = E_{\vartheta|y}(\triangle\mathscr{L}(H_0,\theta,\psi)) = \int_{\Theta} \left(\theta-\bar{\theta}\right)' C_{\theta\theta}\left(\bar{\vartheta}_0\right)\left(\theta-\bar{\theta}\right) p(\vartheta|y)\mathrm{d}\vartheta. \tag{3.2.2}$$

Under repeated sampling, LLY showed that $T_{LLY}(y,\theta_0)$ follows a χ^2 distribution asymptotically, providing an asymptotically pivotal quantity. This statistic is well-defined under improper priors and immune to Jeffreys-Lindley's paradox. Clearly, $T_{LLY}(y,\theta_0)$ requires evaluating the first-order derivative of the (observed-data) likelihood function. In some models, especially in latent variable models, this first-order derivative is not easy to evaluate since the observed-data likelihood function may not have an analytical expression. Another feature of $T_{LLY}(y,\theta_0)$ is that it requires estimating both the null model and the alternative model although, under H_0, it was shown to be asymptotically equivalent to the Lagrange Multiplier (LM) test which requires estimating the null model only.

3.3 A Posterior Wald-type Statistic

3.3.1 The statistic based on a quadratic loss function

For any $\tilde{\vartheta} \in \Theta$, denote

$$V\left(\tilde{\vartheta}\right) = E\left[\left(\vartheta-\tilde{\vartheta}\right)\left(\vartheta-\tilde{\vartheta}\right)'|y,H_1\right] = \int \left(\vartheta-\tilde{\vartheta}\right)\left(\vartheta-\tilde{\vartheta}\right)' p(\vartheta|y)\mathrm{d}\vartheta.$$

We propose the following net loss function for hypothesis testing:

$$\triangle\mathscr{L}[H_0,\theta,\psi] = (\theta-\theta_0)'\left[V_{\theta\theta}(\bar{\vartheta})\right]^{-1}(\theta-\theta_0),$$

where $V_{\theta\theta}\left(\tilde{\vartheta}\right)$ is the submatrix of $V\left(\tilde{\vartheta}\right)$ corresponding to θ, $\left[V_{\theta\theta}\left(\tilde{\vartheta}\right)\right]^{-1}$ the inverse of $V_{\theta\theta}\left(\tilde{\vartheta}\right)$, and $\bar{\vartheta}$ the posterior mean of ϑ under H_1. Then, the new test statistic can be defined as:

$$T(y,\theta_0) = \int (\theta-\theta_0)' \left[V_{\theta\theta}\left(\bar{\vartheta}\right)\right]^{-1}(\theta-\theta_0)\,p(\vartheta|y)\mathrm{d}\vartheta = tr\left[\left[V_{\theta\theta}\left(\bar{\vartheta}\right)\right]^{-1} V_{\theta}(\theta_0)\right], \tag{3.3.1}$$

where $V_\theta(\theta_0) := \int (\theta-\theta_0)(\theta-\theta_0)' p(\vartheta|y)\mathrm{d}\vartheta$.

Remark 3.3.1. *It is easy to see show that $T(y,\theta_0)$ is well-defined under improper priors. An improper prior $p(\vartheta)$ satisfies that $p(\vartheta) = af(\vartheta)$ where $f(\vartheta)$ is a non-integrable function and a is an arbitrary positive constant. Since the posterior distribution $p(\vartheta|y)$ is independent of a, $V_{\theta\theta}\left(\tilde{\vartheta}\right)$, being the posterior covariance matrix of θ, is also independent of a. Hence, the proposed statistic does not depend on a.*

Remark 3.3.2. *To see how the new statistic can avoid Jeffreys-Lindley's paradox, consider the example used in LLY (2015). Let $y_1, y_2, \ldots, y_n \sim N(\theta, \sigma^2)$ with a known σ^2, the null hypothesis be $H_0 : \theta = 0$, the prior distribution of θ be $N(0, \tau^2)$. Denote $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. It is easy to show that the posterior distribution of θ is $N(\mu(y), \omega^2)$ with*

$$\mu(y) = \frac{n\tau^2\bar{y}}{\sigma^2+n\tau^2}, \omega^2 = \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2},$$

and

$$\begin{aligned} 2\log BF_{10} &= \frac{n\tau^2}{n\tau^2+\sigma^2}\frac{n\bar{y}^2}{\sigma^2} + \log\frac{\sigma^2}{n\tau^2+\sigma^2}, \\ T(y,\theta_0) &= \frac{n\tau^2}{n\tau^2+\sigma^2}\frac{n\bar{y}^2}{\sigma^2} + 1. \end{aligned}$$

Thus, when $\tau^2 \to +\infty$ (the prior information becomes more and more uninformative), $\log BF_{10} \to -\infty$ which suggest that the BF supports H_0 regardless how much $\bar{y}$ is. This is exactly what Jeffreys-Lindley's paradox predicts. On the other hand, $T(y,\theta_0) \to \frac{n\bar{y}^2}{\sigma^2} + 1$ as $\tau^2 \to +\infty$. Hence, $T(y,\theta_0)$ is distributed asymptotically as $\chi^2(1)+1$ when H_0 is true, suggesting that $T(y,\theta_0)$ is immune to Jeffreys-Lindley's paradox.

3.3.2 Large-sample theory for $T(y, \theta_0)$

In this subsection, we establish large-sample properties for $T(y, \theta_0)$ under repeated sampling. Let $y^t := (y_0, y_1, \ldots, y_t)$ for any $0 \le t \le n$ and $l_t(y^t, \vartheta) = \log p(y^t|\vartheta) - \log p(y^{t-1}|\vartheta)$ be the conditional log-likelihood for the t^{th} observation for any $1 \le t \le n$. When there is no confusion, we just write $l_t(y^t, \vartheta)$ as $l_t(\vartheta)$ so that the log-likelihood function $\mathscr{L}_n(\vartheta)$ ($:= \log p(y|\vartheta)$ conditional on the initial observation), can be written as $\sum_{t=1}^n l_t(\vartheta)$. Let $l_t^{(j)}(\vartheta)$ be the j^{th} derivative of $l_t(\vartheta)$ and $l_t^{(0)}(\vartheta) = l_t(\vartheta)$. Moreover, let

$$
\begin{aligned}
& s(y^t, \vartheta) := \frac{\partial \log p(y^t|\vartheta)}{\partial \vartheta} = \sum_{i=1}^{t} l_i^{(1)}(\vartheta),\ h(y^t, \vartheta) := \frac{\partial^2 \log p(y^t|\vartheta)}{\partial \vartheta \partial \vartheta'} = \sum_{i=1}^{t} l_i^{(2)}(\vartheta), \\
& s_t(\vartheta) := l_t^{(1)}(\vartheta) = s(y^t, \vartheta) - s(y^{t-1}, \vartheta),\ h_t(\vartheta) := l_t^{(2)}(\vartheta) = h(y^t, \vartheta) - h(y^{t-1}, \vartheta), \\
& \bar{H}_n(\vartheta) := \frac{1}{n}\sum_{t=1}^{n} h_t(\vartheta), \bar{J}_n(\vartheta) := \frac{1}{n}\sum_{t=1}^{n} \left[s_t(\vartheta) - \bar{s}_t(\vartheta)\right]\left[s_t(\vartheta) - \bar{s}_t(\vartheta)\right]', \bar{s}_t(\vartheta) = \frac{1}{n}\sum_{t=1}^{n} s_t(\vartheta), \\
& \mathscr{L}_n^{[j]}(\vartheta) := \partial^j \log p(\vartheta|y)/\partial \vartheta^j, H_n(\vartheta) := \int \bar{H}_n(\vartheta) g(y)\, dy,\ J_n(\vartheta) := \int \bar{J}_n(\vartheta) g(y)\, dy.
\end{aligned}
$$

$H_n(\vartheta)$ and $J_n(\vartheta)$ are generally known as the Hessian matrix and the Fisher information matrix; $\bar{H}_n(\vartheta)$ and $\bar{J}_n(\vartheta)$ are the empirical Hessian matrix and empirical Fisher information matrix.

In this paper, we first impose the following regularity conditions. A similar set of assumptions was used in Li, et al (2017).

Assumption 1: $\Theta \subset \mathbb{R}^q$ where $q = q_\theta + q_\psi$ is compact.

Assumption 2: $\{y_t\}_{t=1}^{\infty}$ satisfies the strong mixing condition with the mixing coefficient $\alpha(m) = O\left(m^{\frac{-2r}{r-2} - \varepsilon}\right)$ for some $\varepsilon > 0$ and $r > 2$.

Assumption 3: For all t, $l_t(\boldsymbol{\vartheta})$ satisfies the standard measurability and continuity condition, and the eight-times differentiability condition on $F_{-\infty}^t \times \Theta$ where $F_{-\infty}^t = \sigma(y_t, y_{t-1}, \cdots)$.

Assumption 4: For $j = 0, 1, 2$, for any $\vartheta, \vartheta' \in \Theta$, $\left\| l_t^{(j)}(\vartheta) - l_t^{(j)}(\vartheta') \right\| \le c_t^j(y^t) \left\| \vartheta - \vartheta' \right\|$ in probability, where $c_t^j(y^t)$ is a positive random variable with $\sup_t E\left\| c_t^j(y^t) \right\| < \infty$ and $\frac{1}{n}\sum_{t=1}^n \left(c_t^j(y^t) - E\left(c_t^j(y^t) \right) \right) \xrightarrow{p} 0$.

Assumption 5: For $j = 0, 1, 2, 3$, there exists a function $M_t(y^t)$ such that for all $\vartheta \in \Theta$, $l_t^{(j)}(\vartheta)$ exists, $\sup_{\vartheta \in \Theta} \left\| l_t^{(j)}(\vartheta) \right\| \leqslant M_t(y^t)$, and $\sup_t E \|M_t(y^t)\|^{r+\delta} \leq M < \infty$ for some $\delta > 0$, where r is the same as that in Assumption 2.

Assumption 6: $\left\{ l_t^{(j)}(\vartheta) \right\}$ is L_2-near epoch dependent with respect to $\{y_t\}$ of size -1 for $0 \leqslant j \leqslant 1$ and $-\frac{1}{2}$ for $j = 2$ uniformly on Θ.

Assumption 7: Let ϑ_n^0 be the value that minimizes the KL loss between the DGP and the candidate model

$$\vartheta_n^0 = \arg\min_{\vartheta \in \Theta} \frac{1}{n} \int \log \frac{g(y)}{p(y|\vartheta)} g(y) dy,$$

where $\left\{ \vartheta_n^0 \right\}$ is the sequence of minimizers interior to Θ uniformly in n. For all $\varepsilon > 0$,

$$\lim_{n \to \infty} \sup \sup_{\Theta \backslash N\left(\vartheta_n^0, \varepsilon\right)} \frac{1}{n} \sum_{t=1}^{n} \left\{ E\left[l_t(\vartheta) \right] - E\left[l_t\left(\vartheta_n^0\right) \right] \right\} < 0, \tag{3.3.2}$$

where $N\left(\vartheta_n^0, \varepsilon\right)$ is the open ball of radius ε around ϑ_n^0.

Assumption 8: The sequence $\left\{ H_n\left(\vartheta_n^0\right) \right\}$ is negative definite.

Assumption 9: The prior density $p(\vartheta)$ is three-times continuously differentiable, $p\left(\vartheta_n^0\right) > 0$ and $\int \|\vartheta\|^2 p(\vartheta) d\vartheta < \infty$.

Remark 3.3.3. *An important condition for the asymptotic posterior normality is the consistency condition which means that, for each $\varepsilon > 0$, there exists $K(\varepsilon) > 0$ such that*

$$\lim_{n \to \infty} P\left(\sup_{\Theta \backslash N\left(\vartheta_n^0, \varepsilon\right)} \frac{1}{n} \sum_{t=1}^{n} \left[l_t(\vartheta) - l_t\left(\vartheta_n^0\right) \right] < -K(\varepsilon) \right) = 1; \tag{3.3.3}$$

see Heyde and Johnstone (1979), Schervish (2012), Ghosh and Ramamoorthi (2003). If Assumptions 1-7 hold true, then (3.3.3) holds, as shown in Li et al. (2017).

Remark 3.3.4. *According to Li et al. (2017), if Assumptions 1-9 hold true, then for* *each $\varepsilon > 0$, there exists $K(\varepsilon) > 0$ such that*

$$\lim_{n\to\infty} P\left(\sup_{\Theta\backslash N\left(\widehat{\vartheta}_{m},\varepsilon\right)} \frac{1}{n}\left[\sum_{t=1}^{n} l_t(\vartheta) - \sum_{t=1}^{n} l_t\left(\vartheta_n^0\right)\right] < -K(\varepsilon)\right) = 1, \tag{3.3.4}$$

where $\widehat{\vartheta}_m$ is the posterior mode of ϑ. Li et al. (2017) showed that this is sufficient to ensure that the concentration condition around the posterior mode given by Chen (1985).

Lemma 3.3.1. *Let $\widehat{\vartheta}$ be the MLE of ϑ and $N_0(\delta) = \left\{\boldsymbol{\vartheta} : \left\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_n^0\right\| \le \delta\right\}$. If Assumptions 1-7 hold true, then for any $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ such that*

$$P\left(\sup_{N_0(\delta(\varepsilon))} \left|\bar{H}_n(\vartheta) - \bar{H}_n\left(\widehat{\vartheta}\right)\right| < \varepsilon\right) \to 1. \tag{3.3.5}$$

and

$$P\left(\sup_{N_0(\delta(\varepsilon)),\|r_0\|=1} \left|1 - r_0'\bar{H}_n^{-1/2}\left(\widehat{\boldsymbol{\vartheta}}\right)\bar{H}_n(\boldsymbol{\vartheta})\bar{H}_n^{-1/2}\left(\widehat{\boldsymbol{\vartheta}}\right)r_0\right| < \varepsilon\right) \to 1.$$

where r_0 is q-dimension vector.

Let $\Sigma_n = -\frac{1}{n}\bar{H}_n^{-1}(\widehat{\vartheta})$ and $z_n = \Sigma_n^{-1/2}\left(\vartheta - \widehat{\vartheta}\right)$. Lemma 3.3.2 below gives the order of the difference between the first k moments of the posterior distribution of z_n under H_1 and those of a standard multivariate normal distribution. To establish the closeness of higher order moments between the two distribution, we have to strengthen Assumption 9 by Assumption 9B. In Assumption 9B, the k-th order moment of the prior distribution is assumed to be finite.

Assumption 9B: The prior density $p(\vartheta)$ is three-times continuously differentiable, $p\left(\vartheta_n^0\right) > 0$ and $\int \|\vartheta\|^k p(\vartheta)\,d\vartheta < \infty$ for integer some $k \ge 1$.

Lemma 3.3.2. *Under Assumptions 1-8 and Assumption 9B, it can be shown that*

$$E\left[z_n^{\{k\}}|y,H_1\right] = MN_q^{\{k\}} + o_p(1),$$

where $E\left[z_n^{\{k\}}|y,H_1\right]$ *is the k-th order moments of the posterior distribution of* z_n *under* H_1 *(i.e.* $z_n|y,H_1$*), and* $MN_q^{\{k\}}$ *is the k-th order moments of a standard multivariate normal distribution with dimension q. When* $k=1,2$*, i.e., Assumption 9 holds, we can have*

$$\overline{\vartheta} = E[\vartheta|y,H_1] = \widehat{\vartheta} + o_p(n^{-1/2}), \tag{3.3.6}$$

$$V\left(\widehat{\vartheta}\right) = E\left[\left(\vartheta-\widehat{\vartheta}\right)\left(\vartheta-\widehat{\vartheta}\right)'|y,H_1\right] = -\frac{1}{n}\bar{H}_n^{-1}\left(\widehat{\vartheta}\right) + o_p(n^{-1}) \tag{3.3.7}$$

Remark 3.3.5. *Under different regularity conditions, the Bernstein-von Mises theorem shows that the posterior distribution converges to a normal distribution with the MLE as its mean and the inverse of the empirical Hessian matrix evaluated at the MLE as its covariance. Based on the Bernstein-von Mises theorem, when the parameter is one-dimension, Ghosh and Ramamoorthi (2003) developed the same results as Lemma 3.3.2 for the i.i.d. case. Hence, Lemma 3.3.2 extends the results of Ghosh and Ramamoorthi (2003) in three aspects: (1) to the weakly dependent case; (2) to the multivariate case; (3) to show that the order of the difference in high-order moments between the posterior distribution and a normal distribution.*

Remark 3.3.6. *Assumptions 1-9 are weaker than those used in Li, et al. (2017) where a high order Laplace expansion was developed. With the high order Laplace expansion, Li, et al. (2018) derived the exact order for the difference in the first and second moments*

$$\bar{\vartheta} = E[\vartheta|y,H_1] = \widehat{\vartheta} + O_p(n^{-1}), \tag{3.3.8}$$

$$V\left(\widehat{\vartheta}\right) = E\left[\left(\vartheta-\widehat{\vartheta}\right)\left(\vartheta-\widehat{\vartheta}\right)'|y,H_1\right] = -\frac{1}{n}\bar{H}_n^{-1}\left(\widehat{\vartheta}\right) + O_p(n^{-2}) \tag{3.3.9}$$

Clearly, (3.3.8) and (3.3.9) are a stronger set of results than (3.3.6) and (3.3.7). Lemma 3.3.2 is sufficient to develop large-sample properties of the proposed statistic. Hence, we can relax the assumptions of Li, et al (2017).

Let $\widehat{\theta}$ be the subvector of $\widehat{\vartheta}$ corresponding to θ. The Wald statistic is

$$\text{Wald} = n\left(\widehat{\theta}-\theta_0\right)'\left[-\bar{H}_{n,\theta\theta}^{-1}\left(\widehat{\vartheta}\right)\right]^{-1}\left(\widehat{\theta}-\theta_0\right), \tag{3.3.10}$$

where $\bar{H}_{n,\theta\theta}^{-1}\left(\widehat{\vartheta}\right)$ is the submatrix of $\bar{H}_n^{-1}\left(\widehat{\vartheta}\right)$ corresponding to θ and $\bar{H}_n^{-1}\left(\widehat{\vartheta}\right)$ is the inverse of $\bar{H}_n\left(\widehat{\vartheta}\right)$.

Theorem 3.3.1. *Under Assumptions 1-9, we can show that, under the null hypothesis,*

$$T\left(y,\theta_0\right)-q_\theta = Wald + o_p(1),$$

and

$$T\left(y,\theta_0\right)-q_\theta \xrightarrow{d} \chi^2(q_\theta).$$

Remark 3.3.7. *From Theorem 3.3.1, $T\left(y,\theta_0\right)-q_\theta$ may be regarded as the posterior version of the Wald statistic. It shares the same asymptotic distribution as the Wald test under the null hypothesis. However, the Wald statistic is based on the MLE of the alternative model, whereas the proposed test is based on the posterior mean and variance under the alternative hypothesis.*

Corollary 3.3.2. *Under Assumptions 1-9, we have, under the null hypothesis,*

$$T\left(y,\theta_0\right)-q_\theta = T_{LLY}(y,\theta_0)+o_p(1) \xrightarrow{d} \chi^2(q_\theta).$$

Remark 3.3.8. *LLY (2015) has established the relationship between $T_{LLY}(y,\theta_0)$ and the LM test statistic, i.e., $T_{LLY}(y,\theta_0)$ =LM+$o_p(1)$ under the null hypothesis. It is noted in Engle (1984) that under the null hypothesis LM =Wald+$o_p(1)$. So Corollary 3.3.2 is the posterior version of this asymptotic equivalence between the Wald and LM statistics.*

Remark 3.3.9. *Theorem 3.3.1 suggests that the asymptotic distribution of $T\left(y,\theta_0\right)$ is pivotal. To implement the proposed test, we can choose the threshold value, c, to* *be the critical value of $\chi^2(q_\theta)$ distribution, i.e.,*

$$Accept\ \ H_0 \text{ if } T(y,\theta_0) - q_\theta \le c;\ Reject\ \ H_0 \ \text{ if } T(y,\theta_0) - q_\theta > c.$$

Remark 3.3.10. *It is obvious that $T(y,\theta_0)$ only requires evaluating the inverse of the submatrix of the covariance matrix corresponding to θ and, thus, it is very easy to compute. In contrast, the Wald statistic in (3.3.10) requires evaluating the inverse of the entire empirical Hessian matrix and then use the submatrix corresponding to θ. When ϑ is high-dimensional, this inversion is numerically more involved than the inversion of the submatrix. For example,, consider the case where the dimension of ϑ is 100, but the null hypothesis involves only one of the parameters. To use the Wald statistic, one has to evaluate the inverse of a 100×100 dimensional Hessian matrix. Whereas, to use $T(y,\theta_0)$, one only needs to evaluate the inverse of a scalar.*

Remark 3.3.11. *Compared with the Wald statistic, the proposed statistic can incorporate the prior information through the posterior distribution. To illustrate the influence of prior distribution, let $y_1,...,y_n \sim N(\theta,\sigma^2)$ with a known variance $\sigma^2 = 1$. The true value of θ is set at $\theta_0 = 0.10$. The prior distribution of θ is set as $N(\mu_0,\tau^2)$. We wish to test $H_0 : \theta = 0$. It can be shown that*

$$\begin{aligned} 2\log BF_{10} &= \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)^2 + \log\frac{\sigma^2}{\sigma^2+n\tau^2}, \\ T(y,\theta_0)-1 &= \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)^2, \\ Wald &= \frac{n\bar{y}^2}{\sigma^2}, \end{aligned}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$. When $n \to \infty$, $T(y,\theta_0) - 1 - Wald \xrightarrow{p} 0$ and the asymptotic distribution for both $T(y,\theta_0) - 1$ and the Wald statistic is $\chi^2(1)$. Suppose two prior distributions are used, a highly informative prior $N(0.10,10^{-3})$ and a very vague prior $N(0,10^{50})$. Table 1 reports $2\log BF_{10}$, $T(y,\theta_0) - 1$, and Wald when $n = 10,100,1000,10000$ under these two priors. It can be seen that $T(y,\theta_0) - 1$ and Wald take identical values when the vague prior is used. It is consistent with *the prediction of our asymptotic theory. Moreover, both the BF and the new statistic depend on the prior (although the BFs tend to choose the wrong model under the vague prior even when the sample size is very large) while the Wald test is independent of the prior. When $n = 10, 100$, $T(y, \theta_0) - 1$ correctly rejects the null hypothesis when the prior is informative but fails to reject it when the prior is vague under the 5% significance level. In this case, the Wald test fails to reject the null hypothesis.*

Table 3.1: Comparison of $2\log BF_{10}$, $T(y, \theta_0) - 1$, and the Wald statistic

Prior	$N(0.10, 10^{-3})$				$N(0, 10^{50})$			
n	10	100	1000	10000	10	100	1000	10000
$2\log BF_{10}$	9.96	11.12	20.60	93.58	-117.42	-118.50	-110.72	-38.00
$T(y, \theta_0) - 1$	9.96	11.22	21.30	95.98	0.01	1.23	11.32	86.03
Wald	0.01	1.23	11.32	86.03	0.01	1.23	11.32	86.03

Remark 3.3.12. *Assumption 9 requires finiteness of the first and second moments of the posterior distribution. When improper priors satisfies this assumption, Theorem 3.3.1 holds. In practice, however, many improper priors do not have finite first and second moments and hence Assumption is violated. In addition, Assumption 9 excludes the Jeffreys prior (Jeffreys, 1961) since the Jeffreys prior depends on the sample size n. If informative priors are not available, we suggest using vague non-informative priors (a prior with large variance spread) to implement our proposed tests. For more details about vague noninformative priors, one can refer to Kass and Raftery (1995).*

3.3.3 Extension to hypotheses in a general form

In this subsection, we extend the point null hypothesis to the following nonlinear form,

$$\begin{cases} H_0: & R(\theta_0) = r \\ H_1: & R(\theta_0) \neq r \end{cases}, \tag{3.3.11}$$

where $R(\cdot) : \Theta_\theta \to \mathbb{R}^m$, $m \leq q$, and $r \in \mathbb{R}^m$. Here R is a set of m nonlinear functions/restrictions. We can test for a single hypothesis on multiple parameters, as well as a jointly multiple hypotheses on single/multiple parameters. While this hypothesis problem is in the standard form for the Wald test, it makes BFs difficult to implement due to nonlinear relationships among parameters. To develop large-sample properties of the proposed test, we need to impose the following assumption on $R(\theta)$.

Assumption 10: $R(\theta)$ is second-order continuously differentiable with respect to θ on Θ and full rank at θ_0 .

For the hypothesis defined in (3.3.11), the classical Wald statistic and its asymptotic theory are

$$\text{Wald} = \left[R\left(\widehat{\theta}\right) - r\right]' \left\{ \frac{\partial R\left(\widehat{\theta}\right)}{\partial \theta'} \left[-\bar{H}_{n,\theta\theta}^{-1}\left(\widehat{\vartheta}\right)\right] \frac{\partial R\left(\widehat{\theta}\right)}{\partial \theta} \right\}^{-1} \left[R\left(\widehat{\theta}\right) - r\right] \xrightarrow{d} \chi^2(m).$$

Based on the statistical decision theory, we can define the following net loss function

$$\Delta \mathscr{L}(H_0, \theta, \psi) = (R(\theta) - r)' \left[\frac{\partial R\left(\overline{\theta}\right)}{\partial \theta'} V_{\theta\theta}\left(\overline{\vartheta}\right) \frac{\partial R\left(\overline{\theta}\right)}{\partial \theta}\right]^{-1} (R(\theta) - r),$$

and introduce our test statistic as:

$$\begin{aligned} T(y,r) &= \int_{\Theta} \Delta \mathscr{L}(H_0, \theta, \psi)\, p(\vartheta|y)\, d\vartheta \\ &= \int_{\Theta} (R(\theta) - r)' \left[\frac{\partial R\left(\overline{\theta}\right)}{\partial \theta'} V_{\theta\theta}\left(\bar{\vartheta}\right) \frac{\partial R\left(\overline{\theta}\right)}{\partial \theta}\right]^{-1} (R(\theta) - r)\, p(\vartheta|y)\, d\vartheta \\ &= tr\left[\left(\frac{\partial R\left(\overline{\theta}\right)}{\partial \theta'} V_{\theta\theta}\left(\bar{\vartheta}\right) \frac{\partial R\left(\overline{\theta}\right)}{\partial \theta}\right)^{-1} V_{\theta}(r)\right], \end{aligned} \tag{3.3.12}$$

where $V_{\theta}(r) = \int (R(\theta) - r)(R(\theta) - r)' p(\vartheta|y)\, d\vartheta$.

Theorem 3.3.3. *Under Assumptions 1-10, , we can show that, under the null hypothesis,*

$$T(y,r) - m = \mathit{Wald} + o_p(1) \xrightarrow{d} \chi^2(m).$$

3.3.4 Calculating the proposed statistic

As noted in Sections 3.2 and 3.3, the proposed statistics are only dependent on the posterior mean and the posterior variance of ϑ, i.e., $\overline{\vartheta}$ and $V\left(\overline{\vartheta}\right)$. In practice, $\overline{\vartheta}$ and $V\left(\overline{\vartheta}\right)$ are often unknown analytically. Fortunately, when random samples from the posterior distribution $p(\vartheta|y)$ are obtained via posterior simulation (such as MCMC or importance sampling), we can consistently estimate $\overline{\vartheta}$ and $V\left(\overline{\vartheta}\right)$ arbitrarily well. Specifically, let $\{\vartheta^{[j]}, j=1,2,\ldots,J\}$ be effective samples generated from $p(\vartheta|y)$, consistent estimates of $\overline{\vartheta}$ and $V\left(\overline{\vartheta}\right)$ are given by

$$\bar{\bar{\vartheta}} = \frac{1}{J}\sum_{j=1}^{J}\vartheta^{[j]},\ \bar{V}\left(\bar{\bar{\vartheta}}\right) = \frac{1}{J}\sum_{j=1}^{J}\left(\vartheta^{[j]}-\bar{\bar{\vartheta}}\right)\left(\vartheta^{[j]}-\bar{\bar{\vartheta}}\right)'.$$

By plugging $\bar{\bar{\vartheta}}$ and $\bar{V}\left(\bar{\bar{\vartheta}}\right)$ into the proposed statistics, we obtain a consistent estimate of $T\left(y,\theta_0\right)$ or $T\left(y,r\right)$ as

$$\begin{aligned}\widehat{T}\left(y,\theta_0\right) &:= tr\left[\left(\bar{V}_{\theta\theta}\left(\bar{\bar{\vartheta}}\right)\right)^{-1}\bar{V}_{\theta}\left(\theta_0\right)\right],\\ \widehat{T}\left(y,r\right) &:= tr\left[\left(\frac{\partial R\left(\bar{\bar{\theta}}\right)}{\partial\theta'}\bar{V}_{\theta\theta}\left(\bar{\bar{\vartheta}}\right)\frac{\partial R\left(\bar{\bar{\vartheta}}\right)}{\partial\theta}\right)^{-1}\bar{V}_{\theta}(r)\right],\end{aligned} \tag{3.3.13}$$

where

$$\bar{V}_{\theta}\left(\theta_0\right) = \frac{1}{J}\sum_{j=1}^{J}\left(\theta^{[j]}-\theta_0\right)\left(\theta^{[j]}-\theta_0\right)',$$

and

$$\bar{V}_{\theta}(r) = \frac{1}{J}\sum_{j=1}^{J}\left(R\left(\theta^{[j]}\right)-r\right)\left(R\left(\theta^{[j]}\right)-r\right)'.$$

Remark 3.3.13. *Various approaches have been developed for posterior simulation. Examples include Monte Carlo (MC) integration, important sampling, MCMC techniques such as the Gibbs sampler and the Metropolis-Hastings algorithm. For more details about posterior simulation, one can refer to Geweke (2005). All these approaches can be used to generate the random observations from $p(\vartheta|y)$. From (3.3.13), the proposed statistics are by-products of posterior simulation. Further-*

more, the test statistics can be applied in a variety of models.

When $\widehat{T}(y,\theta_0)$ and $\widehat{T}(y,r)$ are calculated from posterior simulation, it is important to obtain their numerical standard error (NSE) which measures the magnitude of simulation errors. The following theorem provides formulae to calculate the NSE of $\widehat{T}(y,\theta_0)$ and $\widehat{T}(y,r)$.

Theorem 3.3.4. *Let* $\bar{v}_1 = \frac{1}{J}\sum_{j=1}^{J}\theta^{[j]}$, $\bar{V}_2 = \frac{1}{J}\sum_{j=1}^{J}\left(\theta^{[j]} - \bar{\bar{\theta}}\right)\left(\theta^{[j]} - \bar{\bar{\theta}}\right)'$, $\bar{v}_2 = vech(\bar{V}_2)$, $\bar{v} = (\bar{v}_1', \bar{v}_2')'$, $Var(\bar{v})$ *be the NSE of* $\bar{v}$, *where vech denotes the columnwise vectorization of a matrix. The NSE of* $\widehat{T}(y,\theta_0)$ *is given by*

$$NSE\left(\widehat{T}(y,\theta_0)\right) = \sqrt{\left(\frac{\partial \widehat{T}(y,\ \theta_0)}{\partial \bar{v}}\right)' Var(\bar{v}) \frac{\partial \widehat{T}(y,\theta_0)}{\partial \bar{v}}},$$

where

$$\begin{aligned}\frac{\partial \widehat{T}(y,\theta_0)}{\partial \bar{v}} =& vech\left(I_{q_\theta}\right)' \left[\left(\left((\bar{v}_1-\theta_0)'\bar{V}_2^{-1}\right)' \otimes I_{q_\theta} + \bar{V}_2^{-1}\otimes(\bar{v}_1-\theta_0)\right)\frac{\partial \bar{v}_1}{\partial \bar{v}}\right. \\ &\left. - \left[I_{q_\theta}\otimes(\bar{v}_1-\theta_0)(\bar{v}_1-\theta_0)'\right]\left(\bar{V}_2^{-1}\otimes\bar{V}_2^{-1}\right)\frac{\partial \bar{V}_2}{\partial \bar{v}}\right].\end{aligned}$$

and

$$\frac{\partial \bar{v}_1}{\partial \bar{v}} = \frac{\partial \bar{v}_1'}{\partial \bar{v}} = \left[I_{q_\theta}, 0_{q_\theta\times q^*}\right], \frac{\partial \bar{V}_2}{\partial \bar{v}} = \left[0_{q_\theta^2\times q_\theta}, \left(\frac{\partial vech(\bar{V}_2)}{\partial \bar{v}_2}\right)_{q_\theta^2\times q^*}\right].$$

Furthermore, if $R(\theta)$ *is second-order continuously differentiable, the NSE of* $\widehat{T}(y,r)$ *is given by*

$$NSE\left(\widehat{T}(y,r)\right) = \sqrt{\left(\frac{\partial \widehat{T}(y,r)}{\partial \bar{v}}\right)' Var(\bar{v}) \frac{\partial \widehat{T}(y,r)}{\partial \bar{v}}},$$

where

$$\begin{aligned}\frac{\partial \widehat{T}(y,r)}{\partial \bar{v}} =& vech(I_m)' \left\{ \left[\left((\bar{v}_3 - r)' \left(\bar{V}_4' \bar{V}_2 \bar{V}_4 \right)^{-1} \right)' \otimes I_m \right] \frac{\partial \bar{v}_3}{\partial \bar{v}_1} \frac{\partial \bar{v}_1}{\partial \bar{v}} \right. \\ & + \left[\left(\bar{V}_4' \bar{V}_2 \bar{V}_4 \right)^{-1} \otimes (\bar{v}_3 - r) \right] \frac{\partial \bar{v}_3'}{\partial \bar{v}_1} \frac{\partial \bar{v}_1}{\partial \bar{v}} \\ & + \left[I_m \otimes (\bar{v}_3 - r)(\bar{v}_3 - r)' \right] \left[\left(\bar{V}_4' \bar{V}_2 \bar{V}_4 \right)^{-1} \otimes \left(\bar{V}_4' \bar{V}_2 \bar{V}_4 \right)^{-1} \right] \\ & \left. \times \frac{\partial vech \left(\bar{V}_4' \bar{V}_2 \bar{V}_4 \right)}{\partial \bar{v}} \right\},\end{aligned}$$

$$\bar{v}_3 = R\left(\frac{1}{J}\sum_{j=1}^{J}\theta^{[j]}\right) = R(\bar{v}_1),\ \bar{V}_4 = \frac{\partial R\left(\frac{1}{J}\sum_{j=1}^{J}\theta^{[j]}\right)}{\partial \theta} = \frac{\partial R(\theta)}{\partial \theta}|_{\theta=\bar{v}_1},$$

$$\begin{aligned}\frac{\partial vech \left(\bar{V}_4' \bar{V}_2 \bar{V}_4 \right)}{\partial \bar{v}} =& \left((\bar{V}_2 \bar{V}_4)' \otimes I_m \right) \frac{\partial \bar{V}_4'}{\partial \bar{v}_1} \frac{\partial \bar{v}_1}{\partial \bar{v}} + \left(\bar{V}_4 \otimes \bar{V}_4' \right) \frac{\partial \bar{V}_2}{\partial \bar{v}} \\ & + \left(I_m \otimes \bar{V}_4' \bar{V}_2 \right) \frac{\partial \bar{V}_4}{\partial \bar{v}_1} \frac{\partial \bar{v}_1}{\partial \bar{v}},\end{aligned}$$

and the derivatives of $\bar{V}_4$ and $\bar{v}_3$ depend on the form of $R(\theta)$.

Remark 3.3.14. *Following Newey and West (1987), a consistent estimator of the NSE of $\bar{v}$ is given by*

$$Var(\bar{v}) = \frac{1}{J}\left[\Omega_0 + \sum_{k=1}^{K}\left(1 - \frac{k}{K+1}\right)\left(\Omega_k + \Omega_k'\right)\right],$$

where

$$\Omega_k = J^{-1}\sum_{j=k+1}^{J}\left(v^{[j]} - \bar{v}\right)\left(v^{[j]} - \widehat{v}\right)'.$$

3.4 Hypothesis Testing for Latent Variable Models

Latent variable models have found a wide range of applications in microeconometrics, macroeconometrics and financial econometrics; see Stern (1997), Norets (2009), Koop and Korobilis (2009), Yu (2011). Without loss of generality, let

$y = (y_1, y_2, \ldots, y_n)'$ denote the observed variables and $z = (z_1, z_2, \ldots, z_n)'$ the latent variables. The set of parameters in the model is denoted by ϑ. Let $p(y|\vartheta)$ be the likelihood function of the observed data, and $p(y,z|\vartheta)$ be the complete-data likelihood function. The relationship between these two likelihood functions is

$$p(y|\vartheta) = \int p(y,z|\vartheta)dz. \tag{3.4.1}$$

In many latent variable models, especially dynamic latent variable models, the number of latent variables is often the same as the sample size. Hence, the integral in (3.4.1) is high-dimensional when the sample size is large. If the integral does not have an analytical expression, it will be very difficult to evaluate numerically. Consequently, statistical inferences, including estimation and hypothesis testing, are difficult to implement if they are based on the MLE.

In recent years, it has been documented that latent variable models can be efficiently analyzed using MCMC techniques; see Geweke, et al. (2011). Let $p(\vartheta)$ be the prior distribution of ϑ. To alleviate the difficulty in maximum likelihood, the data-augmentation strategy (Tanner and Wong, 1987) is often employed where the latent variables are treated as additional parameters. Then, the Gibbs sampler can be used to generate random samples from the joint posterior distribution $p(\vartheta, z|y)$, denoted by $\left\{\vartheta^{[j]}, z^{[j]}\right\}_{j=1}^{J}$, after a burn-in phase. The Bayesian estimates of ϑ and the estimates of the covariance matrix can be obtained as,

$$\bar{\bar{\vartheta}} = \frac{1}{J}\sum_{j=1}^{J}\vartheta^{[j]},\ \bar{V}\left(\bar{\bar{\vartheta}}\right) = \frac{1}{J}\sum_{j=1}^{J}\left(\vartheta^{[j]} - \bar{\bar{\vartheta}}\right)\left(\vartheta^{[j]} - \bar{\bar{\vartheta}}\right)'.$$

Similarly, the proposed test can be easily computed from $\left\{\vartheta^{[j]}\right\}_{j=1}^{J}$ and hence it is very easy to implement.

Remark 3.4.1. *As noted before, the test statistic of LLY in (3.2.2) requires the evaluation of the first derivative of the observed-data likelihood function. For many latent variable model, this is difficult to evaluate when the observed-data likelihood* *function does not have a closed-form expression. In addition, it requires estimating both the null model and the alternative model. However, the proposed test does not require evaluating the first derivative and only estimate the model under the alterative hypothesis. Clearly, the proposed test is easier to implement and faster to compute.*

3.5 Simulation Studies

In this section, we first design two experiments to examine the finite-sample performance of the proposed test with simulated data. In the first experiment, we test different null hypotheses in a linear regression model. This study aims to compare the finite sample behavior between $T(y, \theta_0)$ and the Wald statistic in terms of size and power. In the second experiment, we test the point null hypothesis in a discrete choice model. It is a simultaneous equation model with ordered probit and two-limit censored regression. Li (2006) applied this microeconometric model to study the relationship between high school completion and future youth unemployment.

3.5.1 Hypothesis testing in a linear regression model

The linear regression model we consider is specified as

$$y_i = x_i'\beta + \varepsilon_i, \varepsilon_i \sim N\left(0, \sigma^2\right), i = 1, \ldots, n.$$

with $x_{i1} = 1$. Let $\beta = \left(\beta_1', \beta_2'\right)'$. We consider two different null hypotheses, both concerning β_1. The first one is to test $H_0 : \beta_1 = \beta_1^*$ against $H_1 : \beta_1 \neq \beta_1^*$. The other is to test $H_0 : R\beta_1 = r$ against $H_1 : R\beta_1 \neq r$. To do Bayesian analysis, the conjugate priors for β and σ^2 can be specified as the normal distribution and the inverse gamma distribution, respectively,

$$\beta|\sigma^2 \sim N\left(\mu_0, \sigma^2 V_0\right), \sigma^2 \sim IG(a, b),$$

where μ_0, V_0 and a, b are hyperparameters. As a result, the posterior distributions are available analytically.

For simplicity, we consider the case in which $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$, $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})'$, where $x_{i1} = 1$, $x_{i2}, x_{i3}, x_{i4} \sim N(0,1)$. The true parameter values used to simulate data are given as $\sigma^2 = 0.01, \beta_1 = 0.3, \beta_2 = 0.2, \beta_3 = 0.1C, \beta_4 = 0.5C$ for $C = 0, 0.1, 0.3, 0.5$, where C is used to control the difference between the true value and zero. The number of replications is set at 1000 while three sample sizes are considered, $n = 50, 100, 150$. Each of four null hypotheses is tested, $\beta_3 = 0$, or $\beta_4 = 0$, or $\beta_3 = \beta_4 = 0$, or $\beta_3 + \beta_4 = 0$, in every replication. To make the priors vague, the hyperparameters are specified at

$$\mu_0 = (0,0,0,0)', V_0 = 1000 \times I_4, a = 0.0001, b = 0.0001,$$

with I_4 being the 4×4 identity matrix. In each replication, we draw 5000 i.i.d. random samples from the posterior distribution and then use the posterior samples to compute the proposed statistic. Also computed is the Wald statistic for the purpose of comparison. The Wald test is feasible because MLE is easy to obtain in this application.

Table 3.2 reports the size and the power of the proposed test and the Wald test for a nominal size of 5%. In all cases, the size distortion for the new statistic is very small and the two tests perform similarly in terms of size. The size approaches 5% as the sample size increases. Moreover, in all cases, the power of the proposed test is comparable to that of the Wald statistic. As C increases, the power of the proposed statistic approaches 100%. Similarly, as the sample size increases, the power of the proposed statistic approaches 100%.

3.5.2 Hypothesis testing in a discrete choice model

The second model in the simulation study is a simplified version of the model of Li (2006) where the effects of attendance on high school completion and future youth

Table 3.2: The size and power of the proposed test and the Wald test for different null hypothesis in a linear regression model

		Empirical Size		Empirical Power					
		$C=0$		$C=0.1$		$C=0.3$		$C=0.5$	
n	H_0	$T(y,\beta_{10})$	Wald	$T(y,\beta_{10})$	Wald	$T(y,\beta_{10})$	Wald	$T(y,\beta_{10})$	Wald
50	$\beta_3=0$	4.50%	5.10%	10.40%	11.00%	55.80%	57.30%	92.00%	92.20%
	$\beta_4=0$	6.50%	7.10%	92.00%	92.5%	100%	100%	100%	100%
	$\beta_3=\beta_4=0$	6.60%	7.50%	88.80%	89.70%	100%	100%	100%	100%
	$\beta_3+\beta_4=0$	6.20%	6.70%	83.30%	84.00%	100%	100%	100%	100%
100	$\beta_3=0$	5.50%	5.80%	20.20%	20.40%	82.00%	82.80%	99.90%	100%
	$\beta_4=0$	4.60%	5.00%	99.70%	99.70%	100%	100%	100%	100%
	$\beta_3=\beta_4=0$	5.70%	6.00%	99.50%	99.50%	100%	100%	100%	100%
	$\beta_3+\beta_4=0$	6.00%	6.20%	98.60%	98.60%	100%	100%	100%	100%
150	$\beta_3=0$	5.30%	5.40%	24.40	24.60%	95.90%	95.90%	100%	100%
	$\beta_4=0$	5.20%	5.30%	100%	100%	100%	100%	100%	100%
	$\beta_3=\beta_4=0$	5.40%	5.60%	100%	100%	100%	100%	100%	100%
	$\beta_3+\beta_4=0$	4.20%	4.20%	99.80%	99.80%	100%	100%	100%	100%

unemployment were studied. As noted in Li (2006), the likelihood function involves multiple integrals and discrete and censor variables. Consequently, the likelihood function and the corresponding derivatives are not easy to evaluate. Consequently, Li introduced a MCMC approach to do statistical analysis. We perform hypothesis testing in the discrete choice model with latent variables.

Let $z_i = 1,2,3,4$ denote the high school grade completed by individual i which is by definition an ordered integer. Let y_i denote the latent outcome corresponding to z_i. The first part of the model is an ordered probit defined as

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, & \varepsilon_i \sim N\left(0,\sigma^2\right), \gamma_{z_i} < y_i < \gamma_{z_i+1}, \\ \gamma_1 = -\infty, \gamma_2 = 0, & \gamma_2 < \gamma_3 < \gamma_4, \gamma_4 = 1, \gamma_5 = \infty, \end{cases}$$

where $i = 1,\ldots,n$ with n being the total number of individuals, ε_i is an individual level random error term, σ^2 is the variance of the error term, $\left\{\gamma_j\right\}_{j=1}^{5}$ are the cutoff points, x_i contains some covariates which are assumed to be exogenous. For the purpose of simulating data, we simply assume x_i is univariate and $x_i \sim N(0,1)$.

Furthermore, let ω_i denote the proportion of time during which individual i is unemployed, $\tilde{y}_i$ is the latent outcome corresponding to ω_i, and $\tilde{y}_i$ is limited as,

$$\tilde{y}_i \begin{cases} \leq 0, & \omega_i = 0, \\ = \omega_i, & 0 < \omega_i < 1, \\ \geq 1, & \omega_i = 1. \end{cases}$$

Then the censored regression is,

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\varepsilon}_i, \tilde{\varepsilon}_i \sim N\left(0, \tilde{\sigma}^2\right). \tag{3.5.1}$$

The two error terms are correlated, that is,

$$\begin{pmatrix} \varepsilon_i \\ \tilde{\varepsilon}_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma_{12} \\ \sigma_{12} & \tilde{\sigma}^2 \end{pmatrix}\right) := N(0, \Sigma).$$

In the simulation study, the null and alternative hypotheses are,

$$H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0.$$

To calculate the size and power of the proposed statistic, three sample sizes are considered, n = 100, 250 and 500. In each case, we compute the empirical size when $\beta_1 = 0$ at a nominal size of 5%. We also compute the power when $\beta_1 = 0.1, 0.2$ and 0.4. The number of replications is 500. The true values of other parameters are set at,

$$\beta_0 = 1, \tilde{\beta}_0 = 0.01, \tilde{\beta}_1 = 0.1, \Sigma = \begin{pmatrix} 1 & -0.01 \\ -0.01 & 0.1 \end{pmatrix}, \gamma_3 = 0.67.$$

These values are close to those reported in Li (2006) based on actual data.

Following Li (2006), we use the following vague priors to do Bayesian analysis,

$$\beta = \left(\beta_0, \beta_1.\tilde{\beta}_0, \tilde{\beta}_1\right)' \sim N(0, 1000 \times I_4),\ \Sigma \sim IW(6, 6 \times I_2), \gamma_3 \sim Beta(1,1),$$

Table 3.3: The size and power of the proposed test in a discrete choice model

	Empirical Size	Empirical Power		
	$\beta_1 = 0$	$\beta_1 = 0.1$	$\beta_1 = 0.2$	$\beta_1 = 0.4$
$n = 100$	4.2%	11.0%	23.0%	75.8%
$n = 250$	5.2%	24.0%	65.0%	100%
$n = 500$	4.6%	49.4%	97.2%	100%

where IW denotes the inverted Wishart distribution and $Beta$ denotes the Beta distribution.

We run MCMC to obtain 10,000 random samples. After dropping the first 4,000 samples, we treat the remaining 6,000 sample as effective draws from the posterior distribution. Let $\left\{\beta_1^{[j]}\right\}_{j=1}^{J}$ denote the effective posterior draws. From (3.3.13), the proposed statistic can be simply calculated as

$$\widehat{T}\left(y, \beta_1 = 0\right) = \frac{\frac{1}{J}\sum_{j=1}^{J}\left(\beta_1^{[j]}\right)^2}{\frac{1}{J}\sum_{j=1}^{J}\left(\beta_1^{[j]} - \overline{\overline{\beta}}_1\right)^2}, \overline{\overline{\beta}}_1 = \frac{1}{J}\sum_{j=1}^{J}\beta_1^{[j]}.$$

Other test statistics, such as BFs and the Wald statistic, are harder to obtain due to the presence of latent variables.

The empirical size and power of the proposed test are reported in Table 3.3 for a nominal size of 5%. It is obvious that the empirical size is close the nominal size in all cases, even when the sample size is only 100. When β_1 becomes further and further away from 0, the power increases and approaches 100%. Furthermore, as the sample size increase, the power increases in all cases.

3.6 Empirical Examples

We then consider two empirical studies using real data. The first model is the full version of the discrete choice model of Li (2006). The second model is the stochastic volatility model with leverage effect. For both models, it is well-known that the observed-data likelihood function is intractable due to the presence of latent variables. As a result, the observed-data likelihood function and its derivatives are very difficult to evaluate and hence it is advantageous to use the proposed statistic over existing statistics for hypothesis testing.

3.6.1 Hypothesis testing in a discrete choice model

In the first empirical study, we consider the same model and use the same data set as in Li (2006). Let z_{hi} denote the high school grade completed by individual i, and y_{hi} denote the latent outcome corresponding to z_{hi}, where h labels the schooling outcome. Let $z_{hi} = 1$ if individual i dropped out of high school after completing the ninth grade, $z_{hi} = 2$ if he dropped out after completing the tenth grade, $z_{hi} = 3$ if he dropped out after completing the eleventh grade, and $z_{hi} = 4$ if he completed high school. An ordered probit is specified as

$$\begin{cases} y_{hi} = \beta_h' x_{hi} + \varepsilon_{hi}, & \varepsilon_{hi} \sim N\left(0, \sigma_h^2\right), \gamma_{z_{hi}} < y_{hi} < \gamma_{z_{hi}+1}, \\ \gamma_1 = -\infty, \gamma_2 = 0, & \gamma_2 < \gamma_3 < \gamma_4, \gamma_4 = 1, \gamma_5 = \infty, \end{cases} \tag{3.6.1}$$

where x_{hi} is a $k_h \times 1$ vector incorporating individual level variables, including base year cognitive test score, parental income, parental education, number of siblings, gender, race, county level employment growth rate between 1980 and 1982, a fourth-order polynomial in age and a fourth-order polynomial in the time eligible to drop out.

Furthermore, let ω_{ui} represent the proportion of time when individual i is unemployed, y_{ui} the latent outcome corresponding to ω_{ui}, and y_{ui} is limited as,

$$y_{ui} \begin{cases} \leq 0, & \omega_{ui} = 0, \\ = \omega_{ui}, & 0 < \omega_{ui} < 1, \\ \geq 1, & \omega_{ui} = 1. \end{cases} \tag{3.6.2}$$

Thus, the censored regression is,

$$y_{ui} = \beta_u' x_{ui} + s_i' \eta + \varepsilon_{ui}, \varepsilon_{ui} \sim N\left(0, \sigma_u^2\right), \tag{3.6.3}$$

where x_{ui} is a $k_u \times 1$ vector incorporating observed variables, including base year cognitive test score, parental income, parental education, number of siblings, gender, race, age and a dummy variable indicating any post-secondary education.

In Equation (3.6.3), s_i is a 4×1 vector consisting of dummy variables indicating the high school grade completed by individual i. In other words, $s_i = (s_{i,1}, s_{i,2}, s_{i,3}, s_{i,4})'$, and if $s_{i,z_{hi}} = 1$ then $s_{i,j} = 0,\ j \neq z_{hi}$. Besides, η indicates the 4×1 vector of the effect of high school completion on unemployment. For simplicity, η is assumed to be the same across schools. This assumption is different from that in Li (2006) although our empirical results are almost the same as those in Li. The random terms are assumed to be correlated,

$$\begin{pmatrix} \varepsilon_{hi} \\ \varepsilon_{ui} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_h^2 & \sigma_{hu} \\ \sigma_{hu} & \sigma_u^2 \end{pmatrix}\right) = N(0, \Sigma).$$

In total, there are 34 parameters in the model.

As noted in Li (2006), the MLE is difficult to obtain. Hence, the MCMC technique is implemented. We adopted the same priors as Li which are listed in the following,

$$\beta = (\beta_h', \beta_u')' \sim N(0_{k\times 1}, 1000 \times I_k),\ \Sigma \sim IW(6, 6 \times I_2),$$

$$\eta \sim N(0, I_4),\ \gamma_3 \sim Beta(1,1),$$

where $k = k_h + k_u$.

The dataset contains 5,238 students from 871 schools. For more details about the data, one can refer to Li (2006). We run MCMC for 20,000 times. After dropping the first 4,000 samples, we treat the remaining 16,000 as effective draws. Posterior means and posterior standard errors are reported in Table 3.4, all of which are very close to those reported in Li.

Suppose one is interested in testing that the marginal effects of father's education level and mother's education level on the completion of high school can be ignored

Table 3.4: Posterior means and posterior standard errors of parameters in a discrete choice model of Li (2006)

	$E(\cdot\|Data)$	$SE(\cdot\|Data)$
	High school completion y_h	
Constant	0.9474	0.2119
Parental income	0.0110	0.0262
Base year cognitive test	0.4413	0.0370
Father's education	0.0456	0.0131
Mother's education	0.0627	0.0159
Number of siblings	-0.0370	0.0153
Female	-0.0694	0.0534
Minority	0.3840	0.0664
County employment growth	-0.0132	0.0047
Age	-0.4150	0.0853
Age^2	-0.1887	0.0766
Age^3	-0.0333	0.0468
Age^4	0.0311	0.0148
Time eligible to drop out	0.0932	0.0696
Time^2	0.0905	0.0473
Time^3	-0.0090	0.0106
Time^4	-0.0094	0.0053
	Proportion of time unemployed ω_u	
Parental income	-0.0275	0.0056
Base year cognitive test	-0.0392	0.0071
Father's education	-0.0020	0.0025
Mother's education	-0.0043	0.0030
Number of siblings	0.0049	0.0034
Post-secondary education	-0.0113	0.0138
Female	0.0621	0.0112
Minority	0.0826	0.0131
Age	-0.0058	0.0126
Completing ninth grade(η_1)	0.1925	0.0705
Completing tenth grade(η_2)	0.1211	0.0530
Completing eleventh grade(η_3)	0.1187	0.0492
Completing high school(η_4)	0.0083	0.0416
	Covariance matrix Σ	
σ_h^2	0.9450	0.0914
σ_u^2	0.1215	0.0039
σ_{hu}	-0.0099	0.0191
	Cutoff point	
γ_3	0.6684	0.0220

Table 3.5: The proposed statistic, $\widehat{T}_{LLY}(y,\theta_0)$, $\widehat{\log BF_{10}}$, the CPU time (in seconds), and their NSEs in the discrete choice model of Li (2006).

	$\beta_4=\beta_5=0$		
	Value	NSE	CPU Time (seconds)†
$\widehat{T}(y,\theta_0)-1$	44.39	1.59	22.54
$\widehat{T}_{LLY}(y,\theta_0)$	2502.00	89.57	39,096.79
$\widehat{\log BF_{10}}$	5.2019	1.03	292,886.45

† The CPU time for computing each statistic is obtained from a laptop with an Intel i5 CPU and 8 GB memory after MCMC outputs are available.

or not. The null hypothesis can be written as $H_0:\beta_{4h}=\beta_{5h}=0$. With the MCMC output, we can very easily compute the statistic. We also compute $\widehat{\log BF_{10}}$ and $\widehat{T}_{LLY}(y,\theta_0)$. The three test statistics and their numerical standard errors are reported in Table 3.5.[2]

According to Table 3.5, both $\widehat{T}(y,\theta_0)-1$ and $\widehat{T}_{LLY}(y,\theta_0)$ take very large values, indicating that the null hypothesis is overwhelmingly rejected. This conclusion is consistent with that by $\widehat{\log BF_{10}}$, which strongly supports the alternative hypothesis. Furthermore, their numerical standard errors are all small relative to the values of the statistics. Finally, in spite of the same conclusion reached, the CPU time required to compute the test statistics is vastly different. The proposed statistic is more than 1700 times and nearly 13000 times faster to compute than $\widehat{T}_{LLY}(y,\theta_0)$ and $\widehat{\log BF_{10}}$ after MCMC outputs are available. An additional advantage that does not reflect in the CPU time is that the proposed statistic only needs MCMC output from the alternative model while the other two statistics require MCMC output for both the null and alternative models.

Hypothesis testing in a stochastic volatility model

Stochastic volatility (SV) models with leverage effect have been widely used in finance; see Harvey and Shephard (1996) and Aït-Sahalia, et al (2017). Following

[2]We use the marginal likelihood method of Chib (1995) to compute the BF and its NSE.

Yu (2005), the stochastic volatility model with leverage effect is defined as,

$$\begin{cases} r_t = \exp(h_t/2)\,\varepsilon_t, \\ h_{t+1} = \mu + \phi(h_t - \mu) + \sigma\varepsilon_{t+1}, h_0 = \mu, \end{cases}$$

with

$$\begin{pmatrix} \varepsilon_t \\ \varepsilon_{t+1} \end{pmatrix} \overset{i.i.d.}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

where r_t is the return at time t, h_t the latent volatility at period t. In this model, ρ is the parameter that captures the leverage effect when it is negative. Hence, we test $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$. In this example, we use two different datasets for hypothesis testing. For each dataset, we compute the proposed statistic, $T_{LLY}(y, \theta_0)$ and $\widehat{\log BF_{10}}$.[3]

Let $\left\{\rho^{[j]}\right\}_{j=1}^{J}$ denote the effective posterior draws for ρ under H_1. The proposed statistic is simply calculated as

$$\widehat{T}(y, \rho = 0) = \frac{\frac{1}{J}\sum_{j=1}^{J}\left(\rho^{[j]}\right)^2}{\frac{1}{J}\sum_{j=1}^{J}\left(\rho^{[j]} - \overline{\overline{\rho}}\right)^2}, \overline{\overline{\rho}} = \frac{1}{J}\sum_{j=1}^{J}\rho^{[j]}.$$

On the contrary, computing $\widehat{T}_{LLY}(y, \theta_0)$ and $\widehat{\log BF_{10}}$ require substantially higher coding efforts and extra CPU time.

The first dataset consists of daily returns on Pound/Dollar exchange rates from 01/10/81 to 28/06/85 with sample size 945. The series r_t is the daily mean-corrected returns. The following vague priors are used:

$$\mu \sim N(0, 100), \phi \sim Beta(1, 1), \sigma^{-2} \sim \Gamma(0.001, 0.001), \rho \sim U(-1, 1).$$

We draw 50,000 from the posterior distribution and discard the first 20,000 as build-in samples. Then we store every 5th value of the remaining samples as effective random samples. The estimation results are reported in Table 3.6.

[3]Again we use the marginal likelihood method of Chib (1995) to compute the BF.

Table 3.6: Posterior means of parameters in the SV model with and without leverage effect for the Pound/Dollar returns.

	H_1		H_0	
Parameter	Mean	SE	Mean	SE
μ	-0.5776	0.3487	-0.6608	0.3164
ϕ	0.9849	0.0097	0.9793	0.0127
ρ	-0.0941	0.1507	-	-
τ	0.1553	0.0243	0.1618	0.0360

Table 3.7: The proposed statistic, $\widehat{T}_{LLY}(y,\ \theta_0)$, $\widehat{\log BF_{10}}$, the CPU time (in seconds), and the NSEs of the first two statistics for the Pound/Dollar returns.

	$\widehat{T}(y,\theta_0)-1$	$\widehat{T}_{LLY}(y,\theta_0)$	$\widehat{\log BF_{10}}$
Value	0.3893	0.2883	-10.1235
NSE	0.0255	0.2028	-
CPU Time (seconds)	0.9411	549.0631	3,701.2241

Table 3.7 reports the proposed statistic, $\widehat{T}_{LLY}(y,\theta_0)$ and $\widehat{\log BF_{10}}$ and the NSEs for the first two statistics. Since the observed-data likelihood function is expensive to compute, the NSE of BF is too difficult to obtain and not report. $\widehat{\log BF_{10}}$ strongly supports the null hypothesis, that is, the SV model without leverage effect. $\widehat{T}_{LLY}(y,\theta_0)$ takes a very small value, suggesting that we cannot reject the null hypothesis. When the null hypothesis is true, we know that $T(y,\theta_0)-1 \xrightarrow{d} \chi^2(1)$. It can be found that $\widehat{T}(y,\theta_0)-1$ is very closed to $\widehat{T}_{LLY}(y,\theta_0)$, also suggesting that we cannot reject the null hypothesis. Finally, our proposed statistic has a smaller NSE than $\widehat{T}_{LLY}(y,\theta_0)$.

The second dataset contains 1,822 daily returns of the Standard & Poor (S&P) 500 index, covering the period between January 3, 2005 and March 28, 2012. We use the same priors and method as before to estimate the model with and without leverage effect. The estimation results are reported in Table 3.8.

The three test statistics and the NSEs for the first two statistics are reported in Table 3.9. Contrary to the case of Pound/Dollar returns, all three statistics strongly support the alternative hypothesis. Both $\widehat{T}(y,\theta_0)-1$ and $\widehat{T}_{LLY}(y,\theta_0)$ reject the null hypothesis under the 99% significance level. Similarly, $\widehat{\log BF_{10}}$ strongly supports the alternative hypothesis. However, the proposed statistic is nearly 1000 times and

Table 3.8: Posterior means of parameters in the SV model with and without leverage effect for the S&P500 returns.

	H_1		H_0	
Parameter	Mean	SE	Mean	SE
μ	-10.8800	0.1751	-11.2200	0.3349
ϕ	0.9804	0.0039	0.9897	0.0042
ρ	-0.7151	0.0422	-	-
τ	0.2057	0.0178	0.1705	0.0169

more than 6000 times faster to compute than $\widehat{T}_{LLY}(y, \theta_0)$ and $\widehat{\log BF_{10}}$ after MCMC outputs are available.

Table 3.9: The proposed statistic, $\widehat{T}_{LLY}(y,\ \theta_0)$, $\widehat{\log BF_{10}}$, the CPU time (in seconds), and the NSEs of the first two statistics for the S&P500 returns.

	$\widehat{T}(y, \theta_0) - 1$	$\widehat{T}_{LLY}(y, \theta_0)$	$\widehat{\log BF_{10}}$
Value	286.7944	8.2419	51.9582
NSE	0.6915	0.6849	-
CPU Time (seconds)	1.2922	1,256.7768	7,785.6888

3.7 Conclusion

In this paper, a new test statistic is proposed to test for a point null hypothesis which can be treated as the posterior version of the Wald test. Compared with existing methods, the proposed statistic has many important advantages. First, it is well-defined under improper prior distributions. Second, it avoids Jeffreys-Lindley's paradox. Third, its asymptotic distribution is a χ^2 distribution under the null hypothesis and repeated sampling. This property is the same as the Wald statistic so that the critical values can be easily obtained. Fourth, it is very easy to compute as it is based on the posterior mean and posterior variance of the parameters of interest. Fifth, it can be used to test hypotheses that imposes nonlinear relationships among the parameters of interest, for which the BF is difficult to use. Sixth, for latent variable models for which the MLE and the Wald test are more difficult to obtain, the proposed statistic is the by-product of posterior sampling. Finally, only posterior sampling for the alternative hypothesis is needed for the proposed statistic.

The finite sample properties of the proposed statistic is examined in a linear regression model and in a discrete choice model with latent variables. In the linear regression models, the Wald statistics is feasible and compared with the proposed test. Simulation results show that the proposed test has little size distortion even when the sample size is small and its size and power are very similar to those of the Wald test when a vague prior is used. In the discrete choice model, the proposed test has little size distortion even when the sample size is small. The power increases rapidly when the sample size increases or when the difference between the null and alternative hypotheses increases.

We apply the method to two models using real data. The first one is a discrete choice model and the second is a SV model. In both models there are latent variables. Due to the presence of latent variables, the Wald statistic is very difficult to obtain and because the maximum likelihood method is difficult to use. While both the BF and the test proposed by LLY (2015) are feasible to compute based on MCMC output, they are much more expensive to compute than the proposed statistic with longer CPU time after MCMC output is available. The empirical conclusion obtained by these three methods is the same in both empirical applications.

Chapter 4 Estimating Finite-Horizon Life-Cycle Models: A Quasi-Bayesian Approach

4.1 Introduction

Life-cycle models (also known as dynamic structural models) have been used extensively in macroeconomics, labor economics, industrial organizations, demographics, household finance, and many other fields; see Pakes (1994) and Rust (1994) for excellent reviews. The life-cycle model with finite-horizon is a subclass that has been found to have a great number of applications. For a sample of references, see Gourinchas and Parker (2002), Jørgensen (2017), Cagetti (2003), Browning and Ejrnæs (2009), Kaplan and Violante (2014), Li et al. (2016), Fagereng, Gottlieb and Guiso (2017), Koijen, Nijman and Irker (2009), and Fischer and Stamos (2013).

A popular technique used to estimate finite-horizon life-cycle models in the literature is based on the log-linearized approximations to Euler equations. However, it has been argued that this approach can result in estimation bias; see Ludvigson and Paxson (2001), Carroll (2001) and Jøgensen (2016). To deal with this bias, empirical researchers have increasingly adopted the method of simulated moments (MSM) introduced by Duffie and Singleton (1993). Gourinchas and Parker (2002), hereafter GP, were the first to using MSM to estimate the preference parameters in a life-cycle model. Li et al. (2016) studied optimal life-cycle housing and nonhousing consumption using MSM. Fagereng, Gottlieb and Guiso (2017) applied MSM to estimate structural parameters and studied portfolio choice over the life-cycle. In these papers, the estimation procedure was divided into two stages. During the first stage, GMM or calibration was used to estimate parameters of exogenous processes such as the income process. During the second stage the structural parameters were estimated using MSM.

However, since MSM uses iterative optimization algorithms, there are four challenges to its use for estimating finite-horizon life-cycle models. First, the model has to be solved numerically at each iteration. Solving finite-horizon life-cycle models is time consuming and inconvenient because of the nonstationary policy functions. Second, one has to use numerical differentiation to evaluate the gradient of the objective function for parameter updating. Numerical differentiation requires more restrictive assumptions on the objective function and the computation is also cumbersome. Third, due to the complexity of the models, there may exist local optimums. Fourth, typically two-step estimation is necessary, which complicates the asymptotic behavior of the estimator.

The present paper develops a quasi-Bayesian method for estimating structural parameters in finite-horizon life-cycle models during the second stage. Following Chernozhukov and Hong (2003), hereafter CH, we build the quasi-posterior density function based on first-stage estimates and the GMM objective function. The new estimator is obtained by minimizing the Bayesian risk function consisting of the quasi-posterior density function and a net loss function. By doing this, the optimization problem is converted into a sampling one, which avoids the numerical evaluation for the gradient of the objective function and alleviates the local optimum problem; see CH for examples where the local optimum problem was carefully explained.

The asymptotic behavior of the proposed estimator is studied in two cases. First, when the policy functions are analytically available, the asymptotic normality of this estimator is derived. There is a bias in the asymptotic mean that depends on the net loss function. We also show that the estimator reaches the efficiency bound in the framework of GMM. When the net loss function is symmetric, the bias term becomes zero. In particular, if the net loss function is quadratic, the estimator becomes the posterior mean and the associated asymptotic covariance can be approximated by the posterior covariance. This is advantageous in computation since the posterior mean and posterior covariance can be simultaneously computed from the quasi-posterior samples.

Second, when the policy functions are not analytically available, we propose to approximate them over a set of grid points. We show that the magnitude of approximation errors depends both on the number of grid points (j) and the number of observations (N). While the approximation errors associated with a numerical method accumulate as the number of observations grows, it is found that they decrease as the number of grid points (j) increases. Interestingly, the results obtained for the case with analytical solutions still hold true in this case when the approximation errors decrease at a speed faster than the number of observations. This result shows that, even in the presence of approximation errors, the estimation approach is attractive from both the theoretical and computational viewpoints. In practice, most finite-horizon life-cycle models require numerical solutions, making the proposed estimation method useful in practical applications.

In terms of the computational effort, the new estimate requires extensive sampling. It should be noted that Markov Chain Monte Carlo (MCMC) does not work well here. This is because, to use MCMC, such as the Gibbs-sampler and Metropolis-Hasting sampler, one needs to update samples sequentially many times and at each updating the objective function has to be numerically evaluated. Instead of using MCMC, the importance sampling strategy is employed. The algorithm used by Creel and Kristensen (2016) is extended to construct a proposal distribution for important sampling. There are two computational advantages in the proposed algorithm. First, it is easy to parallelize and hence GPU can be used. Second, it is made to be adaptive to the dataset.

This paper makes four contributions to the literature. First, a quasi-Bayesian estimation approach is proposed for finite-horizon life-cycle models. The quasi-Bayesian estimator has desirable properties both in terms of asymptotic behavior and computation. Second, the method extends the seminal work of CH to life-cycle models and is related to a growing strand of literature on approximate Bayesian computation. Third, the econometric problem in the presence of approximation errors caused by numerical methods is carefully studied. The results complement Fernández-Villaverde, Rubio-Ramírez and Santos (2006), hereafter FRS, and Ackerberg, Geweke and Hahn (2009). The present paper considers the problem in the GMM framework while FRS and Ackerberg, Geweke and Hahn (2009) consider the problem in the likelihood setting. If an empirical researcher would like to be agnostic about the error distribution, a GMM framework will be more attractive than the full likelihood approach. Finally, the proposed adaptive algorithm makes use of GPU to enhance computational efficiency and is applicable to other complicated models with moment conditions.

Throughout the paper, a version of the model in GP is used for illustration, but other types of life-cycle models can also be considered. As long as the assumptions listed in the paper are satisfied, the theoretical results can be applied and the estimation algorithm remains useful.

The rest of the paper proceeds as follows. Section 2 introduces the illustrative model in detail. Section 3 presents the first-stage estimation for parameters of the exogenous process and the latent dynamic state variable filtering. Section 4 examines the second-stage estimation, including the definition of the estimator, the asymptotic behavior and the related algorithm to compute the estimator. Section 5 reports results from Monte Carlo studies, including models with and without dynamic latent state. Section 6 concludes. Appendices contain the details of proofs, numerical method used and other related computations.

4.2 An Illustrative Model

Let us first define a discrete-time life-cycle model for households. Households work until an exogenously given retirement age, T_r. At each working age, the utility function is the constant relative risk aversion (CRRA) utility function, i.e.,

$$u(C;\rho_0) = \begin{cases} \frac{C^{1-\rho_0}}{1-\rho_0} & \rho_0 \neq 1 \\ \log C & \rho_0 = 1 \end{cases},$$

where C is the consumption level and ρ_0 is the risk aversion. The number of household is N^{obs}. By forward looking from the initial working age t_i, household i ($\in \{1,...,N^{obs}\}$) chooses the level of consumption $C_{i,t}$ to solve the optimization problem

$$\max_{C_{i,\tau}} E_{t_i}\left[\sum_{\tau=t_i}^{T_r} \beta_0^{\tau-t_i} v(z_{i,\tau};\eta_0)\, u(C_{i,\tau};\rho_0) + \beta_0^{T_r+1-t} \tilde{V}_{T_r+1}(M_{i,T_r+1}, z_{i,T_r+1};\eta_0,\rho_0,\kappa_0)\right] \tag{4.2.1}$$

$$s.t.\ M_{i,t+1} = R(M_{i,t} - C_{i,t}) + Y_{i,t+1}, t_i \leq t \leq T_r - 1, \tag{4.2.2}$$

$$M_{i,T_r+1} = R(M_{i,T_r} - C_{i,T_r}), \tag{4.2.3}$$

$$C_{i,t} \in (0, M_{i,t}], \tag{4.2.4}$$

$$M_{i,t_i} \text{ given,}$$

where the subscript τ indicates that the associated variable realizes at age τ and the subscript i indicates that the variable belongs to household i, β_0 the subject discount factor, $C_{i,\tau}$ the consumption level, $M_{i,\tau}$ the liquid wealth, R the gross interest rate, $z_{i,\tau}$ a vector of characteristics and $v(z;\eta_0)$ a shifter in utility, which can be interpreted as a taste shifter in which the individual characteristic information z plays a role. In many applications, $v(z;\eta_0)$ is a specific function that summarizes the impact of the individual characteristics z.

The equations (4.2.2) and (4.2.3) are wealth accumulation equations before and after retirement. As in GP, the income process, $Y_{i,t+1}$, is assumed to follow the following stochastic process.

Income process: Income process is defined as

$$\begin{cases} Y_{i,t} = P_{i,t}\varepsilon_{i,t}, \\ P_{i,t} = G_t P_{i,t-1}\varsigma_{i,t}, \end{cases} \qquad t_i \le t \le T_r, \tag{4.2.5}$$

where $P_{i,t}$ denotes the latent permanent component of $Y_{i,t}$ and $P_{i,T_r+1} = P_{i,T_r}$ since there is no income at age $T_r + 1$, $\varepsilon_{i,t}$ the transitory component, G_t the real gross permanent income growth, $\varsigma_{i,t}$ the permanent income shock. Specifically,

$$\varepsilon_{i,t} = \begin{cases} \mu_0, & \text{with probability } p_0, \\ \xi_{i,t}, & \text{with probability } 1 - p_0, \end{cases} \qquad \text{where } \log \xi_{i,t} \overset{i.i.d.}{\sim} N\left(0, \sigma^2_{\varepsilon 0}\right),$$

$$\log \varsigma_{i,t} \overset{i.i.d.}{\sim} N\left(0, \sigma^2_{\varsigma 0}\right),$$

where μ_0 can be zero or some other small values, $\log \varsigma_{i,t}$ and $\log \xi_{i,t}$ independent across i and t. The parameters for the income process are denoted as $\chi_0^{inc} = \left(\mu_0, p_0, \sigma^2_{\varepsilon 0}, \sigma^2_{\varsigma 0}, \{G_t\}_{t=t_{min}}^{T_r}\right)'$, where $t_{min} = \min_{1 \le i \le N^{obs}} \{t_i\}$.

Characteristics information vector: The characteristics vector at age t of household i, $z_{i,t}$, can be deterministic or stochastic. The parameters involved in $z_{i,t}$ are denoted as χ_0^{cha}. According to Jøgensen (2017) and GP, researchers can examine the impact of different characteristics such as the number of children or family size on the marginal utility.

Retirement: When household i retires at T_r, for the tractability of the problem (4.2.1), following GP, the retirement value function is assumed to be

$$\tilde{V}_{T_r+1}\left(M_{i,T_r+1}, z_{i,T_r+1};\ \eta_0, \rho_0, \kappa_0\right) = \kappa_0 v\left(z_{i,T_r+1}; \eta_0\right) \frac{\left(M_{i,T_r+1} + H_{i,T_r+1}\right)^{1-\rho_0}}{1-\rho_0},$$

where κ_0 is the motivation to retire, $M_{i,T_{r+1}}$ the liquid wealth at age $T_r + 1$, H_{i,T_r+1} the illiquid wealth after retirement and $H_{i,T_r+1} = hP_{i,T_r+1}$, i.e., H_{i,T_r+1} is proportional to the permanent component at T_r+1. Since there is no income at T_r+1, we let $P_{i,T_r+1}=P_{i,T_r}$.

The Bellman equation for model (4.2.1) is

$$\begin{aligned}\tilde{V}_t\left(M_{i,t},P_{i,t},z_{i,t};\theta_0,\chi_0\right)=&\max_{C_{i,t}\in\left(0,M_{i,t}\right]}\{v\left(z_{i,t};\eta_0\right)u\left(C_{i,t};\rho_0\right)\\&+\beta_0E_t\left[\tilde{V}_{t+1}\left(M_{i,t+1},P_{i,t+1},z_{i,t+1};\theta_0,\chi_0\right)\right]\}\qquad(4.2.6)\\ s.t.\ &M_{i,t+1}=R\left(M_{i,t}-C_{i,t}\right)+Y_{i,t+1},t_i\leq t\leq T_r-1,\\&M_{i,T_r+1}=R\left(M_{i,T_r}-C_{i,T_r}\right),\\&C_{i,t}\in\left(0,M_{i,t}\right]\text{ with }M_{i,t_i}\text{ given,}\end{aligned}$$

where $\chi_0=\left(\left(\chi_0^{inc}\right)',\left(\chi_0^{cha}\right)',R\right)'$, $\theta_0=\left(\eta_0',\rho_0,\beta_0,\kappa_0,h\right)'\in\Theta\subset R^d$. At age T_r+1,

$$\tilde{V}_{T_r+1}\left(M_{i,T_r+1},P_{i,T_r+1},z_{i,T_r+1};\theta_0,\chi_0\right)=\tilde{V}_{T_r+1}\left(M_{i,T_r+1},z_{i,T_r+1};\eta_0,\rho_0,\kappa_0,h\right).$$

According to the model setup, the data that economists obtain are $\{M_{i,t},C_{i,t},Y_{i,t},z_{i,t}\}_{t=t_i}^{T_r+1}$ for household i. Therefore, for the Bellman equation (4.2.6), economists cannot directly solve it since it involves latent state variable $P_{i,t}$, which is only observed by household i. Thus, we instead study the ratio form of the Bellman equation (4.2.6).

The setup of the problem, combined with the retirement value function, makes the problem homogeneous of degree $1-\rho_0$ in $P_{i,t}$. Thus, we define the normalized value functions as follows.

$$V_t\left(m_{i,t},z_{i,t};\theta_0,\chi_0\right)=\frac{1}{P_{i,t}^{1-\rho_0}}\tilde{V}_t\left(M_{i,t},P_{i,t};\theta_0,\chi_0\right),$$

$$\begin{aligned}V_{T_r+1}\left(m_{i,T_r+1},z_{i,T_r+1};\theta_0,\chi_0\right)&=\frac{1}{P_{i,T_r+1}^{1-\rho_0}}\tilde{V}_{T_r+1}\left(M_{i,T_r+1},z_{i,T_r+1};\eta_0,\rho_0,\kappa_0\right)\\&=\kappa_0v\left(z_{i,T_r+1};\eta_0\right)\frac{\left(m_{i,T_r+1}+h\right)^{1-\rho_0}}{1-\rho_0}.\end{aligned}$$

We also normalize the variables of household i at age t by $P_{i,t}$, denoted by lowercase letters, e.g., $m_{i,t} \equiv M_{i,t}/P_{i,t}$, $c_{i,t} \equiv C_{i,t}/P_{i,t}$. Accordingly, the wealth accumulation equations can be expressed as

$$m_{i,t+1} = (m_{i,t} - c_{i,t}) \frac{R}{G_{t+1} \varsigma_{i,t+1}} + \varepsilon_{i,t+1}, t_i \leq t \leq T_r - 1,$$

$$m_{i,T_r+1} = R(m_{i,T_r} - c_{i,T_r}).$$

The ratio-form Bellman equation (4.2.6) is

$$\begin{aligned} V_t(m_{i,t}, z_{i,t}; \theta_0, \chi_0) = \max_{c_{i,t}} \{ & v(z_{i,t}; \eta_0) u(c_{i,\tau}; \rho_0) \\ & + \beta_0 E_t \left[(G_{t+1} \varsigma_{i,t+1})^{1-\rho_0} V_{t+1}(m_{i,t+1}, z_{i,t+1}; \theta_0, \chi_0) \right] \} \end{aligned} \tag{4.2.7}$$

$$\begin{aligned} s.t.\ & m_{i,t+1} = (m_{i,t} - c_{i,t}) \frac{R}{G_{i,t+1} \varsigma_{i,t+1}} + \varepsilon_{i,t+1}, t_i \leq t \leq T_r - 1, \\ & m_{i,T_r+1} = R(m_{i,T_r} - c_{i,T_r}), \\ & c_{i,t} \in (0, m_{i,t}]. \end{aligned}$$

Therefore, economists can solve the model (4.2.7) without the knowledge of latent state variable $P_{i,t}$.

Remark 4.2.1. *In the Bellman equation (4.2.7), the structural parameter θ_0 is the same as that in the original problem (4.2.6). We can solve the model by deriving the analytical solutions or using numerical methods conditional on the value of θ_0 and χ_0. The Euler equations for problem (4.2.7) are*

$$c_{i,t}^{-\rho_0} = \beta_0 R E_{\varsigma_{i,t+1}, \varepsilon_{i,t+1}, z_{i,t+1}} \left[\frac{v(z_{i,t+1}; \eta_0)}{v(z_{i,t}; \eta_0)} (G_{t+1} \varsigma_{i,t+1})^{-\rho_0} c_{i,t+1}^{-\rho_0} \right], t_i \leq t \leq T_r - 1,$$

which are necessary to derive the optimal policies at each age by backward optimization. In particular, the endogenous grid method (EGM) described in detail in Appendix .4.2 can be applied here.

4.3 First-Stage Estimation and Latent State Filtering

Following GP and based on the discussion in the previous section, the parameters are divided into to two parts, the nuisance parameters $\chi_0 = \left((\chi_0^{inc})', (\chi_0^{cha})', R \right)'$ and structural parameters θ_0. Data include a panel dataset used during the second stage estimation, $\left\{ C_{i,t}^d, M_{i,t}^d, Y_{i,t}^d, z_{i,t}^d, \right\}_{t=t_i}^{T_r}, i = 1, \ldots, N^{obs}$ and an additional one with sample size J used during the first stage. In the panel dataset with N^{obs} households, $C_{i,t}^d$, $M_{i,t}^d$, $Y_{i,t}^d$ and $z_{i,t}^d$ are respectively the consumption level, liquid wealth, income level and characteristic information vector of household i at age t, respectively.

At the first stage, conditional on the additional dataset, GMM or calibration is used to estimate χ, denoted as $\widehat{\chi}$. The following assumption is imposed for the first-stage estimator.

ASSUMPTION 1. *In the first-stage estimation, the nuisance parameters $\chi_0 = \left((\chi_0^{inc})', (\chi_0^{cha})', R \right)' \in \Psi$ can be obtained by GMM based on the additional dataset. The estimator $\widehat{\chi}$ satisfies,*

$$\sqrt{J}\left(\widehat{\chi} - \chi_0\right) \xrightarrow{d} N\left(0, \Sigma_\chi\right), \tag{4.3.1}$$

where Σ_χ is the covariance matrix.

Remark 4.3.1. *If the calibration approach is used in the first stage, then we simply treat $\widehat{\chi} = \chi_0$ without considering the dispersion caused by estimation, i.e., $\Sigma_\chi = 0$. This approach is frequently used in empirical literature such as Li et al. (2016) and Jøgensen (2017).*

Define $F_{i,t}$ as the information set up to age t for household i. The income process (4.2.5) can be rewritten as

$$\begin{cases} \log Y_{i,t} = \log P_{i,t} + \log \varepsilon_{i,t}, \\ \log P_{i,t} = \log \widehat{G}_t + \log P_{i,t-1} + \log \varsigma_{i,t}, \end{cases} \quad t_i \le t \le T_r - 1,$$

where $\log \varepsilon_{i,t} \overset{i.i.d.}{\sim} N\left(0, \hat{\sigma}_{\varepsilon}^2\right)$ and $\log \varsigma_{i,t} \overset{i.i.d.}{\sim} N\left(0, \hat{\sigma}_{\varsigma}^2\right)$. This is the standard linear state-space model with Gaussian errors so that the Kalman filter can be used to obtain the distribution of $P_{i,t}$ conditional on $F_{i,t}$ and $\widehat{\chi}$. When $\mu = 0$, the observations with zero income level can be considered as missing variables since the estimate $\hat{p}$ for p_0 is very small and thus zero-valued observation is rare. If $\mu \neq 0$ and is very small, then we can set up the threshold value to check whether there exists a shock. Via the Kalman filter, the mean and variance of $P_{i,t}$ conditional on $F_{i,t}$ are obtained. Denote the expectation of a random variable with respect to $P_{i,t}$ up to the information at age t as $E_{P_{i,t}}\left(\cdot|F_{i,t}\right)$.

4.4 Second-stage Estimation

4.4.1 Estimator

In this section, given $\widehat{\chi}$ from the frist stage, the estimator for θ_0 will be constructed. In this subsection we deal with the case in which there exists a close-form solution for optimal policy at each age. In the next subsection we deal with the case where optimal policies are not analytically available.

Given any generic $\theta \in \Theta$ and $\chi \in \Psi$, the analytical solutions for the optimal policy functions for the Bellman equation (4.2.7) is assumed to exist and denoted as $c_t\left(m_{i,t}^d, z_{i,t}^d; \theta, \chi\right)$ for household i at age t, where $m_{i,t}^d \equiv M_{i,t}^d / P_{i,t}$. For economists, $P_{i,t}$ is unobservable. Hence, taking $P_{i,t}$ into account, conditional on the information up to age t, it is natural to assume that the household i chooses the optimal consumption level according to

$$C_t\left(M_{i,t}^d, z_{i,t}^d; \theta, \chi\right) = E_{P_{i,t}}\left[c_t\left(\frac{M_{i,t}^d}{P_{i,t}}, z_{i,t}^d; \theta, \chi\right) P_{i,t} \middle| F_{i,t}\right], \tag{4.4.1}$$

where $E_{P_{i,t}}\left(\cdot|F_{i,t}\right)$ is the expectation with respect to $P_{i,t}$ based on the filtering at the first-stage estimation.

Remark 4.4.1. *The conditional expectation of equation (4.4.1) is more natural than* *the unconditional expectation used in GP, in which the Monte Carlo method was used based on the paths simulated from the initial working age and hence the information up to age t was discarded. Jøgensen (2017) treated the mean of* $\log P_{i,t}$ *obtained by the Kalman filter as the true value of* $\log P_{i,t}$*, which also ignored the variance information of* $\log P_{i,t}$*. In Appendix .4.5, these two approaches are compared with that based on equation (4.4.1). The evidence shows that equation (4.4.1) is superior to the other two approaches.*

In the following assumption, a moment condition is introduced.

ASSUMPTION 2. ***(Identification)*** *The unique parameter* θ_0 *is in the interior of a compact convex subset* Θ *of the Euclidean space* R^d*. For household i, assume*

$$\begin{aligned} E\left[C_{i,t}^d - C_t\left(M_{i,t}^d, z_{i,t}^d; \theta_0, \chi_0\right)\right] =& E\left[g_t\left(M_{i,t}^d, z_{i,t}^d; \theta_0, \chi_0\right)\right] \\ =& E\left[g_{i,t}\left(\theta_0; \chi_0\right)\right] = 0, \end{aligned} \tag{4.4.2}$$

where $t = t_i, \ldots, T_r$*,* $C_{i,t}^d$ *is the observed consumption level and* $C_t\left(M_{i,t}^d, z_{i,t}^d; \theta_0, \chi_0\right)$ *is defined in equation (4.4.1).*

Remark 4.4.2. *Assumption 2 is the identification assumption of the structural parameters* θ_0*. The assumption ensures the parameters are point-identified, which is also adopted by Hansen (1982) and Duffie and Singleton (1993).*

According to equation (4.4.2), we can have at most T_m moment conditions, where $T_m = T_r - t_{min} + 1$ and $t_{min} = \min\{t_i\}_{i=1}^{N^{obs}}$. Based on $\widehat{\chi}$ from the first stage, the objective function is

$$L_N(\theta) = L_N(\theta; \widehat{\chi}) = -\frac{N}{2}\left[\lambda_N \bar{g}_N(\theta; \widehat{\chi})\right]' W_N(\theta; \widehat{\chi}) \lambda_N \bar{g}_N(\theta; \widehat{\chi}), \tag{4.4.3}$$

where the *total number of observations* $N = \sum_{t=t_{min}}^{T_r} N_t$ with N_t the sample size at age t from $t = t_{min}$ to $t = T_r$,

$$\begin{aligned}\bar{g}_N(\theta;\widehat{\chi}) &= \left(\bar{g}_{t_{min}}(\theta;\widehat{\chi}),\ldots,\bar{g}_{T_r}(\theta;\widehat{\chi})\right)' \\ &= \left(\frac{1}{N_{t_{min}}}\sum_{i=1}^{N_{t_{min}}} g_{i,t_{min}}(\theta;\widehat{\chi}),\ldots,\frac{1}{N_{T_r}}\sum_{i=1}^{N_{T_r}} g_{i,T_r}(\theta;\widehat{\chi})\right)',\end{aligned}$$

$$W_N(\theta;\widehat{\chi}) = V_N^{-1}(\theta;\widehat{\chi}),$$

where,

$$\begin{aligned}V_N(\theta;\widehat{\chi}) =& \zeta_N \sum_{i=1}^{N^{obs}} \lambda_N \tilde{g}_i(\theta;\widehat{\chi})\,\tilde{g}_i(\theta;\widehat{\chi})'\,\lambda_N'\zeta_N' \\ &+ \frac{N}{J}\lambda_N \bar{g}_{N,\chi}(\theta;\widehat{\chi})\,\widehat{\Sigma}_\chi \bar{g}_{N,\chi}(\theta;\widehat{\chi})'\,\lambda_N',\end{aligned} \tag{4.4.4}$$

in which $\bar{g}_{N,\chi}(\theta;\widehat{\chi})$ is the first-order derivative of $\bar{g}_N(\theta;\chi)$ with respect to χ,

$$\tilde{g}_i(\theta;\chi) = \underbrace{\left(0,\ldots,0,g_{i,t_i}(\theta;\chi),\ldots,g_{i,T_r}(\theta;\chi)\right)'}_{T_m \text{ elements}},$$

$$\lambda_N = diag\left(\sqrt{\frac{N_{t_{min}}}{N}},\ldots,\sqrt{\frac{N_{T_r}}{N}}\right) = diag\left(\sqrt{\lambda_{N,t_{min}}},\ldots,\sqrt{\lambda_{N,T_r}}\right),$$

$$\zeta_N = diag\left(\sqrt{\frac{1}{N_{t_{min}}}},\ldots,\sqrt{\frac{1}{N_{T_r}}}\right).$$

The use of the weighting matrices λ_N and ζ_N is because households may have different initial working ages.

Following CH, the quasi-Bayesian estimators (QBE), also called Laplace type estimators (LTE), is constructed. Although the objective function in (4.4.3) is not a probability density function, it is transformed into a proper one by

$$p_N(\theta) = \frac{e^{L_N(\theta)}\pi(\theta)}{\int_\Theta e^{L_N(\theta)}\pi(\theta)\,d\theta}, \tag{4.4.5}$$

where $\pi(\theta)$ is the prior information. The $p_N(\theta)$ in equation (4.4.5) is called the *quasi-posterior* density function. Based on $p_N(\theta)$, given the penalty or loss function

$\rho_N(u)$, the corresponding risk function is

$$R_N(\xi) = \int_{\Theta} \rho_N(\theta - \xi)\, p_N(\theta)\, d\theta. \tag{4.4.6}$$

Following CH, the following assumptions are imposed on the loss function $\rho_N(u)$.

ASSUMPTION 3. *The loss function $\rho_N : R^d \to R_+$ satisfies:*

(i) $\rho_N(u) = \rho\left(\sqrt{N}u\right)$*, where* $\rho(u) \geq 0$ *and* $\rho(u) = 0$ *if and only if* $u = 0$*;*

(ii) ρ *is convex and* $\rho(h) \leq 1 + |h|^p$ *for some* $p \geq 1$*;*

(iii) $\varphi(\xi) = \int_{R^d} \rho(u - \xi)\, e^{-u'au} du$ *is minimized uniquely at some* $\tau \in R^d$ *for any finite* $a > 0$.

Given the loss function $\rho_N(u)$, based on risk function (4.4.6), the QBE for θ_0 is defined below.

Definition 4.4.1. *The QBE is the one minimizing the risk function* $R_N(\xi)$ *in (4.4.6):*

$$\widehat{\theta} = \arg \inf_{\xi \in \Theta} R_N(\xi). \tag{4.4.7}$$

4.4.2 Asymptotic Theory for analytical Solution for Optimal Policy

In this subsection, the asymptotic behavior of the estimator $\widehat{\theta}$ defined in (4.4.7) is studied. The following assumptions are imposed.

ASSUMPTION 4. *The function* $g_t\left(M_{i,t}^d, z_{i,t}^d; \theta, \chi\right)$ *defined in (4.4.2) satisfies the following conditions: (i)* $g_t(\cdot; \theta, \chi)$ *and* $\nabla_\theta g_t(\cdot; \theta, \chi)$ *are Borel measurable for each* $\theta \in \Theta$ *and* $\chi \in \Psi$*; (ii) given* $\chi \in \Psi$*,* $\nabla_\theta g_t\left(M_{i,t}^d, z_{i,t}^d; \theta, \chi\right)$ *is continuously differentiable on* Θ*; (iii)* $\nabla_{\theta\theta} g_t(\cdot; \theta, \chi)$ *is Borel measurable for each* $\theta \in \Theta$ *and* $\chi \in \Psi$.

ASSUMPTION 5. $G(\theta,\chi) = \nabla_\theta E\left[g_t\left(M_{i,t}^d, z_{i,t}^d;\theta,\chi\right)\right]$ *is continuous on* Θ *and* χ. $G(\theta_0,\chi_0)$ *is finite and has full rank.*

ASSUMPTION 6. $\lim_{N\to\infty}\lambda_N = \lambda$, $\lim_{N\to\infty} N/J = \gamma$ *for some constants* $\lambda,\gamma \in R^+$,

Remark 4.4.3. *Assumptions 4 and 5 are similar to those in Hansen (1982). The assumptions on the moment vector are essential for the study of asymptotic behavior of the estimator. Assumption 6 implies* N_t *is proportional to the total number of observations N. Assumption 6 also implicates that N is proportional to the number of households in the dataset,* N^{obs}.

When GMM is adopted during the first stage, the following two assumptions are imposed.

ASSUMPTION 7. *The first-order derivative of* $g_t\left(M_{i,t}^d, z_{i,t}^d;\theta,\chi\right)$ *with respect to* χ, $g_{t,\chi}\left(M_{i,t}^d, z_{i,t}^d;\theta,\chi\right)$ *satisfies the following conditions: (i)* $g_{t,\chi}(\cdot;\theta,\chi)$ *and* $\nabla_\theta g_{t,\chi}(\cdot;\theta,\chi)$ *are Borel measurable for each* $\theta\in\Theta$ *and* $\chi\in\Psi$*; (ii) given* $\chi\in\Psi$, $\nabla_\theta g_{t,\chi}\left(M_{i,t}^d, z_{i,t}^d;\theta,\chi\right)$ *is continuously differentiable on* Θ*; (iii)* $\nabla_{\theta\theta} g_{t,\chi}(\cdot;\theta,\chi)$ *is Borel measurable for each* $\theta\in\Theta$ *and* $\chi\in\Psi$.

ASSUMPTION 8. $G_\chi(\theta,\chi) = \nabla_\chi E\left[g_t\left(M_{i,t}^d, z_{i,t}^d;\theta,\chi\right)\right]$ *is continuous on* Θ *and* χ. $G_\chi(\theta_0,\chi_0)$ *is finite and full rank.*

Remark 4.4.4. *Assumptions 7 and 8 are similar to Assumptions 5 and 6. They are associated with* $g_{t,\chi}\left(M_{i,t}^d,\ z_{i,t}^d;\theta,\ \chi\right)$ *and necessary because the estimation error due to GMM must be taken into account. These two assumptions are not required if the calibration is used during the first stage.*

Finally, there are also some restrictions on the prior information $\pi(\theta)$.

ASSUMPTION 9. $\pi(\theta)$ *is continuous and uniformly positive over* Θ

In this paper, only GMM is used during the first stage because the calibration is a special case of GMM as explained in Remark 4.3.1. Based on the discussion above, we define

$$g_i(\theta;\chi) = \underbrace{\left(g_{i,t_{min}}(\theta;\chi), \dots, g_{i,T_r}(\theta;\chi)\right)'}_{T_m \text{ elements}}.$$

Furthermore, according to the standard assumption that households are independent across i, we have the following lemma and theorems.

Lemma 4.4.1. *Under Assumptions 5–8, $V_N(\theta;\widehat{\chi})$ defined in equation (4.4.4) has the following property, uniformly over Θ,*

$$\begin{aligned} V_N(\theta;\widehat{\chi}) \xrightarrow{p} &\lambda E\left[g_i(\theta;\chi_0)\, g_i(\theta;\chi_0)'\right]\lambda' \\ &+ \gamma\lambda E\left[g_{i,\chi}(\theta;\chi_0)\right]\Sigma_\chi E\left[g_{i,\chi}(\theta;\chi_0)'\right]\lambda' = V(\theta). \end{aligned}$$

Theorem 4.4.1. *Under Assumptions 1–9, for the estimator $\widehat{\theta}$ defined in (4.4.7),*

$$\sqrt{N}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \tau + \mathbb{N}(0, \Sigma_\theta),$$

where

$$\Sigma_\theta = \left[G_\theta'\lambda'\left(\lambda\Sigma_g\lambda' + \gamma\lambda G_\chi'\Sigma_\chi G_\chi\lambda'\right)^{-1}\lambda G_\theta\right]^{-1},$$

$$\tau = \arg\inf_{\alpha\in R^d}\left\{\int_{R^d}\rho(\alpha - u)\, f\left(u; 0, G_\theta'\lambda' W(\theta_0)\lambda G_\theta\right) du\right\},$$

where $f(\cdot,\mu,\Omega)$ is the multivariate normal density with mean μ and covariance Ω, $G_\theta = \nabla_\theta E[g_i(\theta_0;\chi_0)]$, $G_\chi = \nabla_\chi E[g_i(\theta_0;\chi_0)]$, $\Sigma_g = E\left[g_i(\theta_0;\chi_0)\, g_i(\theta_0;\chi_0)'\right]$.

Remark 4.4.5. *If the calibration is used during the first stage, then we have*

$$\sqrt{N}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \tau + \mathbb{N}\left(0, \left(G_\theta'\lambda'\left(\lambda\Sigma_g\lambda'\right)^{-1}\lambda G_\theta\right)^{-1}\right).$$

Since there is no need to take estimation error into account, the second term in the optimal weighting matrix disappears in the calibration.

Usually τ is difficult to evaluate at θ_0 since the value of θ_0 is unknown. However, if we choose the quadratic loss function, according to CH and the Bayesian literature, the estimator in Definition 4.4.1 becomes the mean of the quasi-posterior distribution in (4.4.5), which is called the *quasi-posterior mean* and defined as

$$\bar{\theta} = E_{p_N}[\theta] = \int_{\Theta} \theta p_N(\theta)\, d\theta. \tag{4.4.8}$$

The corollary below follows Theorem 4.4.1.

Corollary 4.4.2. *Under Assumptions 1–9, given $\rho_N(\cdot) = N \cdot u^2$ and the estimator $\bar{\theta}$ defined in (4.4.8),*

$$\sqrt{N}\left(\bar{\theta} - \theta_0\right) \xrightarrow{d} \mathbb{N}\left(0, \Sigma_\theta\right),$$

with $\Sigma_\theta = \left[G_\theta' \lambda' \left(\lambda \Sigma_g \lambda' + \gamma \lambda G_\chi' \Sigma_\chi G_\chi \lambda'\right)^{-1} \lambda G_\theta\right]^{-1}$, where the variables are the same as in Theorem 4.4.1. Meanwhile, Σ_θ has the following property.

$$N \cdot E_{p_N}\left[\left(\theta - \bar{\theta}\right)\left(\theta - \bar{\theta}\right)'\right] = \Sigma_\theta + o_p(1).$$

Remark 4.4.6. *From Corollary 4.4.2 with samples from $p_N(\theta)$, both the estimator and the asymptotic covariance, which are the mean and covariance of quasi-posterior distribution, can be simultaneously calculated. This is in contrast to extremum estimators where the estimator and the asymptotic covariance are obtained separately.*

4.4.3 Asymptotic Theory for Numerical Solution for Optimal Policy

In most cases, there is no analytical solution for the Bellman equation (4.2.7). Numerical methods are needed to solve the model inevitably introducing approximation errors. In this subsection, we develop conditions under which the results obtained in the last subsection continue to hold when numerical solutions are used.

Given the values of θ and χ, the (infeasible) exact solution for the policy function at age t for household i is denoted as $c_t\left(m_{i,t}^d, z_{i,t}^d; \theta, \chi\right)$. Denote the numerical approximation by $c_t^j\left(m_{i,t}^d, z_{i,t}^d; \theta, \chi\right)$ where j is the number of grid points in the finite range of $m_{i,t}^d$ based on which we can evaluate other optimal policies by using interpolation methods. The numerical solution $c_t^j\left(m_{i,t}^d, z_{i,t}^d; \theta, \chi\right)$ is indexed by j because the approximation admits refinements, i.e., when j goes to infinity, $c_t^j\left(m_{i,t}^d, z_{i,t}^d; \theta, \chi\right)$ converges to $c_t\left(m_{i,t}^d, z_{i,t}^d; \theta, \chi\right)$.

With the numerical solution, neither the exact objective function (4.4.3) nor the quasi-posterior density in (4.4.5) can be evaluated. Before we introduce our estimation procedure, let us first fix some new notations.

The approximated optimal consumption level for household i at age t is

$$C_t^j\left(M_{i,t}^d, z_{i,t}^d; \theta, \chi\right) = E_{P_{i,t}}\left[c_t^j\left(\frac{M_{i,t}^d}{P_{i,t}}, z_{i,t}^d; \theta, \chi\right)P_{i,t}\middle| F_{i,t}\right]. \tag{4.4.9}$$

The sample moment becomes

$$\begin{aligned} C_{i,t}^d - C_t^j\left(M_{i,t}^d, z_{i,t}^d; \theta_0, \chi_0\right) =& g_t^j\left(M_{i,t}^d,\ z_{i,t}^d; \theta_0,\ \chi_0\right) \\ =& g_{i,t}^j\left(\theta_0;\ \chi_0\right), \end{aligned} \tag{4.4.10}$$

for household i at age t, where $t = t_i, ..., T_r$. Then the approximate objective function is defined as

$$L_N^j(\theta) = -\frac{N}{2}\left[\lambda_N \bar{g}_N^j(\theta; \widehat{\chi})\right]' W_N^j(\theta; \widehat{\chi})\lambda_N \bar{g}_N^j(\theta; \widehat{\chi}), \tag{4.4.11}$$

where

$$\begin{aligned} \bar{g}_N^j(\theta; \widehat{\chi}) &= \left(\bar{g}_{t_{min}}^j(\theta; \widehat{\chi}), \ldots, \bar{g}_{T_r}^j(\theta; \widehat{\chi})\right)' \\ &= \left(\frac{1}{N_{t_{min}}}\sum_{i=1}^{N_{t_{min}}} g_{i,t_{min}}^j(\theta; \widehat{\chi}), \ldots, \frac{1}{N_{T_r+1}}\sum_{i=1}^{N_{T_r+1}} g_{i,T_r}^j(\theta; \widehat{\chi})\right)', \end{aligned}$$

$$W_N^j(\theta; \widehat{\chi}) = \left[V_N^j(\theta; \widehat{\chi})\right]^{-1},$$

$$\begin{aligned} V_N^j(\theta;\widehat{\chi}) = & \zeta_N \sum_{i=1}^{N^{obs}} \lambda_N \tilde{g}_i^j(\theta;\widehat{\chi}) \tilde{g}_i^j(\theta;\widehat{\chi})' \lambda_N' \zeta_N' \\ & + \frac{N}{J} \lambda_N \bar{g}_{N,\chi}^j(\theta;\widehat{\chi}) \widehat{\Sigma}_\chi \bar{g}_{N,\chi}^j(\theta;\widehat{\chi})' \lambda_N', \end{aligned} \tag{4.4.12}$$

$$\tilde{g}_i^j(\theta;\chi) = \underbrace{\left(0,\ldots,0,g_{i,t_i}^j(\theta;\chi),\ldots,g_{i,T_r}^j(\theta;\chi)\right)'}_{T_m \text{ elements}}.$$

Remark 4.4.7. *Based on the approximated objective function (4.4.11), one can use MSM to obtain the extremum estimator. If so, one must implement an iterative optimization algorithm in which the value and gradient of the objective function have to be numerically evaluated for each parameter updating. These computational efforts and their cost are demanding. Further, as pointed out in CH, sometimes the maximum estimator is the local optimum, not the global one.*

Based on equation (4.4.11), we can define the *approximated quasi-posterior* as

$$p_N^j(\theta) = \frac{e^{L_N^j(\theta)}\pi(\theta)}{\int_\Theta e^{L_N^j(\theta)}\pi(\theta)\,d\theta}. \tag{4.4.13}$$

Given the loss function $\rho_N(u)$, the risk function and estimator corresponding to the approximated quasi-posterior is

$$R_N^j(\xi) = \int_\Theta \rho_N(\theta-\xi)\, p_N^j(\theta)\, d\theta, \tag{4.4.14}$$

$$\widehat{\theta}^j = \arg\inf_{\xi\in\Theta} R_N^j(\xi). \tag{4.4.15}$$

Other variables remain the same as those in the case with the analytical solution.

Following FRS and Ackerberg, Geweke and Hahn (2009), the following assumption is imposed on numerical methods.

ASSUMPTION 10. *For all j, χ and z, over a finite range of m, $c_t^j(m,z;\,\theta,\chi)$ is continuous on m and continuously differentiable at all points except at a finite number of points.*

Remark 4.4.8. *Assumption 10 ensures the continuity of* $c_t^j(m,z;\theta,\widehat{\chi})$ *at all points and differentiability except at a finite number of points in the finite range of m. The lack of differentiability makes it possible to use numerical methods with kinks at a finite number of points. Such methods include the linear interpolation or the approximation within space spanned by linear basis functions. This assumption is satisfied naturally by most solution methods for dynamic economic models.*

FRS studied the econometric problem of computed dynamic models. They found that under some mild conditions, as the approximated policy functions converged to the exact ones, the approximated likelihood also converged to the exact likelihood. Meanwhile, as more data are included, a better approximation is required. Ackerberg, Geweke and Hahn (2009) examined the impact of approximation errors on a classical estimate of a simple time series model. They found the approximation errors are required to vanish at a certain speed as the sample size goes to infinity. Following Ackerberg, Geweke and Hahn (2009), the approximation error is defined as

$$\begin{aligned}\Delta_j = \sup_{\theta\in\Theta,\chi\in\Psi} & \left\{ \max_{z,m,t} \left\{ \left\| c_t^j(m,z;\theta,\chi) - c_t(m,z;\theta,\chi) \right\|, \right.\right. \\ & \left.\left. \left\| C_{t,\chi}^j(M,z;\theta,\chi) - C_{t,\chi}(M,z;\theta,\chi) \right\| \right\} \right\}. \end{aligned} \tag{4.4.16}$$

Remark 4.4.9. *Unlike Ackerberg, Geweke and Hahn (2009), we do not need to consider the approximation error associated with the first and second-order derivatives of the objective function. Note that* $t \in [t_{min}, T_r+1]$ *and from the dataset, the normalized wealth m and characteristic vector z are all bounded. Thus, given any generic* θ *and* χ*,* Δ_j *is controlled by the number of grid points j. Furthermore, if the calibration is adopted during the first stage, we do not have to consider the approximation error of* $C_{t,\chi}^j(m,z;\theta,\chi)$*.*

In accordance with Ackerberg, Geweke and Hahn (2009), the approximation error should disappear asymptotically, i.e., $j \to \infty$, as $N \to \infty$. Given Assumptions 1–10, the following theorem hold.

Theorem 4.4.3. *Under Assumptions 1–10, for the estimator $\widehat{\theta}^j$ defined in (4.4.15), if as $N \to \infty$,*

$$N\Delta_j \to 0,$$

then,

$$\sqrt{N}\left(\widehat{\theta}^j - \theta_0\right) \xrightarrow{d} \tau + \mathbb{N}\left(0, \Sigma_\theta\right),$$

with

$$\Sigma_\theta = \left[G_\theta' \lambda' \left(\lambda \Sigma_g \lambda' + \gamma \lambda G_\chi' \Sigma_\chi G_\chi \lambda'\right)^{-1} \lambda G_\theta\right]^{-1}.$$

Remark 4.4.10. *An approximate optimal policy for every household at every age inevitably introduces the approximation error. As the total number of observations increases, the error will accumulate. Theorem 4.4.3 requires that the accumulative approximation error be smaller than the sampling error, and thus is negligible. The detailed relationship between j and N in different numerical methods is left for future studies.*

Similarly, given the quadratic loss function, the approximated quasi-posterior mean is defined as

$$\bar{\theta}^j = E_{p_N^j}[\theta] = \int_\Theta \theta p_N^j(\theta)\, d\theta. \tag{4.4.17}$$

Corollary 4.4.4. *Under Assumptions 1–10, given the quadratic loss function $\rho_N(\cdot)$ and the estimator $\bar{\theta}^j$ defined in (4.4.17), if $N\Delta_j \to 0$ as $N \to \infty$, then,*

$$\sqrt{N}\left(\bar{\theta}^j - \theta_0\right) \xrightarrow{d} \mathbb{N}\left(0, \Sigma_\theta\right),$$

with $\Sigma_\theta = \left[G_\theta' \lambda' \left(\lambda \Sigma_g \lambda' + \gamma \lambda G_\chi' \Sigma_\chi G_\chi \lambda'\right)^{-1} \lambda G_\theta\right]^{-1}$, where the variables are the same as in Theorem 4.4.1. Meanwhile, Σ_θ has the following property.

$$N \cdot E_{p_N^j}\left[N\left(\theta - \bar{\theta}^j\right)\left(\theta - \bar{\theta}^j\right)'\right] = \Sigma_\theta + o_p(1), \tag{4.4.18}$$

where $E_{p_N^j}$ is the expectation with respect to $p_N^j(\theta)$.

Theorem 4.4.3 and Corollary 4.4.4 are important because they show that when the approximation errors disappears at a speed faster than the total number of observations, the approximated estimator shares the desirable properties of the estimator when policy functions are analytically available.

This result is related to that in FRS and Ackerberg, Geweke and Hahn (2009) with two differences. First, both papers focus on the likelihood inference, whereas the estimation framework is GMM in the present paper. Second, the disappearance rate in Theorem 4.4.3 is also different. In Ackerberg, Geweke and Hahn (2009), a static simple time series model is studied and the rate of the approximation errors is required to be faster than the square root of the time span, i.e., $o\left(T^{1/2}\right)$. The present paper focuses on the life-cycle model with finite horizon and the speed of the approximation error is required to be faster than the total number of the observations, i.e., $o(N)$.

Remark 4.4.11. *Theorem 4.4.3 and Corollary 4.4.4 show that only the approximation error of* $c_t^j\left(m_{i,t}^d, z_{i,t}^d; \theta, \widehat{\chi}\right)$ *and* $C_{t,\chi}^j\left(M_{i,t}^d, z_{i,t}^d; \theta, \widehat{\chi}\right)$ *need to be considered. If the calibration is used at the first stage, the approximation error of* $C_{t,\chi}^j\left(M_{i,t}^d, z_{i,t}^d; \theta, \widehat{\chi}\right)$ *can be ignored. However, if an optimization approach is used, other types of approximation errors, such as those in calculating the first- and second-order derivatives of the objective function, require careful attention, which may be very complicated and difficult to control in practice.*

Remark 4.4.12. *Equation (4.4.18) can be used to compute the asymptotic covariance. On the one hand, it is the by-product of samples from the approximated quasi-posterior distribution. On the other hand, it avoids numerical evaluations of* G_θ *and* G_χ*.*

4.4.4 Estimation

The theoretical results in previous subsections are attractive. However, sampling from the quasi-posterior distribution remains a difficult problem. The MCMC method does not work well here since it requires sampling sequentially many times and numerically evaluating the objective function at each updating. Instead of MCMC, importance sampling is used together with GPU to enhance the computational speed.

In practice, it is very hard to find a good proposal distribution for the importance sampling. Direct sampling from the prior can be computationally inefficient. Recognizing this problem, we adapt the algorithm proposed in Creel and Kristensen (2016) to estimate finite-horizon life-cycle models. The algorithm for the estimation is summarized in Algorithms 1 and 2. Both algorithms request a great number of quasi-posterior density evaluations. The usual CPU time will be high. Thanks to the availability of GPU, we can solve the model numerically given a great number of parameter values and do the interpolation in parallel.

In Algorithm 1, δ and $\exp(L)$ are close to zero. They are threshold values for the search of area and selection of particles with significant quasi-posterior density values, respectively. Specifically, steps 10–24 ensure that the shrinking sampling area is sufficiently narrow given K_1 and δ, and that they are adaptive to different datasets. Besides, step 25 selects particles in S with significant quasi-posterior density values, denoted as $\tilde{S}$. Step 26–29 uniformly draw K particles from $\tilde{S}$ and construct the proposal distribution for important sampling, which is a mixture of normal distributions.

In Algorithm 2, when $K_3 \to \infty$, $\widehat{\bar{\theta}} \to \bar{\theta}$, $\widehat{Var(\theta)} \to Var(\theta)$, where $Var(\theta)$ is the quasi-posterior covariance with respect to $p_N(\theta)$, since

$$\widehat{\bar{\theta}} = \frac{\sum_{k=1}^{K_3} \omega^{(k)} \theta^{(k)}}{\sum_{k=1}^{K_3} \omega^{(k)}} \to \int_{\Theta} \theta p_n(\theta)\, d\theta = \bar{\theta},$$

$$\begin{aligned} \widehat{Var(\theta)} &= \frac{1}{\sum_{k=1}^{K_3} \omega^{(k)}} \sum_{k=1}^{K_3} \omega^{(k)} \left(\theta^{(k)} - \widehat{\bar{\theta}}\right) \left(\theta^{(k)} - \widehat{\bar{\theta}}\right)' \\ &\to \int_{\Theta} \theta \theta' p_n(\theta)\, d\theta + \bar{\theta}\bar{\theta}' \\ &\equiv \int_{\Theta} \left(\theta - \bar{\theta}\right) \left(\theta - \bar{\theta}\right)' p_N(\theta)\, d\theta. \end{aligned}$$

Algorithm 1 Construction of Proposal Distribution

Input: The number of samples K_1, the selected number of particles K_2, the covariance for the random perturbation Σ, the tolerance level δ, the threshold value L, the number of component in proposal distribution K.
Set up $i = 0$
for $k = 1$ to K_1 **do**
  Draw $\theta_k^i \sim \pi(\theta)$.
  Compute $\omega_k^i = L_N\left(\theta_k^i\right) + \log \pi\left(\theta_k^i\right)$.
end for
Set up the set of particles $S = \emptyset$.
Compute $V_1 = \max \omega_k^0$
Compute $V_2 = V_1 + 2\delta$
while $|V_1 - V_2| < \delta$ **do**
  Sort $\left\{\omega_k^i\right\}_{k=1}^{K_1}$ in descending order.
  Select the first K_2 of the sorted ω_k^i and associated θ_k^i, obtain $\left\{\tilde{\omega}_k^i\right\}_{k=1}^{K_2}$ and $\left\{\tilde{\theta}_k^i\right\}_{k=1}^{K_2}$
  $S = S \bigcup \left\{\tilde{\theta}_k^i\right\}_{k=1}^{K_2}$.
  for $k = 1$ to K_2 **do**
    Compute $\omega_{Norm}^k = \frac{e^{\tilde{\omega}_k^i}}{\sum_{k=1}^{K} e^{\tilde{\omega}_k^i}}$.
  end for
  for $k = 1$ to K_1 **do**
    Draw $\tilde{\tilde{\theta}}_k^i \sim Multinomial\left(\left\{\tilde{\theta}_k^i\right\}_{k=1}^{K_2}, \left\{\omega_{Norm}^k\right\}_{k=1}^{K_2}\right)$
    Compute $\theta_k^{i+1} = \tilde{\tilde{\theta}}_k^i + \varepsilon_k^{i+1}, \varepsilon_k^{i+1} \sim N(0, \Sigma)$.
    Compute $\omega_k^{i+1} = L_N\left(\theta_k^{i+1}\right) + \log \pi\left(\theta_k^{i+1}\right)$.
  end for
  Compute $V_1 = V_2$.
  Compute $V_2 = \max \omega_k^{i+1}$.
end while
Select the particle points in S that satisfies $\omega_k^i - V_2 > L$, obtain $\tilde{S}$.
for $k = 1$ to K **do**
  Draw θ_k^{IS} from $\tilde{S}$ uniformly.
end for
Define the importance sampling density as the mixture of densities associated with each drawn θ_k^{IS}:

$$q(\theta) = \sum_{k=1}^{K} p_k q_k\left(\theta | \theta_k^{IS}\right),$$

where $p_k = e^{\omega_k} / \sum_{k=1}^{K} e^{\omega_k}$, $\omega_k == L_N\left(\theta_k^{IS}\right) + \log \pi\left(\theta_k^{IS}\right)$, and $q_k\left(\theta | \theta_k^{IS}\right) = N\left(\theta_k^{IS}, \Sigma\right)$. Or $p_k = \frac{1}{K}$, for $k = 1, \ldots, K$.
Output: $q(\theta)$.

Algorithm 2 Estimator Calculation

Input: The number of samples K_3, the proposal distribution $q(\theta)$.
for $k = 1$ to K_3 **do**
  Draw $\theta^{(k)} \sim q(\theta)$.
  Compute $\tilde{\omega}^{(k)} = e^{L_N\left(\theta^{(k)}\right)} \pi\left(\theta^{(k)}\right)$.
end for
Compute the estimator

$$\widehat{\bar{\theta}} = \frac{\sum_{k=1}^{K_3} \omega^{(k)} \theta^{(k)}}{\sum_{k=1}^{K_3} \omega^{(k)}},$$

$$\widehat{Var(\theta)} = \frac{1}{\sum_{k=1}^{K_3} \omega^{(k)}} \sum_{k=1}^{K_3} \omega^{(k)} \left(\theta^{(k)} - \widehat{\bar{\theta}}\right)\left(\theta^{(k)} - \widehat{\bar{\theta}}\right)',$$

where $\omega^{(k)} = \tilde{\omega}^{(k)} / q\left(\theta^{(k)}\right)$.
Output: $\widehat{\bar{\theta}}$, $\widehat{Var(\theta)}$.

Remark 4.4.13. *The numerical evaluation of the quasi-posterior density values is costly computationally. GPU can enhance the computational speed greatly since it can solve the model and compute the density values in parallel given a great number of sampled parameters. Steps 10–24 are adaptive since the area with the largest posterior density values will be automatically found given the dataset, δ and K_1.*

4.5 Monte Carlo Studies

In this section, two models are studied to examine the performance of the new approach. One is the life-cycle model without exogenous dynamic latent state. The other one is a simplified version of the illustrative model.

4.5.1 The Case without Dynamic Latent State

The households are faced with the same utility maximization problem, i.e.,

$$\max_{\{c_t\}_{t=0}^{T}} E_0\left[\sum_{t=0}^{T}\beta^t\frac{c_t^{1-\rho}}{1-\rho}\right], \tag{4.5.1}$$

$$s.t.\ m_{t+1}=R(m_t-c_t)+y\varepsilon_{t+1}, 0\le t<T,$$

$$c_t\in(0,m_t], \text{with } m_0 \text{ given},$$

where β is the subjective discount factor, ρ the risk aversion of the households, R the gross interest rate, y the income level for the households from period $t=0$ to $t=T$, ε_{t+1} the income shock associated with the income at each period and $\varepsilon_{t+1}\overset{i.i.d.}{\sim}\log N\left(-\frac{\sigma_\varepsilon^2}{2},\sigma_\varepsilon^2\right)$, m_t the liquid wealth at the beginning of period t and c_t the consumption level that chosen by the households, which is in the budget constraint $(0,m_t]$. Thus, the Euler equations for the life-cycle model are

$$c_t^{-\rho}=R\beta E_t\left[c_{t+1}(m_{t+1})^{-\rho}\right], m_{t+1}=R(m_t-c_t)+y\varepsilon_{t+1}, 0\le t\le T-1, \tag{4.5.2}$$

where at period T, $c_T=m_T$, which results from the households seeking to consume all their wealth at the last period. There are no close-form solutions for the optimal consumptions, thus a numerical method is required. Conditional on the values of parameters, EGM is used to construct the grid of the optimal consumption at each period. The detail is illustrated in Appendix .4.2.

In this study, the true values of the parameters are reported in Table 4.1. Conditional on the values listed in Table 4.1, we solve the model numerically and simulate a data set $\left\{c_{i,t}^*, m_{i,t}^d\right\}_{t=0}^{T}$ for each household i, where the initial wealth $m_{0,i}^d$ is drawn from a truncated normal distribution with mean 5 and variance 100 ranging from 0 to infinity, i.e., $N(5,100)I\{x>0\}$, where I is the indicator function. The optimal consumption $c_{i,t}^*$ is interpolated based on the consumption grid obtained from numerical solving. The measurement error is added, $c_{i,t}^d=c_{i,t}^*+\varepsilon_{i,t}$,

Table 4.1: The Values of Parameters Used to Simulate Data

T	β	ρ	R	y	σ_ε^2
10	0.96	2	1.03	0.5	0.04

Table 4.2: The Bias and RMSE of the Estimator for β and ρ

	β		ρ	
	Bias	RMSE	Bias	RMSE
$N^{obs} = 1000$	-1.3602×10^{-3}	3.89×10^{-3}	0.2311	0.6780
$N^{obs} = 1500$	-1.4685×10^{-3}	3.407×10^{-3}	0.2535	0.6008
$N^{obs} = 2000$	-6.2943×10^{-4}	2.683×10^{-3}	0.1081	0.4692
$N^{obs} = 3000$	4.3860×10^{-4}	2.2×10^{-3}	0.0715	0.3926

$\varepsilon_{i,t} \overset{i.i.d.}{\sim} N\left(0, \sigma_\varepsilon^2\right)$, where $\sigma_\varepsilon^2 = 0.005^2$. The numbers of households simulated are $N^{obs} = 1000, 1500, 2000, 3000$, respectively and the number of replications for each case is 200. For each replication, the simulated noisy data $\left\{c_{i,t}^d, m_{i,t}^d\right\}_{t=1}^T$ are used to estimate the parameters ρ and β.

In order to estimate the parameters, the priors for the two parameters are set to

$$\beta \sim U(0.5, 1), \rho \sim U(0, 15),$$

where $U(a,b)$ is the uniform distribution ranging from a to b. For β, based on the economic theory, it should satisfy $\beta \in (0,1)$ and usually it is assumed to be around 0.9. Thus the prior for β is uninformative. Besides, for the risk averse parameter, ρ, the range between 0 and 15 is also quite uninformative.

Algorithms 1 and 2 are applied to estimate the model (for more details of the estimation, please refer to Appendix .4.3) and the bias and root mean square error (RMSE) are computed for each parameter in every scenario. The bias and RMSE are defined in Appendix .4.1. The results are listed in Table 4.2. It is obvious that as the sample size increases, the bias of both parameters decreases. Further, the

[2] Jøgensen (2017) estimated the variance of measurement error, which was approximately 0.46. But the sample size he used ranged from 150,000 to 800,000. Since the sample sizes in Monte Carlo studies are between 1000 and 3000, the variance of measurement error is proportionally set as 0.005 in terms of the variance of sample moments.

RMSE of both parameters also decreases and the magnitude of all the RMSE is proportional to the square root of the sample size approximately. This simulation study justifies the asymptotic theory and the usefulness of the algorithm.

4.5.2 The Case with Dynamic Latent State

In this subsection, a simplified life-cycle model in GP is considered to examine the performance of the new approach. The model is defined in the following. The household i is faced with the following optimization problem,

$$\max_{C_{i,\tau}} E_{t_i}\left[\sum_{\tau=t_0}^{T_r}\beta^{\tau-t_0}\frac{C_{i,\tau}^{1-\rho}}{1-\rho}+\kappa\beta^{T_r+1-t_0}\frac{\left(M_{i,T_r+1}+H_{i,T_{r+1}}\right)^{1-\rho}}{1-\rho}\right] \tag{4.5.3}$$

$$s.t. M_{i,t+1}=R\left(M_{i,t}-C_{i,t}\right)+Y_{i,t+1}, t_0\leq t\leq T_r-1$$

$$M_{i,T_r+1}=R\left(M_{i,T_r}-C_{i,T_r}\right), t=T_r,$$

$$C_{i,t}\in\left(0,M_{i,t}\right], \text{with } M_{i,t_0} \text{ given.}$$

The model specification is almost the same as the illustrative model except that all households start to work at the same age and the marginal utility shifter is not included. The income process is also the same and is defined as,

$$\begin{cases} Y_{i,t}=P_{i,t}\varepsilon_{i,t}, \\ P_{i,t}=G_tP_{i,t-1}\varsigma_{i,t}, \end{cases} \quad t_i\leq t\leq T_r,$$

$$\varepsilon_{i,t}=\begin{cases}\mu, & \text{with probability } p,\\ \xi_{i,t}, & \text{with probability } 1-p,\end{cases} \quad \text{where } \log\xi_{i,t}\overset{i.i.d.}{\sim}N\left(0,\sigma_{\varepsilon}^2\right),$$

$$\log\varsigma_{i,t}\overset{i.i.d.}{\sim}N\left(0,\sigma_{\varsigma}^2\right).$$

The parameters of the income process are given and the ratio-form Bellman

Table 4.3: Parameter values used to simulate data.

$\{G_t\}_{t=1}^{10}$	R	σ_ε^2	σ_ς^2	p	μ	β	ρ	T_r	γ_1	t_0
Figure 4.1	1.03	0.04	0.02	0.03	10^{-6}	0.96	2	10	0.07	1

equation is now,

$$V_t(m_{i,t};\theta) = \max_{c_{i,t}} \left\{ \frac{c_{i,\tau}^{1-\rho}}{1-\rho} + +\beta E_t\left[(G_{t+1}N_{i,t+1})^{1-\rho} V_{t+1}(m_{i,t+1};\theta) \right] \right\} \tag{4.5.4}$$

$$s.t. m_{i,t+1} = (m_{i,t} - c_{i,t}) \frac{R}{G_{i,t+1}\varsigma_{i,t+1}} + \varepsilon_{i,t+1}, t_i \le t \le T_r - 1,$$

$$m_{i,T_r+1} = R(m_{i,T_r} - c_{i,T_r}), t = T_r,$$

$$c_{i,t} \in (0, m_{i,t}],$$

with

$$\begin{aligned} V_{T_r+1}(m_{i,T_r+1};\theta) &= \kappa \frac{(m_{i,T_r+1}+h)^{1-\rho}}{1-\rho} \\ &= \frac{1}{(1-\rho)\kappa^{-\frac{1}{\rho}}} \left(\kappa^{-\frac{1}{\rho}} m_{i,T_r+1} + \kappa^{-\frac{1}{\rho}} h \right)^{1-\rho} \\ &= \frac{1}{(1-\rho)\gamma_1} (\gamma_1 m_{i,T_r+1} + \gamma_0)^{1-\rho}, \end{aligned}$$

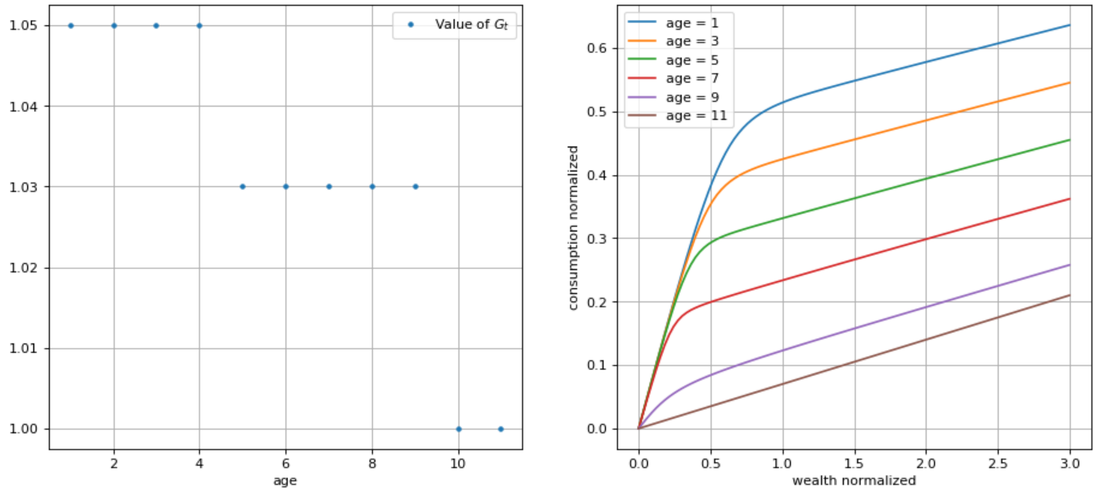
where $c_{i,t}$ and $m_{i,t}$ are the normalized values of consumption level $C_{i,t}$ and wealth $M_{i,t}$, respectively. For simplicity, γ_0 is equal to 0, which is consistent with the result obtained by GP. The value function after retirement becomes

$$V_{T_r+1}(m_{i,T_r+1};\theta) = \frac{1}{(1-\rho)\gamma_1} (\gamma_1 m_{i,T_r+1})^{1-\rho}. \tag{4.5.5}$$

The structural parameter is now $\theta = \{\beta, \rho, \gamma_1\}$. The values of parameters for the simulation are listed in Table 4.3.

The values of $\{G_t\}_{t=1}^{10}$ are described in the left panel of Figure 4.1, which is the same as Jøgensen (2016). The discount factor β, gross interest rate R, income shock probability p, variance of transitory shock σ_ε^2, retirement rule parameter γ_1

Figure 4.1: The values of G_t and the policy functions for Bellman equation in ratio form



Notes: The left panel presents the plots of the value of G_t at different ages. The right panel is the numerical solution of the ratio-form model (4.5.4).

and variance of the shock to permanent income σ_ς^2 are approximately equal to those in GP. Following Jøgensen (2016), the risk aversion ρ equals 2 and the value of μ is very close to zero.

For this model, the corresponding ratio-form Euler equations are

$$c_{i,t}^{-\rho} = \max\left\{m_{i,t}^{-\rho}, \beta R E_{\varsigma_{i,t+1},\varepsilon_{i,t+1}}\left[\left(G_{t+1}\varsigma_{i,t+1}\right)^{-\rho} c_{i,t+1}\left(m_{i,t+1}\right)^{-\rho}\right]\right\}, t_0 \leq t \leq T_r - 1,$$

$$c_{i,T_r}^{-\rho} = \max\left\{m_{i,T_r}^{-\rho}, \beta R\left(\gamma_1 m_{i,T_r+1}\right)^{-\rho}\right\}, \text{ at age } T_r.$$

EGM is used to solve the model (for more details, one can refer to Appendix .4.2). The solution of the ratio-form model is presented in the right panel of Figure 4.1.

In the simulation, we assume at age $t = 1$, the corresponding permanent component of income $P_{i,1}^d$ for every household is drawn from a log-normal distribution, i.e.,

$$\log P_{i,1}^d \sim N\left(0, \sigma_\varsigma^2\right), \forall i = 1, \ldots, N^{obs},$$

where N^{obs} is the number of simulated households. We then simulate an income panel dataset $\left\{Y_{i,t}^{d}, P_{i,t}^{d}\right\}_{t=1}^{10}$ for each household i. Meanwhile, household's initial wealth at age 1, $M_{i,1}^{d}$, is sampled from a truncated normal distribution with mean 1 and variance 1 ranging from 0 to infinity, i.e., $M_{i,1}^{d} \sim N(1,1) I\{x>0\}$, for $i = 1,\ldots,N^{obs}$, where I is the indicator function.

The Bellman equation in ratio form is solved by EGM and we obtain the consumption grid at each period. At each t, we normalize the wealth $m_{i,t}^{d} = \frac{M_{i,t}^{d}}{P_{i,t}^{d}}$ and use the grid to interpolate the corresponding optimal ratio-form consumption $c_{i,t}^{*}$. We then compute the optimal consumption level as $C_{i,t}^{*} = c_{i,t}^{*} P_{i,t}^{d}$ and obtain $\left\{C_{i,t}^{*}, M_{i,t}^{d}, Y_{i,t}^{d}, P_{i,t}^{d}\right\}_{t=1}^{10}$ for each household i. Following the simulation procedure in the last subsection, we add the measurement error, $C_{i,t}^{d} = C_{i,t}^{*} + \varepsilon_{i,t}$, $\varepsilon_{i,t} \overset{i.i.d.}{\sim} N\left(0, \sigma_{\varepsilon}^{2}\right)$, $\sigma_{\varepsilon}^{2} = 0.008$. Finally we have $\left\{C_{i,t}^{d}, M_{i,t}^{d}, Y_{i,t}^{d}\right\}_{t=1}^{10}$, for $i = 1,\ldots,N^{obs}$, which is used for estimation.

In order to obtain the sample moment vector, the Kalman filter is used to filter the income observations to obtain the mean and variance for $P_{i,t}$ at each t for household i. The Kalman filter for income process is documented in detail in Appendix .4.4.

To estimate the parameters ρ, β, γ_1, the following priors are used,

$$\rho \sim U(0,15), \beta \sim U(0.5,1), \gamma_1 \sim U(0,1).$$

It is quite intuitive that households must use their wealth to support their lives after retirement and they would not consume all their liquid wealth in the first year after retirement. Thus, the bound is quite reasonable and uninformative. For the priors for ρ and β, they are also uninformative as argued earlier.

We use Algorithms 1 and 2 to do the estimation. In the estimation, we set $K_1 = K_3 = 38400$, $K_2 = 1280$, $\Sigma = diag(0.0001, 0.04, 0.0001)$, $\delta = 0.5$, $L = -10$, $K = 7680$ and the number of grids in EGM is 100. The sample sizes considered here are $N^{obs} = 1500, 2000, 3000$, respectively. The number of replications is 50. The biases and RMSE of the estimation are reported in Table 4.4.

The results in Table 4.4 have similar patterns to the outputs in the preceding

Table 4.4: The bias and RMSE of the estimator

		$N^{obs} = 1500$	$N^{obs} = 2000$	$N^{obs} = 3000$
β	Bias	2.8583×10^{-4}	-3.0266×10^{-5}	-3.7344×10^{-5}
	RMSE	2.8394×10^{-3}	2.7242×10^{-3}	1.9599×10^{-3}
ρ	Bias	-3.4112×10^{-2}	-1.8676×10^{-3}	-1.5472×10^{-2}
	RMSE	0.1726	0.1676	0.1321
γ_1	Bias	5.7411×10^{-5}	2.2704×10^{-5}	-6.8096×10^{-6}
	RMSE	2.0429×10^{-4}	1.6755×10^{-4}	1.3224×10^{-4}

subsection. The bias for all parameters decreases as the sample size increases. Further, the RMSE is approximately proportional to the square root of sample size as predicted by theory. In summary, the results in Table 4.4 still justify the asymptotic theory.

4.6 Conclusion

In this paper, a quasi-Bayesian estimator is introduced for structural parameters in finite-horizon life-cycle models. The asymptotic normality of the estimator is derived when an analytical solution for the model exists. When the policy functions are not analytically available, it is shown that if the approximation errors caused by numerical solving vanish fast enough, the estimator remains to be asymptotically normal. Further, it is shown that the estimator reaches the efficiency bound in the GMM framework. In the proposed method, the usual optimization procedure is converted into a sampling procedure, thereby avoiding the numerical evaluation for the gradient of objective function and alleviating the local optimum problem. The estimator and associated asymptotic covariance can be computed simultaneously. The estimation procedure is also easy to parallelize, facilitating a GPU-based and adaptive algorithm to enhance computational efficiency. The estimation procedure is also illustrated based on a variant of the model in GP.

In general our estimator is less efficient than the full likelihood-based procedures, such as those proposed by FRS and Ackerberg, Geweke, and Hahn (2009).

However, our procedure is less stringent about the model specification. For example, the distribution is left unspecified in our approach. Hence, our set up may be more appealing to empirical researchers who are agnostic about distributional behaviors of the errors.

There are many possible extensions for this method. For example, finite-horizon life-cycle models with endogenous discrete choices can be considered since these models have received considerable attention recently; see Iskhakov et.al. (2017), Kaplan and Violante (2014) and references therein. Meanwhile, the present paper only focuses on the estimation. There also remains plenty of work related to inference. These topics are left for future research.

Chapter 5 Summary of Conclusions

In Chapter 2, we propose a new Bayesian test statistic to test a point null hypothesis based on a quadratic loss function. The main advantages of the proposed test statistic are as follows. Relative to the BF, first, it is well-defined under improper prior distributions; second, it is immune to Jeffreys-Lindley's paradox; third, it is easy to compute, even for the latent variable models; fourth, its asymptotic distribution is pivotal so that the threshold values are easy to obtain; fifth, its NSE can be easily obtained. Relative to the LM test, first, it can incorporate the prior information to improve hypothesis testing when the sample size is small; second, it does not suffer from the problem of taking negative values; third,it does not need to invert any matrix.

In Chapter 3, we propose a new test statistic to test for a point null hypothesis which can be treated as the posterior version of the Wald test. Compared with existing methods, the proposed statistic has many important advantages. First, it is well-defined under improper prior distributions. Second, it avoids Jeffreys-Lindley's paradox. Third, its asymptotic distribution is a χ^2 distribution under the null hypothesis and repeated sampling. This property is the same as the Wald statistic so that the critical values can be easily obtained. Fourth, it is very easy to compute as it is based on the posterior mean and posterior variance of the parameters of interest. Fifth, it can be used to test hypotheses that imposes nonlinear relationships among the parameters of interest, for which the BF is difficult to use. Sixth, for latent variable models for which the MLE and the Wald test are more difficult to obtain, the proposed statistic is the by-product of posterior sampling. Finally, only posterior sampling for the alternative hypothesis is needed for the proposed statistic.

In Chapter 4, we introduce a quasi-Bayesian estimator for structural parameters in finite-horizon life-cycle models. The asymptotic normality of the estimator is derived no matter whether there exists analytical solution for the model. Further, it is shown that the estimator reaches the efficiency bound in the GMM framework. In the proposed method, the usual optimization procedure is converted into a sampling procedure, thereby avoiding the numerical evaluation for the gradient of objective function and alleviating the local optimum problem. The estimator and associated asymptotic covariance can be computed simultaneously. The estimation procedure is also easy to parallelize, facilitating a GPU-based and adaptive algorithm to enhance computational efficiency. The estimation procedure is also illustrated based on a variant of the model in Gourinchas and Parker (2002)

Bibliography

ACKERBERG, D., J. GEWEKE, AND J. HAHN (2009): "Comments on convergence properties of the likelihood of computed dynamic models," *Econometrica*, 77, 2009–2017.

AÏT-SAHALIA, Y., J. FAN, R. J. LAEVEN, C. D. WANG, AND X. YANG (2017): "Estimation of the continuous and discontinuous leverage effects," *Journal of the American Statistical Association*, 112, 1744–1758.

AN, S. AND F. SCHORFHEIDE (2007): "Bayesian analysis of DSGE models," *Econometric reviews*, 26, 113–172.

ANDREWS, D. W. (1987): "Consistency in nonlinear econometric models: A generic uniform law of large numbers," *Econometrica: Journal of the Econometric Society*, 1465–1471.

BAUWENS, L. AND D. VEREDAS (2004): "The stochastic conditional duration model: a latent variable model for the analysis of financial durations," *Journal of econometrics*, 119, 381–412.

BERGER, J. O. (2013): *Statistical decision theory and Bayesian analysis*, Springer Science & Business Media.

BERGER, J. O. AND J. M. BERNARDO (1992): "On the development of the reference prior method," *Bayesian statistics*, 4, 35–60.

BERGER, J. O. AND L. R. PERICCHI (1996): "The intrinsic Bayes factor for model selection and prediction," *Journal of the American Statistical Association*, 91, 109–122.

BERNARDO, J. M. AND R. RUEDA (2002): "Bayesian hypothesis testing: A reference approach," *International Statistical Review*, 70, 351–372.

BLACK, F. (1976): "Studies of Sock Market Volatility Changes," *Proceedings of the American*.

BREUSCH, T. S. AND A. R. PAGAN (1980): "The Lagrange multiplier test and its applications to model specification in econometrics," *The Review of Economic Studies*, 47, 239–253.

BROWNING, M. AND M. EJRNÆS (2009): "Consumption and children," *The Review of Economics and Statistics*, 91, 93–111.

CAGETTI, M. (2003): "Wealth accumulation over the life cycle and precautionary savings," *Journal of Business & Economic Statistics*, 21, 339–353.

CAMPBELL, J. Y. (2006): "Household finance," *The journal of finance*, 61, 1553–1604.

CARROLL, C. D. (1997): "Buffer-stock saving and the life cycle/permanent income hypothesis," *The Quarterly journal of economics*, 112, 1–55.

——— (2001): "Death to the log-linearized consumption Euler equation!(And very poor health to the second-order approximation)," *Advances in Macroeconomics*, 1.

——— (2006): "The method of endogenous gridpoints for solving dynamic stochastic optimization problems," *Economics letters*, 91, 312–320.

CHEN, C.-F. (1985): "On asymptotic normality of limiting density functions with Bayesian implications," *Journal of the Royal Statistical Society. Series B (Methodological)*, 540–546.

CHEN, X., T. CHRISTENSEN, AND E. T. TAMER (2016): "MCMC confidence sets for identified sets," .

CHERNOZHUKOV, V. AND H. HONG (2003): "An MCMC approach to classical estimation," *Journal of Econometrics*, 115, 293–346.

CHIB, S. (1995): "Marginal likelihood from the Gibbs output," *Journal of the american statistical association*, 90, 1313–1321.

CHIB, S. AND I. JELIAZKOV (2001): "Marginal likelihood from the Metropolis–Hastings output," *Journal of the American Statistical Association*, 96, 270–281.

CHRIS, B. (2008): "Introductory econometrics for finance," *Cambrige, Cambrige University*.

CREAL, D. (2012): "A survey of sequential Monte Carlo methods for economics and finance," *Econometric reviews*, 31, 245–296.

CREEL, M., J. GAO, H. HONG, AND D. KRISTENSEN (2015): "Bayesian Indirect Inference and the ABC of GMM," *arXiv preprint arXiv:1512.07385*.

CREEL, M. AND D. KRISTENSEN (2015): "ABC of SV: Limited information likelihood inference in stochastic volatility jump-diffusion models," *Journal of Empirical Finance*, 31, 85–108.

——— (2016): "On selection of statistics for approximate Bayesian computing (or the method of simulated moments)," *Computational Statistics & Data Analysis*, 100, 99–114.

DEMPSTER, A. P., N. M. LAIRD, AND D. B. RUBIN (1977): "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, 1–38.

DICICCIO, T. J., R. E. KASS, A. RAFTERY, AND L. WASSERMAN (1997): "Computing Bayes factors by combining simulation and asymptotic approximations," *Journal of the American Statistical Association*, 92, 903–915.

DOUCET, A. AND A. M. JOHANSEN (2009): "A tutorial on particle filtering and smoothing: Fifteen years later," *Handbook of nonlinear filtering*, 12, 3.

DOUCET, A., N. SHEPHARD, ET AL. (2012): "Robust inference on parameters via particle filters and sandwich covariance matrices," *University of Oxford, Department of Economics*.

DUFFIE, D. AND K. J. SINGLETON (1990): "Simulated moments estimation of Markov models of asset prices," .

EDWARDS, W., H. LINDMAN, AND L. J. SAVAGE (1963): "Bayesian statistical inference for psychological research." *Psychological review*, 70, 193.

EFRON, B., A. GOUS, R. KASS, G. DATTA, AND P. LAHIRI (2001): "Scales of evidence for model selection: Fisher versus Jeffreys," *Lecture Notes-Monograph Series*, 208–256.

ENGLE, R. F. (1984): "Wald, likelihood ratio, and Lagrange multiplier tests in econometrics," *Handbook of econometrics*, 2, 775–826.

ENGLE, R. F. AND J. R. RUSSELL (1998): "Autoregressive conditional duration: a new model for irregularly spaced transaction data," *Econometrica*, 1127–1162.

FAGERENG, A., C. GOTTLIEB, AND L. GUISO (2017): "Asset market participation and portfolio choice over the life-cycle," *The Journal of Finance*, 72, 705–750.

FERNÁNDEZ-VILLAVERDE, J., J. F. RUBIO-RAMÍREZ, AND M. S. SANTOS (2006): "Convergence properties of the likelihood of computed dynamic models," *Econometrica*, 74, 93–119.

FISCHER, M. AND M. Z. STAMOS (2013): "Optimal life cycle portfolio choice with housing market cycles," *The Review of Financial Studies*, 26, 2311–2352.

GALLANT, A. R. AND H. WHITE (1988): *A unified theory of estimation and inference for nonlinear dynamic models*, Blackwell.

GEWEKE, J. (2005): *Contemporary Bayesian econometrics and statistics*, vol. 537, John Wiley & Sons.

——— (2007): "Bayesian model comparison and validation," *American Economic Review*, 97, 60–64.

GEWEKE, J., G. KOOP, AND H. VAN DIJK (2011): *The Oxford handbook of Bayesian econometrics*, Oxford University Press.

GHOSH, J. AND R. RAMAMOORTHI (????): "Bayesian Nonparametrics. 2003," .

GILKS, W. R., N. BEST, AND K. TAN (1995): "Adaptive rejection Metropolis sampling within Gibbs sampling," *Applied Statistics*, 455–472.

GOURINCHAS, P.-O. AND J. A. PARKER (2002): "Consumption over the life cycle," *Econometrica*, 70, 47–89.

GRIBKOVA, N. (1995): "Bounds for absolute moments of order statistics," *Exploring Stochastic Laws: Festschrift in honor of the 70th birthday of Acad. VS Korolyuk*, 129–134.

HAN, C. AND B. P. CARLIN (2001): "Markov chain Monte Carlo methods for computing Bayes factors: A comparative review," *Journal of the American Statistical Association*, 96, 1122–1132.

HANSEN, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the Econometric Society*, 1029–1054.

HEIDELBERGER, P. AND P. D. WELCH (1983): "Simulation run length control in the presence of an initial transient," *Operations Research*, 31, 1109–1144.

HEYDE, C. AND I. JOHNSTONE (1979): "On asymptotic posterior normality for stochastic processes," *Journal of the Royal Statistical Society. Series B (Methodological)*, 184–189.

IMAI, S., N. JAIN, AND A. CHING (2009): "Bayesian estimation of dynamic discrete choice models," *Econometrica*, 77, 1865–1899.

ISKHAKOV, F., T. JØRGENSEN, J. RUST, AND B. SCHJERNING (2015): "Estimating discrete-continuous choice models: The endogenous grid method with taste shocks," .

JEFFREYS, H. (1998): *The theory of probability*, OUP Oxford.

JØRGENSEN, T. H. (2016): "Euler equation estimation: Children and credit constraints," *Quantitative Economics*, 7, 935–968.

——— (2017): "Life-Cycle Consumption and Children: Evidence from a Structural Estimation," *Oxford Bulletin of Economics and Statistics*, 79, 717–746.

KAPLAN, G. AND G. L. VIOLANTE (2014): "A model of the consumption response to fiscal stimulus payments," *Econometrica*, 82, 1199–1239.

KASS, R. E. AND A. E. RAFTERY (1995): "Bayes factors," *Journal of the american statistical association*, 90, 773–795.

KIM, J.-Y. (1994): "Bayesian asymptotic theory in a time series model with a possible nonstationary process," *Econometric Theory*, 10, 764–773.

——— (1998): "Large sample properties of posterior densities, Bayesian information criterion and the likelihood principle in nonstationary time series models," *Econometrica*, 359–380.

KLINE, B. AND E. TAMER (2016): "Bayesian inference in a class of partially identified models," *Quantitative Economics*, 7, 329–366.

KOIJEN, R. S., T. E. NIJMAN, AND B. J. WERKER (2009): "When can life cycle investors benefit from time-varying bond risk premia?" *The review of financial studies*, 23, 741–780.

KOOP, G., D. KOROBILIS, ET AL. (2010): "Bayesian multivariate time series methods for empirical macroeconomics," *Foundations and Trends® in Econometrics*, 3, 267–358.

LI, M. (2006): "High school completion and future youth unemployment: New evidence from High School and Beyond," *Journal of Applied Econometrics*, 21, 23–53.

LI, W., H. LIU, F. YANG, AND R. YAO (2016): "Housing over time and over the life cycle: a structural estimation," *International Economic Review*, 57, 1237–1260.

LI, Y., Y. JUN, AND T. ZENG (2017): "Deviance information criterion for Bayesian model selection: Justification and variation," .

LI, Y. AND J. YU (2012): "Bayesian hypothesis testing in latent variable models," *Journal of Econometrics*, 166, 237–246.

LI, Y., J. YU, AND T. ZENG (2018): "Integrated deviance information criterion for latent variable models," .

LI, Y., T. ZENG, AND J. YU (2012): "Robust deviation information criterion for latent variable models," *Research Collection School of Economics Paper*, 1403.

——— (2014): "A new approach to Bayesian hypothesis testing," *Journal of Econometrics*, 178, 602–612.

LIAO, Y. AND A. SIMONI (2015): "Posterior properties of the support function for set inference," Tech. rep., Working paper, University of Maryland and CREST.

LUDVIGSON, S. AND C. H. PAXSON (2001): "Approximation bias in linearized Euler equations," *Review of Economics and Statistics*, 83, 242–256.

MAGNUS, J. R. AND H. NEUDECKER (2002): "Matrix differential calculus with applications in statistics and econometrics," *Wiley series in probability and mathematical statistics*.

McCulloch, R. E. (1989): "Local model influence," *Journal of the American Statistical Association*, 84, 473–478.

Moon, H. R. and F. Schorfheide (2012): "Bayesian and frequentist inference in partially identified models," *Econometrica*, 80, 755–782.

Mroz, T. A. (1987): "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions," *Econometrica: Journal of the Econometric Society*, 765–799.

Newey, W. K. and K. D. West (????): "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, 55, 703–708.

——— (1987): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 777–787.

Norets, A. (2009): "Inference in dynamic discrete choice models with serially orrelated unobserved state variables," *Econometrica*, 77, 1665–1682.

Norets, A. and X. Tang (2013): "Semiparametric inference in dynamic binary choice models," *Review of Economic Studies*, 81, 1229–1262.

O'Hagan, A. (1995): "Fractional Bayes factors for model comparison," *Journal of the Royal Statistical Society. Series B (Methodological)*, 99–138.

Pitt, M. K. and N. Shephard (1999): "Filtering via simulation: Auxiliary particle filters," *Journal of the American statistical association*, 94, 590–599.

Poirier, D. J. (1995): *Intermediate statistics and econometrics: a comparative approach*, Mit Press.

——— (1997): "A predictive motivation for loss function specification in parametric hypothesis testing," *Economics Letters*, 56, 1–3.

POYIADJIS, G., A. DOUCET, AND S. S. SINGH (2011): "Particle approximations of the score and observed information matrix in state space models with application to parameter estimation," *Biometrika*, 98, 65–80.

ROBERT, C. (2007): *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Springer Science & Business Media.

ROBERT, C. P. (1993): "A note on Jeffreys-Lindley paradox," *Statistica Sinica*, 601–608.

SCHERVISH, M. J. (2012): *Theory of statistics*, Springer Science & Business Media.

SHEPHARD, N. (2005): *Stochastic volatility: selected readings*, Oxford University Press on Demand.

SPIEGELHALTER, D. J., N. G. BEST, B. P. CARLIN, AND A. VAN DER LINDE (2002): "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583–639.

STERN, S. (1997): "Simulation-based estimation," *Journal of economic Literature*, 35, 2006–2039.

TANNER, M. A. AND W. H. WONG (1987): "The calculation of posterior distributions by data augmentation," *Journal of the American statistical Association*, 82, 528–540.

WHITE, H. (1996): *Estimation, inference and specification analysis*, 22, Cambridge university press.

WOOLDRIDGE, J. M. (1994): "Estimation and inference for dependent processes," *Handbook of econometrics*, 4, 2639–2738.

——— (2015): *Introductory econometrics: A modern approach*, Nelson Education.

YU, J. (2005): "On leverage in a stochastic volatility model," *Journal of Econometrics*, 127, 165–178.

——— (2012): "Simulation-based estimation methods for financial time series models," in *Handbook of Computational Finance*, Springer, 401–435.

Appendix

.1 Proofs in Chapter 2

.1.1 Proof of Lemma 2.3.1

When the likelihood information dominates the prior information, the posterior mean $\bar{\vartheta}$ reduces to the ML estimator $\widehat{\vartheta}$, under the alternative hypothesis. When H_0 is true, let $\vartheta_0 = (\theta_0, \psi_0)$ be the true value of ϑ. According to the standard ML theory and the central limit theorem, it can be shown that

$$\sqrt{n}(\widehat{\vartheta} - \vartheta_0) \xrightarrow{d} N\left[0, F(\vartheta_0)\right],$$

where $F(\vartheta_0) = n\mathscr{I}^{-1}(\vartheta_0)$, $\mathscr{I}(\vartheta_0) = -E\left[\mathbf{I}(\vartheta_0)\right]$ is the Fisher information matrix, and

$$\mathbf{I}(\vartheta) = \frac{\partial^2 \log p(\mathbf{y}|\vartheta)}{\partial\vartheta\partial\vartheta'} = L_n^{(2)}(\vartheta).$$

Under the standard regularity conditions, as $n \to \infty$, we have

$$-n\mathbf{J}(\vartheta_0) \xrightarrow{p} F(\vartheta_0),$$

where $\mathbf{J}(\vartheta_0)$ is the inverse matrix of $\mathbf{I}(\vartheta_0)$. Therefore, it can be shown that

$$\widehat{\vartheta} - \vartheta_0 = O_p(n^{-\frac{1}{2}}),$$
$$\mathbf{J}(\vartheta_0) = O_p(n^{-1}), \mathbf{I}(\vartheta_0) = O_p(n).$$

For any consistent estimator of ϑ, say $\tilde{\vartheta}$, there exists a positive sequence $k_n^* \to 0$ such that $p(||\tilde{\vartheta} - \vartheta_0|| \leq k_n^*) \geq 1 - k_n^*$. Hence, when n is large enough, we can find some $N > 0$, and $n > N$ to make $||\tilde{\vartheta} - \vartheta_0|| \leq k_n^*$. Under Assumption 5, we have

$$\frac{1}{n}||\mathbf{I}(\tilde{\vartheta}) - \mathbf{I}(\vartheta_0)|| \leq \sup_{||\vartheta - \vartheta_0|| < k_n} \frac{1}{n}||\mathbf{I}(\vartheta) - \mathbf{I}(\vartheta_0)|| \xrightarrow{p} 0.$$

Hence, for any consistent estimator $\tilde{\vartheta}$, $\frac{1}{n}\left[\mathbf{I}(\tilde{\vartheta}) - \mathbf{I}(\vartheta_0)\right] = o_p(1)$ so that $\mathbf{I}(\tilde{\vartheta}) = \mathbf{I}(\vartheta_0) + o_p(n)$ and that $\mathbf{I}(\tilde{\vartheta}) = O_p(n)$. Similarly, $\mathbf{J}(\tilde{\vartheta}) = \mathbf{J}(\vartheta_0) + o_p(n^{-1})$ and $\mathbf{J}(\tilde{\vartheta}) = O_p(n^{-1})$.

.1.2 Proof of Lemma 2.3.2

When the likelihood information dominates the prior information, the posterior mode $\widehat{\vartheta}_0$ of ϑ under the null hypothesis reduces to the ML estimator of ϑ under the null hypothesis. Similar to Lemma 2.3.1, when H_0 is true, according to the standard ML theory, we have

$$\frac{1}{\sqrt{n}}s(\vartheta_0) \sim N[0, F(\vartheta_0)],$$
$$\sqrt{n}(\widehat{\psi}_0 - \psi_0) \sim N[0, F_{\psi\psi}(\vartheta_0)],$$

where $F_{\psi\psi}(\vartheta_0)$ is the submatrix of $F(\vartheta_0)$ corresponding to ψ. Hence, we have

$$s(\vartheta_0) = O_p(n^{1/2}), \widehat{\psi}_0 - \psi_0 = O_p(n^{-1/2}), \widehat{\vartheta}_0 - \vartheta_0 = O_p(n^{-1/2}).$$

Furthermore, based on Remark 3.7, it can be shown that

$$\bar{\psi}_0 - \widehat{\psi}_0 = o_p(n^{-1/2}), \bar{\vartheta}_0 - \widehat{\vartheta}_0 = o_p(n^{-1/2}),$$
$$\bar{\psi}_0 - \psi_0 = \bar{\psi}_0 - \widehat{\psi}_0 + \widehat{\psi}_0 - \psi_0 = o_p(n^{-1/2}) + O_p(n^{-1/2}) = O_p(n^{-1/2}),$$
$$\bar{\vartheta}_0 - \vartheta_0 = O_p(n^{-1/2}).$$

Using the first-order Taylor expansion, we have

$$s(\widehat{\vartheta}_0) = s(\vartheta_0) + \mathbf{I}(\tilde{\vartheta}_0)(\widehat{\vartheta}_0 - \vartheta_0),$$

where $\tilde{\theta}_0$ lies on the segment between $\widehat{\vartheta}_0$ and ϑ_0. Since $\widehat{\vartheta}_0 - \vartheta_0 = O_p(n^{-1/2})$, it means that $\widehat{\vartheta}_0$ is a consistent estimator of ϑ_0 so that $\tilde{\vartheta}_0$ is also a consistent estimator of ϑ_0. Hence, we get

$$\begin{aligned}
&s(\widehat{\vartheta}_0) = s(\vartheta_0) + \mathbf{I}(\tilde{\vartheta}_0)(\widehat{\vartheta}_0 - \vartheta_0) \\
&= s(\vartheta_0) + [\mathbf{I}(\vartheta_0) + o_p(n)](\widehat{\vartheta}_0 - \vartheta_0) \\
&= s(\vartheta_0) + \mathbf{I}(\vartheta_0)(\widehat{\vartheta}_0 - \vartheta_0) + o_p(n)(\widehat{\vartheta}_0 - \vartheta_0) \\
&= s(\vartheta_0) + \mathbf{I}(\vartheta_0)(\widehat{\vartheta}_0 - \vartheta_0) + o_p(n)O_p(n^{-1/2}) \\
&= s(\vartheta_0) + \mathbf{I}(\vartheta_0)(\widehat{\vartheta}_0 - \vartheta_0) + o_p(n^{1/2}) \\
&= O_p(n^{1/2}) + O_p(n)O_p(n^{-1/2}) + o_p(n^{1/2}) = O_p(n^{1/2}), \\
&C(\widehat{\vartheta}_0) = s(\widehat{\vartheta}_0)s(\widehat{\vartheta}_0)' = O_p(n^{1/2})O_p(n^{1/2}) = O_p(n).
\end{aligned}$$

Similarly, since $\bar{\vartheta}_0 - \vartheta_0 = O_p(n^{-1/2})$, it means that $\bar{\vartheta}_0$ is a consistent estimator of ϑ_0 so that $\tilde{\vartheta}_0$ is also a consistent estimator of ϑ_0. Hence, we can get

$$\begin{aligned}
&s(\bar{\vartheta}_0) = O_p(n^{1/2}), \\
&C(\bar{\vartheta}_0) = s(\bar{\vartheta}_0)s(\bar{\vartheta}_0)' = O_p(n).
\end{aligned}$$

Furthermore, we can show that

$$s(\bar{\vartheta}_0) = s(\widehat{\vartheta}_0) + \mathbf{I}(\tilde{\vartheta}_0)(\bar{\vartheta}_0 - \widehat{\vartheta}_0),$$

where $\tilde{\vartheta}_0$ lies on the segment between $\bar{\vartheta}_0$ and $\widehat{\vartheta}_0$. Because both $\widehat{\vartheta}_0$ and $\bar{\vartheta}_0$ are consistent estimators of ϑ_0, $\tilde{\vartheta}_0$ is also a consistent estimator of ϑ_0. Using Lemma

2.3.1, we get

$$
\begin{aligned}
& C(\bar{\vartheta}_0) = s(\bar{\vartheta}_0)s(\bar{\vartheta}_0)' = [s(\widehat{\vartheta}_0) + \mathbf{I}(\tilde{\vartheta}_0)(\bar{\vartheta}_0 - \widehat{\vartheta}_0)][s(\widehat{\vartheta}_0) + \mathbf{I}(\tilde{\vartheta}_0)(\bar{\vartheta}_0 - \widehat{\vartheta}_0)]' \\
= \; & s(\widehat{\vartheta}_0)s(\widehat{\vartheta}_0)' + 2\mathbf{I}(\tilde{\vartheta}_0)(\bar{\vartheta}_0 - \widehat{\vartheta}_0)s(\widehat{\vartheta}_0) + \mathbf{I}(\tilde{\vartheta}_0)(\bar{\vartheta}_0 - \widehat{\vartheta}_0)(\bar{\vartheta}_0 - \widehat{\vartheta}_0)'\mathbf{I}(\tilde{\vartheta}_0) \\
= \; & s(\widehat{\vartheta}_0)s(\widehat{\vartheta}_0)' + 2O_p(n)o_p(n^{-1/2})O_p(n^{1/2}) + O_p(n)o_p(n^{-1/2})o_p(n^{-1/2})O_p(n) \\
= \; & s(\widehat{\vartheta}_0)s(\widehat{\vartheta}_0)' + o_p(n) = C(\widehat{\vartheta}_0) + o_p(n).
\end{aligned}
$$

.1.3 Proof of Theorem 2.3.1

Using the Bayesian large sample theory, we have

$$
\begin{aligned}
& E\left[(\vartheta - \bar{\vartheta})(\vartheta - \bar{\vartheta})'|\mathbf{y}\right] = E\left[(\vartheta - \widehat{\vartheta} + \widehat{\vartheta} - \bar{\vartheta})(\vartheta - \widehat{\vartheta} + \widehat{\vartheta} - \bar{\vartheta})'|\mathbf{y}\right] \\
= \; & E\left[(\vartheta - \widehat{\vartheta})(\vartheta - \widehat{\vartheta})'|\mathbf{y}\right] + 2E\left[(\vartheta - \widehat{\vartheta})|\mathbf{y}\right](\widehat{\vartheta} - \bar{\vartheta}) + (\widehat{\vartheta} - \bar{\vartheta})(\widehat{\vartheta} - \bar{\vartheta})' \\
= \; & E\left[(\vartheta - \widehat{\vartheta})(\vartheta - \widehat{\vartheta})'|\mathbf{y}\right] - 2(\widehat{\vartheta} - \bar{\vartheta})(\widehat{\vartheta} - \bar{\vartheta}) + (\widehat{\vartheta} - \bar{\vartheta})(\widehat{\vartheta} - \vartheta_0)' \\
= \; & E\left[(\vartheta - \widehat{\vartheta})(\vartheta - \widehat{\vartheta})'|\mathbf{y}\right] - (\widehat{\vartheta} - \bar{\vartheta})(\widehat{\vartheta} - \bar{\vartheta}) \\
= \; & -L_n^{-(2)}(\widehat{\vartheta}) + o_p(n^{-1}) + o_p(n^{-1/2})o_p(n^{-1/2}).
\end{aligned}
$$

The last equality $E\left[(\vartheta - \widehat{\vartheta})(\vartheta - \widehat{\vartheta})'|\mathbf{y}\right] = -L_n^{-(2)}(\widehat{\vartheta}) + o_p(n^{-1})$ follows Li, Zeng and Yu (2012) based on the assumptions listed in Section 3.2. Hence, we have

$$
\begin{aligned}
& \mathbf{T}(\mathbf{y}, \theta_0) = \int (\theta - \bar{\theta})' C_{\theta\theta}(\bar{\vartheta}_0)(\theta - \bar{\theta}) p(\vartheta|\mathbf{y}) \mathrm{d}\vartheta \\
= \; & \mathbf{tr}\left[C_{\theta\theta}(\bar{\vartheta}_0) E[(\theta - \bar{\theta})(\theta - \bar{\theta})'|\mathbf{y}]\right] \\
= \; & \mathbf{tr}\left[C_{\theta\theta}(\bar{\vartheta}_0)[-L_{n,\theta\theta}^{-(2)}(\widehat{\vartheta}) + o_p(n^{-1})]\right] \\
= \; & \mathbf{tr}\left[\left(C_{\theta\theta}(\widehat{\vartheta}_0) + o_p(n)\right)[-L_{n,\theta\theta}^{-(2)}(\widehat{\vartheta})]\right] + \mathbf{tr}\left[C_{\theta\theta}(\bar{\vartheta}_0) o_p(n^{-1})\right] \\
= \; & \mathbf{tr}\left[C_{\theta\theta}(\widehat{\vartheta}_0)[-L_{n,\theta\theta}^{-(2)}(\widehat{\vartheta})]\right] + o_p(n)[-L_{n,\theta\theta}^{-(2)}(\widehat{\vartheta})] + O_p(n)o_p(n^{-1}) \\
= \; & \mathbf{tr}\left[s_\theta(\widehat{\vartheta}_0)s_\theta(\widehat{\vartheta}_0)'[-L_{n,\theta\theta}^{-(2)}(\widehat{\vartheta})]\right] + o_p(n)O_p(n^{-1}) + o_p(1) \\
= \; & \mathbf{tr}\left[s_\theta(\widehat{\vartheta}_0)s_\theta(\widehat{\vartheta}_0)'[-L_{n,\theta\theta}^{-(2)}(\widehat{\vartheta})]\right] + o_p(1) \\
= \; & s_\theta(\widehat{\vartheta}_0)'[-L_{n,\theta\theta}^{-(2)}(\widehat{\vartheta})]s_\theta(\widehat{\vartheta}_0) + o_p(1).
\end{aligned}
$$

This proves Equation (2.3.5) in the theorem.

When the likelihood information dominates the prior information, the posterior mode $\widehat{\vartheta}$ reduces to the ML estimator of ϑ under the alternative hypothesis, the posterior mode $\widehat{\psi}_0$ to the ML estimator of ψ under the null hypothesis, and $L_n^{(2)}(\vartheta)$ to $\mathbf{I}(\vartheta)$. Under H_0, let $\vartheta_0 = (\theta_0, \psi_0)$ be the true value of ϑ, and $\widehat{\vartheta}_0 = (\theta_0, \widehat{\psi})$ be the ML estimator of ϑ. Then, when the null hypothesis is true, $\widehat{\vartheta}$ and $\widehat{\vartheta}_0$ are both consistent estimators of ϑ. Hence, based on Lemma 3.1 and Lemma 3.2, we get

$$\begin{aligned}
\mathbf{J}(\widehat{\vartheta}) &= \mathbf{I}^{-1}(\widehat{\vartheta}) = [\mathbf{I}(\vartheta_0) + o_p(n)]^{-1} + o_p(n^{-1}) = \mathbf{J}(\vartheta_0) + o_p(n^{-1}), \\
\mathbf{J}(\widehat{\vartheta}_0) &= \mathbf{I}^{-1}(\vartheta_0) = [\mathbf{I}(\vartheta_0) + o_p(n)]^{-1} + o_p(n^{-1}) = \mathbf{J}(\vartheta_0) + o_p(n^{-1}).
\end{aligned}$$

Then, we can further derive that

$$\begin{aligned}
& \mathbf{T}(\mathbf{y}, \theta_0) = \int (\theta - \bar{\theta})' C_{\theta\theta}(\bar{\vartheta}_0)(\theta - \bar{\theta}) p(\vartheta|\mathbf{y}) \mathrm{d}\vartheta \\
= \;& s_\theta(\widehat{\vartheta}_0)'[-L_{n,\theta\theta}^{-(2)}(\widehat{\vartheta})] s_\theta(\widehat{\vartheta}_0) + o_p(1) \\
= \;& -s_\theta(\widehat{\vartheta}_0)' \mathbf{J}_{\theta\theta}(\widehat{\vartheta}) s_\theta(\widehat{\vartheta}_0) + o_p(1) \\
= \;& -s_\theta(\widehat{\vartheta}_0)' \mathbf{J}_{\theta\theta}(\widehat{\vartheta}) s_\theta(\widehat{\vartheta}_0) + o_p(1) \\
= \;& -s_\theta(\widehat{\vartheta}_0)' \left[\mathbf{J}_{\theta\theta}(\vartheta_0) + o_p(n^{-1})\right] s_\theta(\widehat{\vartheta}_0) + o_p(1) \\
= \;& -s_\theta(\widehat{\vartheta}_0)' [\mathbf{J}_{\theta\theta}(\vartheta_0)] s_\theta(\widehat{\vartheta}_0) + s_\theta(\widehat{\vartheta}_0)' o_p(n^{-1}) s_\theta(\widehat{\vartheta}_0) + o_p(1) \\
= \;& -s_\theta(\widehat{\vartheta}_0)' [\mathbf{J}_{\theta\theta}(\vartheta_0)] s_\theta(\widehat{\vartheta}_0) + O_p(n^{1/2}) o_p(n^{-1}) O_p(n^{1/2}) + o_p(1) \\
= \;& -s_\theta(\widehat{\vartheta}_0)' [\mathbf{J}_{\theta\theta}(\vartheta_0)] s_\theta(\widehat{\vartheta}_0) + o_p(1) \\
= \;& -s_\theta(\widehat{\vartheta}_0)' \left[\mathbf{J}_{\theta\theta}(\widehat{\vartheta}_0) + o_p(n^{1/2})\right] s_\theta(\widehat{\vartheta}_0) + o_p(1) \\
= \;& -s_\theta(\widehat{\vartheta}_0)' \mathbf{J}_{\theta\theta}(\widehat{\vartheta}_0) s_\theta(\widehat{\vartheta}_0) + O_p(n^{1/2}) o_p(n^{-1}) O_p(n^{1/2}) + o_p(1) \\
= \;& -s_\theta(\widehat{\vartheta}_0)' \mathbf{J}_{\theta\theta}(\widehat{\vartheta}_0) s_\theta(\widehat{\vartheta}_0) + o_p(1) \\
= \;& LM + o_p(1).
\end{aligned}$$

According to the standard ML theory, under the null hypothesis, LM$\xrightarrow{d} \chi^2(p)$. Therefore, $\mathbf{T}(\mathbf{y}, \theta_0) \xrightarrow{d} \chi^2(p)$ and the theorem is proved.

.1.4 Derivation of $\mathbf{T}(\mathbf{y},\ \theta_0)$ and the BF in linear regression model

It is known that the $\log BF_{10}$ can be expressed as

$$\log BF_{10} = \log p\left(\mathbf{y}|M_1\right) - \log p\left(\mathbf{y}|M_0\right).$$

In the simple linear regression model, under H_0, the marginal likelihood $p(\mathbf{y}|M_0)$ is given by

$$\begin{aligned}
& p(\mathbf{y}|M_0) = \int\int p\left(\mathbf{y}|\alpha,\beta_0\right) p\left(\alpha|\sigma^2\right) p\left(\sigma^2\right) d\alpha d\sigma^2 \\
= \; & \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)}\int\int \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta_0x_i)^2\right)\frac{1}{\sqrt{2\pi V_\alpha}\sigma}\exp\left(-\frac{(\alpha-\mu_\alpha)^2}{2\sigma^2V_\alpha}\right) \\
& \times\left(\sigma^2\right)^{-a-\frac{n}{2}-1}\exp\left(-\frac{b}{\sigma^2}\right)d\alpha d\sigma^2 \\
= \; & \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)}\int\int \frac{1}{\sqrt{2\pi V_\alpha}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left[-2\alpha\sum_{i=1}^{n}(y_i-\beta_0x_i)+n\alpha^2\right]\right\} \\
& \times\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0x_i)^2\right)\exp\left[-\frac{1}{2\sigma^2V_a}\left(\alpha^2-2\mu_\alpha\alpha\right)\right]\exp\left(-\frac{\mu_\alpha^2}{2\sigma^2V_\alpha}\right)d\alpha d\sigma^2 \\
= \; & \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)}\int\int \frac{1}{\sqrt{2\pi V_\alpha}\sigma}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0x_i)^2\right)\exp\left(-\frac{\mu_\alpha^2}{2\sigma^2V_\alpha}\right) \\
& \times\exp\left\{-\frac{1}{2\sigma^2}\left[\left(n+\frac{1}{V_\alpha}\right)\alpha^2-2\alpha\left(\sum_{i=1}^{n}(y_i-\beta_0x_i)+\frac{\mu_\alpha}{V_\alpha}\right)\right]\right\}d\alpha d\sigma^2 \\
= \; & \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)}\sqrt{\frac{1}{nV_\alpha+1}} \\
& \times\int_0^{+\infty}\left(\sigma^2\right)^{-a-\frac{n}{2}-1}\exp\left\{-\frac{1}{\sigma^2}\left[b+\frac{1}{2}\left(\sum_{i=1}^{n}(y_i-\beta_0x_i)^2+\frac{\mu_\alpha^2}{V_\alpha}-\frac{\mu_\alpha^{*2}}{V_\alpha^*}\right)\right]\right\}d\sigma^2 \\
= \; & \frac{b^a\Gamma\left(a+\frac{n}{2}\right)}{(2\pi)^{\frac{n}{2}}\Gamma(a)}\sqrt{\frac{1}{nV_\alpha+1}}\left[b+\frac{1}{2}\left(\sum_{i=1}^{n}(y_i-\beta_0x_i)^2+\frac{\mu_\alpha^2}{V_\alpha}-\frac{\mu_\alpha^*}{V_\alpha^*}\right)\right]^{-\left(a+\frac{n}{2}\right)} \\
= \; & \frac{b^a\Gamma\left(a+\frac{n}{2}\right)}{(2\pi)^{\frac{n}{2}}\Gamma(a)}\sqrt{\frac{1}{nV_\alpha+1}}\left[b+\frac{1}{2}\left((\mathbf{y}-\beta_0\mathbf{x})'(\mathbf{y}-\beta_0\mathbf{x})+\frac{\mu_\alpha^2}{V_\alpha}-\frac{\mu_\alpha^*}{V_\alpha^*}\right)\right]^{-\left(a+\frac{n}{2}\right)}.
\end{aligned}$$

Under the alternative hypothesis H_1, we rewrite the equation in a matrix form:

$$\mathbf{y} = X\gamma + \varepsilon,$$

where $\gamma = (\alpha, \beta)'$, $X = (\iota, \mathbf{x})$. The prior for γ is $N\left(\tilde{\mu}, \sigma^2 \tilde{V}\right)$, where $\tilde{\mu} = \left(\mu_\alpha, \mu_\beta\right)'$, $\tilde{V} = diag\left(V_\alpha, V_\beta\right)$. Similarly, the marginal likelihood $p(\mathbf{y}|M_1)$ is

$$
\begin{aligned}
& p(\mathbf{y}|M_1) = \int\int p\left(\mathbf{y}|\beta,\alpha\right) p\left(\gamma|\sigma^2\right) p\left(\sigma^2\right) d\gamma d\sigma^2 \\
= \; & \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)} \int\int \left(\sigma^2\right)^{-a-\frac{n}{2}-1} \exp\left(-\frac{b}{\sigma^2}\right) \\
& \times \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{y}-X\gamma\right)'\left(\mathbf{y}-X\gamma\right)\right) \frac{1}{2\pi|\tilde{V}|^{\frac{1}{2}}\sigma^2} \exp\left(-\frac{1}{2\sigma^2}\left(\gamma-\tilde{\mu}\right)'\tilde{V}^{-1}\left(\gamma-\tilde{\mu}\right)\right) d\gamma d\sigma^2 \\
= \; & \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)\sqrt{|\tilde{V}|}} \int\int \frac{1}{2\pi\sigma^2}\left(\sigma^2\right)^{-a-\frac{n}{2}-1} \left\{\left(-\frac{1}{\sigma^2}\left[b + \frac{1}{2}\left(\mathbf{y}'\mathbf{y} + (\tilde{\mu})'\tilde{V}^{-1}\tilde{\mu}\right)\right]\right)\right\} \\
& \times \exp\left\{-\frac{1}{2\sigma^2}\left(\gamma'\left(X'X+\tilde{V}^{-1}\right)\gamma - \gamma'\left(X'\mathbf{y}+\tilde{V}^{-1}\tilde{\mu}\right) - \left(X'\mathbf{y}+\tilde{V}^{-1}\tilde{\mu}\right)'\gamma\right)\right\} d\gamma d\sigma^2 \\
= \; & \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)\sqrt{|\tilde{V}|}} \int\int \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left(\gamma-\mu^*\right)'V^{*-1}\left(\gamma-\mu^*\right)\right\} \\
& \times \exp\left(-\frac{1}{2\sigma^2}\left((\tilde{\mu})'\tilde{V}^{-1}\tilde{\mu} + \mathbf{y}'\mathbf{y} - (\mu^*)'V^{*-1}\mu^*\right)\right)\left(\sigma^2\right)^{-a-\frac{n}{2}-1} \exp\left(-\frac{b}{\sigma^2}\right) d\gamma d\sigma^2 \\
= \; & \frac{b^a\sqrt{|V^*|}}{(2\pi)^{\frac{n}{2}}\Gamma(a)\sqrt{|\tilde{V}|}} \int \left(\sigma^2\right)^{-a-\frac{n}{2}-1} \exp\left\{-\frac{1}{\sigma^2}\left[b + \frac{1}{2}\left((\tilde{\mu})'\tilde{V}^{-1}\tilde{\mu} + \mathbf{y}'\mathbf{y} - (\mu^*)'V^{*-1}\mu^*\right)\right]\right\} d\sigma^2 \\
= \; & \frac{b^a\sqrt{|V^*|}\Gamma\left(a+\frac{n}{2}\right)}{(2\pi)^{\frac{n}{2}}\Gamma(a)\sqrt{|\tilde{V}|}} \left[b + \frac{1}{2}\left((\tilde{\mu})'\tilde{V}^{-1}\tilde{\mu} + \mathbf{y}'\mathbf{y} - (\mu^*)'V^{*-1}\mu^*\right)\right]^{-\left(a+\frac{n}{2}\right)}.
\end{aligned}
$$

In the following, we show how to calculate $\mathbf{T}\left(\mathbf{y}, \theta_0\right)$. It is noted that the log-likelihood function is:

$$\log p\left(\mathbf{y}|\vartheta\right) = -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \alpha - \beta x_i\right)^2.$$

Hence, given $\vartheta = \left(\alpha, \beta, \sigma^2\right)'$, for H_0 of $\theta = \beta$, we have

$$s\left(\vartheta\right) = \left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \alpha - \beta x_i\right), \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i\left(y_i - \alpha - \beta x_i\right), -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(y_i - \alpha - \beta x_i\right)^2\right)',$$

and

$$C_{\theta\theta}\left(\bar{\vartheta}_0\right) = \frac{1}{\bar{\sigma}_0^4}\left[\sum_{i=1}^{n} x_i\left(y_i - \bar{\alpha}_0 - \beta_0 x_i\right)\right]^2 = \frac{1}{\bar{\sigma}_0^4}\left[\mathbf{x}'\left(\mathbf{y} - \bar{\alpha}_0\iota - \beta_0\mathbf{x}\right)\right]^2,$$

where $\bar{\sigma}_0^4 = \left(\bar{\sigma}_0^2\right)^2$, $\bar{\alpha}_0$ and $\bar{\sigma}_0^2$ are the posterior means of α and σ^2 under H_0.

Since the likelihood and the prior are both in the Normal-Gamma form, based on the previous derivation of $p(\mathbf{y}|M_1)$, if we integrate the σ^2, we can have the posterior density of $\gamma = (\alpha, \beta)'$

$$\begin{aligned}
\pi(\gamma|\mathbf{y}) &\propto \left[b + \frac{1}{2}\left((\tilde{\mu})'\tilde{V}^{-1}\tilde{\mu} + \mathbf{y}'\mathbf{y} - (\mu^*)'V^{*-1}\mu^*\right) + \frac{1}{2}(\gamma - \mu^*)'V^{*-1}(\gamma - \mu^*)\right]^{\frac{2a+n}{2}+1} \\
&\propto \left[1 + \frac{1}{2\nu s}(\gamma - \mu^*)'V^{*-1}(\gamma - \mu^*)\right]^{\frac{\nu}{2}+1},
\end{aligned}$$

which is a density function of multivariate t distribution with degrees of freedom $\nu = 2a + n$, mean μ^*, and a positive definite symmetric matrix, V^*. That is,

$$\gamma|\mathbf{y} \sim t(\mu^*, 2sV^*, \nu).$$

Let

$$\mu^* = \begin{pmatrix} \mu_1^* \\ \mu_2^* \end{pmatrix}, V^* = \begin{pmatrix} V_{11}^* & V_{12}^* \\ V_{21}^* & V_{22}^* \end{pmatrix}.$$

It is easy to show that $\beta|\mathbf{y} \sim t(\mu_2^*, 2sV_{22}^*, \nu)$. Then, the posterior variance of β is $Var(\beta|\mathbf{y}) = \frac{2sV_{22}^*}{\nu-2}$. Hence, the proposed test statistic can be calculated analytically as

$$\mathbf{T}(\mathbf{y}, \theta_0) = C_{\theta\theta}\left(\bar{\vartheta}_0\right)Var(\beta|\mathbf{y}) = \frac{2sV_{22}^*}{\nu-2}C_{\theta\theta}\left(\bar{\vartheta}_0\right).$$

.1.5 Derivation of the BF and $\mathbf{T}(\mathbf{y},\ \theta_0)$ in the probit model

In the binary probit model, for each y_i, $i = 1, 2, ..., n$, there is a corresponding latent variable z_i that satisfies:

$$\begin{cases} y_i = 1 & \text{if } z_i \geq 0 \\ y_i = 0 & \text{if } z_i < 0 \end{cases},$$

and

$$z_i = \mathbf{x}_i'\vartheta + e_i,$$

where ϑ is the $(p+q)\times 1$ parameter vector measuring the marginal effects and $e_i \sim N(0,1)$ for $i=1,...,n$.

Rewrite the above equation as:

$$z_i = \mathbf{x}_{i1}'\psi + \mathbf{x}_{i2}'\theta + e_i.$$

For each i, we have

$$\begin{cases} p(y_i=1|\vartheta) = p(z_i \geq 0|\ \vartheta) = p\left(e_i \geq -\left(\mathbf{x}_{i1}'\psi + \mathbf{x}_{i2}'\ \theta\right)|\vartheta\right) = \Phi\left[(2y_i-1)\left(\mathbf{x}_{i1}'\psi + \mathbf{x}_{i2}'\theta\right)\right] \\ p(y_i=0|\vartheta) = p(z_i < 0|\ \vartheta) = p\left(e_i < -\left(\mathbf{x}_{i1}'\psi + \mathbf{x}_{i2}'\ \theta\right)|\vartheta\right) = \Phi\left[(2y_i-1)\left(\mathbf{x}_{i1}'\psi + \mathbf{x}_{i2}'\theta\right)\right] \end{cases},$$

where the $\Phi(\cdot)$ is the standard normal cumulative distribution function. Note that the log-likelihood function is:

$$\log p(\mathbf{y}|\vartheta) = \sum_{i=1}^{n} \log \Phi\left[q_i\left(\mathbf{x}_{i1}'\psi + \mathbf{x}_{i2}'\theta\right)\right],$$

where $q_i = 2y_i - 1$.

- The estimator of $\mathbf{T}(\mathbf{y},\theta_0)$ and its NSE.

  For H_0 of $\theta = \mathbf{0}$, note that,

  $$\frac{\partial \log p(\mathbf{y}|\vartheta)}{\partial \theta} = \sum_{i=1}^{n} q_i \frac{\phi\left[q_i\left(\mathbf{x}_{i1}'\psi + \mathbf{x}_{i2}'\theta\right)\right]\mathbf{x}_{i2}}{\Phi\left[q_i\left(\mathbf{x}_{i1}'\psi + \mathbf{x}_{i2}'\theta\right)\right]},$$

  where $\phi(\cdot)$ is the pdf of the standard normal distribution. The proposed test statistic is

  $$\mathbf{T}(\mathbf{y},\theta_{\mathbf{0}}) = \int \left(\theta - \bar{\theta}\right)' C_{\theta\theta}\left(\bar{\vartheta}_0\right)\left(\ \theta - \bar{\theta}\right) p(\vartheta|\mathbf{y})\, d\ \vartheta,$$

where

$$C_{\theta\theta}\left(\bar{\vartheta}_0\right) = \left(\frac{\partial \log p\left(\mathbf{y}|\ \vartheta\right)}{\partial\ \theta}\right)\left(\frac{\partial \log p\left(\mathbf{y}|\vartheta\right)}{\partial\ \theta}\right)'\Bigg|_{\vartheta=\bar{\vartheta}_0}$$
$$= \left(\sum_{i=1}^{n}\frac{\phi\left[q_i\left(\mathbf{x}_{i1}'\ \bar{\psi}_0\right)\right]q_i\mathbf{x}_{i2}}{\Phi\left[q_i\left(\mathbf{x}_{i1}'\ \bar{\psi}_0\right)\right]}\right)\times\left(\sum_{i=1}^{n}\frac{\phi\left[q_i\left(\mathbf{x}_{i1}'\ \bar{\psi}_0\right)\right]q_i\mathbf{x}_{i2}}{\Phi\left[q_i\left(\mathbf{x}_{i1}'\ \bar{\psi}_0\right)\right]}\right)',$$

where $\bar{\vartheta}_0 = (\theta_0, \bar{\psi}_0)$ and $\bar{\psi}_0$ is the posterior mean of ψ under H_0.

To sum up, to compute the $\widehat{\mathbf{T}}(\mathbf{y}, \theta_0)$, we firstly draw MCMC samples for the model under H_0 and calculate $C_{\theta\theta}\left(\bar{\vartheta}_0\right)$. We then draw MCMC samples for the model under H_1 to obtain $\left\{\vartheta^{(g)}\right\}_{g=1}^{G} = \left\{\theta^{(g)}, \psi^{(g)}\right\}_{g=1}^{G}$. Naturally, the estimator of the statistic is

$$\widehat{\mathbf{T}}(\mathbf{y}, \theta_0) = \frac{1}{M}\sum_{g=1}^{G} f\left(\theta^{(g)}\right),$$

where,

$$f\left(\theta^{(g)}\right) = \left(\theta^{(g)} - \bar{\theta}\right)' C_{\theta\theta}\left(\bar{\vartheta}_0\right)\left(\theta^{(g)} - \bar{\theta}\right),$$

where $\bar{\theta}$ is the posterior mean of θ for the model under H_1.

Following the discussion about the NSE in Section 3, the numerical variance of $\widehat{\mathbf{T}}(\mathbf{y}, \theta_0)$ is

$$Var\left(\widehat{\mathbf{T}}(\mathbf{y}, \theta_0)\right) = \frac{1}{G}\left[\Omega_0 + 2\sum_{k=1}^{q}\left(1 - \frac{k}{q+1}\right)\Omega_k\right],$$

where

$$\Omega_k = \frac{1}{G}\sum_{g=k+1}^{G}\left(f\left(\theta^{(g)}\right) - \widehat{\mathbf{T}}(\mathbf{y}, \theta_0)\right)^2.$$

- The estimator of the BF and its NSE.

  We know that the logarithmic observed likelihood function, $\log p(\mathbf{y}|\vartheta)$, is given by

$$\log p(\mathbf{y}|\vartheta) = \sum_{i=1}^{n}\log\Phi\left[q_i\left(\mathbf{x}_{i1}'\psi + \mathbf{x}_{i2}'\theta\right)\right],$$

which is easy to compute.

Based on Chib (1995), the logarithmic marginal likelihood under H_1, $\log p(\mathbf{y}|M_1)$, is given by

$$\log p(\mathbf{y}|M_1) = \log p\left(\mathbf{y}|\bar{\vartheta}\right) + \log p\left(\bar{\vartheta}\right) - \log p\left(\bar{\vartheta}|\mathbf{y}\right),$$

where $p\left(\bar{\vartheta}\right)$ is the pdf of the prior evaluated at $\bar{\vartheta}$, $p\left(\bar{\vartheta}|\mathbf{y}\right)$ is the pdf of the posterior distribution evaluated at $\bar{\vartheta}$. The posterior quantity can be approximated by

$$\widehat{p}\left(\bar{\vartheta}|\mathbf{y}\right) = \frac{1}{G}\sum_{g=1}^{G} p\left(\bar{\vartheta}|\mathbf{z}_1^{(g)}\right),$$

where $\{\mathbf{z}_1^{(g)}, g = 1, 2, \cdots, G\}$ are efficient random draws from $p\left(\mathbf{z}_1|\mathbf{y}, \bar{\vartheta}\right)$ and the posterior distribution $p(\vartheta|\mathbf{z})$ has a closed-form expression in this model. The logarithmic marginal likelihood under H_0, $\log p(\mathbf{y}|M_0)$, is given by

$$\log p(\mathbf{y}|M_0) = \log p\left(\mathbf{y}|\bar{\vartheta}_0\right) + \log p(\bar{\psi}_0) - \log p(\bar{\psi}_0|\mathbf{y}, \theta_0).$$

Similarly, $\widehat{p}(\bar{\psi}_0|\mathbf{y}, \theta_0) = \frac{1}{G}\sum_{g=1}^{G} p\left(\bar{\psi}_0|\mathbf{z}_0^{(g)}, \theta_0\right)$, and $\{\mathbf{z}_0^{(g)}, g = 1, 2, \cdots, G\}$ are efficient random draws from $p\left(\mathbf{z}_0|\mathbf{y}, \bar{\vartheta}_0\right)$.

Hence, the logarithmic BF can be estimated by

$$\begin{aligned} \widehat{\log BF_{10}} &= \left[\log p\left(\mathbf{y}|\bar{\vartheta}\right) + \log p\left(\bar{\vartheta}\right) - \log \widehat{p}\left(\bar{\vartheta}|\mathbf{y}\right)\right] \\ &- \left[\log p\left(\mathbf{y}|\bar{\vartheta}_0\right) + \log p(\bar{\psi}_0) - \log \widehat{p}(\bar{\psi}_0|\mathbf{y}, \theta_0)\right]. \end{aligned}$$

To calculate the NSE, following Chib (1995), let $h_1^{(g)} = p\left(\bar{\vartheta}|\mathbf{z}^{(g)}\right)$, $h_0^{(g)} = p\left(\bar{\psi}_0|\mathbf{z}_0^{(g)}, \theta_0\right)$, $h^{(g)} = \left(h_1^{(g)}, h_0^{(g)}\right)'$, $\widehat{h} = \left(\widehat{h}_1, \widehat{h}_0\right)$, $\widehat{h}_0 = \frac{1}{G}\sum_{g=1}^{G} h_0^{(g)}$, $\widehat{h}_1 = \frac{1}{G}\sum_{g=1}^{G} h_1^{(g)}$. Then the numerical variance is

$$Var\left(\widehat{\log BF_{10}}\right) = \left(\frac{\partial \widehat{\log BF_{10}}}{\partial \widehat{h}}\right)' Var(h) \left(\frac{\partial \widehat{\log BF_{10}}}{\partial \widehat{h}}\right),$$

$$Var(h) = \frac{1}{G}\left[\Omega_0 + \sum_{k=1}^{q}\left(1 - \frac{k}{q+1}\right)\left(\Omega_k + \Omega_k'\right)\right],$$

$$\Omega_k = \frac{1}{G}\sum_{g=k+1}^{G}\left(h^{(g)} - \widehat{h}\right)\left(h^{(g)} - \widehat{h}\right)',$$

$$\frac{\partial \widehat{\log BF_{10}}}{\partial \widehat{h}} = \begin{pmatrix} -\widehat{p}\left(\bar{\vartheta}|\mathbf{y}\right)^{-1} \\ \widehat{p}\left(\bar{\psi}_0|\mathbf{y},\theta_0\right)^{-1} \end{pmatrix}.$$

.1.6 Derivation of the BF and $\mathbf{T}(\mathbf{y},\ \theta_0)$ in the stochastic conditional duration model

To save the space, here we only discuss the most specification corresponding to H_1. For the SDC model under H_1, denoted as M_1, given by

$$\begin{cases} d_t = \exp\left(\varphi_t\right)\varepsilon_t, & \varepsilon_t \sim Exp\left(1\right), \\ \varphi_t = \phi\varphi_{t-1} + \alpha + x_t'\beta + \sigma\varepsilon_t, & \varepsilon_t \sim N\left(0,1\right), \\ \varphi_1 \sim N\left(\frac{\alpha + x_1'\beta}{1-\phi}, \frac{\sigma^2}{1-\phi^2}\right), & \end{cases}$$

we want to test whether $\beta = \mathbf{0}$ (hence $\theta = \beta$ in this case). As a result, the nuisance parameter $\psi = \left(\alpha, \phi, \sigma^2\right)'$ and $\vartheta = \left(\theta', \psi'\right)'$.

- The estimator of $\mathbf{T}(\mathbf{y},\theta_0)$ and its NSE.

  The proposed statistic is given by:

$$\begin{aligned} \mathbf{T}\left(\mathbf{d},\theta_0\right) &= \int \left(\beta - \bar{\beta}\right)' C_{\theta\theta}\left(\bar{\vartheta}_0\right)\left(\beta - \bar{\beta}\right) p\left(\vartheta|\mathbf{d}\right) d\vartheta \\ &= \mathbf{tr}\left[C_{\theta\theta}\left(\bar{\vartheta}_0\right) E\left(\left(\beta - \bar{\beta}\right)\left(\beta - \bar{\beta}\right)'|\mathbf{y}\right)\right], \end{aligned}$$

  where $\mathbf{d} = \{d_t\}_{t=1}^{T}$, $\bar{\vartheta}_0 = (\mathbf{0}, \bar{\psi}_0)$, $\bar{\psi}_0$ is the posterior mean of ψ under H_0, $\bar{\beta}$ is the posterior mean of β under H_1, and

$$C_{\theta\theta}\left(\bar{\vartheta}_0\right) = \left[\frac{\partial \log p\left(\mathbf{d}|\vartheta\right)}{\partial\theta}\left(\frac{\partial \log p\left(\mathbf{d}|\vartheta\right)}{\partial\theta}\right)'\right]\Bigg|_{\vartheta=\bar{\vartheta}_0} = s_\theta\left(\bar{\vartheta}_0\right) s_\theta\left(\bar{\vartheta}_0\right)'.$$

According to Remark 3.4, the partial derivative of log-likelihood function with respect to θ can be approximated based on the $\mathscr{Q}$-function. That is,

$$s_\theta\left(\bar{\vartheta}_0\right) \approx \frac{1}{G}\sum_{g=1}^{G}\left[-\frac{1}{2\bar{\sigma}_0^2}\tilde{X}'\left(\tilde{\mathbf{y}}^{(g)}-\bar{\alpha}_0\iota\right)\right] = \frac{1}{G}\sum_{g=1}^{G}\mathbf{h}_1^{(g)} = \widehat{\mathbf{h}}_1,$$

where $\iota=(1,...,1)'$, $\tilde{\mathbf{y}}^{(g)}=\left(\sqrt{1-\bar{\phi}_0^2}\varphi_1^{(g)},\varphi_2^{(g)}-\bar{\phi}_0\varphi_1^{(g)},...,\varphi_T^{(g)}-\bar{\phi}_0\varphi_{T-1}^{(g)}\right)'$, $\tilde{X}=\left(\sqrt{\frac{1+\bar{\phi}_0}{1-\bar{\phi}_0}}x_1',x_2',...,x_T'\right)$, $(\bar{\alpha}_0,\bar{\phi}_0,\bar{\sigma}_0^2)$ is the Bayesian estimator under H_0, $\{\varphi_t^{(g)},g=1,2,\cdots,G,t=1,2,\cdots,T\}$ are effective draws of the latent variables from the posterior distribution $p(\varphi|\mathbf{d},\bar{\vartheta}_0)$. Hence, $\mathbf{T}(\mathbf{d},\theta_0)$ can be approximated by

$$\widehat{\mathbf{T}}(\mathbf{d},\theta_0)=\mathbf{tr}\left\{\left[\widehat{C}_{\theta\theta}\left(\bar{\vartheta}_0\right)\right]\left[\frac{1}{G}\sum_{g=1}^{G}\left(\beta^{(g)}-\bar{\beta}\right)\left(\beta^{(g)}-\bar{\beta}\right)'\right]\right\}=\mathbf{tr}\left(\widehat{\mathbf{h}}_1\widehat{\mathbf{h}}_1'\widehat{\mathbf{H}}_2\right),$$

where $\bar{\beta}$ is the posterior mean of β under H_1 , $\left\{\beta^{(g)}\right\}_{g=1}^{G}$ are the MCMC draws from the posterior distribution $p(\vartheta|\mathbf{y})$, and

$$\widehat{C}_{\theta\theta}\left(\bar{\vartheta}_0\right)=\widehat{\mathbf{h}}_1\widehat{\mathbf{h}}_1',\widehat{\mathbf{H}}_2=\frac{1}{G}\sum_{g=1}^{G}\left(\beta^{(g)}-\bar{\beta}\right)\left(\beta^{(g)}-\bar{\beta}\right)'=\frac{1}{G}\sum_{g=1}^{G}\mathbf{H}_2^{(g)}.$$

To calculate the NSE, let $\mathbf{h}_2^{(g)}=vech\left(\mathbf{H}_2^{(g)}\right)$, $\mathbf{h}^{(g)}=\left(\mathbf{h}_1^{(g)\prime},\mathbf{h}_2^{(g)\prime}\right)'$, $\widehat{\mathbf{h}}=\left(\widehat{\mathbf{h}}_1',\widehat{\mathbf{h}}_2'\right)'$. We have

$$\frac{\partial\widehat{\mathbf{h}}_1}{\partial\widehat{\mathbf{h}}}=\begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\end{bmatrix},\frac{\partial\widehat{\mathbf{H}}_2}{\partial\widehat{\mathbf{h}}}=\begin{bmatrix}0&0&1&0&0\\0&0&0&1&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix}.$$

Hence,

$$\frac{\partial\widehat{\mathbf{T}}(\mathbf{d},\theta_0)}{\partial\widehat{\mathbf{h}}}=vec(I_p)'\left[\left(\widehat{H}_2'\widehat{h}_1\otimes I_p+\widehat{H}_2'\otimes\widehat{h}_1\right)\frac{\partial\widehat{h}_1}{\partial\widehat{h}}+\left(I_p\otimes\widehat{h}_1\widehat{h}_1'\right)\frac{\partial\widehat{H}_2}{\partial\widehat{\mathbf{h}}}\right],$$

$$Var\left(\widehat{\mathbf{T}}(\mathbf{d},\theta_0)\right)=\frac{\partial\widehat{\mathbf{T}}(\mathbf{y},\ \theta_0)}{\partial\widehat{\mathbf{h}}}Var\left(\widehat{h}\right)\left(\frac{\partial\widehat{\mathbf{T}}(\mathbf{y},\theta_0)}{\partial\widehat{h}}\right)',$$

$$Var(\widehat{\mathbf{h}})=\frac{1}{G}\left[\Omega_0+\sum_{k=1}^{q}\left(1-\frac{k}{q+1}\right)\left(\Omega_k+\Omega_k'\right)\right],$$

$$\Omega_k=G^{-1}\sum_{g=k+1}^{G}\left(\mathbf{h}^{(g)}-\widehat{\mathbf{h}}\right)\left(\mathbf{h}^{(g)}-\widehat{\mathbf{h}}\right)'.$$

- The estimator of the BF.

  Let $\log p(\mathbf{d}|M_0)$ and $\log p(\mathbf{d}|M_1)$ be the marginal likelihood under H_0 and H_1 respectively. Hence,

  $$\log BF_{10}=\log p(\mathbf{d}|M_1)-\log p(\mathbf{d}|M_0).$$

  The marginal likelihood under H_1 is

  $$\log p(\mathbf{d}|M_1)=\log p\left(\mathbf{d}|\bar{\vartheta}\right)+\log p\left(\bar{\vartheta}\right)-\log p\left(\bar{\vartheta}|\mathbf{y}\right),$$

  where $p\left(\bar{\vartheta}\right)$ is the prior density function evaluated at $\bar{\vartheta}$, $p\left(\bar{\vartheta}|\mathbf{y}\right)$ is the posterior density function evaluated at $\bar{\vartheta}$. The marginal likelihood under H_0 is

  $$\log p(\mathbf{d}|M_0)=\log p\left(\mathbf{y}|\bar{\vartheta}_0\right)+\log p(\bar{\psi}_0)-\log p(\bar{\psi}_0|\mathbf{d},\theta_0).$$

  Following Chib (1995), we can approximate the quantities at the right hand side of the marginal likelihood equations as follows,

  - We use the auxiliary particle filter method proposed by Pitt and Shephard (1999) to estimate $\log p\left(\mathbf{y}|\bar{\vartheta}\right)$ and $\log p\left(\mathbf{y}|\bar{\vartheta}_0\right)$. The code is provided by Creal (2009).
  - $\log p\left(\bar{\vartheta}\right)$ and $\log p(\bar{\psi}_0)$ are easy to evaluate since the prior distributions are standard statistical distributions.
  - $\log p\left(\bar{\vartheta}|\mathbf{y}\right)$ and $\log p(\bar{\psi}_0|\mathbf{d},\theta_0)$ can be estimated via the approach of

Chib (1995).

However, since the NSE of the logarithmic observed likelihood function dominates that of the logarithmic marginal likelihood which is estimated by particle filters, the NSE of the BF cannot be obtained.

.2 Proofs in Chapter 3

.2.1 Proof of Lemma 3.3.1

First, we can show that

$$\begin{aligned}
& \left\| \bar{H}_n(\boldsymbol{\vartheta}) - \bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right) \right\| \\
= & \left\| \bar{H}_n(\boldsymbol{\vartheta}) - \bar{H}_n\left(\boldsymbol{\vartheta}_n^0\right) + \bar{H}_n\left(\boldsymbol{\vartheta}_n^0\right) - H_n\left(\boldsymbol{\vartheta}_n^0\right) + H_n\left(\boldsymbol{\vartheta}_n^0\right) - \bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right) \right\| \\
\leq & \left\| \bar{H}_n(\boldsymbol{\vartheta}) - \bar{H}_n\left(\boldsymbol{\vartheta}_n^0\right) \right\| + \left\| \bar{H}_n\left(\boldsymbol{\vartheta}_n^0\right) - H_n\left(\boldsymbol{\vartheta}_n^0\right) \right\| \\
& + \left\| H_n\left(\boldsymbol{\vartheta}_n^0\right) - \bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right) \right\|. \qquad (.2.1)
\end{aligned}$$

For any ε, there exists a $\delta(\varepsilon) > 0$ such that

$$P\left(\sup_{N\left(\boldsymbol{\vartheta}_n^0, \delta(\varepsilon)\right)} \left\| \bar{H}_n(\boldsymbol{\vartheta}) - \bar{H}_n\left(\boldsymbol{\vartheta}_n^0\right) \right\| < \frac{\varepsilon}{3} \right) \to 1. \qquad (.2.2)$$

From Assumption 3 that $l_t^{(2)}(\boldsymbol{\vartheta})$ is almost surely continuous at $\boldsymbol{\vartheta}_n^0$. We also have

$$P\left(\left\| \bar{H}_n\left(\boldsymbol{\vartheta}_n^0\right) - H_n\left(\boldsymbol{\vartheta}_n^0\right) \right\| < \frac{\varepsilon}{3} \right) \to 1, \qquad (.2.3)$$

$$P\left(\left\| H_n\left(\boldsymbol{\vartheta}_n^0\right) - \bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right) \right\| < \frac{\varepsilon}{3} \right) \to 1, \qquad (.2.4)$$

because of the uniform convergence of $l_t^{(2)}(\boldsymbol{\vartheta})$ and $\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_n^0 \xrightarrow{p} 0$ by Assumptions 1-7 (Gallant and White, 1988). Define events $A_n(\varepsilon) = \left\{ \sup_{N\left(\boldsymbol{\vartheta}_n^0, \delta(\varepsilon)\right)} \left\| \bar{H}_n(\boldsymbol{\vartheta}) - \bar{H}_n\left(\boldsymbol{\vartheta}_n^0\right) \right\| < \frac{\varepsilon}{3} \right\}$, $B_n(\varepsilon) = \left\{ \left\| \bar{H}_n\left(\boldsymbol{\vartheta}_n^0\right) - H_n\left(\boldsymbol{\vartheta}_n^0\right) \right\| < \frac{\varepsilon}{3} \right\}$ and $C_n(\varepsilon) = \left\{ \left\| H_n\left(\boldsymbol{\vartheta}_n^0\right) - \bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right) \right\| < \frac{\varepsilon}{3} \right\}$.

Then we have

$$P(D_n) \geq P(A_n(\varepsilon) \cap B_n(\varepsilon) \cap C_n(\varepsilon)),$$

where

$$D_n = \left\{ \sup_{N\left(\boldsymbol{\vartheta}_n^0, \delta(\varepsilon)\right)} \left\{ \left\| \bar{H}_n(\boldsymbol{\vartheta}) - \bar{H}_n\left(\boldsymbol{\vartheta}_n^0\right) \right\| + \left\| \bar{H}_n\left(\boldsymbol{\vartheta}_n^0\right) - H_n\left(\boldsymbol{\vartheta}_n^0\right) \right\| + \left\| H_n\left(\boldsymbol{\vartheta}_n^0\right) - \bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right) \right\| \right\} < \varepsilon \right\}$$

From (.2.2), (.2.3)and (.2.4), the probability of the complementary event of $A_n(\varepsilon) \cap A_n(\varepsilon) \cap A_n(\varepsilon)$ is

$$\begin{aligned} & P\left((A_n(\varepsilon) \cap B_n(\varepsilon) \cap C_n(\varepsilon))^c\right) \\ = \; & P\left(A_n(\varepsilon)^c \cup B_n(\varepsilon)^c \cup C_n(\varepsilon)^c\right) \leq P\left(A_n(\varepsilon)^c\right) + P\left(B_n(\varepsilon)^c\right) + P\left(C_n(\varepsilon)^c\right) \to 0. \end{aligned}$$

Then

$$P(A_n(\varepsilon) \cap B_n(\varepsilon) \cap C_n(\varepsilon)) \to 1.$$

Hence, by (.2.1), for any $\varepsilon > 0$

$$\begin{aligned} & P\left(\sup_{N\left(\boldsymbol{\vartheta}_n^0, \delta(\varepsilon)\right)} \left\| \bar{H}_n(\boldsymbol{\vartheta}) - \bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right) \right\| < \varepsilon \right) \\ \geq \; & P\left(\sup_{N\left(\boldsymbol{\vartheta}_n^0, \delta(\varepsilon)\right)} \left\| \bar{H}_n(\boldsymbol{\vartheta}) - \bar{H}_n\left(\boldsymbol{\vartheta}_n^0\right) \right\| + \left\| \bar{H}_n\left(\boldsymbol{\vartheta}_n^0\right) - H_n\left(\boldsymbol{\vartheta}_n^0\right) \right\| + \left\| H_n\left(\boldsymbol{\vartheta}_n^0\right) - \bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right) \right\| < \varepsilon \right) \\ \to \; & 1. \end{aligned} \tag{.2.5}$$

It is noted that

$$
\begin{aligned}
& \sup_{N\left(\boldsymbol{\vartheta}_n^0,\delta(\varepsilon)\right),\|r_0\|=1} \left|1-r_0'\bar{H}_n^{-1/2}\left(\widehat{\boldsymbol{\vartheta}}\right)\bar{H}_n\left(\boldsymbol{\vartheta}\right)\bar{H}_n^{-1/2}\left(\widehat{\boldsymbol{\vartheta}}\right)r_0\right| \\
= & \sup_{N\left(\boldsymbol{\vartheta}_n^0,\delta(\varepsilon)\right),\|r_0\|=1} \left|1+r_0'\left(-\bar{H}_n^{-1/2}\left(\widehat{\boldsymbol{\vartheta}}\right)\right)\left(-\bar{H}_n\left(\boldsymbol{\vartheta}\right)\right)\left(-\bar{H}_n^{-1/2}\left(\widehat{\boldsymbol{\vartheta}}\right)\right)r_0\right| \\
= & \sup_{N\left(\boldsymbol{\vartheta}_n^0,\delta(\varepsilon)\right),\|r_0\|=1} \left|r_0'\left(-\bar{H}_n^{-1/2}\left(\widehat{\boldsymbol{\vartheta}}\right)\right)\left[-\bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right)+\bar{H}_n\left(\boldsymbol{\vartheta}\right)\right]\left(-\bar{H}_n^{-1/2}\left(\widehat{\boldsymbol{\vartheta}}\right)\right)r_0\right| \\
\leq & \lambda_n \sup_{N\left(\boldsymbol{\vartheta}_n^0,\delta(\varepsilon)\right),\|r_0\|=1} \left|r_0'\left(\bar{H}_n\left(\boldsymbol{\vartheta}\right)-\bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right)\right)r_0\right| \\
\leq & \lambda_n \sup_{N\left(\boldsymbol{\vartheta}_n^0,\delta(\varepsilon)\right),\|r_0\|=1} \left\|r_0'\right\|\left\|\bar{H}_n\left(\boldsymbol{\vartheta}\right)-\bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right)\right\|\left\|r_0\right\| \\
= & \lambda_n \sup_{N\left(\boldsymbol{\vartheta}_n^0,\delta(\varepsilon)\right),\|r_0\|=1} \left\|\bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right)-\bar{H}_n\left(\boldsymbol{\vartheta}\right)\right\|,
\end{aligned}
$$

where λ_n is the smallest eigenvalue of $-\bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right)$. Then from (.2.5), for any $\varepsilon > 0$

$$
P\left(\sup_{N\left(\boldsymbol{\vartheta}_n^0,\delta(\varepsilon)\right),\|r_0\|=1} \left|1-r_0'\bar{H}_n^{-1/2}\left(\widehat{\boldsymbol{\vartheta}}\right)\bar{H}_n\left(\boldsymbol{\vartheta}\right)\bar{H}_n^{-1/2}\left(\widehat{\boldsymbol{\vartheta}}\right)r_0\right| < \varepsilon\right) \to 1. \quad (.2.6)
$$

.2.2 Proof of Lemma 3.3.2

Lemma .2.1. *Let $X_1, X_2, \ldots, X_q$ be independently and identically distributed, then the following inequality for the order statistic $\max_i X_i$ holds*

$$
E\left[\left(\max_i |X_i|\right)^k\right] < \sqrt{2}\exp\left(\frac{5}{3}\right)\frac{q+1}{\sqrt{q}}\left[E|X_1|^{2k}\right]^{1/2},
$$

under the condition that $E|X_1|^{2k} < \infty$ and $k > 0$.

Proof. Let $\delta = k\rho^{-1}$, $0 < \rho \leq 1/2$, then from Gribkova (1995), the following inequality

$$
E\left[\left|\max_i X_i\right|^k\right] < C(\rho)\left\{E|X_1|^{\delta} g^{-1}\left(\frac{q}{q+1}\right)\right\}^{\rho},
$$

holds for $q \geq 2\rho+1$, where $C(\rho) = 2\sqrt{\rho}\exp(\rho+7/6)$ and $g(u) = u(1-u)$. By setting $\rho = 1/2$, it can be shown that

$$\begin{aligned} E\left[\left|\max_i X_i\right|^k\right] &< C\left(\frac{1}{2}\right)\left\{E\left|X_1\right|^\delta g^{-1}\left(\frac{q}{q+1}\right)\right\}^{1/2} \\ &= \sqrt{2}\exp\left(\frac{5}{3}\right)\frac{q+1}{\sqrt{q}}\left[E\left|X_1\right|^{2k}\right]^{1/2}, \end{aligned} \tag{.2.7}$$

for $q \geq 2$.

For $q = 1$, by Jensen's Inequality,

$$E\left[\left|\max_i X_i\right|^k\right] = E\left[\left|X_1\right|^k\right] \leq \left[E\left|X_1\right|^{2k}\right]^{1/2},$$

then

$$E\left[\left|\max_i X_i\right|^k\right] < \sqrt{2}\exp\left(\frac{5}{3}\right)\frac{1+1}{\sqrt{1}}\left[E\left|X_1\right|^{2k}\right]^{1/2}. \tag{.2.8}$$

From (.2.7) and (.2.8), we can get

$$E\left[\left|\max_i X_i\right|^k\right] < \sqrt{2}\exp\left(\frac{5}{3}\right)\frac{q+1}{\sqrt{q}}\left[E\left|X_1\right|^{2k}\right]^{1/2}, \tag{.2.9}$$

for $k > 0$ and $q \geq 1$.

Let $Y_i = |X_i|$, then it is easy to show that

$$\begin{aligned} E\left[\left(\max_i |X_i|\right)^k\right] &= E\left[\left|\max_i |X_i|\right|^k\right] = E\left[\left|\max_i Y_i\right|^k\right] \\ &< \sqrt{2}\exp\left(\frac{5}{3}\right)\frac{q+1}{\sqrt{q}}\left[E\left|Y_1\right|^{2k}\right]^{1/2} \\ &= \sqrt{2}\exp\left(\frac{5}{3}\right)\frac{q+1}{\sqrt{q}}\left[E\left|X_1\right|^{2k}\right]^{1/2}, \end{aligned}$$

by (.2.9).

$\square$

Lemma .2.2. *Suppose the posterior density of $\boldsymbol{\vartheta}$ can be written as*

$$p(\boldsymbol{\vartheta}|y) = \frac{p(\boldsymbol{\vartheta})\,p(y|\boldsymbol{\vartheta})}{p(y)},$$

where

$$p(y)=\int_{\Theta}p(\boldsymbol{\vartheta})\,p(y|\boldsymbol{\vartheta})\,d\boldsymbol{\vartheta}.$$

Then

$$\lim_{n\to\infty}P\left(\int_{A_n}\|z_n\|^k\left|p(z_n|y)-(2\pi)^{-q/2}\exp\left(-\frac{z_n'z_n}{2}\right)\right|dz_n>\varepsilon\right)=0, \qquad (.2.10)$$

where $A_n=\left\{z_n:\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\in\Theta\right\}$ *is the support of* z_n $\left(:=\Sigma_n^{-1/2}\left(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}}\right)\right)$, $\Sigma_n^{-1}=-\frac{\partial^2\log p\left(y|\widehat{\boldsymbol{\vartheta}}\right)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}'}$.

Proof. The posterior density of z_n, $p(z_n|y)$, can be written as

$$p(z_n|y)=\frac{|\Sigma_n|^{1/2}p(y|\vartheta)\,p(\vartheta)}{p(y)}=\frac{|\Sigma_n|^{1/2}p\left(y|\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)}{p(y)} \qquad (.2.11)$$

Then, we take the Taylor expansion to $\log p\left(y|\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)$ at $\widehat{\boldsymbol{\vartheta}}$ so that we can have

$$\begin{aligned}
&\log p\left(y|\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)\\
=\;&\log p\left(y|\widehat{\boldsymbol{\vartheta}}\right)+\frac{1}{2}z_n'\Sigma_n^{1/2}\frac{\partial^2\log p\left(y|\widetilde{\boldsymbol{\vartheta}}_1\right)}{\partial\theta\partial\boldsymbol{\vartheta}'}\Sigma_n^{1/2}z_n\\
=\;&\log p\left(y|\widehat{\boldsymbol{\vartheta}}\right)-\frac{1}{2}z_n'\Sigma_n^{1/2}\left[-\frac{\partial^2\log p\left(y|\widehat{\boldsymbol{\vartheta}}\right)}{\partial\theta\partial\boldsymbol{\vartheta}'}-\frac{\partial^2\log p\left(y|\widetilde{\boldsymbol{\vartheta}}_1\right)}{\partial\theta\partial\boldsymbol{\vartheta}'}+\frac{\partial^2\log p\left(y|\widehat{\boldsymbol{\vartheta}}\right)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}'}\right]\Sigma_n^{1/2}z_n\\
=\;&\log p\left(y|\widehat{\boldsymbol{\vartheta}}\right)-\frac{1}{2}z_n'\Sigma_n^{1/2}\left[\Sigma_n^{-1}-\frac{\partial^2\log p\left(y|\widetilde{\boldsymbol{\vartheta}}_1\right)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}'}-\Sigma_n^{-1}\right]\Sigma_n^{1/2}z_n\\
=\;&\log p\left(y|\widehat{\boldsymbol{\vartheta}}\right)-\frac{1}{2}z_n'\left[I_q-R_n(\boldsymbol{\vartheta},y)\right]z_n, \qquad (.2.12)
\end{aligned}$$

where I_q is a q-dimension identity matrix and

$$R_n(\boldsymbol{\vartheta},y)=I_q+\Sigma_n^{1/2}\frac{\partial^2\log p\left(y|\widetilde{\boldsymbol{\vartheta}}_1\right)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}'}\Sigma_n^{1/2},$$

with $\widetilde{\boldsymbol{\vartheta}}_1$ lies between $\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n$ and $\widehat{\boldsymbol{\vartheta}}$.

To prove (.2.10), note that

$$
\begin{aligned}
& p\left(z_n|y\right)-(2\pi)^{-q/2}\exp\left(-\frac{z_n'z_n}{2}\right)\\
= & \; p\left(y\right)^{-1}\left|\Sigma_n\right|^{1/2}p\left(y|\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)-(2\pi)^{-q/2}\exp\left(-\frac{z_n'z_n}{2}\right)\\
= & \; p\left(y\right)^{-1}\left|\Sigma_n\right|^{1/2}p\left(y|\widehat{\boldsymbol{\vartheta}}\right)p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)\frac{p\left(y|\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)}{p\left(y|\widehat{\boldsymbol{\vartheta}}_n\right)}-(2\pi)^{-q/2}\exp\left(-\frac{z_n'z_n}{2}\right),
\end{aligned}
$$

and that

$$
p\left(y\right)^{-1}\left|\Sigma_n\right|^{1/2}p\left(y|\widehat{\boldsymbol{\vartheta}}\right)\xrightarrow{p}\frac{(2\pi)^{-q/2}}{p\left(\boldsymbol{\vartheta}_n^0\right)},
$$

by Chen (1985) and Schervish (2012). Hence, according to (.2.12), to verify (.2.10), it is sufficient to show

$$
P\left(\int_{A_n}\|z_n\|^k\left|\frac{p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)}{p\left(\boldsymbol{\vartheta}_n^0\right)}\exp\left[-\frac{z_n'[I_q-R_n(\boldsymbol{\vartheta},y)]z_n}{2}\right]-\exp\left(-\frac{z_n'z_n}{2}\right)\right|dz_n<\varepsilon\right)\to 1 \tag{.2.13}
$$

Hence, to ensure (.2.13), by assumption 9, it is enough to prove

$$
P\left(\int_{A_n}\|z_n\|^k\left|p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)\exp\left[-\frac{z_n'[I_q-R_n(\boldsymbol{\vartheta},y)]z_n}{2}\right]-p\left(\boldsymbol{\vartheta}_n^0\right)\exp\left(-\frac{z_n'z_n}{2}\right)\right|dz_n<\varepsilon\right)\to 1 \tag{.2.14}
$$

In the following, we prove that (.2.14) holds. Since the prior density function is continuous at $\boldsymbol{\vartheta}_n^0$, that is, given any $\varepsilon>0$, for any $\eta\in(0,1)$ satisfying

$$
\varepsilon\geq\eta\left(\frac{q^2(1+\eta)\sqrt{(2k+1)(2k+3)}}{2(1-\eta)^{\frac{q+k+2}{2}}}+1\right),
$$

$\exists\delta_1>0$, so that for any $\boldsymbol{\vartheta}$ satisfying $\left\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_n^0\right\|\leq\delta_1$, that is, $\boldsymbol{\vartheta}\in N_0(\delta_1)=\left\{\boldsymbol{\vartheta}:\left\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_n^0\right\|\leq\delta_1\right\}$,

$$
\left|p(\boldsymbol{\vartheta})-p\left(\boldsymbol{\vartheta}_n^0\right)\right|=\left|p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)-p\left(\boldsymbol{\vartheta}_n^0\right)\right|\leq\eta p\left(\boldsymbol{\vartheta}_n^0\right). \tag{.2.15}
$$

Furthermore, by Lemma 3.3.1, $\forall \eta > 0$, $\exists \delta_2 > 0$,

$$\lim_{n\to\infty} P\left(\sup_{\boldsymbol{\vartheta}\in N_0(\delta_2),\|r_0\|=1}\left|1+r_0'\Sigma_n^{1/2}\frac{\partial^2 \log p\left(y|\boldsymbol{\vartheta}\right)}{\partial \boldsymbol{\vartheta}\partial \boldsymbol{\vartheta}'}\Sigma_n^{1/2}r_0\right|<\eta\right)=1, \qquad (.2.16)$$

where $N_0\left(\delta\right)=\left\{\boldsymbol{\vartheta}:\left\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_n^0\right\|\leq\delta\right\}$, see Schervish (2012).

Let $\delta=\min\left\{\delta_1,\delta_2\right\}$ and define

$$A_{1n}=\left\{z_n:\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\in N_0\left(\delta\right)\right\}, A_{2n}=\left\{z_n:\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\in\Theta\backslash N_0\left(\delta\right)\right\},$$

and

$$C_n=\|z_n\|^k\left|p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)\exp\left[-\frac{1}{2}z_n'\left[I_q-R_n(\tilde{\boldsymbol{\vartheta}}_1)\right]z_n\right]-p\left(\boldsymbol{\vartheta}_n^0\right)\exp\left(-\frac{z_n'z_n}{2}\right)\right|. \qquad (.2.17)$$

The integration of C_n in the space A_n can be decomposed into two areas, A_{1n} and A_{2n}, i.e.,

$$J=\int_{A_n}C_ndz_n=J_1+J_2,$$

where $J_1=\int_{A_{1n}}C_ndz_n$, $J_2=\int_{A_{2n}}C_ndz_n$. In the following, we try to prove

$$J_1=\int_{A_{1n}}C_ndz_n\xrightarrow{p}0,\ J_2=\int_{A_{2n}}C_ndz_n\xrightarrow{p}0.$$

For J_1, we note that

$$C_n\leq C_{1n}+C_{2n}$$

where

$$C_{1n}=\|z_n\|^k\left|p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)\right|\left|\exp\left[-\frac{1}{2}z_n'\left[I_q-R_n\left(\boldsymbol{\vartheta},y\right)\right]z_n\right]-\exp\left(-\frac{z_n'z_n}{2}\right)\right|,$$

$$C_{2n}=\|z_n\|^k\left|p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)-p\left(\boldsymbol{\vartheta}_n^0\right)\right|\exp\left(-\frac{z_n'z_n}{2}\right).$$

Then we have

$$0\leq J_1\leq J_{11}+J_{12},$$

where

$$J_{11}=\int_{A_{1n}} C_{1n}dz_n,\ \ J_{12}=\int_{A_{1n}} C_{2n}dz_n.$$

It is noted that since $\delta \leq \delta_1$, from (.2.15), we can know that $\left|p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)\right| \leq (1+\eta)\,p\left(\boldsymbol{\vartheta}_n^0\right)$. Hence, we can have

$$J_{11}\leq(1+\eta)\,p\left(\boldsymbol{\vartheta}_n^0\right)\int_{A_{1n}}\|z_n\|^k\left|\exp\left[-\frac{z_n'\left[I_q-R_n\left(\boldsymbol{\vartheta},y\right)\right]z_n}{2}\right]-\exp\left(-\frac{z_n'z_n}{2}\right)\right|dz_n.$$

Let $r_0=z_n/\|z_n\|$, so $\|r_0\|=1$, then, we can get that

$$r_0'R_n\left(\tilde{\boldsymbol{\vartheta}}_1\right)r_0=r_0'r_0+r_0'\Sigma^{1/2}\frac{\partial^2\log p\left(y|\widetilde{\boldsymbol{\vartheta}}_1\right)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}'}\Sigma^{1/2}r_0=1+r_0'\Sigma^{1/2}\frac{\partial^2\log p\left(y|\widetilde{\boldsymbol{\vartheta}}_1\right)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}'}\Sigma^{1/2}r_0,$$

where $\widetilde{\boldsymbol{\vartheta}}_1$ lies between $\boldsymbol{\vartheta}$ and $\widehat{\boldsymbol{\vartheta}}$. Since $\widehat{\boldsymbol{\vartheta}}\xrightarrow{p}\boldsymbol{\vartheta}_n^0$, we can get that with probability 1, $\hat{\vartheta}\in N_0(\delta)$, and hence, $\tilde{\vartheta}_1\in N_0(\delta)$ with probability 1.

Following (.2.16), with probability 1, when $\theta\in N_0(\delta)$, we can further get that

$$\begin{aligned}
& \|z_n\|^k\left|\exp\left[-\frac{1}{2}z_n'\left[I_q-R_n\left(\boldsymbol{\vartheta},y\right)\right]z_n\right]-\exp\left(-\frac{z_n'z_n}{2}\right)\right| \\
= \ & \|z_n\|^k\left|\exp\left[\frac{1}{2}z_n'R_n\left(\boldsymbol{\vartheta},y\right)z_n\right]-1\right|\exp\left(-\frac{z_n'z_n}{2}\right) \\
\leq \ & \|z_n\|^k\exp\left[\left|\frac{1}{2}z_n'R_n\left(\boldsymbol{\vartheta},y\right)z_n\right|\right]\left|\frac{1}{2}z_n'R_n\left(\boldsymbol{\vartheta},y\right)z_n\right|\exp\left(-\frac{z_n'z_n}{2}\right) \\
= \ & \|z_n\|^k\exp\left[\left|\frac{1}{2}z_n'z_n\right|\left|r_0'R_n\left(\boldsymbol{\vartheta},y\right)r_0\right|\right]\left|\frac{1}{2}z_n'z_n\right|\left|r_0'R_n\left(\boldsymbol{\vartheta},y\right)r_0\right|\exp\left(-\frac{z_n'z_n}{2}\right) \\
\leq \ & \frac{\eta}{2}\|z_n\|^k\exp\left[\left|\frac{\eta}{2}z_n'z_n\right|\right]\left|z_n'z_n\right|\exp\left(-\frac{z_n'z_n}{2}\right) \\
= \ & \frac{\eta}{2}\|z_n\|^{k+2}\exp\left(-\frac{(1-\eta)z_n'z_n}{2}\right) \qquad (.2.18)
\end{aligned}$$

Let

$$J_{11}^*=\int_{A_{1n}}\|z_n\|^k\left|\exp\left[-\frac{1}{2}z_n'\left[I_q-R_n\left(\boldsymbol{\vartheta},y\right)\right]z_n\right]-\exp\left(-\frac{z_n'z_n}{2}\right)\right|dz_n,$$

It follows from (.2.18), we can get that

$$\lim_{n\to\infty} P\left\{J_{11}^{*} \leq \frac{\eta}{2}\int_{A_{1n}} \|z_n\|^{k+2} \exp\left(-\frac{(1-\eta)\,z_n' z_n}{2}\right) dz_n\right\} = 1. \qquad (.2.19)$$

It is noted that, by Lemma .2.1, we have

$$\begin{aligned}
& \int_{A_{1n}} \|z_n\|^{k+2} \exp\left(-\frac{(1-\eta)\,z_n' z_n}{2}\right) dz_n \\
\leq\ & \int_{\mathbb{R}^q} \|z_n\|^{k+2} \exp\left(-\frac{1-\eta}{2} z_n' z_n\right) dz_n \leq \int_{\mathbb{R}^q} \left(\sum_{i=1}^{q} |z_{ni}|^2\right)^{\frac{k+2}{2}} \exp\left(-\frac{1-\eta}{2} z_n' z_n\right) dz_n \\
\leq\ & (2\pi)^{q/2}(1-\eta)^{-q/2} q^{k+2} \int_{\mathbb{R}^q} \left(\max_i |z_{ni}|\right)^{k+2} (2\pi)^{-q/2}(1-\eta)^{q/2} \exp\left(-\frac{1-\eta}{2} z_n' z_n\right) dz_n \\
\leq\ & \sqrt{2}\exp\left(\frac{5}{3}\right)\left(\frac{q+1}{\sqrt{q}}\right) q^{k+2} (2\pi)^{q/2} (1-\eta)^{-q/2} \\
& \times \left[\int_{\mathbb{R}} |t|^{2(k+2)} \sqrt{\frac{1-\eta}{2\pi}} \exp\left(-\frac{1-\eta}{2} t^2\right) dt\right]^{1/2} \\
=\ & \sqrt{2}\exp\left(\frac{5}{3}\right)(q+1)\, q^{k+\frac{3}{2}} (2\pi)^{q/2} (1-\eta)^{-q/2} (1-\eta)^{-(k+2)/2} 2^{(k+2)/2} \left(\frac{\Gamma\left(\frac{2k+5}{2}\right)}{\sqrt{\pi}}\right)^{1/2} \\
=\ & 2^{\frac{k+q+3}{2}} \exp\left(\frac{5}{3}\right)(q+1)\, q^{k+\frac{3}{2}} \sqrt{\Gamma\left(\frac{2k+5}{2}\right)}\, \pi^{\frac{2q-1}{4}} \left(\frac{1}{1-\eta}\right)^{(k+q+2)/2} \\
=\ & 2^{\frac{k+q+3}{2}} \exp\left(\frac{5}{3}\right)(q+1)\, q^{k+\frac{3}{2}} \pi^{\frac{2q-1}{4}} \sqrt{\frac{2k+3}{2}\frac{2k+1}{2}\Gamma\left(\frac{2k+1}{2}\right)} \left(\frac{1}{1-\eta}\right)^{(k+q+2)/2},
\end{aligned}$$

where z_{ni} is the ith element of z_n and the penultimate equation results from the fact that the central absolute moment of a scalar normal random variable X with mean μ and variance σ^2 is

$$E\left\{|X-\mu|^{\nu}\right\} = \sigma^{\nu} 2^{\nu/2} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}}.$$

Hence, we have,

$$\lim_{n\to\infty} P\left(\frac{J_{11}}{C_{J_1}} \leq \frac{q^2 \eta\,(1+\eta)\sqrt{(2k+1)(2k+3)}}{2\,(1-\eta)^{\frac{q+k+2}{2}}}\right) = 1, \qquad (.2.20)$$

where

$$C_{J_1} = \exp\left(\frac{5}{3}\right) p\left(\boldsymbol{\vartheta}_n^0\right) 2^{\frac{q+k+1}{2}} \pi^{\frac{2q-1}{4}} (q+1) q^{k-\frac{1}{2}} \sqrt{\Gamma\left(\frac{2k+1}{2}\right)}.$$

In the following, we deal with J_{12}. From (.2.15) and Lemma .2.1, we have

$$\begin{aligned}
J_{12} &\leq \int_{A_{1n}} \|z_n\|^k \left| p\left(\widehat{\boldsymbol{\vartheta}} + \Sigma_n^{1/2} z_n\right) - p\left(\boldsymbol{\vartheta}_n^0\right)\right| \exp\left(-\frac{z_n' z_n}{2}\right) dz_n \\
&\leq \eta p\left(\boldsymbol{\vartheta}_n^0\right) \int_{A_{1n}} \|z_n\|^k \exp\left(-\frac{z_n' z_n}{2}\right) dz_n \\
&\leq \eta p\left(\boldsymbol{\vartheta}_n^0\right) \int_{\mathbb{R}^q} \|z_n\|^k \exp\left(-\frac{z_n' z_n}{2}\right) dz_n \\
&= \eta p\left(\boldsymbol{\vartheta}_n^0\right) (2\pi)^{q/2} \int_{\mathbb{R}^q} \|z_n\|^k (2\pi)^{-q/2} \exp\left(-\frac{z_n' z_n}{2}\right) dz_n \\
&\leq \eta p\left(\boldsymbol{\vartheta}_n^0\right) (2\pi)^{q/2} q^k \int_{\mathbb{R}^q} \left(\max_i |z_{ni}|\right)^k (2\pi)^{-q/2} \exp\left(-\frac{1-\eta}{2} z_n' z_n\right) dz_n \\
&\leq \sqrt{2} \exp\left(\frac{5}{3}\right) \left(\frac{q+1}{\sqrt{q}}\right) \eta p\left(\boldsymbol{\vartheta}_n^0\right) (2\pi)^{q/2} q^k \left[\int_{\mathbb{R}} |t|^{2k} (2\pi)^{-1/2} \exp\left(-\frac{t^2}{2}\right) dt\right]^{1/2} \\
&= \eta \sqrt{2} \exp\left(\frac{5}{3}\right) p\left(\boldsymbol{\vartheta}_n^0\right) (2\pi)^{q/2} (q+1) q^{k-\frac{1}{2}} 2^{k/2} \left(\frac{\Gamma\left(\frac{2k+1}{2}\right)}{\sqrt{\pi}}\right)^{1/2} \\
&= \eta \exp\left(\frac{5}{3}\right) p\left(\boldsymbol{\vartheta}_n^0\right) 2^{\frac{q+k+1}{2}} \pi^{\frac{2q-1}{4}} (q+1) q^{k-\frac{1}{2}} \sqrt{\Gamma\left(\frac{2k+1}{2}\right)} \\
&= C_{J_1} \eta.
\end{aligned}$$

Similarly, we have

$$\lim_{n\to\infty} P\left\{\frac{J_{12}}{C_{J_1}} \leq \eta\right\} = 1. \tag{.2.21}$$

And from (.2.20) and (.2.21),

$$\lim_{n\to\infty} P\left\{\frac{J_{11}+J_{12}}{C_{J_1}} \leq \eta \left(\frac{q^2 (1+\eta) \sqrt{(2k+1)(2k+3)}}{2(1-\eta)^{\frac{q+k+2}{2}}} + 1\right)\right\} = 1. \tag{.2.22}$$

By the way how η and ε are chosen, we can get from (.2.22) that

$$\lim_{n\to\infty} P\left\{\frac{J_1}{C_{J_1}} \leq \varepsilon\right\} = 1. \tag{.2.23}$$

Since ε is chosen arbitrarily and $J_1 \geq 0$, we have

$$J_1 \xrightarrow{p} 0.$$

Next we show that

$$J_2 \xrightarrow{p} 0. \tag{.2.24}$$

Using (.2.17), we can write

$$0 \leq J_2 = \int_{A_{2n}} C_n dz_n \leq J_{21} + J_{22},$$

where

$$J_{21} = \int_{A_{2n}} \|z_n\|^k p\left(\widehat{\boldsymbol{\vartheta}} + \Sigma_n^{1/2} z_n\right) \exp\left[-\frac{1}{2} z_n' \left[I_P - R_n(\boldsymbol{\vartheta}, y)\right] z_n\right] dz_n,$$

$$J_{22} = \int_{A_{2n}} \|z_n\|^k p\left(\boldsymbol{\vartheta}_n^0\right) \exp\left(-\frac{z_n' z_n}{2}\right) dz_n.$$

For J_{21}, in terms of (.2.12), we have

$$\begin{aligned}
J_{21} &= \int_{A_{2n}} \|z_n\|^k p\left(\widehat{\boldsymbol{\vartheta}} + \Sigma_n^{1/2} z_n\right) \exp\left[-\frac{1}{2} z_n' \left[I_q - R_n(\boldsymbol{\vartheta}, y)\right] z_n\right] dz_n \\
&= \int_{A_{2n}} \|z_n\|^k p\left(\widehat{\boldsymbol{\vartheta}} + \Sigma_n^{1/2} z_n\right) \exp\left[\log p\left(y|\widehat{\boldsymbol{\vartheta}} + \Sigma_n^{1/2} z_n\right) - \log p\left(y|\widehat{\boldsymbol{\vartheta}}\right)\right] dz_n \\
&= \int_{A_{2n}} \|z_n\|^k p\left(\widehat{\boldsymbol{\vartheta}} + \Sigma_n^{1/2} z_n\right) \exp\left[\log p\left(y|\widehat{\boldsymbol{\vartheta}} + \Sigma_n^{1/2} z_n\right) - \log p\left(y|\boldsymbol{\vartheta}_n^0\right)\right] dz_n \\
&\quad \times \exp\left[\log p\left(y|\boldsymbol{\vartheta}_n^0\right) - \log p\left(y|\widehat{\boldsymbol{\vartheta}}\right)\right]. \qquad (.2.25)
\end{aligned}$$

According to Lemma 3.1 in Li et al. (2017), if $z_n \in A_{2n}$, $\log p\left(y|\widehat{\boldsymbol{\vartheta}} + \Sigma^{1/2} z_n\right) - \log p\left(y|\boldsymbol{\vartheta}_n^0\right) < -nK(\delta)$ with probability approaching 1. It is noted that

$$\exp\left[\log p\left(y|\boldsymbol{\vartheta}_n^0\right) - \log p\left(y|\widehat{\boldsymbol{\vartheta}}\right)\right] \leq 1.$$

Hence, the integral on the right-hand side of (.2.25) is less than

$$\exp\left[-nK(\delta)\right]\int_{A_{2n}}\|z_n\|^k p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)dz_n,$$

with probability approaching 1. Then, we can have

$$\begin{aligned}
& \exp\left[-nK(\delta)\right]\int_{A_{2n}}\|z_n\|^k p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2}z_n\right)dz_n \\
= \;& \exp\left[-nK(\delta)\right]\int_{\Theta\backslash N_0(\delta)}\left\|\Sigma_n^{-1/2}\left(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}}\right)\right\|^k p(\boldsymbol{\vartheta})\left|\Sigma_n\right|^{-1/2}d\boldsymbol{\vartheta} \\
\leq \;& \exp\left[-nK(\delta)\right]\int_{\Theta\backslash N_0(\delta)}\left\|\Sigma_n^{-1/2}\right\|^k\left\|\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}}\right\|^k p(\boldsymbol{\vartheta})\left|\Sigma_n\right|^{-1/2}d\boldsymbol{\vartheta} \\
\leq \;& \exp\left[-nK(\delta)\right]\left\|\Sigma_n^{-1/2}\right\|^k\left|\Sigma_n\right|^{-1/2}\int_{\boldsymbol{\Theta}}\left\|\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}}\right\|^k p(\boldsymbol{\vartheta})d\boldsymbol{\vartheta} \\
\leq \;& \exp\left[-nK(\delta)\right]\left\|\Sigma_n^{-1/2}\right\|^k\left|\Sigma_n\right|^{-1/2}\int_{\boldsymbol{\Theta}}\left(\|\boldsymbol{\vartheta}\|+\left\|\widehat{\boldsymbol{\vartheta}}\right\|\right)^k p(\boldsymbol{\vartheta})d\boldsymbol{\vartheta} \\
\leq \;& \exp\left[-nK(\delta)\right]\left\|\Sigma_n^{-1/2}\right\|^k\left|\Sigma_n\right|^{-1/2}\sum_{s=0}^{k}\begin{pmatrix} k \\ s \end{pmatrix}\left\|\widehat{\boldsymbol{\vartheta}}\right\|^{k-s}\int_{\boldsymbol{\Theta}}\|\boldsymbol{\vartheta}\|^s p(\boldsymbol{\vartheta})d\boldsymbol{\vartheta}.
\end{aligned}$$

Note that

$$\begin{aligned}
& \exp\left[-nK(\delta)\right]\left\|\Sigma_n^{-1/2}\right\|^k\left|\Sigma_n\right|^{-1/2} \\
= \;& \exp\left[-nK(\delta)\right]n^{(k+1)/2}\left\|-\bar{H}_n^{-1/2}\right\|^k\left|\bar{H}_n\right|^{-1/2}\xrightarrow{p}0,
\end{aligned}$$

Furthermore, $\int_{\boldsymbol{\Theta}}\|\boldsymbol{\vartheta}\|^k p(\boldsymbol{\vartheta})d\boldsymbol{\vartheta}<\infty$ and $\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}_n^0\xrightarrow{p}0$ by the Assumptions 1-8, then we have

$$J_{21}\xrightarrow{p}0. \tag{.2.26}$$

For J_{22}, we can show that

$$
\begin{aligned}
J_{22} &= \int_{A_{2n}} \|z_n\|^k p(\boldsymbol{\vartheta}_n^p) \exp\left(-\frac{z_n' z_n}{2}\right) dz_n \\
&= p(\boldsymbol{\vartheta}_n^p) \int_{A_{2n}} \|z_n\|^k \exp\left(-\frac{z_n' z_n}{2}\right) dz_n \\
&\leq p\left(\boldsymbol{\vartheta}_n^0\right) \int_{\|z_n\| > \sqrt{n\lambda_n}\delta} \|z_n\|^k \exp\left(-\frac{z_n' z_n}{2}\right) dz_n \\
&\leq (2\pi)^{q/2} p\left(\boldsymbol{\vartheta}_n^0\right) \int_{\cap_{i=1}^q \{|z_{ni}| > \sqrt{\frac{n\lambda_n}{q+1}}\delta\}} \|z_n\|^k (2\pi)^{-q/2} \exp\left(-\frac{z_n' z_n}{2}\right) dz_n \\
&\leq (2\pi)^{q/2} p\left(\boldsymbol{\vartheta}_n^0\right) q^k \int_{\cap_{i=1}^q \{|z_{ni}| > \sqrt{\frac{n\lambda_n}{q+1}}\delta\}} \left(\max_i |z_{ni}|\right)^k (2\pi)^{-q/2} \exp\left(-\frac{z_n' z_n}{2}\right) dz_n \\
&= (2\pi)^{q/2} p\left(\boldsymbol{\vartheta}_n^0\right) q^k \int_{R^q} \left(\max_i |z_{ni}|\right)^k 1\left(\cap_{i=1}^q \{|z_{ni}| > \sqrt{\frac{n\lambda_n}{q+1}}\delta\}\right) (2\pi)^{-q/2} \exp\left(-\frac{z_n' z_n}{2}\right) dz_n \\
&= (2\pi)^{q/2} p\left(\boldsymbol{\vartheta}_n^0\right) q^k \int_{R^q} \left(\max_i |z_{ni}|\right)^k \prod_{i=1}^q 1\left(|z_{ni}| > \sqrt{\frac{n\lambda_n}{q+1}}\delta\right) (2\pi)^{-q/2} \exp\left(-\frac{z_n' z_n}{2}\right) dz_n \\
&\leq (2\pi)^{q/2} p\left(\boldsymbol{\vartheta}_n^0\right) q^k \left[\int_{R^q} \left(\max_i |z_{ni}|\right)^{2k} (2\pi)^{-q/2} \exp\left(-\frac{z_n' z_n}{2}\right) dz_n\right]^{1/2} \\
&\quad \times \left\{\int_{R^q} \left[\prod_{i=1}^q 1\left(|z_{ni}| > \sqrt{\frac{n\lambda_n}{q+1}}\delta\right)\right]^2 (2\pi)^{-q/2} \exp\left(-\frac{z_n' z_n}{2}\right) dz_n\right\}^{1/2}
\end{aligned}
$$

where z_{ni} is the ith element of z_n and λ_n is the smallest eigenvalue of $-\bar{H}_n\left(\widehat{\boldsymbol{\vartheta}}\right)$.

From (.2.9), we have

$$
\int_{R^q} \left(\max_i |z_{ni}|\right)^{2k} (2\pi)^{-q/2} \exp\left(-\frac{z_n' z_n}{2}\right) dz_n < \infty. \qquad (.2.27)
$$

It can be shown that

$$\begin{aligned}
& \int_{R^q}\left[\prod_{i=1}^q 1\left(|z_{ni}|>\sqrt{\frac{n\lambda_n}{q+1}}\delta\right)\right]^2(2\pi)^{-q/2}\exp\left(-\frac{z_n'z_n}{2}\right)dz_n \\
= \; & \int_{R^q}\prod_{i=1}^q 1\left(|z_{ni}|>\sqrt{\frac{n\lambda_n}{q+1}}\delta\right)(2\pi)^{-q/2}\exp\left(-\frac{z_n'z_n}{2}\right)dz_n \\
= \; & \prod_{i=1}^q\left[\int_R 1\left(|z_{ni}|>\sqrt{\frac{n\lambda_n}{q+1}}\delta\right)(2\pi)^{-1/2}\exp\left(-\frac{z_{ni}^2}{2}\right)dz_{ni}\right] \\
= \; & \prod_{i=1}^q\left[\int_{|z_{ni}|>\sqrt{\frac{n\lambda_n}{q+1}}\delta}(2\pi)^{-1/2}\exp\left(-\frac{z_{ni}^2}{2}\right)dz_{ni}\right] \\
\leq \; & \left(\sqrt{q+1}\frac{\exp\left(-n\lambda_n\delta^2/2(q+1)\right)}{\sqrt{n\lambda_n 2\pi}\delta}\right)^q \\
= \; & 2^{-\frac{q}{2}}(q+1)^{\frac{q}{2}}\left(\frac{1}{\sqrt{\pi}\delta}\right)^q(n\lambda_n)^{-\frac{q}{2}}\exp\left(-\frac{n\lambda_n q\delta^2}{q+1}\right)\xrightarrow{p}0, \qquad (.2.28)
\end{aligned}$$

where the last inequality results from

$$\int_x^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \leq \int_x^\infty \frac{1}{\sqrt{2\pi}}\frac{t}{x}e^{-\frac{t^2}{2}}dt = \frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}}.$$

From (.2.27) and (.2.28), we have

$$J_{22}\xrightarrow{p}0. \qquad (.2.29)$$

From (.2.26) and (.2.29), we can get (.2.24). And from (.2.23) and (.2.24), we have

$$J\xrightarrow{p}0.$$

□

To prove $E\left[\left(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}}\right)\left(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}}\right)'|y\right]-\Sigma_n=o_p\left(\frac{1}{n}\right)$, it is sufficient to show that, for any $\varepsilon>0$,

$$\lim_{n\to\infty}P\left(\int_{A_n}\|z_nz_n'\|\left|p\left(z_n|y\right)-(2\pi)^{-q/2}\exp\left(-\frac{z_n'z_n}{2}\right)\right|dz_n>\varepsilon\right)=0, \qquad (.2.30)$$

where $\|\cdot\|$ is the matrix norm for a matrix A defined as $\|A\| = \sup_{\|x\|=1} \|Ax\|$. It is because that by (.2.30),

$$\int_{A_n} \|z_n z_n'\| \left| p(z_n|y) - (2\pi)^{-q/2} \exp\left(-\frac{z_n' z_n}{2}\right)\right| dz_n \xrightarrow{p} 0.$$

Thus, we have $\left|\int_{A_n} z_n z_n' \left[p(z_n|y) - (2\pi)^{-q/2} \exp\left(-\frac{z_n' z_n}{2}\right)\right] dz_n\right| \xrightarrow{p} 0_{q\times q}$, which implies that

$$\int_{A_n} z_n z_n' p(z_n|y)\, dz_n - \int_{A_n} z_n z_n' (2\pi)^{-q/2} \exp\left(-\frac{z_n' z_n}{2}\right) dz_n \xrightarrow{p} 0_{q\times q}. \qquad (.2.31)$$

So from (.2.11) we can get

$$\begin{aligned}
\int_{A_n} z_n z_n' p(z_n|y)\, dz_n &= \int_{A_n} z_n z_n' p(y)^{-1} |\Sigma_n|^{1/2} p\left(y|\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2} z_n\right) p\left(\widehat{\boldsymbol{\vartheta}}+\Sigma_n^{1/2} z_n\right) dz_n \\
&= \int_{\boldsymbol{\Theta}} \Sigma_n^{-1/2} \left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)\left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)' \Sigma_n^{-1/2} p(y)^{-1} |\Sigma_n|^{1/2} p(y|\boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta}) |\Sigma_n|^{-1/2} d\boldsymbol{\vartheta} \\
&= \int_{\boldsymbol{\Theta}} \Sigma_n^{-1/2} \left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)\left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)' \Sigma_n^{-1/2} p(\boldsymbol{\vartheta}|y)\, d\boldsymbol{\vartheta} \\
&= \Sigma_n^{-1/2} E\left[\left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)\left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)' |y\right] \Sigma_n^{-1/2}, \qquad (.2.32)
\end{aligned}$$

by changing of variables. From (.2.31) and (.2.32), using Assumptions 1-9, we have

$$\Sigma_n^{-1/2} E\left[\left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)\left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)' |y\right] \Sigma_n^{-1/2} - I_q \xrightarrow{p} 0_{q\times q}.$$

Hence, we can have that

$$\begin{aligned}
E\left[\left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)\left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)' |y\right] - \Sigma_n &= E\left[\left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)\left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)' |y\right] + \left(\frac{\partial^2 \log p\left(y|\widehat{\boldsymbol{\vartheta}}\right)}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}'}\right)^{-1} \\
&= o_p\left(\frac{1}{n}\right).
\end{aligned}$$

Since $\|z_n z_n'\| \leq \|z_n\|^2$, when $k = 2$, the formula (.2.10) holds so that (.2.30) is also held. Similarly, from this Lemma with $k = 1$, it is also easy to derive that $\sqrt{n}\left(\bar{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}}\right) \xrightarrow{p} 0$.

.2.3 Proof of Theorem 3.3.1

According to Lemma 3.3.2, we have

$$E\left[\left(\vartheta-\widehat{\vartheta}\right)|y\right]=o_p\left(n^{-\frac{1}{2}}\right)$$
$$V\left(\widehat{\vartheta}\right)=E\left[\left(\vartheta-\widehat{\vartheta}_n\right)\left(\vartheta-\widehat{\vartheta}\right)'|y\right]=-\frac{1}{n}\bar{H}_n^{-1}(\widehat{\vartheta})+o_p\left(n^{-1}\right)=O_p(\frac{1}{n}).$$

Hence, based on Lemma 3.3.2, we have

$$\begin{aligned}
&V\left(\bar{\vartheta}\right)=E\left[(\vartheta-\bar{\vartheta})(\vartheta-\bar{\vartheta})'|y\right]\\
=\;&E\left[\left(\vartheta-\widehat{\vartheta}+\widehat{\vartheta}-\bar{\vartheta}\right)\left(\vartheta-\widehat{\vartheta}+\widehat{\vartheta}-\bar{\vartheta}\right)'|y\right]\\
=\;&E\left[\left(\vartheta-\widehat{\vartheta}\right)\left(\vartheta-\widehat{\vartheta}\right)'|y\right]+2E\left[\left(\widehat{\vartheta}-\bar{\vartheta}\right)\left(\vartheta-\widehat{\vartheta}\right)'|y\right]+E\left[\left(\widehat{\vartheta}-\bar{\vartheta}\right)\left(\widehat{\vartheta}-\bar{\vartheta}\right)'|y\right]\\
=\;&V\left(\widehat{\vartheta}\right)-E\left[\left(\widehat{\vartheta}-\bar{\vartheta}\right)\left(\widehat{\vartheta}-\bar{\vartheta}\right)'|y\right]\\
=\;&V\left(\widehat{\vartheta}\right)+o_p(n^{-1/2})o_p(n^{-1/2})\\
=\;&V\left(\widehat{\vartheta}\right)+o_p(n^{-1})\\
=\;&-\frac{1}{n}\bar{H}_n^{-1}(\widehat{\vartheta})+o_p\left(n^{-1}\right)=O_p(\frac{1}{n}).
\end{aligned}$$

According to the maximum likelihood theory (White, 1996), $\widehat{\theta}-\theta_0=O_p(n^{-1/2})$ under the null hypothesis. Thus, we can show that

$$\begin{aligned}
&(\widehat{\theta}_n-\theta_0)'\left[V_{\theta\theta}(\bar{\vartheta})\right]^{-1}(\widehat{\theta}_n-\theta_0)=(\widehat{\theta}-\theta_0)'\left[-n^{-1}\bar{H}_{n,\theta\theta}^{-1}(\widehat{\vartheta}_n)+o_p(n^{-1})\right]^{-1}(\widehat{\theta}-\theta_0)\\
=\;&\sqrt{n}(\widehat{\theta}-\theta_0)'\left[-\bar{H}_{n,\theta\theta}^{-1}(\widehat{\vartheta})+o_p(1)\right]^{-1}\sqrt{n}(\widehat{\theta}-\theta_0)\\
=\;&\sqrt{n}(\widehat{\theta}_n-\theta_0)'\left[-\bar{H}_{n,\theta\theta}^{-1}(\widehat{\vartheta})\right]^{-1}\sqrt{n}(\widehat{\theta}-\theta_0)+o_p(1)\sqrt{n}(\widehat{\theta}-\theta_0)'\sqrt{n}(\widehat{\theta}-\theta_0)\\
=\;&\sqrt{n}(\widehat{\theta}-\theta_0)'\left[-\bar{H}_{n,\theta\theta}^{-1}(\widehat{\vartheta})\right]^{-1}\sqrt{n}(\widehat{\theta}-\theta_0)+o_p(1)\sqrt{n}O_p(n^{-1/2})\sqrt{n}O_p(n^{-1/2})\\
=\;&\sqrt{n}(\widehat{\theta}-\theta_0)'\left[-\bar{H}_{n,\theta\theta}^{-1}(\widehat{\vartheta})\right]^{-1}\sqrt{n}(\widehat{\theta}-\theta_0)+o_p(1)\\
=\;&\text{Wald}+o_p(1). \qquad (.2.33)
\end{aligned}$$

Furthermore, we can simply derive that

$$\begin{aligned}
& V(\boldsymbol{\vartheta}_0) = E\left[(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_0)(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_0)'|y\right] \\
= \; & E\left[\left(\boldsymbol{\vartheta}-\bar{\boldsymbol{\vartheta}}+\bar{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}_0\right)\left(\boldsymbol{\vartheta}-\bar{\boldsymbol{\vartheta}}+\bar{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}_0\right)'|y\right] \\
= \; & V\left(\bar{\boldsymbol{\vartheta}}\right)+2\left(\bar{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}_0\right)\left(\bar{\boldsymbol{\vartheta}}-\bar{\boldsymbol{\vartheta}}\right)'+\left(\bar{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}_0\right)\left(\bar{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}_0\right)' \\
= \; & V\left(\bar{\boldsymbol{\vartheta}}\right)+\left(\bar{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}_0\right)\left(\bar{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}_0\right)', \qquad (.2.34)
\end{aligned}$$

under the null hypothesis.

Hence, we can further prove that

$$\begin{aligned}
T(y,\theta_0) &= \int_{\Theta_{\boldsymbol{\theta}}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\left[V_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)\right]^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\,d\boldsymbol{\theta} \\
&= tr\left\{\left[V_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)\right]^{-1}E\left[(\boldsymbol{\theta}-\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'|y\right]\right\} \\
&= tr\left\{\left[V_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)\right]^{-1}\left[V_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)+\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\right]\right\} \\
&= q_\theta+tr\left\{\left[V_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)\right]^{-1}\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\right\} \\
&= q_\theta+\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\left[V_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)\right]^{-1}\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \\
&= q_\theta+\left(\bar{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}+\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\left[V_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)\right]^{-1}\left(\bar{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}+\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \\
&= q_\theta+\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\left[V_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)\right]^{-1}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \\
&\quad +2\left(\bar{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)'\left[V_{\theta\theta}(\bar{\boldsymbol{\vartheta}})\right]^{-1}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)+\left(\bar{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)'\left[V_{\theta\theta}(\bar{\boldsymbol{\vartheta}})\right]^{-1}\left(\bar{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right) \\
&= q_\theta+\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\left[V_{\theta\theta}(\bar{\boldsymbol{\vartheta}})\right]^{-1}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)+o_p\left(\frac{1}{\sqrt{n}}\right)O_p(n)\,O_p\left(\frac{1}{\sqrt{n}}\right) \\
&\quad +o_p\left(\frac{1}{\sqrt{n}}\right)O_p(n)\,o_p\left(\frac{1}{\sqrt{n}}\right) \\
&= q_\theta+\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\left[V_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)\right]^{-1}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)+o_p(1) \\
&= q_\theta+\text{Wald}+o_p(1),
\end{aligned}$$

form (.2.33) and (.2.34).

From the above derivation, it is easy to show that

$$T(y,\theta_0)-q_\theta=\text{Wald}+o_p(1)\xrightarrow{d}\chi^2(q),$$

under the null hypothesis.

.2.4 Proof of Theorem 3.3.3

Note that

$$\begin{aligned}
& T(y,r) = \int_{\boldsymbol{\Theta}} \Delta \mathscr{L}(H_0, \vartheta)\, p(\vartheta|y)\, d\vartheta \\
= & \int_{\boldsymbol{\Theta}} (R(\theta)-r)' \left[\frac{\partial R(\overline{\theta})}{\partial \theta'} V_{\theta\theta}(\bar{\boldsymbol{\vartheta}}) \frac{\partial R(\overline{\theta})}{\partial \theta} \right]^{-1} (R(\theta)-r)\, p(\vartheta|y)\, d\vartheta \\
= & \; tr \left\{ \int_{\boldsymbol{\Theta}} [R(\theta)-r]\,[R(\theta)-r]'\, p(\vartheta|y)\, d\vartheta \left[\frac{\partial R(\overline{\theta})}{\partial \theta'} V_{\theta\theta}(\bar{\boldsymbol{\vartheta}}) \frac{\partial R(\overline{\theta})}{\partial \theta} \right]^{-1} \right\} \\
= & \; tr \left\{ E\left[n(R(\theta)-r)(R(\theta)-r)' \middle| y, H_1 \right] \left[\frac{\partial R(\overline{\theta})}{\partial \theta'} n V_{\theta\theta}(\bar{\boldsymbol{\vartheta}}) \frac{\partial R(\overline{\theta})}{\partial \theta} \right]^{-1} \right\}.
\end{aligned}$$

We have

$$\begin{aligned}
& E\left[n(R(\theta)-r)(R(\theta)-r)' \middle| y, H_1 \right] \\
= & \; E\left[n\left(R(\theta)-R\left(\widehat{\theta}\right)+R\left(\widehat{\theta}\right)-r\right)\left(R(\theta)-R\left(\widehat{\theta}\right)+R\left(\widehat{\theta}\right)-r\right)' \middle| y, H_1 \right] \\
= & \; E\left[n\left(R(\theta)-R\left(\widehat{\theta}\right)\right)\left(R(\theta)-R\left(\widehat{\theta}\right)\right)' \middle| y, H_1 \right] \\
& +2E\left[n\left(R(\theta)-R\left(\widehat{\theta}\right)\right)\left(R\left(\widehat{\theta}\right)-\ r\right)' \middle| y, H_1 \right] \\
& +n\left(R\left(\widehat{\theta}\right)-r\right)\left(R\left(\widehat{\theta}\right)-r\right)'. \qquad (.2.35)
\end{aligned}$$

By the Taylor expansion, we can show that

$$\sqrt{n}\left(R(\theta)-R\left(\widehat{\theta}\right)\right) = \frac{\partial R\left(\widehat{\theta}\right)}{\partial \theta} \sqrt{n}\left(\theta-\widehat{\theta}\right) + \left[\sqrt{n}\left(\theta-\widehat{\theta}\right)' \otimes I_m\right] \frac{\partial R^2\left(\widetilde{\theta}\right)}{\partial \theta \partial \theta'} \left(\theta-\widehat{\theta}\right),$$

where $\widetilde{\theta}$ lies between θ and $\widehat{\theta}$. Note that $\frac{\partial^2 R(\theta)}{\partial \theta \partial \theta'}$ is continuous and Θ is compact. Thus, we have

$$\left\| \frac{\partial^2 R\left(\widetilde{\theta}\right)}{\partial \theta \partial \theta'} \right\| \leq M', \qquad (.2.36)$$

for some $0 < M' < \infty$. Furthermore, by Bayesian large-sample theory, $\sqrt{n}(\vartheta - \hat{\vartheta}) = O_p(1)$. Hence, from (.2.36), we can further derive that

$$\begin{aligned}
& \left[\sqrt{n}\left(\theta - \widehat{\theta}\right)' \otimes I_m\right] \frac{\partial R^2\left(\widetilde{\theta}\right)}{\partial\theta\partial\theta'}\left(\theta - \widehat{\theta}\right) \\
= \ & O_p(1)O(1)\frac{1}{\sqrt{n}} = O_p(\frac{1}{\sqrt{n}}) = o_p(1), \qquad (.2.37)
\end{aligned}$$

Since $\int \sqrt{n}\left(\theta - \widehat{\theta}\right) p(\vartheta|y)\,\mathrm{d}\vartheta = \sqrt{n}(\bar{\theta} - \hat{\theta}) = o_p(1)$ according to Lemma 3.2 and also $\sqrt{n}\left(R\left(\widehat{\theta}\right) - R(\theta_0)\right) = O_p(1)$ by using Delta method and the consistency property of MLE, the second term of (.2.35) is

$$\begin{aligned}
& E\left[n\left(R(\theta) - R\left(\widehat{\theta}\right)\right)\left(R\left(\widehat{\theta}\right) - r\right)'\middle| y, H_1\right] \\
= \ & \int_{\mathbf{\Theta}} \sqrt{n}\left[R(\theta) - R\left(\widehat{\theta}\right)\right] p(\vartheta|y)\,d\vartheta \times \sqrt{n}\left(R\left(\widehat{\theta}\right) - R(\theta_0)\right)' \\
= \ & \int_{\mathbf{\Theta}} \sqrt{n}\left[R(\theta) - R\left(\widehat{\theta}\right)\right] p(\vartheta|y)\,d\vartheta \times \sqrt{n}\left(R\left(\widehat{\theta}\right) - R(\theta_0)\right)' \\
= \ & \frac{\partial R\left(\widehat{\theta}\right)}{\partial\theta}\int_{\mathbf{\Theta}} \sqrt{n}\left(\theta - \widehat{\theta}\right) p(\vartheta|y)\,d\vartheta\sqrt{n}\left(R\left(\widehat{\theta}\right) - R(\theta_0)\right) + o_p(1) \\
= \ & O_p(1)o_p(1)\,O_p(1) + o_p(1) = o_p(1).
\end{aligned}$$

For the first term of (.2.35), after integrating out the nuisance parameters, we have,

$$\begin{aligned}
& E\left[n\left(R(\theta) - R\left(\widehat{\theta}\right)\right)\left(R(\theta) - R\left(\widehat{\theta}\right)\right)'\middle| y, H_1\right] \\
= \ & \int_{\mathbf{\Theta}_{\boldsymbol{\theta}}} n\left(R(\theta) - R\left(\widehat{\theta}\right)\right)\left(R(\theta) - R\left(\widehat{\theta}\right)\right)' p(\theta|y)\,d\theta \\
= \ & \frac{\partial R\left(\widehat{\theta}_n\right)}{\partial\theta'}\int_{\mathbf{\Theta}_{\boldsymbol{\theta}}} n\left(\theta - \widehat{\theta}\right)\left(\theta - \widehat{\theta}\right)' p(\theta|y)\,d\theta\frac{\partial R\left(\widehat{\theta}\right)}{\partial\theta} + o_p(1),
\end{aligned}$$

by the Taylor expansion. Based on Lemma 3.3.2,

$$\int_{\mathbf{\Theta}_{\boldsymbol{\theta}}} \left(\theta - \widehat{\theta}\right)\left(\theta - \widehat{\theta}\right)' p(\theta|y)\,d\theta = -\frac{1}{n}\bar{H}^{-1}_{n,\theta\theta}\left(\widehat{\boldsymbol{\vartheta}}\right) + o_p\left(n^{-1}\right).$$

Therefore, we have

$$
\begin{aligned}
& E\left[n\left(R(\theta)-R\left(\widehat{\theta}\right)\right)\left(R(\theta)-R\left(\widehat{\theta}\right)\right)'\middle| y,H_1\right] \\
= & \frac{\partial R\left(\widehat{\theta}\right)}{\partial\theta'}\left[-\bar{H}_{n,\theta\theta}^{-1}\left(\widehat{\boldsymbol{\vartheta}}\right)\right]\frac{\partial R\left(\widehat{\theta}\right)}{\partial\theta}+o_p(1).
\end{aligned}
$$

Since $V_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)=-\frac{1}{n}\bar{H}_{n,\theta\theta}^{-1}\left(\widehat{\boldsymbol{\vartheta}}\right)+o_p\left(n^{-1}\right), \overline{\theta}=\widehat{\theta}+o_p\left(n^{-1/2}\right)=\theta_0+O_p\left(n^{-1/2}\right)$, by (.2.37), we have

$$
\begin{aligned}
& tr\left\{E\left[n\left(R(\theta)-R\left(\widehat{\theta}\right)\right)\left(R(\theta)-R\left(\widehat{\theta}\right)\right)'\middle| y,H_1\right]\left[\frac{\partial R\left(\overline{\theta}\right)}{\partial\theta'}nV_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)\frac{\partial R\left(\overline{\theta}\right)}{\partial\theta}\right]^{-1}\right\} \\
= & \frac{\partial R\left(\widehat{\theta}\right)}{\partial\theta'}\left[-\bar{H}_{n,\theta\theta}^{-1}\left(\widehat{\boldsymbol{\vartheta}}\right)\right]\frac{\partial R\left(\widehat{\theta}\right)}{\partial\theta}\left[\frac{\partial R\left(\overline{\theta}\right)}{\partial\theta'}nV_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)\frac{\partial R\left(\overline{\theta}\right)}{\partial\theta}\right]^{-1}+o_p(1) \\
= & \frac{\partial R\left(\widehat{\theta}\right)}{\partial\theta'}\left[-\bar{H}_{n,\theta\theta}^{-1}\left(\widehat{\boldsymbol{\vartheta}}\right)\right]\frac{\partial R\left(\widehat{\theta}\right)}{\partial\theta}\left[\frac{\partial R\left(\overline{\theta}\right)}{\partial\theta'}nV_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)\frac{\partial R\left(\overline{\theta}\right)}{\partial\theta}\right]^{-1}+o_p(1) \\
\xrightarrow{p} & m.
\end{aligned}
$$

Finally, the third term of (.2.35) can be expressed as

$$
\begin{aligned}
& tr\left\{n\left(R\left(\widehat{\theta}\right)-r\right)\left(R\left(\widehat{\theta}\right)-r\right)'\left[\frac{\partial R\left(\overline{\theta}\right)}{\partial\theta'}nV_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}\right)\frac{\partial R\left(\overline{\theta}\right)}{\partial\theta}\right]^{-1}\right\}+o_p(1) \\
= & \left[R\left(\widehat{\theta}\right)-r\right]'\left\{\frac{\partial R\left(\widehat{\theta}\right)}{\partial\theta'}\left[-\bar{H}_{n,\theta\theta}^{-1}\left(\widehat{\boldsymbol{\vartheta}}\right)\right]\frac{\partial R\left(\widehat{\theta}\right)}{\partial\theta}\right\}^{-1}\left[R\left(\widehat{\theta}\right)-r\right]+o_p(1) \\
= & \text{Wald}+o_p(1).
\end{aligned}
$$

Therefore, under the null hypotheses, we have

$$
T(y,r)-m=\text{Wald}+o_p(1)\xrightarrow{d}\chi^2(m).
$$

.2.5 Proof of Theorem 3.3.4

Let $\{\vartheta^{[j]}, j = 1, 2, \cdots, J\}$ be the efficient random draws from $p(\vartheta|y)$. Then, we have

$$\bar{V}_2 = \frac{1}{J}\sum_{j=1}^{J}\left(\theta^{[j]} - \bar{\bar{\theta}}\right)\left(\theta^{[j]} - \bar{\bar{\theta}}\right)' = \frac{1}{J}\sum_{j=1}^{J} V_2^{[j]} = \bar{V}_{\theta\theta}(\bar{\bar{\vartheta}}),$$

$$\bar{v}_1 = \bar{\bar{\theta}} = \frac{1}{J}\sum_{j=1}^{J}\theta^{[j]},$$

Hence, $\widehat{T}(y, \theta_0)$ in (3.3.13) can be rewritten as

$$\begin{aligned}
\widehat{T}(y, \theta_0) &= tr\left[\left(\bar{V}_{\theta\theta}(\bar{\bar{\vartheta}})\right)^{-1}\bar{V}_{\theta}(\theta_0)\right] \\
&= tr\left\{\left[\bar{V}_{\theta\theta}(\bar{\bar{\vartheta}})\right]^{-1}\left[\frac{1}{J}\sum_{j=1}^{J}\left(\theta^{[j]} - \theta_0\right)\left(\theta^{[j]} - \theta_0\right)\right]\right\} \\
&= tr\left\{\left[\bar{V}_{\theta\theta}(\bar{\bar{\vartheta}})\right]^{-1}\left[\bar{V}_{\theta\theta}\left(\bar{\bar{\vartheta}}\right) + \left(\bar{\bar{\theta}} - \theta_0\right)\left(\bar{\bar{\theta}} - \theta_0\right)'\right]\right\} \\
&= q_{\theta} + tr\left[(\bar{v}_1 - \theta_0)(\bar{v}_1 - \theta_0)'\bar{V}_2^{-1}\right],
\end{aligned}$$

which is a consistent estimator of $T(y, \theta_0)$.

Following the notations of Magnus and Neudecker (2002) about matrix derivatives, let

$$v_2^{(j)} = vech\left(V_2^{[j]}\right),\ v_1^{[j]} = \theta^{[j]},$$

$$\bar{v}_2 = vech(\bar{V}_2),\ \bar{v}_1 = \bar{\bar{\theta}},\ \bar{v} = \left(\bar{v}_1', \bar{v}_2'\right)'.$$

Note that the dimension of $\bar{v}_2$ is $q^* \times 1, q^* = q_{\theta}(q_{\theta} + 1)/2$. Hence, we have

$$\begin{aligned}
\frac{\partial \widehat{T}(y, \theta_0)}{\partial \bar{v}} =& vec\left(I_{q_\theta}\right)'\left\{\left[\left((\bar{v}_1 - \theta_0)'\bar{V}_2^{-1}\right)' \otimes I_{q_\theta}\right]\frac{\partial \bar{v}_1}{\partial \bar{v}} + \left[\bar{V}_2^{-1} \otimes (\bar{v}_1 - \theta_0)\right]\frac{\partial \bar{v}_1'}{\partial \bar{v}}\right. \\
&\left. - \left[I_{q_\theta} \otimes (\bar{v}_1 - \theta_0)(\bar{v}_1 - \theta_0)'\right]\left(\bar{V}_2^{-1} \otimes \bar{V}_2^{-1}\right)\frac{\partial vec(\bar{V}_2)}{\partial \bar{v}}\right\} \\
=& vec\left(I_{q_\theta}\right)'\left[\left(\left((\bar{v}_1 - \theta_0)'\bar{V}_2^{-1}\right)' \otimes I_{q_\theta} + \bar{V}_2^{-1} \otimes (\bar{v}_1 - \theta_0)\right)\frac{\partial \bar{v}_1}{\partial \bar{v}}\right. \\
&\left. - \left[I_{q_\theta} \otimes (\bar{v}_1 - \theta_0)(\bar{v}_1 - \theta_0)'\right]\left(\bar{V}_2^{-1} \otimes \bar{V}_2^{-1}\right)\frac{\partial \bar{V}_2}{\partial \bar{v}}\right].
\end{aligned}$$

where

$$\frac{\partial\bar{v}_1}{\partial\bar{v}}=\frac{\partial\bar{v}_1'}{\partial\bar{v}}=\left[I_{q_\theta},0_{q_\theta\times q^*}\right],\frac{\partial\bar{V}_2}{\partial\bar{v}}=\left[0_{q_\theta^2\times q_\theta},\left(\frac{\partial vec\left(\bar{V}_2\right)}{\partial\bar{v}_2}\right)_{q_\theta^2\times q^*}\right].$$

By the Delta method,

$$Var\left(\widehat{T}\left(y,\theta_0\right)\right)=\frac{\partial\widehat{T}\left(y,\theta_0\right)}{\partial\bar{v}}Var\left(\bar{v}\right)\left(\frac{\partial\widehat{T}\left(y,\theta_0\right)}{\partial\bar{v}}\right)'.$$

The expression of the NSE for $\widehat{T}\left(y,r\right)$ can also be obtained in the similar way.

$$\begin{aligned}\widehat{T}\left(y,r\right)=&tr\left[\left(\frac{\partial R\left(\bar{\bar{\theta}}\right)}{\partial\theta'}\bar{V}_{\theta\theta}\left(\bar{\bar{\vartheta}}\right)\frac{\partial R\left(\bar{\bar{\vartheta}}\right)}{\partial\theta}\right)^{-1}\bar{V}_\theta(r)\right]\\&=m+tr\left\{\left(R\left(\bar{\bar{\theta}}\right)-r\right)\left(R\left(\bar{\bar{\theta}}\right)-r\right)'\left(\frac{\partial R\left(\bar{\bar{\theta}}\right)}{\partial\theta'}\bar{V}_{\theta\theta}\left(\bar{\bar{\vartheta}}\right)\frac{\partial R\left(\bar{\bar{\vartheta}}\right)}{\partial\theta}\right)^{-1}\right\}\\&=m+tr\left[\left(\bar{v}_3-r\right)\left(\bar{v}_3-r\right)'\left(\bar{V}_4'\bar{V}_2\bar{V}_4\right)^{-1}\right],\end{aligned}$$

where similarly,

$$\bar{v}_3=R\left(\frac{1}{J}\sum_{j=1}^{J}\theta^{[j]}\right)=R\left(\bar{v}_1\right),\ \bar{V}_4=\frac{\partial R\left(\frac{1}{J}\sum_{j=1}^{J}\theta^{[j]}\right)}{\partial\theta}=\frac{\partial R\left(\bar{v}_1\right)}{\partial\theta},\ \bar{v}=\left(\bar{v}_1',\bar{v}_2'\right)'.$$

So that,

$$\begin{aligned}\frac{\partial\widehat{T}\left(y,r\right)}{\partial\bar{v}}=&vec\left(I_m\right)'\left\{\left[\left(\left(\bar{v}_3-r\right)'\left(\bar{V}_4'\bar{V}_2\bar{V}_4\right)^{-1}\right)'\otimes I_m\right]\frac{\partial\bar{v}_3}{\partial\bar{v}_1}\frac{\partial\bar{v}_1}{\partial\bar{v}}\right.\\&+\left[\left(\bar{V}_4'\bar{V}_2\bar{V}_4\right)^{-1}\otimes\left(\bar{v}_3-r\right)\right]\frac{\partial\bar{v}_3'}{\partial\bar{v}_1}\frac{\partial\bar{v}_1}{\partial\bar{v}}\\&+\left[I_m\otimes\left(\bar{v}_3-r\right)\left(\bar{v}_3-r\right)'\right]\left[\left(\bar{V}_4'\bar{V}_2\bar{V}_4\right)^{-1}\otimes\left(\bar{V}_4'\bar{V}_2\bar{V}_4\right)^{-1}\right]\\&\left.\times\frac{\partial vec\left(\bar{V}_4'\bar{V}_2\bar{V}_4\right)}{\partial\bar{v}}\right\},\end{aligned}$$

where

$$\frac{\partial vec\left(\bar{V}_4'\bar{V}_2\bar{V}_4\right)}{\partial\bar{v}} = \left((\bar{V}_2\bar{V}_4)'\otimes I_m\right)\frac{\partial\bar{V}_4'}{\partial\bar{v}_1}\frac{\partial\bar{v}_1}{\partial\bar{v}} + \left(\bar{V}_4\otimes\bar{V}_4'\right)\frac{\partial\bar{V}_2}{\partial\bar{v}}$$
$$+\left(I_m\otimes\bar{V}_4'\bar{V}_2\right)\frac{\partial\bar{V}_4}{\partial\bar{v}_1}\frac{\partial\bar{v}_1}{\partial\bar{v}},$$

where the derivatives of $\bar{V}_4$ and $\bar{v}_3$ depend on the form of the function $R(\theta)$. By the Delta method, we have

$$Var\left(\widehat{T}(y,r)\right) = \frac{\partial\widehat{T}(y,r)}{\partial\bar{v}}Var(\bar{v})\left(\frac{\partial\widehat{T}(y,r)}{\partial\bar{v}}\right)'.$$

.3 Proofs in Chapter 4

.3.1 The Proof of Lemma 4.4.1

As in (4.4.4),

$$V_N(\theta;\widehat{\chi}) = \zeta_N\sum_{i=1}^{N^{obs}}\lambda_N\tilde{g}_i(\theta;\widehat{\chi})\tilde{g}_i(\theta;\widehat{\chi})'\lambda_N'\zeta_N'$$
$$+\frac{N}{J}\lambda_N\bar{g}_{N,\chi}(\theta;\widehat{\chi})\widehat{\Sigma}_\chi\bar{g}_{N,\chi}(\theta;\widehat{\chi})'\lambda_N',$$

For the first term, by Assumption 2 and Assumption 7, as $N\to\infty$, $\widehat{\chi}\to\chi_0$. And in the framework of the structural model, $\{\tilde{g}_i(\theta;\widehat{\chi})\}_{i=1}^{N^{obs}}$ are independent across i. Combined with Assumption 1, 5 and 6, we have

$$\zeta_N\sum_{i=1}^{N^{obs}}\lambda_N\tilde{g}_i(\theta;\widehat{\chi})\tilde{g}_i(\theta;\widehat{\chi})'\lambda_N\zeta_N \xrightarrow{p} E\left[\lambda g_i(\theta;\chi_0)g_i(\theta;\chi_0)'\lambda'\right],$$

where

$$g_i(\theta;\chi) = \underbrace{\left(g_{i,t_{min}}(\theta;\chi),\ldots,g_{i,T_r}(\theta;\chi)\right)'}_{T_m \text{ elements}}.$$

Similarly, by Assumption 2 and 7, as $N \to \infty$, $\widehat{\Sigma}_\chi \xrightarrow{p} \Sigma_\chi$, $\widehat{\chi} \to \chi_0$. Combined with Assumption 1, 8 and 9, we can have

$$\frac{N}{J}\lambda_N \bar{g}_{N,\chi}(\theta;\widehat{\chi})\widehat{\Sigma}_\chi \bar{g}_{N,\chi}(\theta;\widehat{\chi})'\lambda_N', \xrightarrow{p} \gamma\lambda E\left[g_{i,\chi}(\theta;\chi_0)\right]\Sigma_\chi E\left[g_{i,\chi}(\theta;\chi_0)'\right]\lambda'.$$

.3.2 The Proof of Theorem 4.4.1

We define

$$M(\theta) = -\frac{1}{2}E\left[g_i(\theta;\chi_0)\right]'\lambda' W(\theta)\lambda E\left[g_i(\theta;\chi_0)\right],$$

where $W(\theta) = V^{-1}(\theta) = \left\{\lambda E\left[g_i(\theta;\chi_0)g_i(\theta;\chi_0)'\right]\lambda + \gamma\lambda E\left[g_{i,\chi}(\theta;\chi_0)\right]\Sigma_\chi E\left[g_{i,\chi}(\theta;\chi_0)'\right]\lambda'\right\}^{-1}$, where $V(\theta)$ defined in Lemma 4.4.1. From the definition of criterion function (4.4.3), under Assumption 1- 10 , we have

$$\frac{1}{N}L_N(\theta) = -\frac{1}{2}\bar{g}(\theta;\widehat{\chi})'\lambda_N' V_N^{-1}(\theta;\widehat{\chi})\lambda_N \bar{g}(\theta;\widehat{\chi}) \xrightarrow{p} M(\theta).$$

Further, in the framework, we implies that the matrix $V_N(\theta;\widehat{\chi})$ and $V(\theta)$ are positive definite for all $\theta \in \Theta$. Thus, the as $W_N(\theta;\widehat{\chi}) = V_N^{-1}(\theta;\widehat{\chi})$ and $W(\theta) = V^{-1}(\theta)$. Due to $W(\theta) > 0$ and $M(\theta_0) = 0$, by Assumption 3, for any $\delta > 0$, $\theta \in \{\theta : \|\theta - \theta_0\| \geq \delta\} \subset \Theta$, we have $M(\theta) < 0$, so that $M(\theta) - M(\theta_0) < 0$. Therefore, the Lemma 1 in Chernozukov and Hong (2003) is satisfied.

Since $\{g_{i,t}(\theta;\chi)\}$ are independent across i, we have

$$\sqrt{N}\lambda_N \bar{g}_N(\theta_0;\chi_0) \xrightarrow{d} N\left(0, \lambda\Sigma_g\lambda'\right),$$

where $\Sigma_g = E\left[g_i(\theta_0;\chi_0)g_i(\theta_0;\chi_0)'\right]$. If we use the GMM method to estimate the parameter χ_0, for $\sqrt{N}\lambda_N \bar{g}(\theta_0;\widehat{\chi})$, expanding it around χ_0,

$$
\begin{aligned}
\sqrt{N}\lambda_N \bar{g}_N(\theta_0;\widehat{\chi}) =& \sqrt{N}\lambda_N\left[\bar{g}_N(\theta_0;\chi_0)+\bar{g}_\chi(\theta_0;\tilde{\chi})'(\widehat{\chi}-\chi_0)+o_p\left(\frac{1}{\sqrt{J}}\right)\right]\\
=& \sqrt{N}\lambda_N\bar{g}_N(\theta_0;\chi_0)+\sqrt{\frac{N}{J}}\lambda_N\bar{g}_{N,\chi}(\theta_0;\tilde{\chi})'\sqrt{J}(\widehat{\chi}-\chi_0)+o_p\left(\sqrt{\frac{N}{J}}\right).
\end{aligned}
$$

By Assumption 2, from the first-stage estimation,

$$
\sqrt{J}(\widehat{\chi}-\chi_0)\xrightarrow{d}N\left(0,\Sigma_\chi\right).
$$

Following GP, since the first-stage estimator is obtained conditional on exogenous structural models and mostly different data, then we can have

$$
\sqrt{N}\lambda_N\bar{g}_N(\theta_0;\widehat{\chi})\xrightarrow{d}N\left(0,\lambda\Sigma_g\lambda'+\gamma\lambda G_\chi'\Sigma_\chi G_\chi\lambda'\right), \tag{.3.1}
$$

where $G_\chi = E\left[\nabla_\chi g_i(\theta_0;\chi_0)\right]$, $\gamma = \lim_{N\to\infty}\frac{N}{J}$, $\lambda = \lim_{N\to\infty}\lambda_N$, $\Sigma_g = E\left[g_i(\theta_0;\chi_0)\,g_i(\theta_0;\chi_0)'\right]$.

We can rewrite the criterion function as

$$
\begin{aligned}
L_N(\theta) =& -\frac{N}{2}\left[\lambda_N\bar{g}_N(\theta;\widehat{\chi})\right]'W_N(\theta;\widehat{\chi})\lambda_N\bar{g}_N(\theta;\widehat{\chi})\\
=& -\frac{N}{2}\left[\lambda_N\bar{g}_N(\theta;\widehat{\chi})\right]'\left[\zeta_N\sum_{i=1}^{N^{obs}}\lambda_N\tilde{g}_i(\theta;\widehat{\chi})\,\tilde{g}_i(\theta;\widehat{\chi})'\lambda_N'\zeta_N'\right.\\
&\left.+\frac{N}{J}\lambda_N\bar{g}_{N,\chi}(\theta;\widehat{\chi})\widehat{\Sigma}_\chi\bar{g}_{N,\chi}(\theta;\widehat{\chi})'\lambda_N'\right]^{-1}\lambda_N\bar{g}_N(\theta;\widehat{\chi})\\
=& -\frac{N}{2}tr\left\{\lambda_N\bar{g}_N(\theta;\widehat{\chi})\,\bar{g}_N(\theta;\widehat{\chi})'\lambda_N'\left[\zeta_N\sum_{i=1}^{N^{obs}}\lambda_N\tilde{g}_i(\theta;\widehat{\chi})\,\tilde{g}_i(\theta;\widehat{\chi})'\lambda_N'\zeta_N'\right.\right.\\
&\left.\left.\times+\frac{N}{J}\lambda_N\bar{g}_{N,\chi}(\theta;\widehat{\chi})\widehat{\Sigma}_\chi\bar{g}_{N,\chi}(\theta;\widehat{\chi})'\lambda_N'\right]^{-1}\right\}\\
=& -\frac{N}{2}tr\left[C(\theta)D^{-1}(\theta)\right],
\end{aligned}
$$

where $C(\theta)$ and $D(\theta)$ are symmetric. Then following Magnus and Neudecker

(1995), we have

$$
\begin{aligned}
d\left\{tr\left[C(\theta)D^{-1}(\theta)\right]\right\} =& tr\left\{dC(\theta)D^{-1}(\theta)+C(\theta)dD^{-1}(\theta)\right\}\\
=& tr\left\{D^{-1}(\theta)dC(\theta)+C(\theta)D^{-1}(\theta)dD(\theta)D^{-1}(\theta)\right\}\\
=& tr\left\{D^{-1}(\theta)dC(\theta)-D^{-1}(\theta)C(\theta)D^{-1}(\theta)dD(\theta)\right\}.
\end{aligned}
$$

Before we derive the first-order and second-order differentiation of $L_N(\theta)$, we consider the following formula,

$$
\begin{aligned}
& tr\{K_1(\theta)dD(\theta)K_2(\theta)\}\\
=& tr\left\{K_1(\theta)\zeta_N\lambda_N d\left[\sum_{i=1}^{N^{obs}}\tilde{g}_i(\theta;\widehat{\chi})\tilde{g}_i(\theta;\widehat{\chi})'\right]\lambda_N'\zeta_N'K_2(\theta)\right\}\\
& +\frac{N}{J}tr\left\{K_1(\theta)\lambda_N d\left[\bar{g}_{N,\chi}(\theta;\widehat{\chi})\widehat{\Sigma}_\chi\bar{g}_{N,\chi}(\theta;\widehat{\chi})'\right]\lambda_N'K_2(\theta)\right\}\\
=& \sum_{i=1}^{N^{obs}}tr\left\{K_1(\theta)\zeta_N\lambda_N\left[\nabla_\theta\tilde{g}_i(\theta;\widehat{\chi})d\theta\tilde{g}_i(\theta;\widehat{\chi})'+\tilde{g}_i(\theta;\widehat{\chi})d\theta'\nabla_\theta\tilde{g}_i(\theta;\widehat{\chi})'\right]\lambda_N'\zeta_N'K_2(\theta)\right\}+\\
& \frac{N}{J}tr\left\{K_1(\theta)\lambda_N\left[\nabla_\theta\bar{g}_{N,\chi}(\theta;\widehat{\chi})d\theta\widehat{\Sigma}_\chi\bar{g}_{N,\chi}(\theta;\widehat{\chi})'+\bar{g}_{N,\chi}(\theta;\widehat{\chi})\widehat{\Sigma}_\chi d\theta'\nabla_\theta\bar{g}_{N,\chi}(\theta;\widehat{\chi})'\right]\lambda_N'K_2(\theta)\right\}\\
=& \sum_{i=1}^{N^{obs}}tr\left\{\tilde{g}_i(\theta;\widehat{\chi})'\lambda_N'\zeta_N'K_2(\theta)K_1(\theta)\zeta_N\lambda_N\nabla_\theta\tilde{g}_i(\theta;\widehat{\chi})d\theta\right\}+\\
& \sum_{i=1}^{N^{obs}}tr\left\{\tilde{g}_i(\theta;\widehat{\chi})'\lambda_N'\zeta_N'K_1(\theta)'K_2(\theta)'\zeta_N\lambda_N\nabla_\theta\tilde{g}_i(\theta;\widehat{\chi})d\theta\right\}+\\
& \frac{N}{J}tr\left\{\widehat{\Sigma}_\chi\bar{g}_{N,\chi}(\theta;\widehat{\chi})'\lambda_N'K_2(\theta)K_1(\theta)\lambda_N\nabla_\theta\bar{g}_{N,\chi}(\theta;\widehat{\chi})d\theta\right\}+\\
& \frac{N}{J}tr\left\{\widehat{\Sigma}_\chi\bar{g}_{N,\chi}(\theta;\widehat{\chi})'\lambda_N'K_1(\theta)'K_2(\theta)'\lambda_N\nabla_\theta\bar{g}_{N,\chi}(\theta;\widehat{\chi})d\theta\right\}. \qquad (.3.2)
\end{aligned}
$$

Then, for the first term $tr\left[D^{-1}(\theta)\,dC(\theta)\right]$,

$$
\begin{aligned}
& tr\left[D^{-1}(\theta)\,dC(\theta)\right] \\
=& tr\left\{D^{-1}(\theta)\,\lambda_N d\left[\bar{g}_N(\theta;\widehat{\chi})\right]\bar{g}_N(\theta;\widehat{\chi})'\lambda_N' + D^{-1}(\theta)\,\lambda_N\bar{g}_N(\theta;\widehat{\chi})\left[d\bar{g}_N(\theta;\widehat{\chi})\right]'\lambda_N'\right\} \\
=& tr\left\{D^{-1}(\theta)\,\lambda_N\nabla_\theta\bar{g}_N(\theta;\widehat{\chi})\,d\theta\,\bar{g}_N(\theta;\widehat{\chi})'\lambda_N' + D^{-1}(\theta)\,\lambda_N\bar{g}_N(\theta;\widehat{\chi})\left[\nabla_\theta\bar{g}_N(\theta;\widehat{\chi})\,d\theta\right]'\lambda_N'\right\} \\
=& tr\left\{\bar{g}_N(\theta;\widehat{\chi})'\lambda_N'D^{-1}(\theta)\,\lambda_N\nabla_\theta\bar{g}_N(\theta;\widehat{\chi})\,d\theta + \lambda_N\nabla_\theta\bar{g}_N(\theta;\widehat{\chi})\,d\theta\,\bar{g}_N(\theta;\widehat{\chi})'\lambda_N'D^{-1}(\theta)\right\} \\
=& 2tr\left\{\bar{g}_N(\theta;\widehat{\chi})'\lambda_N'D^{-1}(\theta)\,\lambda_N\nabla_\theta\bar{g}_N(\theta;\widehat{\chi})\,d\theta\right\} \\
=& 2tr\left\{\bar{g}_N(\theta;\widehat{\chi})'\lambda_N'W_N(\theta;\widehat{\chi})\,\lambda_N\nabla_\theta\bar{g}_N(\theta;\widehat{\chi})\,d\theta\right\}.
\end{aligned}
$$

By formula (.3.2),

$$
\begin{aligned}
& tr\left\{D^{-1}(\theta)\,C(\theta)\,D^{-1}(\theta)\,dD(\theta)\right\} \\
=& 2\sum_{i=1}^{N^{obs}} tr\left\{\tilde{g}_i(\theta;\widehat{\chi})'\lambda_N'\zeta_N'D^{-1}(\theta)\,C(\theta)\,D^{-1}(\theta)\,\zeta_N\lambda_N\nabla_\theta\tilde{g}_i(\theta;\widehat{\chi})\,d\theta\right\} + \\
& \frac{2N}{J}tr\left\{\widehat{\Sigma}_\chi\bar{g}_{N,\chi}(\theta;\widehat{\chi})'\lambda_N'D^{-1}(\theta)\,C(\theta)\,D^{-1}(\theta)\,\lambda_N\nabla_\theta\bar{g}_{N,\chi}(\theta;\widehat{\chi})\,d\theta\right\} \\
=& 2\sum_{i=1}^{N^{obs}} tr\left\{\tilde{g}_i(\theta;\widehat{\chi})'\lambda_N'\zeta_N'W_N(\theta;\widehat{\chi})\,\lambda_N\bar{g}_N(\theta;\widehat{\chi})\,\bar{g}_N(\theta;\widehat{\chi})'\lambda_N'W_N(\theta;\widehat{\chi})\,\zeta_N\lambda_N\nabla_\theta\tilde{g}_i(\theta;\widehat{\chi})\,d\theta\right\} + \\
& \frac{2N}{J}tr\left\{\widehat{\Sigma}_\chi\bar{g}_{N,\chi}(\theta;\widehat{\chi})'\lambda_N'W_N(\theta;\widehat{\chi})\,\lambda_N\bar{g}_N(\theta;\widehat{\chi})\,\bar{g}_N(\theta;\widehat{\chi})'\lambda_N'W_N(\theta;\widehat{\chi})\,\lambda_N\nabla_\theta\bar{g}_{N,\chi}(\theta;\widehat{\chi})\,d\theta\right\}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& dL_N(\theta) \\
=& -Ntr\left\{\bar{g}_N(\theta;\widehat{\chi})'\lambda_N'W_N(\theta;\widehat{\chi})\,\lambda_N\nabla_\theta\bar{g}_N(\theta;\widehat{\chi})\,d\theta\right\} + \\
& N\sum_{i=1}^{N^{obs}} tr\left\{\tilde{g}_i(\theta;\widehat{\chi})'\lambda_N'\zeta_N'W_N(\theta;\widehat{\chi})\,\lambda_N\bar{g}_N(\theta;\widehat{\chi})\,\bar{g}_N(\theta;\widehat{\chi})'\lambda_N'W_N(\theta;\widehat{\chi})\,\zeta_N\lambda_N\nabla_\theta\tilde{g}_i(\theta;\widehat{\chi})\,d\theta\right\} + \\
& \frac{N^2}{J}tr\left\{\widehat{\Sigma}_\chi\bar{g}_{N,\chi}(\theta;\widehat{\chi})'\lambda_N'W_N(\theta;\widehat{\chi})\,\lambda_N\bar{g}_N(\theta;\widehat{\chi})\,\bar{g}_N(\theta;\widehat{\chi})'\lambda_N'W_N(\theta;\widehat{\chi})\,\lambda_N\nabla_\theta\bar{g}_{N,\chi}(\theta;\widehat{\chi})\,d\theta\right\},
\end{aligned}
$$

which implies,

$$
\begin{aligned}
&\nabla_\theta L_N(\theta)\\
=&-N\nabla_\theta \bar{g}_N(\theta;\widehat{\chi})'\lambda_N' W_N(\theta;\widehat{\chi})\lambda_N\bar{g}_N(\theta;\widehat{\chi})+\\
&N\sum_{i=1}^{N^{obs}}\nabla_\theta \tilde{g}_i(\theta;\widehat{\chi})'\lambda_N'\zeta_N' W_N(\theta;\widehat{\chi})\lambda_N\bar{g}_N(\theta;\widehat{\chi})\bar{g}_N(\theta;\widehat{\chi})'\lambda_N' W_N(\theta;\widehat{\chi})\zeta_N\lambda_N\tilde{g}_i(\theta;\widehat{\chi})+\\
&\frac{N^2}{J}\nabla_\theta \bar{g}_{N,\chi}(\theta;\widehat{\chi})'\lambda_N' W_N(\theta;\widehat{\chi})\lambda_N\bar{g}_N(\theta;\widehat{\chi})\bar{g}_N(\theta;\widehat{\chi})'\lambda_N' W_N(\theta;\widehat{\chi})\lambda_N\bar{g}_{N,\chi}(\theta;\widehat{\chi})\widehat{\Sigma}_\chi.
\end{aligned}
$$

By (.3.1),

$$
\bar{g}_N(\theta_0;\widehat{\chi})=o_p\left(\frac{1}{\sqrt{n}}\right), W_N(\theta_0;\widehat{\chi})=O_p(1), \tag{.3.3}
$$

it is obvious that

$$
\begin{aligned}
&N\sum_{i=1}^{N^{obs}}\nabla_\theta \tilde{g}_i(\theta_0;\widehat{\chi})'\lambda_N'\zeta_N' W_N(\theta_0;\widehat{\chi})\lambda_N\bar{g}_N(\theta_0;\widehat{\chi})\bar{g}_N(\theta_0;\widehat{\chi})'\lambda_N' W_N(\theta_0;\widehat{\chi})\zeta_N\lambda_N\tilde{g}_i(\theta_0;\widehat{\chi})\\
=&N^2 O_p\left(\frac{1}{\sqrt{N}}\right)o_p\left(\frac{1}{\sqrt{N}}\right)o_p\left(\frac{1}{\sqrt{N}}\right)O_p\left(\frac{1}{\sqrt{N}}\right)\\
=&o_p(1).
\end{aligned}
$$

$$
\begin{aligned}
&\frac{N^2}{J}\nabla_\theta \bar{g}_{N,\chi}(\theta_0;\widehat{\chi})'\lambda_N' W_N(\theta_0;\widehat{\chi})\lambda_N\bar{g}_N(\theta_0;\widehat{\chi})\bar{g}_N(\theta_0;\widehat{\chi})'\lambda_N' W_N(\theta_0;\widehat{\chi})\lambda_N\bar{g}_{N,\chi}(\theta_0;\widehat{\chi})\widehat{\Sigma}_\chi\\
=&NO_p(1)o_p\left(\frac{1}{\sqrt{N}}\right)o_p\left(\frac{1}{\sqrt{N}}\right)O_p(1)=o_p(1).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\frac{\nabla_\theta L_N(\theta_0)}{\sqrt{N}}=&-\nabla_\theta \bar{g}_N(\theta_0;\widehat{\chi})'\lambda_N' W_N(\theta_0;\widehat{\chi})\sqrt{N}\lambda_N\bar{g}_N(\theta_0;\widehat{\chi})+o_p(1)\\
&\xrightarrow{d}N\left(0,G_\theta'\lambda' V^{-1}(\theta_0)\lambda G_\theta'\right),
\end{aligned}
$$

where $V^{-1}(\theta_0)=\left(\lambda\Sigma_g\lambda'+\gamma\lambda G_\chi'\Sigma_\chi G_\chi\lambda'\right)^{-1}$ and $G_\theta=\nabla_\theta E\left[g_{i,t}(\theta_0;\chi_0)\right]$. This is because from (.3.1),

$$
\sqrt{N}\lambda_N\bar{g}(\theta_0;\widehat{\chi})\xrightarrow{d}N\left(0,\lambda\Sigma_g\lambda'+\gamma\lambda G_\chi'\Sigma_\chi G_\chi\lambda'\right),
$$

where $G_\chi = E\left[\nabla_\chi \bar{g}(\theta_0;\chi_0)\right]$, $\gamma = \lim_{N\to\infty}\frac{N}{J}$, $\lambda = \lim_{N\to\infty}\lambda_N$, $\Sigma_g = E\left[g_i(\theta_0;\chi_0)\, g_i(\theta_0;\chi_0)'\right]$ and

$$\nabla_\theta \bar{g}_N(\theta_0;\widehat{\chi})' \xrightarrow{p} \nabla_\theta E\left[g_{i,t}(\theta_0;\chi_0)\right] = G_\theta,$$

$$\begin{aligned}
W_N(\theta_0;\widehat{\chi}) &\xrightarrow{p} V^{-1}(\theta_0) \\
&= \left\{\lambda E\left[g_i(\theta_0;\chi_0)\, g_i(\theta_0;\chi_0)'\right]\lambda + \gamma\lambda E\left[g_{i,\chi}(\theta_0;\chi_0)\right]\Sigma_\chi E\left[g_{i,\chi}(\theta_0;\chi_0)'\right]\lambda'\right\}^{-1} \\
&= \left(\lambda\Sigma_g\lambda' + \gamma\lambda G_\chi'\Sigma_\chi G_\chi\lambda'\right)^{-1}.
\end{aligned}$$

Now turn to the second derivative of the criterion function, which is the Hessian matrix of $L_n(\theta)$. The second order differentiation,

$$\begin{aligned}
&d^2\left\{tr\left[A(\theta)B^{-1}(\theta)\right]\right\} \\
=&d\left\{-tr\left\{\bar{g}_N(\theta;\widehat{\chi})'\lambda_N' W_N(\theta;\widehat{\chi})\lambda_N\nabla_\theta\bar{g}_N(\theta;\widehat{\chi})\,d\theta\right\}+\right. \\
&\sum_{i=1}^{N^{obs}} tr\left\{\tilde{g}_i(\theta;\widehat{\chi})'\lambda_N'\zeta_N' W_N(\theta;\widehat{\chi})\lambda_N\bar{g}_N(\theta;\widehat{\chi})\,\bar{g}_N(\theta;\widehat{\chi})'\lambda_N' W_N(\theta;\widehat{\chi})\,\zeta_N\lambda_N\nabla_\theta\tilde{g}_i(\theta;\widehat{\chi})\,d\theta\right\}+ \\
&\left.\frac{N^2}{J}tr\left\{\widehat{\Sigma}_\chi\bar{g}_{N,\chi}(\theta;\widehat{\chi})'\lambda_N' W_N(\theta;\widehat{\chi})\lambda_N\bar{g}_N(\theta;\widehat{\chi})\,\bar{g}_N(\theta;\widehat{\chi})'\lambda_N' W_N(\theta;\widehat{\chi})\lambda_N\nabla_\theta\bar{g}_{N,\chi}(\theta;\widehat{\chi})\,d\theta\right\}\right\}.
\end{aligned}$$

Following the preceding procedure to derive the first-order differentiation,, we can obtain the form of $\nabla_{\theta\theta'}L_n(\theta)$. Due to Assumptions 5-9, for any $\delta > 0$, $\nabla_{\theta\theta'}L_n(\theta)$ is continuous when $\|\theta-\theta_0\| \le \delta$ and we can have

$$\frac{\nabla_{\theta\theta'}L_N(\theta_0)}{N} = -\nabla_\theta\bar{g}_N(\theta_0;\widehat{\chi})'\lambda_N' W_N(\theta_0;\widehat{\chi})\lambda_N\nabla_\theta\bar{g}_N(\theta_0;\widehat{\chi}) + o_p(1).$$

Meanwhile, we have

$$M(\theta) = -\frac{1}{2}E\left[g_i(\theta;\chi_0)\right]'\lambda' W(\theta)\lambda E\left[g_i(\theta;\chi_0)\right],$$

where $W(\theta) = V^{-1}(\theta) = \left\{\lambda E\left[g_i(\theta;\chi_0)\, g_i(\theta;\chi_0)'\right]\lambda + \gamma\lambda E\left[g_{i,\chi}(\theta;\chi_0)\right]\Sigma_\chi E\left[g_{i,\chi}(\theta;\chi_0)'\right]\lambda'\right\}^{-1}$.

Then,

$$\begin{aligned}\nabla_{\theta\theta'}M(\theta) = &-E[\nabla_\theta g_i(\theta;\chi_0)]'\lambda' W(\theta)\lambda E[\nabla_\theta g_i(\theta;\chi_0)] - \\ &\{W(\theta)E[g_i(\theta;\chi_0)]\otimes I_d\}E[\nabla_{\theta\theta'}g_i(\theta;\chi_0)] - \\ &-\frac{1}{2}E[g_i(\theta;\chi_0)]'\lambda'\nabla_{\theta\theta'}W(\theta)\lambda E[g_i(\theta;\chi_0)]\end{aligned}$$

$$\nabla_{\theta\theta'}M(\theta_0) = -E[g_i(\theta_0,\chi_0)]'V^{-1}(\theta_0)E[g_i(\theta_0,\chi_0)] + o_p(1).$$

And thus,

$$\frac{\nabla_{\theta\theta'}L_N(\theta_0)}{N} - \nabla_{\theta\theta'}M(\theta_0) \xrightarrow{p} 0.$$

Then for $\varepsilon > 0$, $N > 0$, $\exists\delta_1(\varepsilon,N) > 0$, $\forall\theta \in \{\theta : \|\theta-\theta_0\| < \delta_1(\varepsilon,N)\}$, due to the continuity,

$$\sup_\theta \left\|\frac{\nabla_{\theta\theta'}L_N(\theta)}{N} - \frac{\nabla_{\theta\theta'}L_N(\theta_0)}{N}\right\| < \frac{1}{3}\varepsilon.$$

$\exists\delta_2(\varepsilon) > 0$, $\forall\theta \in \{\theta : \|\theta-\theta_0\| < \delta_2(\varepsilon)\}$, due to continuity,

$$\sup_\theta \|\nabla_{\theta\theta'}M(\theta) - \nabla_{\theta\theta'}M(\theta_0)\| < \frac{1}{3}\varepsilon.$$

And for $\varepsilon > 0$, $\exists N(\varepsilon,\varepsilon) > 0$, $\forall N > N(\varepsilon,\varepsilon)$,

$$P\left\{\left\|\frac{\nabla_{\theta\theta'}L_N(\theta_0)}{N} - \nabla_{\theta\theta'}M(\theta_0)\right\| < \frac{1}{3}\varepsilon\right\} \geq 1-\varepsilon.$$

Therefore, for any $\varepsilon > 0$, $\forall N > N(\varepsilon,\varepsilon)$, let $\delta(\varepsilon,N) = \min\{\delta_1(\varepsilon,N),\delta_2(\varepsilon)\}$, $\forall\theta \in \{\theta : \|\theta-\theta_0\| < \delta(\varepsilon,N)\}$,

$$\begin{aligned}\sup_\theta \left\|\frac{\nabla_{\theta\theta'}L_N(\theta)}{N} - \nabla_{\theta\theta'}M(\theta)\right\| \leq &\sup_\theta \left\|\frac{\nabla_{\theta\theta'}L_N(\theta)}{N} - \frac{\nabla_{\theta\theta'}L_N(\theta_0)}{N}\right\| + \sup_\theta \|\nabla_{\theta\theta'}M(\theta) - \nabla_{\theta\theta'}M(\theta_0) \\ &+ \left\|\frac{\nabla_{\theta\theta'}L_n(\theta_0)}{n} - \nabla_{\theta\theta'}M(\theta_0)\right\| \\ < &\frac{2}{3}\varepsilon + \left\|\frac{\nabla_{\theta\theta'}L_N(\theta_0)}{N} - \nabla_{\theta\theta'}M(\theta_0)\right\|.\end{aligned}$$

Then

$$\left\{\sup_{\theta}\left\|\frac{\nabla_{\theta\theta'}L_N(\theta_0)}{N}-\nabla_{\theta\theta'}M(\theta_0)\right\|<\frac{1}{3}\varepsilon\right\}\subset\left\{\sup_{\theta}\left\|\frac{\nabla_{\theta\theta'}L_N(\theta)}{N}-\nabla_{\theta\theta'}M(\theta)\right\|<\varepsilon\right\},$$

which implies

$$P\left\{\sup_{\|\theta-\theta_0\|<\delta(\varepsilon)}\left\|\frac{\nabla_{\theta\theta'}L_N(\theta)}{N}-\nabla_{\theta\theta'}M(\theta)\right\|<\varepsilon\right\}\geq 1-\varepsilon,$$

in other words, for $\varepsilon>0$,

$$\limsup_{n\to\infty}P\left\{\sup_{\|\theta-\theta_0\|<\delta(\varepsilon)}\left\|\frac{\nabla_{\theta\theta'}L_N(\theta)}{N}-\nabla_{\theta\theta'}M(\theta)\right\|>\varepsilon\right\}=0$$

Therefore, the Lemma 2 in CH (2003) is satisfied. By the Theorem 2 in CH (2003), for the etimator $\widehat{\theta}$ defined in (4.4.7), we can have

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right)\xrightarrow{d}\tau+N(0,\Sigma_\theta),$$

where

$$\Sigma_\theta=\left[G_\theta'\lambda'\left(\lambda\Sigma_g\lambda'+\gamma\lambda G_\chi'\Sigma_\chi G_\chi\lambda'\right)^{-1}\lambda G_\theta\right]^{-1},$$

$G_\theta=\nabla_\theta E[g_i(\theta_0;\chi_0)]$, $G_\chi=E\left[\nabla_\chi g_i(\theta_0;\chi_0)\right]$, $\gamma=\lim_{N\to\infty}\frac{N}{J}$, $\lambda=\lim_{N\to\infty}\lambda_N$, $\Sigma_g=E\left[g_i(\theta_0;\chi_0)g_i(\theta_0;\chi_0)'\right]$, $\tau=\arg\inf_{z\in R^d}\left\{\int_{R^d}\rho(z-u)f\left(u;0,G_\theta'\lambda'W(\theta_0)\lambda G_\theta\right)du\right\}$.

.3.3 The proof of Theorem 4.4.3

Lemma .3.1. *By the definition of Δ_j in (4.4.16), $\forall\theta\in\Theta$,*

$$\bar{g}_N(\theta;\widehat{\chi})-\bar{g}_N^j(\theta;\widehat{\chi})=O_p\left(\Delta_j\right).$$

$$\bar{g}_{N,\chi}(\theta;\widehat{\chi})-\bar{g}_{N,\chi}^j(\theta;\widehat{\chi})=O_p\left(\Delta_j\right),$$

$$V_N^j(\theta;\widehat{\chi})-V_N(\theta;\widehat{\chi})=O_p\left(\Delta_j\right).$$

Proof: By definition, for any $\theta \in \Theta$,

$$\bar{g}_N(\theta;\widehat{\chi}) = (\bar{g}_{t_{min}}(\theta;\widehat{\chi}),\ldots,\bar{g}_{T_r}(\theta;\widehat{\chi}))' = \left(\frac{1}{N_{t_{min}}}\sum_{i=1}^{N_{t_{min}}} g_{i,t_{min}}(\theta;\widehat{\chi}),\ldots,\frac{1}{N_{T_r}}\sum_{i=1}^{N_{T_r}} g_{i,T_r}(\theta;\widehat{\chi})\right)'.$$

Let $t \in [t_{min}, T_r]$, for $g_{i,t}(\theta;\chi) = C_{i,t}^d - C_t\left(M_{i,t}^d, z_{i,t}^d;\theta,\chi\right)$, $g_{i,t}^j(\theta;\chi) = C_{i,t}^d - C_t^j\left(M_{i,t}^d, z_{i,t}^d;\theta,\chi\right)$,

$$\begin{aligned}
\bar{g}_t(\theta;\widehat{\chi}) - \bar{g}_t^j(\theta;\widehat{\chi}) =& \frac{1}{N_t}\sum_{i=1}^{N_t}\left[g_{i,t}(\theta;\widehat{\chi}) - g_{i,t}^j(\theta;\widehat{\chi})\right] \\
=& \frac{1}{N_t}\sum_{i=1}^{N_t}\left[C_t\left(M_{i,t}^d, z_{i,t}^d;\theta,\widehat{\chi}\right) - C_t^j\left(M_{i,t}^d, z_{i,t}^d;\theta,\widehat{\chi}\right)\right] \\
=& \frac{1}{N_t}\sum_{i=1}^{N_t} E_{P_{i,t}}\left\{\left[c_t\left(\frac{M_{i,t}^d}{P_{i,t}}, z_{i,t}^d;\theta,\widehat{\chi}\right) - c_t^j\left(\frac{M_{i,t}^d}{P_{i,t}}, z_{i,t}^d;\theta,\widehat{\chi}\right)\right]P_{i,t}\right\} \\
\leq& \Delta_j \frac{1}{N_t}\sum_{i=1}^{N_t} E_{P_{i,t}}(P_{i,t}) = O_p\left(\Delta_j\right),
\end{aligned}$$

which implies

$$\bar{g}_N(\theta;\widehat{\chi}) - \bar{g}_N^j(\theta;\widehat{\chi}) = O_p\left(\Delta_j\right).$$

And similarly, we can also have

$$\bar{g}_{N,\chi}(\theta;\widehat{\chi}) - \bar{g}_{N,\chi}^j(\theta;\widehat{\chi}) = O_p\left(\Delta_j\right).$$

And thus for,

$$\begin{aligned}
V_N^j(\theta;\widehat{\chi}) =& \zeta_N \sum_{i=1}^{N^{obs}} \lambda_N \tilde{g}_i^j(\theta;\widehat{\chi})\,\tilde{g}_i^j(\theta;\widehat{\chi})'\,\lambda_N'\zeta_N' \\
&+ \frac{N}{J}\lambda_N \bar{g}_{N,\chi}^j(\theta;\widehat{\chi})\,\widehat{\Sigma}_\chi \bar{g}_{N,\chi}^j(\theta;\widehat{\chi})'\,\lambda_N',
\end{aligned}$$

the first term, since $\tilde{g}_i^j(\theta;\widehat{\chi})$ and $\tilde{g}_i(\theta;\widehat{\chi})$ are continuous and Θ is compact by assumptions,

$$
\begin{aligned}
&\zeta_N \sum_{i=1}^{N^{obs}} \lambda_N \tilde{g}_i^j(\theta;\widehat{\chi})\, \tilde{g}_i^j(\theta;\widehat{\chi})' \lambda_N' \zeta_N' - \zeta_N \sum_{i=1}^{N^{obs}} \lambda_N \tilde{g}_i(\theta;\widehat{\chi})\, \tilde{g}_i(\theta;\widehat{\chi})' \lambda_N' \zeta_N' \\
=&\zeta_N \lambda_N \sum_{i=1}^{N^{obs}} \left[\tilde{g}_i^j(\theta;\widehat{\chi})\, \tilde{g}_i^j(\theta;\widehat{\chi})' - \tilde{g}_i(\theta;\widehat{\chi})\, \tilde{g}_i(\theta;\widehat{\chi})'\right] \lambda_N' \zeta_N' \\
=&\zeta_N \lambda_N \sum_{i=1}^{N^{obs}} \left[\tilde{g}_i^j(\theta;\widehat{\chi})\, \tilde{g}_i^j(\theta;\widehat{\chi})' - \tilde{g}_i^j(\theta;\widehat{\chi})\, \tilde{g}_i(\theta;\widehat{\chi})'\right] \lambda_N' \zeta_N' + \\
&\zeta_N \lambda_N \sum_{i=1}^{N^{obs}} \left[\tilde{g}_i^j(\theta;\widehat{\chi})\, \tilde{g}_i(\theta;\widehat{\chi})' - \tilde{g}_i(\theta;\widehat{\chi})\, \tilde{g}_i(\theta;\widehat{\chi})'\right] \lambda_N' \zeta_N' \\
=&O_p\left(\Delta_j\right).
\end{aligned}
$$

And the second term is similar, which means

$$
V_N^j(\theta;\widehat{\chi}) - V_N(\theta;\widehat{\chi}) = O_p\left(\Delta_j\right).
$$

The Proof of Theorem 4.4.3: The criterion function for the case using analytical solution and the one approximated by numerical methods are

$$
L_N(\theta) = -\frac{N}{2} \bar{g}_N(\theta;\widehat{\chi})' \lambda_N' W_N(\theta;\widehat{\chi})\, \lambda_N \bar{g}_N(\theta;\widehat{\chi}),
$$

and

$$
L_N^j(\theta) = -\frac{N}{2} \bar{g}_N^j(\theta;\widehat{\chi})' \lambda_N' W_N^j(\theta;\widehat{\chi})\, \lambda_N \bar{g}_N^j(\theta;\widehat{\chi}),
$$

respectively. By Lemma .3.1, if $N\Delta_j \to 0$, as $N \to \infty$, for all $\theta \in \Theta$,

$$
V_N^j(\theta;\widehat{\chi}) - V_N(\theta;\widehat{\chi}) = O_p\left(\Delta_j\right),
$$

so that

$$\begin{aligned}\left[V_N^j(\theta;\widehat{\chi})\right]^{-1}\left[V_N^j(\theta;\widehat{\chi})-V_N(\theta;\widehat{\chi})\right]V_N^{-1}(\theta;\widehat{\chi}) =& V_N^{-1}(\theta;\widehat{\chi})-\left[V_N^j(\theta;\widehat{\chi})\right]^{-1}\\ =& O_p(1)\,O\left(\Delta_j\right)O_p(1)\\ =& O_p\left(\Delta_j\right).\end{aligned}$$

So that,

$$\begin{aligned}&\sup_{\theta\in\Theta}\left\{L_N^j(\theta)-\tilde{L}_N^j(\theta)\right\}\\ =&\sup_{\theta\in\Theta}\left\{-\frac{N}{2}\bar{g}_N^j(\theta;\widehat{\chi})'\lambda_N' W_N^j(\theta;\widehat{\chi})\lambda_N\bar{g}_N^j(\theta;\widehat{\chi})+\frac{N}{2}\bar{g}_N^j(\theta;\widehat{\chi})'\lambda_N' W_N(\theta;\widehat{\chi})\lambda_N\bar{g}_N^j(\theta;\widehat{\chi})\right\}\\ =&\sup_{\theta\in\Theta}\left\{-\frac{N}{2}\bar{g}_N^j(\theta;\widehat{\chi})'\lambda_N'\left[W_N^j(\theta;\widehat{\chi})-W_N(\theta;\widehat{\chi})\right]\lambda_N\bar{g}_N^j(\theta;\widehat{\chi})\right\}\\ =&NO_p\left(\Delta_j\right)=O_p\left(N\Delta_j\right).\end{aligned}$$

Therefore, denote $\tilde{W}_N(\theta;\widehat{\chi})=\lambda_N' W_N(\theta;\widehat{\chi})\lambda_N$,

$$\begin{aligned}&\sup_{\theta\in\Theta}\left\|L_N(\theta)-L_N^j(\theta)\right\|\\ \le&\sup_{\theta\in\Theta}\left\|L_N(\theta)-\tilde{L}_N^j(\theta)\right\|+\sup_{\theta\in\Theta}\left\|L_N^j(\theta)-\tilde{L}_N^j(\theta)\right\|\\ \le&\sup_{\theta\in\Theta}\left\|\frac{N}{2}\bar{g}_N^j(\theta;\widehat{\chi})'\tilde{W}_N(\theta;\widehat{\chi})\bar{g}_N^j(\theta;\widehat{\chi})-\frac{N}{2}\bar{g}_N^j(\theta;\widehat{\chi})'\tilde{W}_N(\theta;\widehat{\chi})\bar{g}_N(\theta;\widehat{\chi})\right\|+\\ &\sup_{\theta\in\Theta}\left\|\frac{N}{2}\bar{g}_N^j(\theta;\widehat{\chi})'\tilde{W}_N(\theta;\widehat{\chi})\bar{g}_N(\theta;\widehat{\chi})-\frac{N}{2}\bar{g}_N(\theta;\widehat{\chi})'\tilde{W}_N(\theta;\widehat{\chi})\bar{g}_N(\theta;\widehat{\chi})\right\|+O_p\left(N\Delta_j\right)\\ \le&\frac{N}{2}\sup_{\theta\in\Theta}\left\|\bar{g}_N^j(\theta;\widehat{\chi})'\tilde{W}_N(\theta;\widehat{\chi})\right\|\sup_{\theta\in\Theta}\left\|\bar{g}_N^j(\theta;\widehat{\chi})-\bar{g}_N(\theta;\widehat{\chi})\right\|+\\ &\frac{N}{2}\sup_{\theta\in\Theta}\left\|\bar{g}_N(\theta;\widehat{\chi})'\tilde{W}_N(\theta;\widehat{\chi})\right\|\sup_{\theta\in\Theta}\left\|\bar{g}_N^j(\theta;\widehat{\chi})-\bar{g}_N(\theta;\widehat{\chi})\right\|+O_p\left(N\Delta_j\right)\\ =&O_p\left(N\Delta_j\right).\end{aligned}$$

Therefore, when $N\Delta_j\to 0$, as $N\to\infty$, $L_N(\theta)-L_N^j(\theta)\xrightarrow{p}0$ over Θ. Further, due to the compactness of Θ and the Taylor expansion,

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left\|\exp\left[L_N(\theta)\right]-\exp\left[L_N^j(\theta)\right]\right\| \\
=&\sup_{\theta\in\Theta}\left\|\exp\left[L_N(\theta)\right]\right\|\sup_{\theta\in\Theta}\left\|\exp\left[L_N(\theta)-L_N^j(\theta)\right]-1\right\| \\
\leq& C_1\sup_{\theta\in\Theta}\left\|\exp\left(L_N(\tilde{\theta})-L_N^j(\tilde{\theta})\right)\left[L_N(\theta)-L_N^j(\theta)\right]\right\| \\
\leq& C_1\sup_{\theta\in\Theta}\left\|\exp\left(L_N(\tilde{\theta})-L_N^j(\tilde{\theta})\right)\right\|\sup_{\theta\in\Theta}\left\|\left[L_N(\theta)-L_N^j(\theta)\right]\right\| \\
\approx& C_1\left(1+O_p\left(N\Delta_j\right)\right)O_p\left(N\Delta_j\right) \\
=& O_p\left(N\Delta_j\right),
\end{aligned} \tag{.3.4}$$

where $\tilde{\theta}$ is between 0 and θ.

$$\begin{aligned}
&\int_\Theta\exp\left[L_N^j(\theta)\right]\pi(\theta)\,d\theta-\int_\Theta\exp\left[L_N(\theta)\right]\pi(\theta)\,d\theta \\
=&\int_\Theta\exp\left[L_N^j(\theta)-L_N(\theta)\right]\pi(\theta)\,d\theta \\
\leq&\sup_{\theta\in\Theta}\left\|\exp\left[L_N(\theta)\right]-\exp\left[L_N^j(\theta)\right]\right\|\int_\Theta\pi(\theta)\,d\theta \\
=& O_p\left(N\Delta_j\right).
\end{aligned} \tag{.3.5}$$

Following the proof of Theorem 4.4.1, we define

$$J(\theta_0)=-E\left[\nabla_\theta g_i(\theta_0,\chi_0)\right]'V^{-1}(\theta_0)E\left[\nabla_\theta g_i(\theta_0,\chi_0)\right],$$

and

$$h\equiv\sqrt{N}\left(\theta-T_N\right),T_N=\theta_0+\frac{1}{\sqrt{N}}U_N,U_N=\frac{1}{\sqrt{N}}J^{-1}(\theta_0)\nabla_\theta L_N(\theta_0),$$

so that, let $H_N=\left\{\sqrt{N}(\theta-\theta_0)-U_N:\theta\in\Theta\right\}$, $p_N(\theta)$ and $p_N^j(\theta)$ can be transformed into $\frac{1}{\sqrt{N}} p_N^*(h)$ and $\frac{1}{\sqrt{N}} p_N^{*j}(h)$, respectively, where,

$$p_N^{*j}(h) = \frac{\pi\left(T_N + \frac{h}{\sqrt{N}}\right)\exp\left[L_N^j\left(T_N + \frac{h}{\sqrt{N}}\right)\right]}{\int_{H_N} \pi\left(T_N + \frac{h}{\sqrt{N}}\right)\exp\left[L_N^j\left(T_N + \frac{h}{\sqrt{N}}\right)\right] dh} = \frac{\pi\left(T_N + \frac{h}{\sqrt{N}}\right)\exp\left[L_N^j\left(T_N + \frac{h}{\sqrt{N}}\right)\right]}{C^j},$$

$$p_N^{*}(h) = \frac{\pi\left(T_N + \frac{h}{\sqrt{N}}\right)\exp\left[L_N\left(T_N + \frac{h}{\sqrt{N}}\right)\right]}{\int_{H_N} \pi\left(T_N + \frac{h}{\sqrt{N}}\right)\exp\left[L_N\left(T_N + \frac{h}{\sqrt{N}}\right)\right] dh} = \frac{\pi\left(T_N + \frac{h}{\sqrt{N}}\right)\exp\left[L_N\left(T_N + \frac{h}{\sqrt{N}}\right)\right]}{C}.$$

The corresponding transformed risk functions of $R_N^j(\xi)$ and $R_N(\xi)$ are denoted as $Q_N^j(\zeta)$ and $Q_N(\zeta)$, respectively, where

$$Q_N^j(\zeta) = \int_{H_N} \rho(h + U_N - \zeta)\, p_N^{*j}(h)\, dh,$$

$$Q_N(\zeta) = \int_{H_N} \rho(h + U_N - \zeta)\, p_N^{*}(h)\, dh.$$

As in Theorem 4.4.1, the Lemma 1 and Lemma 2 in CH (2003) are satisfied, which implies that the Theorem 1 and Theorem 2 in their paper hold. So that we have for any $0 \leq \alpha < \infty$,

$$\int_{H_N} \|h\|^{\alpha} \left|p_N^*(h) - p_\infty(h)\right| dh \xrightarrow{p} 0,$$

where

$$p_\infty(h) = \sqrt{\frac{|J(\theta_0)|}{(2\pi)^d}} \exp\left(-\frac{1}{2} h' J(\theta_0)\, h\right),$$

and

$$\lim_{N \to \infty} \int_{H_N} \|h\|^{\alpha} p_\infty(h)\, dh = C_\alpha < \infty.$$

$$Q_\infty(\zeta) = \int_{R^d} \rho(h + U_N - \zeta)\, p_\infty(h)\, dh.$$

Therefore,

$$
\begin{aligned}
&\int_{H_N} \|h\|^{\alpha} \left| p_N^{*j}(h) - p_\infty(h) \right| dh \\
\leq &\int_{H_N} \|h\|^{\alpha} \left| p_N^{*j}(h) - p_N^{*}(h) \right| dh + \int_{H_N} \|h\|^{\alpha} \left| p_N^{*}(h) - p_\infty(h) \right| dh \\
= &\int_{H_N} \|h\|^{\alpha} \pi\left(T_N + \frac{h}{\sqrt{N}}\right) \left| \frac{\exp\left[L_N^j\left(T_N + \frac{h}{\sqrt{N}}\right)\right]}{C^j} - \frac{\exp\left[L_N\left(T_N + \frac{h}{\sqrt{N}}\right)\right]}{C} \right| dh + o_p(1) \\
\leq &\int_{H_N} \|h\|^{\alpha} \pi\left(T_N + \frac{h}{\sqrt{N}}\right) \left| \frac{\exp\left[L_N^j\left(T_N + \frac{h}{\sqrt{N}}\right)\right]}{C^j} - \frac{\exp\left[L_N\left(T_N + \frac{h}{\sqrt{N}}\right)\right]}{C^j} \right| dh + \\
&\int_{H_N} \|h\|^{\alpha} \pi\left(T_N + \frac{h}{\sqrt{N}}\right) \left| \frac{\exp\left[L_N\left(T_N + \frac{h}{\sqrt{N}}\right)\right]}{C^j} - \frac{\exp\left[L_N\left(T_N + \frac{h}{\sqrt{N}}\right)\right]}{C} \right| dh + o_p(1).
\end{aligned}
$$

For the second term, it is obvious that

$$
C^j = \int_{H_N} \pi\left(T_N + \frac{h}{\sqrt{N}}\right) \exp\left[L_N^j\left(T_N + \frac{h}{\sqrt{N}}\right)\right] dh = \int_{\Theta} \exp\left[L_N^j(\theta)\right] \pi(\theta)\, d\theta,
$$

$$
C = \int_{H_N} \pi\left(T_N + \frac{h}{\sqrt{N}}\right) \exp\left[L_N\left(T_N + \frac{h}{\sqrt{N}}\right)\right] dh = \int_{\Theta} \exp\left[L_N(\theta)\right] \pi(\theta)\, d\theta.
$$

which implies $C^j - C = O_p\left(N\Delta_j\right)$ by (.3.5) and then for the first term, since $N\Delta_j \to 0$,

$$
\begin{aligned}
&\int_{H_N} \|h\|^{\alpha} \pi\left(T_N + \frac{h}{\sqrt{N}}\right) \left| \frac{\exp\left[L_N\left(T_N + \frac{h}{\sqrt{N}}\right)\right]}{C^j} - \frac{\exp\left[L_N\left(T_N + \frac{h}{\sqrt{N}}\right)\right]}{C} \right| dh \\
= &\left| \frac{1}{C^j} - \frac{1}{C} \right| \int_{H_N} \|h\|^{\alpha} \pi\left(T_N + \frac{h}{\sqrt{N}}\right) \exp\left[L_N\left(T_N + \frac{h}{\sqrt{N}}\right)\right] dh \\
= &\left| \frac{1}{C^j} - \frac{1}{C} \right| \int_{H_N} \|h\|^{\alpha} p_\infty(h)\, dh + o_p(1) \\
= &C_\alpha \left| \frac{1}{C^j} - \frac{1}{C} \right| + o_p(1) = O_p\left(N\Delta_j\right).
\end{aligned}
$$

For the second term, by the Taylor expansion and (.3.5)

$$
\begin{aligned}
&\int_{H_N} \|h\|^{\alpha} \pi\left(T_N+\frac{h}{\sqrt{N}}\right)\left|\frac{\exp\left[L_N^j\left(T_N+\frac{h}{\sqrt{N}}\right)\right]}{C^j}-\frac{\exp\left[L_N\left(T_N+\frac{h}{\sqrt{N}}\right)\right]}{C^j}\right| dh \\
=&\frac{C}{C^j}\int_{H_N}\|h\|^{\alpha}\frac{1}{C}\pi\left(T_N+\frac{h}{\sqrt{N}}\right)\exp\left[L_N\left(T_N+\frac{h}{\sqrt{N}}\right)\right] \\
&\times\left|L_N^j\left(T_N+\frac{h}{\sqrt{N}}\right)-L_N\left(T_N+\frac{h}{\sqrt{N}}\right)+o_p\left(N\Delta_j\right)\right| dh \\
=&O_p(1)\,O_p\left(N\Delta_j\right)\int_{H_N}\|h\|^{\alpha}\frac{1}{C}\pi\left(T_N+\frac{h}{\sqrt{N}}\right)\exp\left[L_N\left(T_N+\frac{h}{\sqrt{N}}\right)\right] dh \\
=&O_p(1)\,O_p\left(N\Delta_j\right)C_{\alpha} \\
=&O_p\left(N\Delta_j\right).
\end{aligned}
$$

Therefore,

$$
\int_{H_N}\|h\|^{\alpha}\left|p_N^{*j}(h)-p_{\infty}(h)\right| dh = O_p\left(N\Delta_j\right).
$$

By the Assumption 3, $\rho(u)\le 1+|u|^p$ and by $|a+b|^p\le 2^{p-1}|a|^p+2^{p-1}|b|^p$ for $p\ge 1$. For any fixed ζ,

$$
\begin{aligned}
\left|Q_N^j(\zeta)-Q_{\infty}(\zeta)\right| \le & \int_{H_N}\left(1+\|h+U_N-\zeta\|^p\right)\left|p_N^{*j}(h)-p_{\infty}(h)\right| dh \\
&+\int_{R^d\setminus H_N}\left(1+\|h+U_N-\zeta\|^p\right)p_{\infty}(h)\,dh \\
\le & \int_{H_N}\left(1+2^{p-1}\|h\|^{p-1}+2^{p-1}\|U_N-\zeta\|^{p-1}\right)\left|p_N^{*j}(h)-p_{\infty}(h)\right| dh \\
&+\int_{R^d\setminus H_N}\left(1+2^{p-1}\|h\|^{p-1}+2^{p-1}\|U_N-\zeta\|^{p-1}\right)p_{\infty}(h)\,dh \\
= & \int_{H_N}\left(1+2^{p-1}\|h\|^{p-1}+O_p(1)\right)\left|p_N^{*j}(h)-p_{\infty}(h)\right| dh \\
&+\int_{R^d\setminus H_N}\left(1+2^{p-1}\|h\|^{p-1}+O_p(1)\right)p_{\infty}(h)\,dh.
\end{aligned}
$$

From above discussions,

$$
\int_{H_N}\left(1+2^{p-1}\|h\|^{p-1}+O_p(1)\right)\left|p_N^{*j}(h)-p_{\infty}(h)\right| dh = O_p\left(N\Delta_j\right),
$$

and by the exponentially small tails of the normal density,

$$\int_{R^d \setminus H_N} \left(1 + 2^{p-1} \|h\|^{p-1} + O_p(1)\right) p_\infty(h)\, dh = o_p(1).$$

Hence, if $N\Delta_j \to 0$, given fixed ζ, $Q_N^j(\zeta) - Q_\infty(\zeta) \xrightarrow{p} 0$.

Then, we show that both $Q_N^j(\zeta)$ and $Q_\infty(\zeta)$ are convex, for any given ζ and $\tilde{\zeta}$, and $\alpha \in [0,1]$,

$$\begin{aligned}
Q_N^j\left(\alpha\zeta + (1-\alpha)\tilde{\zeta}\right) =& \int_{H_N} \rho\left[h + U_N - \alpha\zeta - (1-\alpha)\tilde{\zeta}\right] p_N^{j*}(h)\, dh \\
=& \int_{H_N} \rho\left[\alpha(h + U_N - \zeta) + (1-\alpha)\left(h + U_N - \tilde{\zeta}\right)\right] p_N^{j*}(h)\, dh \\
\leq& \alpha \int_{H_N} \rho(h + U_N - \zeta)\, p_N^{j*}(h)\, dh \\
&+ (1-\alpha) \int_{H_N} \rho\left(h + U_N - \tilde{\zeta}\right) p_N^{j*}(h)\, dh \\
=& \alpha Q_N^j(\zeta) + (1-\alpha) Q_N^j\left(\tilde{\zeta}\right).
\end{aligned}$$

Hence $Q_N^j(\zeta)$ is convex. Similarly, $Q_\infty(\zeta)$ is also convex. Further,

$$\begin{aligned}
Q_\infty(\zeta) \leq& \int_{H_N} \left(1 + 2^{p-1}\|h\|^{p-1} + 2^{p-1}\|U_N - \zeta\|^{p-1}\right) p_\infty(h)\, dh \\
=& 1 + 2^{p-1} \int_{H_N} \|h\|^{p-1} p_\infty(h)\, dh + 2^{p-1} \int_{H_N} \|U_N - \zeta\|^{p-1} p_\infty(h)\, dh \\
=& O_p(1).
\end{aligned}$$

And by the same logic $Q_N^j(\zeta) = O_p(1)$.

If $N\Delta_j \to 0$, by the convexity lemma of Polard (1991), pointwise convergence entails the uniform convergence over the compact set $\mathfrak{B}$,

$$\sup_{\zeta \in \mathfrak{B}} \left| Q_N^j(\zeta) - Q_\infty(\zeta) \right| \xrightarrow{p} 0.$$

For $Q_\infty(\zeta) = \int_{R^d} \rho(h + U_N - \zeta)\, p_\infty(h)\, dh$, it is minimized at $\zeta^* = \tau + U_N = O_p(1)$. And $Q_N^j(\zeta)$ is minimized at $\sqrt{N}\left(\widehat{\theta}^j - \theta_0\right)$. Following CH, the uniform convergence property above as well as the convexity property imply that $\sqrt{N}\left(\widehat{\theta}^{j}-\theta_0\right)=U_N+\tau+o_p(1)$. Combined with the fact that

$$U_N=\frac{1}{\sqrt{N}}J^{-1}(\theta_0)\nabla_\theta L_N(\theta_0)\xrightarrow{d}N(0,\Sigma_\theta),$$

the results in the theorem follows.

.3.4 The Proof of Corollary 4.4.4

The asymptotic theory is easily obtained from Theorem 4.4.3. For

$$E^j\left[N\left(\theta-\bar{\theta}^j\right)\left(\theta-\bar{\theta}^j\right)'\Big|\cdot\right]=\int_\Theta N\left(\theta-\bar{\theta}^j\right)\left(\theta-\bar{\theta}^j\right)'p_N^j(\theta)\,d\theta,$$

we let

$$h\equiv\sqrt{N}(\theta-T_N),T_N=\theta_0+\frac{1}{\sqrt{N}}U_N,U_N=\frac{1}{\sqrt{N}}J^{-1}(\theta_0)\nabla_\theta L_N(\theta_0),$$

then

$$\theta=\frac{h}{\sqrt{N}}+T_N,\bar{\theta}^j=\frac{\bar{h}^j}{\sqrt{N}}+T_N,\bar{h}^j=\int_{H_N}hp_N^{*j}(h)\,dh,$$

so that

$$\theta-\bar{\theta}^j=\frac{1}{\sqrt{N}}\left(h-\bar{h}^j\right).$$

Therefore,

$$\begin{aligned}&\int_\Theta N\left(\theta-\bar{\theta}^j\right)\left(\theta-\bar{\theta}^j\right)'p_N^j(\theta)\,d\theta\\&=\int_{H_N}\left(h-\bar{h}^j\right)\left(h-\bar{h}^j\right)'p_N^{*j}(h)\,dh\\&=\int_{H_N}hh'p_N^{*j}(h)\,dh-\bar{h}^j\bar{h}^{j\prime},\end{aligned}$$

As in Theorem 4.4.3, if $N\Delta_j \to 0$, $\int_{H_N} \|h\|^{\alpha} \left| p_N^{*j}(h) - p_\infty(h) \right| dh = o_p(1)$, which implies

$$
\begin{aligned}
\bar{h}^j \bar{h}^{j\prime} =& \int_{H_N} h p_N^{*j}(h)\, dh \int_{H_N} h' p_N^{*j}(h)\, dh \\
\xrightarrow{p}& \int_{R^d} h p_\infty(h)\, dh \int_{R^d} h' p_\infty(h)\, dh \\
=& \bar{h}\bar{h}',
\end{aligned}
$$

and

$$
\int_{H_N} hh' p_N^{*j}(h)\, dh \xrightarrow{p} \int_{R^d} hh' p_\infty(h)\, dh.
$$

Therefore,

$$
\begin{aligned}
\int_{H_N} \left(h - \bar{h}^j\right)\left(h - \bar{h}^j\right)' p_N^{*j}(h)\, dh \xrightarrow{p}& \int_{H_N} \left(h - \bar{h}\right)\left(h - \bar{h}\right)' p_N^{*}(h)\, dh \\
=& J^{-1}(\theta_0) \\
=& -\nabla_{\theta\theta'} M(\theta_0) \\
=& \Sigma_g.
\end{aligned}
$$

That is, if $N\Delta_j \to 0$ as $N \to \infty$,

$$
\int_{\Theta} N\left(\theta - \bar{\theta}^j\right)\left(\theta - \bar{\theta}^j\right)' p_N^j(\theta)\, d\theta = \Sigma_g + o_p(1).
$$

.4 The Details of Estimation and Computation

.4.1 The Computation of Bias and Root Mean Square Error

This subsection shows how to compute the bias and RMSE. Assume the true value of the target parameter x is x_0 and $\{\hat{x}^m\}_{m=1}^M$ is the set of estimates of x in M Monte Carlo replications. The bias is defined as

$$
Bias(x) = \frac{1}{M}\sum_{m=1}^{M} \hat{x}^m - x_0.
$$

The root mean square error is defined as

$$RMSE\left(x\right)=\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{x}^{m}-x_{0}\right)^{2}}.$$

.4.2 The Endogenous Grid Method for the Model (4.2.7)

The application of EGM for model (4.2.7) is documented in Algorithm 3.

Algorithm 3 The Endogenous Grid Method for Dynamic Model (4.2.7)

1: **Inputs:** Optimal consumption at period $t+1$, $c^j\left(\overrightarrow{m_{t+1}}, z_{t+1}; \theta, \chi\right)$ and the endogenous grid at period $t+1$, $\overrightarrow{m_{t+1}}$.
2: Form an exogenous ascending grid over end-of-period wealth at period t, denoted as $\overrightarrow{A_t} = \left\{A_t^k\right\}_{k=1}^{j}$, where $A_t^k > A_t^{k-1}$, $\forall k \in \{2, \ldots, j\}$.
3: **for** $k = 1$ to j **do**
4: Compute $c_{i,t}^k = \left\{\beta_0 R E_{\varsigma_{t+1},\varepsilon_{t+1},z_{i,t+1}}\left[\frac{v(z_{t+1};\eta_0)}{v(z_t;\eta_0)}\left(G_{t+1}\varsigma_{t+1}\right)^{-\rho} c^j\left(m_{t+1}^k, z_{t+1}; \theta, \chi\right)\right]\right\}^{-\frac{1}{\rho}}$ with $m_{t+1}^k = \frac{RA_t^k}{G_{t+1}\varsigma_{t+1}} + \varepsilon_{t+1}$.
5: Compute $m_t^k = c_t^k + A_t^k$.
6: **end for**
7: Store the endogenous grid. $\overrightarrow{m_t} = \left\{m_t^k\right\}_{k=1}^{j}$.
8: Store the corresponding optimal consumption at period t. $c^j\left(\overrightarrow{m_t}, z_t; \theta, \chi\right) = \left\{c_t^k\right\}_{k=1}^{j}$
9: **Ouputs:** $c^j\left(\overrightarrow{m_t}, z_t; \theta, \chi\right)$, $\overrightarrow{m_t}$.

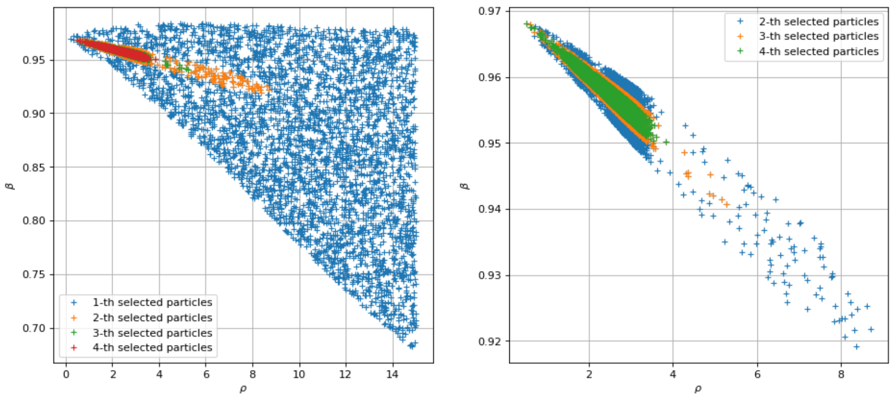
Note:

(i) In Step 4, numerical method is used.

- $E_{\varsigma_{t+1},\varepsilon_{t+1},z_{i,t+1}}$ is the expectation with respect to $\varsigma_{t+1}, \varepsilon_{t+1}$ and z_{t+1}. The expectation is numerically evaluated by using Gauss-Hermite quadrature method.
- The algorithm solves the model backwards, therefore $c^j\left(m_{t+1}^k, z_{t+1}; \theta, \chi\right)$ is the interpolated value of optimal consumption at period $t+1$ to approximate the income shocks.

(ii) During the EGM step, as in Carroll (2006), the credit constraints are dealt with by setting the smallest possible end-of-period resources A_t^1 equal 0. After operating the EGM, due to the monotonicity of saving, m_t^1 is the threshold value so that when $m_t < m_t^1$, the optimal consumption $c_t = m_t$.

Figure 1: The particle points selected during the estimation



.4.3 The details of the estimation procedure for Section 4.5.1

During the estimation for the model (4.5.1), let $K_1 = 12800$, $K_2 = 3840$, $K = 2560$, $\delta = 0.5$ and the cutoff value $L = -10$. The number of grid to solve the model is 100. The perturbation variance is $\Sigma = diag\,(0.0001, 0.04)$, where 0.0001 and 0.04 are for β and ρ, respectively. We use the case where $N^{obs} = 3000$ for illustration.

Figure 1 plots the particle selected during the estimation procedure. As the process goes on, the area shrinks very quickly. The area of the first particle selection is wide but starting from the second selection, the area is very narrow. After the fourth particle points selection, we collect all the particles and select a subset of them based on the threshold value L. Afterwards, we uniformly choose K points from the subset. Based on these K selected particles, we construct a proposal distribution – a mixture normal distribution. At last, we draw K_3 samples from the proposal distribution.

The subset of particles and the contour of the quasi-posterior density are plotted in Figure 2. The left panel is the contour plot and the right panel is the contour plot plus the subset of particles. We can readily find that the particles cover the area with significant density value quite well, which justifies that the proposal distribution is very close to the quasi-density function.

We can see from the left panel of Figure 3. The area with significant weights

Figure 2: The contour of the quasi-posterior density function and finally selected particle points

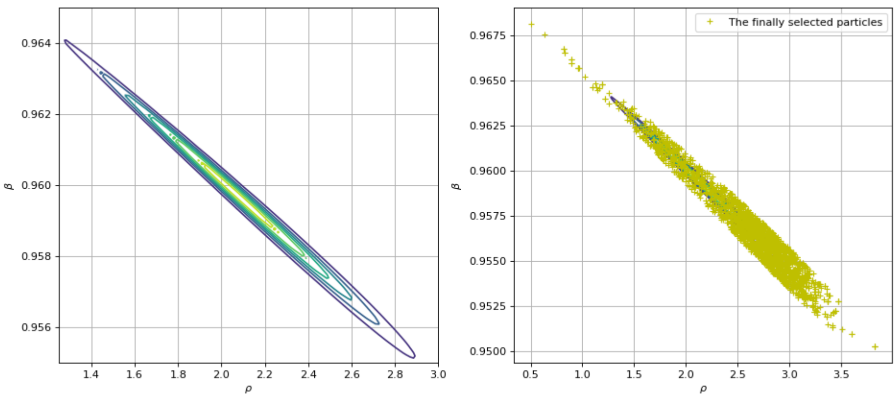
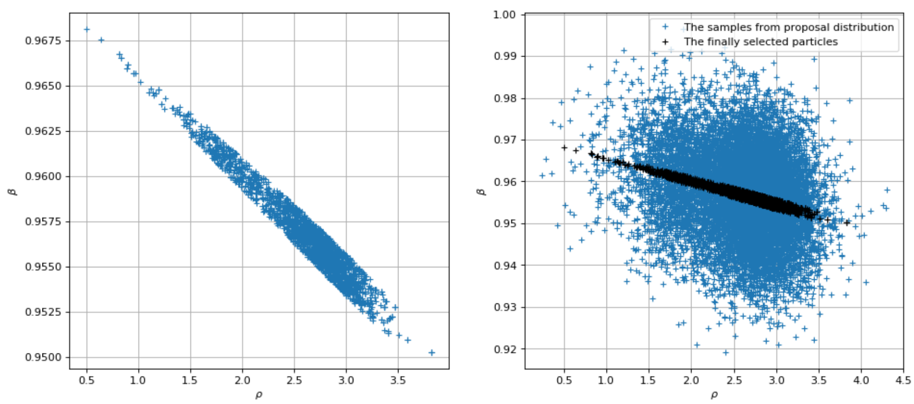


Figure 3: The finally selected particle points and samples from proposal distribution



is very narrow. The algorithm can identify the area quite accurately. After the final selection, we draw K_3 samples from the proposal distribution. From the right panel of Figure 3, we can find that the finally selected particles are almost covered by the samples from the proposal distribution.

.4.4 The Kalman Filter for the Income Process

When there is not income shock, we have

$$z_{it} = \mathfrak{A} + \mathfrak{B} x_{it} + \mathfrak{y}_{it},$$

$$x_{it} = \mathfrak{C}_t + \mathfrak{D} x_{it-1} + \mathfrak{u}_{it},$$

where $\mathfrak{A} = 0$, $\mathfrak{B} = 1$, $z_{it} = \log Y_{it}$, $x_{it} = \log P_{it}$, $\mathfrak{y}_{it} = \log \varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right)$, $\mathfrak{C}_t = \log G_t$, $\mathfrak{D} = 1$, $\mathfrak{u}_{it} = \log \varsigma_{it} \sim N\left(0, \sigma_\varsigma^2\right)$. According to the dataset, Y_{it} is observed household income, G_t , σ_ε^2 and σ_ς^2 are known. The permanent income component P_{it} is the one that we want to recover. In the following, the subscripts i is suppressed.

The Kalman filter consists of following three steps. Since the error terms are all normal and the structure is linear, all the variables in the system are normal distributed. Thus we only need to filter the mean and variance. Initialize the mean and variance at the beginning, $\mu_{0|0} = E\left[x_0|F_0\right]$, $\Sigma_{0|0} = Var\left(x_0|F_0\right)$, where F_0 is the information set known at time 0. Later the details of initialization is discussed.

- Initialize $\mu_{0|0}$ and $\Sigma_{0|0}$. At the beginning of time t, we have $\mu_{t-1|t-1}$, $\Sigma_{t-1|t-1}$.
- One-step-ahead predictive distribution of $x_t|F_{t-1} \sim N\left(\mu_{t|t-1}, \Sigma_{t|t-1}\right)$:

$$\begin{aligned}
\mu_{t|t-1} \equiv E\left[x_t|F_{t-1}\right] =& E\left[\mathfrak{C}_t + \mathfrak{D} x_{it-1}|F_{t-1}\right] \\
=& \mathfrak{C}_t + \mathfrak{D} \mu_{t-1|t-1},
\end{aligned}$$

$$\Sigma_{t|t-1} \equiv Var\left[x_t|F_{t-1}\right] = E\left[Var\left(x_t|F_{t-1}\right)|F_{t-1}\right] + Var\left[E\left(x_t|F_{t-1}\right)|F_{t-1}\right]$$
$$= \sigma_\varsigma^2 + \mathfrak{D}^2\Sigma_{t-1|t-1},$$

where F_t denotes the information known up to time t.

- One-step-ahead predictive distribution of $z_t|F_{t-1} \sim N\left(f_{t|t-1}, Q_{t|t-1}\right)$:

$$f_{t|t-1} \equiv E\left[z_t|F_{t-1}\right] = E\left\{E\left[z_t|x_t, F_{t-1}\right]|F_{t-1}\right\}$$
$$= \mathfrak{U} + \mathfrak{B}\mu_{t|t-1},$$

$$Q_{t|t-1} \equiv Var\left[z_t|F_{t-1}\right] = E\left[Var\left(z_t|F_{t-1}\right)|F_{t-1}\right] + Var\left[E\left(z_t|F_{t-1}\right)|F_{t-1}\right]$$
$$= \sigma_\varepsilon^2 + \mathfrak{B}^2\Sigma_{t|t-1}.$$

- The filtering distribution of x_t given F_t. $x_t|F_t \sim N\left(\mu_{t|t}, \Sigma_{t|t}\right)$:

$$\mu_{t|t} = \mu_{t|t-1} + \Sigma_{t|t-1}\mathfrak{B}Q_{t|t-1}^{-1}\left(z_t - f_{t|t-1}\right),$$

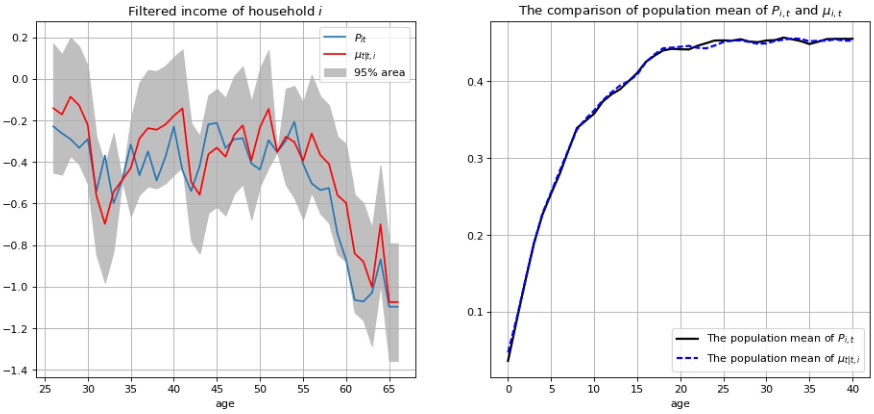
$$\Sigma_{t|t} = \Sigma_{t|t-1} - \Sigma_{t|t-1}\mathfrak{B}^2Q_{t|t-1}^{-1}\Sigma_{t|t-1}.$$

If $p > 0$, and μ is very closed to 0. We can use some threshold value to judge whether there is a shock or not. Once the shock is in presence at any time t, $\log P_t = \log Y_t - \log\mu$, in which case P_t can be directly recovered. Thus, we can set $\mu_{t|t} = \log Y_t - \log\mu$ and $\Sigma_{t|t} = 0$. Otherwise if $p > 0$, and $\mu = 0$, the income here can be treated as missing.

For the values of $\mu_{0|0} = E\left[x_0|F_0\right]$, $\Sigma_{0|0} = Var\left(x_0|F_0\right)$, since $\log Y_{it} = \log P_{it} + \log\varepsilon_{it}$, we simply assume for each household i, the initial value $\mu_{0|0} = \log Y_0 - p\mu$, where $\log Y_0$ is the population mean of income level at time 0, and accordingly $\Sigma_{0|0} = \sigma_\varepsilon^2$.

Figure 4 reports the performance of the income filter where $G_{26:29} = 1.05$, $G_{30:35} = 1.03$, $G_{36:45} = 1.01$, $G_{46:65} = 1$, $T_r = 65$, $p = 0.03$, $\mu = 10^{-6}$, $\sigma_\varsigma^2 = 0.02$,

Figure 4: The performance of income filter



$\sigma_\varepsilon^2 = 0.04$. From the left panel, the 95% area centering at the filtered mean $\mu_{t|t,i}$ and bounded by $\pm 2\Sigma_{t|t}$ can cover $P_{i,t}$ at majority of the life time. Further, the right panel shows that the difference between the population means of $\mu_{t|t}$ and $P_{i,t}$ are quite small.

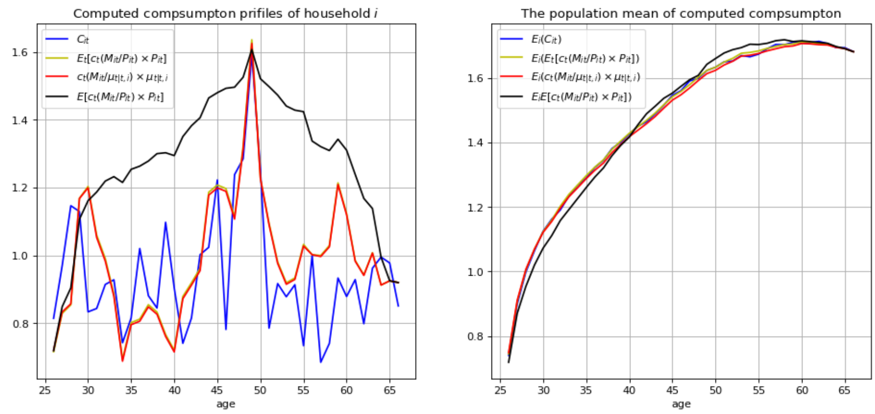
.4.5 The Comparison of Different Computations for Optimal Consumption

Here the second example in the Monte Carlo study section is used with $T_r = 65$ to compare the performance of different computation methods for the optimal consumption level $C_{i,t}$ for household i at age t. One is to simulate numerous income sample paths and compute the optimal consumption at every path at every age. At each age we collect the consumptions of all households and compute sample mean. This is the approach proposed by GP. We call it as 'GP' and it can be expressed by

$$C_{i,t}^{GP} = E\left[c_t\left(\frac{M_{i,t}^d}{P_{i,t}}\right)P_{i,t}\right] = \frac{1}{G}\sum_{g=1}^{G} c_t\left(\frac{M_{i,t}^d}{P_{i,t}^{(g)}}\right)P_{i,t}^{(g)}, \text{ for each } i,t,$$

$$E_i\left\{E\left[c_t\left(\frac{M_{i,t}^d}{P_{i,t}}\right)P_{i,t}\right]\right\} = \frac{1}{N_t}\sum_{i=1}^{N_t} G_{i,t}^{GP},$$

Figure 5: The computed consumption profiles when $N^{obs} = 1500, \rho = 2$



where $\left\{P^{(g)}\right\}_{t=26}^{66}$ is the permanent income component from $t = 26$ to $t = 66$ at g^{th} simulated income path.

The other is to treat the filtered mean $\mu_{t|t,i}$ from the Kalman income filter, as $\log P_{i,t}$, which is used by Jørgensen (2017). We call this approach as 'J' and it can also expressed by

$$C_{i,t}^{J} = c_t \left(\frac{M_{i,t}^{d}}{\mu_{t|t,i}} \right) \mu_{t|t,i}, \text{ for each } i,t,$$

$$E_i \left[c_t \left(\frac{M_{i,t}^{d}}{\mu_{t|t,i}} \right) \mu_{t|t,i} \right] = \frac{1}{N_t} \sum_{i=1}^{N_t} G_{i,t}^{J},$$

The proposed approach in equation (4.4.1) is denoted as 'L'. Given $N^{obs} = 1500$, we compare these three computation approaches, which is reported in Figure 5. The number of simulated paths for 'GP' is 1000. From the following figures, it is obvious 'GP' does not approximate the population mean of consumption profile quite well even when sample path is 1000. 'J' is close to the population mean, similar to 'L'.

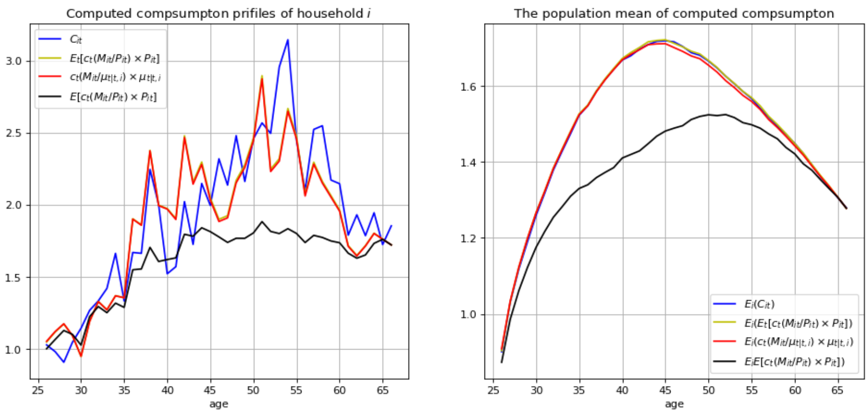
For further comparison, we use the following statistics to compare the three approaches,

$$dist = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} C_{i,t}^{d} - \frac{1}{N} \sum_{i=1}^{N} C_{i,t}^{\mathfrak{a}} \right)^2}, \mathfrak{a} = GP, J, L.$$

Table 1: The values of the statistics for three approaches

	GP	J	L
$N^{obs} = 1500$	6.5382×10^{-4}	7.2569×10^{-5}	2.8096×10^{-5}
$N^{obs} = 3000$	6.8139×10^{-4}	7.1134×10^{-5}	1.3468×10^{-5}
$N^{obs} = 6000$	2.4233×10^{-3}	6.6381×10^{-5}	9.9386×10^{-6}

Figure 6: The computed consumption profiles when $N^{obs} = 6000, \rho = 0.5$



The values of the statistics are reported in Table 1. It is apparent that 'L' has the smallest distance from the population mean of consumption profile in all cases. As the sample size increases, the distance of 'L' decreases dramatically. But the other two approaches remains the same magnitudes.

Besides, we change the value of ρ into 0.5, which is the same as GP. Following Figure 5, we draw the corresponding figures in Figure 6 which shows that 'L' is better.