2013

Three Essays on Bayesian Hypothesis Testing and Model Selection

Zeng TAO
Singapore Management University, tao.zeng.2009@smu.edu.sg

THREE ESSAYS ON BAYESIAN HYPOTHESIS TESTING AND MODEL SELECTION

TAO ZENG

SINGAPORE MANAGEMENT UNIVERSITY

2013

Three Essays on Bayesian Hypothesis Testing and Model Selection

by
Tao Zeng

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

Dissertation Committee:

Peter C.B. Phillips (Supervisor/Co-Chair)
Sterling Professor of Economics and Statistics
Yale University
Distinguished Term Professor of Economics
Singapore Management University

Jun Yu (Supervisor/Co-Chair)
Professor of Economics and Professor of Finance
Director, Sim Kee Boon Institute for Financial Economics
Singapore Management University

Sungbae An
Assistant Professor of Economics
Singapore Management University

Jun Tu
Associate professor of Finance
Singapore Management University

Singapore Management University
2013

Abstract

Three Essays on Bayesian Hypothesis Testing and Model Selection

Tao Zeng

My dissertation consists of three essays which contribute new theoretical results to Bayesian hypothesis and model selection.

Chapter 2 shows that the data augmentation technique undermines the theoretical underpinnings of the deviance information criterion (DIC), a widely used information criterion for Bayesian model comparison, although it facilitates parameter estimation for latent variable models via Markov chain Monte Carlo (MCMC) simulation. Data augmentation makes the likelihood function non-regular and hence invalidates the standard asymptotic arguments. A robust form of DIC, denoted as RDIC, is advocated for Bayesian comparison of latent variable models. RDIC is shown to be a good approximation to DIC without data augmentation. While the later quantity is difficult to compute, the expectation – maximization (EM) algorithm facilitates the computation of RDIC when the MCMC output is available. Moreover, RDIC is robust to nonlinear transformations of latent variables and distributional representations of model specification. The proposed approach is applied to several popular models in economics and finance. While DIC is very sensitive to the nonlinear transformations of latent variables in these models, RDIC is robust to these transformations. As a result, substantial discrepancy has been found between DIC and RDIC.

Chapter 3 proposes a new Bayesian approach to test a point null hypothesis based on the deviance in a decision-theoretical framework. The proposed test statistic may be regarded as the Bayesian version of likelihood ratio test and appeals in

practical applications with three desirable properties. First, it is immune to Bartlett's paradox. Second, it avoids Jeffreys-Lindley's paradox, Third, it is easy to compute and its threshold value is easily derived, facilitating the implementation in practice. The method is applied to three real examples in economics and finance. Empirical results confirm the strength of the test over the Bayes factor and reject the well-known three factor Fama-French model.

Chapter 4 proposes a Bayesian method for assess the model specification of an econometric model after it is estimated by Bayesian MCMC methods. The proposed approach does not required an alternative model be specified and is applicable to a variety of models, including latent variable models for which frequentist's methods are more difficult to use. It is shown that the proposed statistic and its threshold values are easy to compute. The method is illustrated using the Fama-French asset price model and dynamic stochastic general equilibrium (DSGE) model.

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 Sungbae An 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 YiuKuen Tse, Professor Sainan Jin, and Professor Anthony Tay for their kind helps during my stay at SMU.

I would like to give thanks to my friend Professor Yong Li for the stimulating discussions, for the sleepless nights we were working together and for his devotion to our joint works during his visit to Sim Kee Boon Institute for Financial Economics at SMU. I have learnt a lot from him about the Bayesian econometrics. And the Muay Thai training with him in Evolve is a very interesting experience.

I want to give thanks to my friend Shi Cheng and his family. In the last fifteen years, they were always supporting me and encouraging me with their best wishes.

Also I thank my dear friends Wenjian Chen, Ye Chen, Xun Dai, Jing Hu, Liang

Jiang, Shouwei Liu, Hu Lin, Xiaohu Wang, Sen Xue, Tao Yang, Yonghui Zhang, Huaxia Zeng, Haiwei Zeng, and Qiankun Zhou.

Last but not least, I wish to thank my family 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

One of the most important developments in the Bayesian literature in recent years is the deviance information criterion (DIC) of Spiegelhalter et al. (2002). DIC is a Bayesian version of the well known Akaike Information Criterion (AIC) (Akaike (1973)). Like AIC, it trades off a measure of model adequacy against a measure of complexity and is concerned about how replicate data predict the observed data. DIC is constructed based on the posterior distribution of the log-likelihood or the deviance, and has several desirable features. First, DIC is simple to calculate when the likelihood function is available in closed-form and the posterior distributions of the models are obtained by Markov chain Monte Carlo (MCMC) simulation. Second, it is applicable to a wide range of statistical models. Third, unlike Bayes factors (BFs), it can be implemented under noninformative priors.

For latent variable models, Bayesian methods via MCMC simulation have proven to be a powerful alternative to frequentist methods for estimating model parameters. In particular, the *data augmentation* strategy proposed by Tanner and Wong (1987), which expands the parameter space by treating the latent variables as additional model parameters, has been found very useful for simplifying the MCMC computation of posterior distributions. This simplification is achieved because data augmentation leads to a closed-form expression for the likelihood function.

In Chapter 2, we argue that DIC has to be used with care in the context of latent variable models. In particular, we believe DIC, as the way it is commonly implemented in practice, has some conceptual and practical problems. we advocate the use of a robust versio of DIC, denoted by RDIC, to make Bayesian comparison of latent variable models. It is shown that RDIC is a good approximation to DIC with-

out data augmentation and hence is theoretically justified. We then show that the expectation – maximization (EM) algorithm facilitates the computation of RDIC for latent variable models when the MCMC output is available. Moreover, RDIC is robust to nonlinear transformations of latent variables and to distributional representations of model specification. The advantages of the proposed approach are illustrated using three popular models in economics and finance, including a class of asset pricing models, a class of dynamic factor models and a class of stochastic volatility models.

Hypothesis testing plays a fundamental role in making statistical inference about the model specification. After models are estimated, empirical researchers would often like to test a relevant hypothesis to look for evidence to support or to be against a particular theory. An important class of hypotheses involve a single parameter value in the null.

In Chapter 3, we develop a new Bayesian hypothesis testing approach for the point null hypothesis testing. The test statistic is based on the Bayesian deviance and constructed in a decision theoretical framework. It can be regarded as the Bayesian version of the likelihood ratio test. We show that the statistic appeals in four aspects. First, it does not suffer from Bartlett's paradox and, hence, can be used under improper priors. Second, it does not suffer from Jeffreys-Lindley's paradox and, hence, can be used under vague priors. Third, it is easy to compute. Finally, the threshold values can be easily determined and are dependent on the data as well as the candidate models. To show the strength of the proposed method, we apply the test to three real examples in economics and finance. In the first example, we compare the performance of the proposed test with that of the BF in the context of CEO salary determination. It is shown that the new test is much more robust than the BF with respect to the prior. In the second example, we test the validity of the three factor Fama-French model and the new test rejects the well-known specification. Finally, we test the absence of the leverage effect in a stochastic volatility model for exchange rates and the new test suggests that there is no leverage effect in the

exchange rate series.

Economic theory has long been used to justify a particular choice of econometric models. It almost always does do by using a set of economic assumptions. When some of these assumptions are invalid, the corresponding econometric models may be misspecified. In a worse scenario, economic theory may not be available and the choice of econometric model can then be more arbitrary and, hence, the model is more vulnerable to the specification errors. Given the popularity of Bayesian MCMC methods for estimating latent variable models, it is therefore natural to introduce a Bayesian test to assess the goodness-of-fit of the model.

In Chapter 4, we introduce a Bayesian approach to testing model specification without specifying an alternative model. The proposed Bayesian test statistic is the Bayesian version of a *m*-type test. We show how to compute the test statistic from MCMC output in the context of latent variable models. To implement our method, threshold values are needed. We then show that the threshold values can be obtained using Monte Carlo simulations.

Chapter 2 Robust Deviance Information Criterion for Latent Variable Models

2.1 Introduction

One of the most important developments in the Bayesian literature in recent years is the deviance information criterion (DIC) of Spiegelhalter et al. (2002). DIC is a Bayesian version of the well known Akaike Information Criterion (AIC) (Akaike (1973)). Like AIC, it trades off a measure of model adequacy against a measure of complexity and is concerned about how replicate data predict the observed data. DIC is constructed based on the posterior distribution of the log-likelihood or the deviance, and has several desirable features. First, DIC is simple to calculate when the likelihood function is available in closed-form and the posterior distributions of the models are obtained by Markov chain Monte Carlo (MCMC) simulation. Second, it is applicable to a wide range of statistical models. Third, unlike Bayes factors (BFs), it can be implemented under noninformative priors.

An important class of models in economics and finance involves latent variables. Latent variables have figured prominently in stories about consumption decision, investment decision, labor force participation, conducts of monetary policy, indices of economic activity, inflation dynamics and other economic, business and financial activities and decisions. For example, one important class of latent variable models, the state space model, in which the state variable is latent, provides a unified

methodology for treating a wide range of problems in time series analysis. Another example can be found in the values of stocks, bonds, options, futures, and derivatives which are often determined by a small number of factors. Sometimes these factors, such as the level, the slope and the curvature in the term structure of interest rates, are not observed. In microeconometrics, discrete choices can depend on unobserved variables or there may be unobserved individual heterogeneity across economic entities.

For latent variable models, Bayesian methods via MCMC simulation have proven to be a powerful alternative to frequentist methods for estimating model parameters. In particular, the *data augmentation* strategy proposed by Tanner and Wong (1987), which expands the parameter space by treating the latent variables as additional model parameters, has been found very useful for simplifying the MCMC computation of posterior distributions. This simplification is achieved because data augmentation leads to a closed-form expression for the likelihood function.

Comparing alternative latent variable models in the Bayesian paradigm is a daunting and yet important task. The gold standard to carry out Bayesian model comparison is to compute BFs, which basically compare marginal likelihood of alternative models (Kass and Raftery (1995)). Several interesting developments have been made in recent years for computing marginal likelihood from the MCMC output; see for example, Chib (1995), Chib and Jeliazkov (2001). While these methods are very general and widely applicable, for latent variable models, they are difficult to use because the marginal likelihood may be hard to calculate. In addition, BFs cannot be used under improper priors and are subject to the Jeffrey-Lindley paradox. Given that DIC is simple to calculate from the MCMC output with the data augmentation technique and also that data augmentation is often used for Bayesian parameter estimation, DIC has been used widely for comparing alternative latent variable models; see for example, Berg et al. (2004), Huang and Yu (2010).

The first contribution of this paper is that we argue DIC has to be used with care in the context of latent variable models. In particular, we believe DIC, as the way

it is commonly implemented in practice, has some conceptual and practical problems. Firstly, DIC requires a concrete "focus" which is often not easily identified in practice. If the "focus" cannot be identified, using DIC violates the likelihood principle; see Gelfand and Trevisani (2002). Secondly, DIC is not robust to apparently innocuous transformations and distributional representations. This problem is made worse by the data augmentation technique for latent variable models. Data augmentation greatly inflates the number of parameters and hence the "effective" number of parameter used in DIC is very sensitive to transformations and distributional representations. The detail will be explained in Section 3. Finally, DIC requires that the likelihood function has a closed form expression for it to be computationally operational. For latent variable models, this is achieved by data augmentation and, as a consequence, DIC opens up to possible variations. It is unclear which variation should be used in practice; see Celeux et al. (2006) for further discussion of this problem. In this paper we argue that although data augmentation leads to a likelihood function in closed-form and greatly facilitates parameter estimation, DIC should NOT be used in connection to data augmentation. The reason is that data augmentation makes the likelihood function non-regular and hence invalidates the standard asymptotic arguments. Consequently, it undermines the theoretical underpinnings of DIC.

The source of the problem is data augmentation. With data augmentation, a closed-form expression for likelihood is ensured and it is easy to compute DIC, but the asymptotic justification of DIC is invalidated. Without data augmentation, the likelihood function does not have a closed form expression and hence DIC is much harder to compute for latent variable models, although it is asymptotically justified.

The second contribution of this paper is that we advocate the use of a robust versio of DIC, denoted by RDIC, to make Bayesian comparison of latent variable models. It is shown that RDIC is a good approximation to DIC without data augmentation and hence is theoretically justified. We then show that the expectation – maximization (EM) algorithm facilitates the computation of RDIC for latent vari-

able models when the MCMC output is available. Moreover, RDIC is robust to nonlinear transformations of latent variables and to distributional representations of model specification.

The advantages of the proposed approach are illustrated using three popular models in economics and finance, including a class of asset pricing models, a class of dynamic factor models and a class of stochastic volatility models. It is shown that DIC is very sensitive to the nonlinear transformations of latent variables in these models, whereas RDIC is robust to these transformations. As a result, substantial discrepancy is found between DIC and RDIC.

The paper is organized as follows. In Section 2, the latent variable models are introduced. The Bayesian estimation method with data augmentation and the EM algorithm are also reviewed. Section 3 reviews DIC, introduces and justifies RDIC for latent variable models, and discusses how to compute RDIC from the MCMC output. Section 4 illustrates the method using models from economics and finance. Section 5 concludes the paper. The Appendix collects the proof of the theoretical results in the paper.

2.2 Latent Variable Models, EM Algorithm and MCMC

Let $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$ denote observed variables and $\mathbf{z} = (z_1, z_2, \cdots, z_n)'$ the latent variables. The latent variable model is indexed by the a set of P parameters, $\theta = (\theta_1, \ldots, \theta_P)'$. Let $p(\mathbf{y}|\theta)$ be the likelihood function of the observed data (denoted the observed-data likelihood), and $p(\mathbf{y}, \mathbf{z}|\theta)$ be the complete-data likelihood function. The relationship between the two functions is:

$$p(\mathbf{y}|\theta) = \int p(\mathbf{y}, \mathbf{z}|\theta) d\mathbf{z}. \tag{2.2.1}$$

In many cases, the integral does not have a closed-form solution. Consequently, statistical inferences, such as estimation and model comparison, are difficult to make. In the literature, maximum likelihood (ML) analysis using the EM algorithm and

Bayesian analysis using MCMC are two popular approaches for carrying out statistical inference of the latent variable models.

2.2.1 Maximum likelihood via the EM algorithm

The EM algorithm is an iterative numerical method for finding the ML estimates of θ in the latent variable models. It has been widely used in applications since Dempster et al. (1977) gave its name and did the convergence analysis. In this subsection, we briefly review the main idea of the EM algorithm. For more details, one can refer to McLachlan and Krishnan (2008).

Let $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ be the complete data with a density $p(\mathbf{x}|\theta)$ parameterized by a P-dimension parameter vector $\theta \in \Theta \subseteq R^P$. The observed-data log-likelihood $\mathscr{L}_o(\mathbf{y}|\theta) = \ln p(\mathbf{y}|\theta)$ often involves some intractable integral, preventing researchers from directly optimizing $\mathscr{L}_o(\mathbf{y}|\theta)$ with respect to θ. In many cases, however, the complete-data log-likelihood $\mathscr{L}_c(\mathbf{x}|\theta) = \ln p(\mathbf{x}|\theta)$ has a closed-form expression. Instead of maximizing $\mathscr{L}_o(\mathbf{y}|\theta)$ directly, the EM algorithm maximizes $\mathscr{Q}(\theta|\theta^{(r)})$, the conditional expectation of the complete-data log-likelihood function $\mathscr{L}_c(\mathbf{x}|\theta)$ given the observed data $\mathbf{y}$ and a current fit $\theta^{(r)}$ of the parameter.

Generally, a standard EM algorithm has two steps: the *expectation* (E) step and the *maximization* (M) step. The E-step evaluates

$$\mathscr{Q}(\theta|\theta^{(r)}) = E_{\mathbf{z}}\{\mathscr{L}_c(\mathbf{x}|\theta)|\mathbf{y}, \theta^{(r)}\}, \tag{2.2.2}$$

where the expectation is taken with respect to the conditional distribution $p(\mathbf{z}|\mathbf{y}, \theta^{(r)})$. The M-step determines a $\theta^{(r+1)}$ that maximizes $\mathscr{Q}(\theta|\theta^{(r)})$. Under some mild regularity conditions, the sequence $\{\theta^{(r)}\}$ obtained from the EM iterations converges to the ML estimate $\widehat{\theta}$; see Dempster et al. (1977) and Wu (1983) for details about the convergence properties of $\{\theta^{(r)}\}$.

2.2.2 Bayesian analysis using MCMC

Although the EM algorithm is a reasonable statistical approach for analyzing latent variable models, the numerical optimization in the M-step is often unstable. This numerical problem worsens as the dimension of θ increases. It is well recognized that Bayesian methods using MCMC provide a powerful tool to analyze the latent variables models. However, if the posterior analysis is conducted from the observed-data likelihood, $p(\mathbf{y}|\theta)$, one would end up with the same problem as in the ML method as $p(\mathbf{y}|\theta)$ does not have a closed-form expression.

The novelty in the Bayesian methods is to treat the latent variable model as a hierarchical structure of conditional distributions, namely, $p(\mathbf{y}|\mathbf{z},\theta)$, $p(\mathbf{z}|\theta)$, and $p(\theta)$. In other words, one can use the data augmentation strategy of Tanner and Wong (1987) to expand the parameter space from θ to (θ,$\mathbf{z}$). The advantage of data augmentation is that the Bayesian analysis is now based on the new likelihood function, $p(\mathbf{y}|\theta,\mathbf{z})$ which often has a closed-form expression. Then the Gibbs sampler and other MCMC samplers can be used to generate random samples from the joint posterior distribution $p(\theta,\mathbf{z}|\mathbf{y})$. After a sufficiently long period for a burning-in phase, the simulated random samples can be regarded as random observations from the joint distribution. The statistical analysis can be established on the basis of these simulated posterior random observations. As a by-product to the Bayesian analysis, one also obtains Markov chains for the latent variables $\mathbf{z}$ and hence statistical inference can be made about $\mathbf{z}$. For further details about Bayesian analysis of latent variable models via MCMC, including algorithms, examples and references, see Geweke et al. (2011). From the above discussion, it can be seen that data augmentation is the key technique for Bayesian estimation of latent variable models.

Two observations are in order. First, with data augmentation, the parameter space is much bigger. More than often, the dimension of the space increases as the number of observations increases and is larger than the number of observations. In the latter case, the new likelihood function becomes non-regular. Second, it

is difficult to argue that the latent variables can be always treated as the model parameters. Models parameters are typically fixed but the latent variables are often time varying. Consequently, the same treatment of these two types of variables does not seem to be justifiable from the perspective of model selection.

2.3 Bayesian Comparison of Latent Variable Models

2.3.1 DIC

Spiegelhalter et al. (2002) proposed DIC for Bayesian model comparison. The criterion is based on the deviance given by:

$$D(\theta) = -2\ln p(\mathbf{y}|\theta) + 2\ln f(\mathbf{y}),$$

where $f(\mathbf{y})$ is some fully specified standardizing term that is a function of the data alone. Based on the deviance, DIC takes the form of:

$$\text{DIC} = \overline{D(\theta)} + P_D. \tag{2.3.1}$$

The first term, used as a Bayesian measure of model fit, is defined as the posterior expectation of the deviance, that is,

$$\overline{D(\theta)} = E_{\theta|\mathbf{y}}[D(\theta)] = E_{\theta|\mathbf{y}}[-2\ln p(\mathbf{y}|\theta)].$$

The better the model fits the data, the larger the log-likelihood value and hence the smaller the value for $\overline{D(\theta)}$. The second term, used to measure the model complexity and also known as "effective number of parameters", is defined as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters:

$$P_D = \overline{D(\theta)} - D(\bar{\theta}) = -2\int [\ln p(\mathbf{y}|\theta) - \ln p(\mathbf{y}|\bar{\theta})]p(\theta|\mathbf{y})\mathrm{d}\theta, \tag{2.3.2}$$

where $\bar{\theta}$ is the Bayesian estimator, and more precisely the posterior mean, of the parameter θ. Here, P_D can be explained as the expected excess of the true over the estimated residual information conditional on data $\mathbf{y}$. In other words, P_D can be interpreted as the expected reduction in uncertainty due to estimation.

Note that DIC can be rewritten by two equivalent forms:

$$\text{DIC} = D(\bar{\theta}) + 2P_D, \tag{2.3.3}$$

and

$$\text{DIC} = 2\overline{D(\theta)} - D(\bar{\theta}) = -4E_{\theta|\mathbf{y}}[\ln p(\mathbf{y}|\theta)] + 2\ln p(\mathbf{y}|\bar{\theta}). \tag{2.3.4}$$

DIC defined in Equation (2.3.3) bears similarity to AIC of Akaike (1973) and can be interpreted as a classical "plug-in" measure of fit plus a measure of complexity. In Equation (2.3.1) the Bayesian measure, $\overline{D(\theta)}$, is the same as $D(\bar{\theta}) + P_D$ which already includes a penalty term for model complexity and thus could be better thought of as a measure of model adequacy rather than pure goodness of fit.

Remark 2.3.1 *The asymptotic justification of DIC requires that the candidate models nest the true model and that the posterior distribution is approximately normal. These two requirements parallel to those in AIC where the candidate models nest the true model and the ML estimator is asymptotically normally distributed. To see the importance of the asymptotic normality, Spiegelhalter et al. (2002) show that, when the prior is noninformative, P_D is approximately the same as P. In this case DIC is explained as Bayesian version of AIC. However, if the asymptotic normality does not hold true, P_D cannot be approximated by P and DIC is not the Bayesian version of AIC. Furthermore, the decision-theoretical explanation of DIC requires the asymptotic normality of the Bayesian posterior be held true.*

Remark 2.3.2 *If $p(\mathbf{y}|\theta)$ has a closed-form expression, DIC is trivially computable from the MCMC output. This is in sharp contrast to BFs and some other model selection criteria within the classical framework. The computational tractability,*

together with the versatility of MCMC and the fact that DIC is incorporated into a Bayesian software, WinBUGS, allows DIC to enjoy a very wide range of applications.[1] *However, if $p(\mathbf{y}|\boldsymbol{\theta})$ is not available in closed form, such as in random effects models and state space models, computing DIC may become infeasible, or at least, very time consuming.*

Remark 2.3.3 *When an information criterion is used for model selection, the degrees of freedom are typically used to measure the model complexity. In the Bayesian framework, the prior information almost always imposes additional restrictions on the parameter space and hence the degrees of freedom may be reduced by the prior information. A useful contribution of DIC is to provide a way to measure the model complexity when the prior information is incorporated; see Brooks (2002).*

Remark 2.3.4 *Unlike BFs that address how observed data are predicted by the priors, DIC "addresses how well the posterior might predict future data generated by the same mechanism that gave rise to the observed data" (Spiegelhalter et al. (2002)). This predictive perspective for selecting a good model is important in many practical business, economic, and financial decisions.*

Remark 2.3.5 *As acknowledged in Spiegelhalter et al. (2002), DIC requires a concrete specification of a "focus". In the context of random effects models, Vaida and Blanchard (2005) pointed out that the likelihood function used for information criterion depends on the "focus". A different "focus" leads to a different AIC and DIC. In practice, however, the choice of a "focus" is not always easy. Unfortunately, it is well known that Bayesian decisions may depend on the choice of the "focus". For example, in Section 8.2 of Spiegelhalter et al. (2002), where Models 4 and 5 are predictively identical but their DIC values are quite different. In this example, it is unclear what should be the right "focus". The same difficulty also shows up in Model 8 of Berg et al. (2004). If the "focus" is not identified, DIC suffers from*

[1] As of July 8, 2012, Spiegelhalter et al. (2002) has been cited 3396 times according to Google Scholar and 1,984 time according to Science Citation Index.

an incoherent inference problem. That is, when one model is a distributional representation of another model and the same prior is used in the two models, they have different DIC values. For further illustrations of the problem, see Gelfand and Trevisani (2002) and Daniels and Hogan (2008).

For latent variable models, there are alternative ways to define DIC, as discussed in Celeux et al. (2006) (see also, DeIorio and Robert (2002)), two of which are especially important. First, DIC is based on the observed-data likelihood and denoted by DIC_1 in Celeux et al. (2006) as,

$$\text{DIC}_1 = -4E_{\theta|\mathbf{y}}[\ln p(\mathbf{y}|\theta)] + 2\ln p(\mathbf{y}|\bar{\theta}). \tag{2.3.5}$$

For certain mixture models, such as scale mixtures of normals of Andrews and Mallows (1974) , the observed-data likelihood $p(\mathbf{y}|\theta)$ is available in closed form. In this case, DIC_1 is trivially obtained, although its value depends on the choice of the "focus", namely, the hierarchical structure here.

However, for state-space models, including linear Gaussian state space models, the observed-data likelihood $p(\mathbf{y}|\theta)$ is not available in closed form.[2] In this case, computing DIC_1 from the MCMC output is time consuming or even infeasible since $p(\mathbf{y}|\theta)$ has to be computed at each draw from the Markov chain.

Second, DIC is defined based on the data augmentation technique, treating $\mathbf{z}$ as the additional parameters, and denoted by DIC_7 in Celeux et al. (2006) as,

$$\text{DIC}_7 = -4E_{\theta,\mathbf{z}|\mathbf{y}}[\ln p(\mathbf{y}|\mathbf{z},\theta)] + 2\ln p(\mathbf{y}|\bar{\mathbf{z}},\bar{\theta})]. \tag{2.3.6}$$

The corresponding P_D is

$$P_D = -2\int [\ln p(\mathbf{y}|\mathbf{z},\theta) - \ln p(\mathbf{y}|\bar{\mathbf{z}},\bar{\theta})]p(\mathbf{z},\theta|\mathbf{y})\mathrm{d}\mathbf{z}\mathrm{d}\theta. \tag{2.3.7}$$

[2]For linear Gaussian state space models, to do ML, the Kalman filter can be used to obtain the likelihood function numerically. Numeircally more efficient algorithms have been developed in the recent literature; see for example, Chan and Jeliazkov (2009).

For most state space models, including the nonlinear non-Gaussian state space models, $p(\mathbf{y}|\mathbf{z},\theta)$ is available in closed form and hence computing DIC_7 is straightforward.

Remark 2.3.6 *For all random effects models and state space models, applied researchers always calculate DIC based on DIC_7 in (2.3.6) which is also implemented in WinBUGS. Examples that use DIC_7 in applications include Berg et al. (2004) and Wang et al. (2011). Clearly this choice of defining DIC is simple for computational convenience.*

Remark 2.3.7 *From a theoretical viewpoint, DIC_7 has a couple of serious problems. First, due to the data augmentation, the number of the latent variables often increases with the sample size in latent variable models, causing the problem of a non-regular likelihood-based statistical inference; see Gelman (2003). This invalidates the asymptotic justification of DIC because the standard asymptotic theory derived from regular likelihood is not applicable to non-regular likelihood. Second, due to the data augmentation, the dimension of the parameter space becomes larger and hence we expect that DIC_7 is more sensitive to transformations of latent variables than DIC_1.*

To illustrate the second problem, we consider a simple transformation of latent variables in the well-known Clark model (Clark (1973)) which is given by,

$$\text{Model 1}: y_t \sim N(\mu, \exp(h_t)), h_t \sim N(0,\sigma^2), t = 1,\cdots,n. \tag{2.3.8}$$

An equivalent representation of the model is

$$\text{Model 2}: y_t \sim N(\mu,\sigma_t^2), \sigma_t^2 \sim LN(0,\sigma^2), t = 1,\cdots,n, \tag{2.3.9}$$

where *LN* denotes the log-normal distribution. In Model 2 the latent variable is the volatility σ_t^2, while the latent variable is the logarithmic volatility $h_t = \ln\sigma_t^2$ in

Model 1. Suppose the parameters of interest are μ and σ^2. With the same "focus", the two models are identical and hence are expected to have the same DIC and P_D. To calculate the P_D component in DIC_7, we simulate 1000 observations from the model with $\mu = 0, \sigma^2 = 0.5$. Vague priors are selected for the two parameters, namely, $\mu \sim N(0, 100)$, $\sigma^{-2} \sim \Gamma(0.001, 0.001)$. We run Gibbs sampler to make 240,000 simulated draws from the posterior distributions. The first 40,000 are discarded as burn-in samples. The remaining observations with every 10th observation are collected as effective observations for statistical inference. With the data augmentation, the latent variables, h_t and σ_t^2 are regarded as parameters, and we find that $P_D = 89.806$ for Model 1 but $P_D = 59.366$ for Model 2. The difference is very significant. Given that we have the identical models and priors, and use the same dataset, the vast difference suggests that DIC_7 and the corresponding P_D are very sensitive to transformations of latent variables.

For latent variable models, DIC_1 does not suffer from the same theoretical problem as DIC_7. However, computing DIC_1 from the MCMC output is much harder, if not infeasible, since $p(\mathbf{y}|\theta)$ is not available in closed-form and computing $E_{\theta|\mathbf{y}}[\ln p(\mathbf{y}|\theta)]$ necessitates numerical calculation of $p(\mathbf{y}|\theta)$ at each draw from the Markov chain.

To summarize the problems with DIC in the context of latent variable models, while DIC_7 is trivial to calculate but cannot be theoretically justified, DIC_1 is theoretically justified but infeasible to compute.

2.3.2 RDIC

In this section we introduce a robust version of DIC, denoted as RDIC, as follows

$$\text{RDIC} = D(\bar{\theta}) + 2\mathbf{tr}\left\{\mathbf{I}(\bar{\theta})V(\bar{\theta})\right\} = D(\bar{\theta}) + 2P_D^*, \tag{2.3.10}$$

where

$$P_D^* = \mathbf{tr}\left\{\mathbf{I}(\bar{\theta})V(\bar{\theta})\right\}, \tag{2.3.11}$$

with **tr** denoting the trace of a matrix,

$$\mathbf{I}(\theta) = -\frac{\partial^2 \ln p(\mathbf{y}|\theta)}{\partial\theta\partial\theta'}, V(\bar{\theta}) = E\left[\left(\theta-\bar{\theta}\right)\left(\theta-\bar{\theta}\right)'|\mathbf{y}\right].$$

Interestingly, in Equation (15) on Page 590 Spiegelhalter et al. (2002) obtained the expression for P_D^* and claimed that P_D^* approximates the P_D component in DIC_1. Unfortunately, to the best of our knowledge, P_D^* has never been implemented in practice and WinBUGS does not report P_D^*. Moreover, the proof of $P_D^* \approx P_D$ was not given in Spiegelhalter et al. (2002). The conditions under which $P_D^* \approx P_D$ holds true were not specified. The order of the approximation remains unknown.

To justify the choice of RDIC, we will have to establish conditions under which we can show that RDIC approximates DIC_1 and P_D^* approximates P_D that corresponds to DIC_1 with a known order of magnitude. We then show that how the EM algorithm facilitates the computation of RDIC from the MCMC output for latent variable models.

Let $L_n(\theta) = \ln p(\theta|\mathbf{y}), L_n^{(1)}(\theta) = \partial \ln p(\theta|\mathbf{y})/\partial\theta, L_n^{(2)}(\theta) = \partial^2 \ln p(\theta|\mathbf{y})/\partial\theta\theta'$. In this paper, we impose the following regularity conditions.

Assumption 1: There exists a finite sample size n^*, for $n > n^*$, there is a local maximum at $\hat{\theta}_m$ so that $L_n^{(1)}\left(\hat{\theta}_m\right) = 0$ and $L_n^{(2)}\left(\hat{\theta}_m\right)$ is a negative definite matrix. Obviously, $\hat{\theta}_m$ is the posterior mode.

Assumption 2: The largest eigenvalue of $\left[-L_n^{(2)}(\hat{\theta}_m)\right]^{-1}$, σ_n^2, 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 $\theta \in H\left(\hat{\theta}_m, \delta\right) = \left\{\theta : ||\theta - \hat{\theta}_m|| \leq \delta\right\}$, $L_n^{(2)}(\theta)$ exists and satisfies

$$-A(\varepsilon) \leq L_n^{(2)}(\theta) L_n^{-(2)}\left(\hat{\theta}_m\right) - \mathbf{I}_P \leq A(\varepsilon),$$

where $\mathbf{I}_P$ is a $P \times P$ identity matrix, $A(\varepsilon)$ a $P \times P$ semi-definite symmetric matrix whose largest eigenvalue goes to zero as $\varepsilon \to 0$.

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

$$\int_{\Theta - H(\hat{\theta}_m, \delta)} p(\theta|\mathbf{y}) d\theta \to 0,$$

where Θ is the support of θ.

Assumption 5: For any $\delta > 0$, as $n \to \infty$, when $\theta \in H\left(\hat{\theta}_m, \delta\right)$, conditional on the observed data $\mathbf{y}$, $L_n^{(2)}(\theta)/n = O(1)$.

Assumption 6: The likelihood information dominates the prior information, that is, when the sample size goes to infinity, the prior information can be ignored.

Assumption 7: Under the true model, for any $\delta > 0$, as $n \to \infty$, when $\theta \in H\left(\hat{\theta}_m, \delta\right)$, then, $\frac{1}{n}\frac{\partial^3 p(\mathbf{y}|\theta)}{\partial\theta\partial\theta\partial\theta} = O_p(1)$

Lemma 2.3.1 *Under Assumptions 1-5, conditional on the observed data* $\mathbf{y}$*, we have*

$$\begin{aligned}
&\bar{\theta} = E\left[\theta|\mathbf{y}\right] = \hat{\theta}_m + o(n^{-1/2}), \\
&V\left(\hat{\theta}_m\right) = E\left[\left(\theta - \hat{\theta}_m\right)\left(\theta - \hat{\theta}_m\right)' |\mathbf{y}\right] = -L_n^{-(2)}\left(\hat{\theta}_m\right) + o(n^{-1}).
\end{aligned}$$

Remark 2.3.8 *Lemma 2.3.1 establishes Bayesian large sample theory. The regularity conditions 1-4 have been used in the literature to develop Bayesian large sample theory for stationary and nonstationary dynamic models and nondynamic models; see, for example, Chen (1985), Kim (1994), Kim (1998), Geweke (2005). The Bayesian large sample theory was also developed from different sets of regularity conditions in different contexts. For example, Ghosh and Ramamoorthi (2003) developed the asymptotic posterior normality and Lemma 2.3.1 in the iid case.*

Theorem 2.3.1 *Under Assumptions 1-6, it can be shown that, conditional on the observed data* $\mathbf{y}$*,*

$$P_D = P_D^* + o(1),\ DIC_1 = RDIC + o(1),$$

where P_D *is defined in (2.3.2).*

Remark 2.3.9 *Theorem 4.3.1 improves Equation (15) Spiegelhalter et al. (2002) in*

two ways. First, it gives the order of the approximation errors. Second, it specifies the conditions under which P_D approximates P_D^* and DIC_1 approximates RDIC.

Remark 2.3.10 *As DIC_1 is theoretically justified for the latent variable models, Theorem 4.3.1 justifies RDIC asymptotically since RDIC and DIC_1 are asymptotically equivalent.*

Remark 2.3.11 *RDIC maintains all the good features of DIC_1. For example, RDIC incorporates the prior information when measuring the model complexity. As shown in Spiegelhalter et al. (2002),*

$$I\left(\hat{\theta}_m\right) = -\left\{\frac{\partial^2 \ln p(\theta|\mathbf{y})}{\partial\theta\partial\theta'} - \frac{\partial^2 \ln p(\theta)}{\partial\theta\partial\theta'}\right\}|_{\theta=\hat{\theta}_m} = -L_n^{(2)}(\hat{\theta}_m) - \left\{-\frac{\partial^2 \ln p(\theta)}{\partial\theta\partial\theta'}\right\}|_{\theta=\hat{\theta}_m}.$$

Under Assumption 1-5, following Lemma 2.3.1 and the proof of Theorem 4.3.1, we get

$$\begin{aligned} P_D^* &= \mathbf{tr}\left\{I(\hat{\theta}_m)V(\bar{\theta})\right\} + o(1) \\ &= \mathbf{tr}\left\{\left[-L_n^{(2)}(\hat{\theta}_m) - \left\{-\frac{\partial^2 \ln p(\theta)}{\partial\theta\partial\theta'}\right\}|_{\theta=\hat{\theta}_m}\right] V(\bar{\theta})\right\} + o(1) \\ &= \mathbf{tr}\left\{-L_n^{(2)}(\hat{\theta}_m)V(\bar{\theta})\right\} - \mathbf{tr}\left\{\left[-\frac{\partial^2 \ln p(\theta)}{\partial\theta\partial\theta'}|_{\theta=\hat{\theta}_m}\right] V(\bar{\theta})\right\} + o(1) \\ &= P - \mathbf{tr}\left\{\left[-\frac{\partial^2 \ln p(\theta)}{\partial\theta\partial\theta'}|_{\theta=\hat{\theta}_m}\right] V(\bar{\theta})\right\} + o(1). \end{aligned} \tag{2.3.12}$$

From (2.3.12), it can be seen clearly that the prior information can reduce the model complexity.

Remark 2.3.12 *Conditional on the observed data* $\mathbf{y}$*, when the likelihood information dominates the prior information (say, for example, if $-\partial^2 \ln p(\theta)/\partial\theta\theta'|_{\theta=\hat{\theta}_m} = O(1)$), from (2.3.12) it can be shown that $P_D = P_D^* + o(1) = P + o(1)$. In addition, as $n \to \infty$ the posterior mode $\hat{\theta}_m$ is reduced to the ML estimator $\hat{\theta}$. Hence,*

$$\ln p(\mathbf{y}|\bar{\theta}) = \ln p(\mathbf{y}|\hat{\theta}) - \frac{1}{2}(\bar{\theta} - \hat{\theta})' I(\tilde{\theta})(\bar{\theta} - \hat{\theta}),$$

where $\tilde{\theta}$ lies in the segment between $\bar{\theta}$ and $\hat{\theta}$. Using Assumption 5 and Lemma 2.3.1, we can show that $\ln p(\mathbf{y}|\bar{\theta}) = \ln p(\mathbf{y}|\hat{\theta}) + o(1)$. *Consequently,*

$$DIC_1 = RDIC + o(1) = -2\ln p(\mathbf{y}|\hat{\theta}) + 2P + o(1) = AIC + o(1).$$

Namely, both RDIC and DIC_1 can be regarded as the Bayesian version of AIC.

Remark 2.3.13 *Like DIC_1, RDIC is justified by the standard Bayesian large sample theory. When the Bayesian large sample theory is not available, RDIC is not justified. These include models in which the number of the parameters increases with the sample size, under-identified models, models with an unbounded likelihood, and models with improper posterior distributions. For more details about the standard Bayesian large sample theory, see Gelman (2003) and Geweke (2005). For the latent variable models, since the number of the latent variables increases with sample size, the standard Bayesian large sample theory is not applicable if the data augmentation technique is used. As a result, when calculating RDIC, data augmentation should NOT be used.*

Remark 2.3.14 *Since RDIC is defined from the observed-data likelihood $p(\mathbf{y}|\theta)$, there is no need to specify a "focus", and hence, RDIC does not suffer from the incoherent inference problem.*

Remark 2.3.15 *For the latent variable models, while the number of the model parameters (P) is fixed and usually not so big, the number of the latent variables increases as the sample size increases. In the definition of RDIC, the latent variables are not regarded as the parameters. Consequently, the problem of parameter transformation is less serious. For example, in the Clark model, with the same setting as before, we get $P_D^* = 1.75$ for Model 1 and $P_D^* = 1.80$ for Model 2. There is no significant difference between them. Moreover, these two values are close to 2, that is the actual number of parameters. This is what we expected given that the vague priors are used and hence $P_D^* \approx P = 2$.*

Remark 2.3.16 *An obvious computational advantage in RDIC is that P_D^* does not involve inverting a matrix. This advantage is not so important when the latent variable model only has a small number of parameters. However, for high dimensional latent variable models where there are many parameters, this computational advantage may be important.*

Suppose a loss function, when using the observed data $\mathbf{y}$ to predict a future replicate dataset, $\mathbf{y}_{rep}$, in a model, is given by $\mathscr{L}(\mathbf{y}_{rep}, \mathbf{y})$. From the decision-theoretic viewpoint, a desirable model selection criterion should choose a model to minimize the risk function, $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\mathscr{L}(\mathbf{y}_{rep}, \mathbf{y})$. The following theorem provides the justification of RDIC from the decision-theoretic viewpoint.

Theorem 2.3.2 *Let $\mathbf{y}_{rep} = (\mathbf{y}_{1,rep}, \mathbf{y}_{2,rep}, \cdots, \mathbf{y}_{n,rep})$ be the future data generated by the same mechanism that gives rise to the observed data $\mathbf{y}$, i.e, $p(\mathbf{y}_{rep}) = p(\mathbf{y})$. The predictive distribution is $p(\mathbf{y}_{rep}|\mathbf{y}) = \int p(\mathbf{y}_{rep}|\theta)p(\theta|\mathbf{y})d\theta$. If $\mathscr{L}(\mathbf{y}_{rep}, \mathbf{y}) = -2\ln p(\mathbf{y}_{rep}|\mathbf{y})$, it can be shown that, conditional on the observed data $\mathbf{y}$ and under Assumptions 1-7,*

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\mathscr{L}(\mathbf{y}_{rep}, \mathbf{y}) = E_{\mathbf{y}}DIC_1 + o(1) = E_{\mathbf{y}}RDIC + o(1).$$

Remark 2.3.17 *Spiegelhalter et al. (2002) gave a heuristic explanation to why DIC is an approximate estimator of a loss function. In Theorem 3.3.2, we complement Spiegelhalter et al. (2002) by providing a formal decision-theoretical justification to DIC_1 and RDIC.*

Remark 2.3.18 *RDIC and DIC_1 are both an unbiased estimator of the risk function asymptotically.*

Remark 2.3.19 *Like DIC_1, RDIC addresses how well the posterior may predict future data generated by the same mechanism that gives rise to the observed data. This posterior predictive feature could be appealing in many applications.*

Remark 2.3.20 *Like AIC, both DIC_1 and RDIC require the candidate models nest the true model. This is of course a strong assumption. Under the iid case, Ando and Tsay (2010) relaxed this assumption and obtained a predictive likelihood information criterion (BPIC) that minimizes the loss function $\eta = E_{\mathbf{y}}E_{\mathbf{y}_{rep}}[-\log p(\mathbf{y}_{rep}|\mathbf{y})]$. The estimator of η is given by*

$$\hat{\eta} = -\log p(\mathbf{y}_{rep}|\mathbf{y})|_{\mathbf{y}_{rep}=\mathbf{y}} + \frac{1}{2}\mathbf{tr}\left[I^{-1}(\hat{\theta})J(\hat{\theta})\right],$$

where $I(\theta)$ and $J(\theta)$ are the Hessian matrix and the Fisher information matrix. In Ando (2007), another BPIC was given as

$$BPIC = -\log p(\mathbf{y}|\hat{\theta}) + \mathbf{tr}[I^{-1}(\hat{\theta})J(\hat{\theta})] + P/2.$$

Ando (2007) showed that BPIC is an estimator of the loss function

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-\int \log p(\mathbf{y}_{rep}|\theta)p(\theta|\mathbf{y})d\theta\right].$$

Like TIC of Takeuchi (1976), these two information criteria involve the inverse of Hessian matrix which is numerically changing when the dimension of the parameter space is large. This is one of the reasons why TIC has not been widely used in practice. Furthermore, the derivation of these two information criteria requires the data be iid. For data in economic and finance, this requirement is often too restrictive. In addition, for many latent variable models, the maximum likelihood estimator, the Hessian matrix and the Fisher information matrix are difficult to obtain. How to develop a good information criterion for comparing latent variable models, without assuming the candidate models nest the true model, will be pursued in future research.

Remark 2.3.21 *It is easy to verify that Assumptions 1-7 hold true for nondynamic models or stationary dynamic models. Hence, Lemma 2.3.1 and Theorem 4.3.1 are applicable to these models. For unit root models, Kim (1994) and Kim (1998)*

showed that the asymptotic normality of posterior distribution can be established. Hence, Lemma 2.3.1 is applicable to models with a unit root. Unfortunately, it is critical for developing Theorem 3.3.2 to require that $\mathbf{y}_{rep}$ *and* $\mathbf{y}$ *have the same data generating process. Hence, it does not hold true for models with a unit or explosive root due to the initial condition. Consequently, Theorem 3.3.2 is not applicable to models with unit or explosive roots. This topic on comparing non-stationary statistical models will be pursued in future studies. Within the classical framework, Phillips and Ploberger (1996) and Phillips (1996) have proposed model selection criteria for models without latent variables.*

Remark 2.3.22 *If the observed-data likelihood function,* $p(\mathbf{y}|\theta)$*, does not have a closed-from expression, its second derivative,* $\partial^2 \log p(\mathbf{y}|\theta)/\partial\theta\partial\theta'$ *and hence RDIC will be difficult to compute. In the following section, we show how the EM algorithm may be used to compute the second derivative and RDIC.*

2.3.3 Computing RDIC by the EM algorithm

The definition of RDIC clearly requires the evaluation of observed-data likelihood at the posterior mean, $p(\mathbf{y}|\bar{\theta})$, as well as the information matrix and the second derivative of the observed-data likelihood function. For most latent variable models, the observed-data likelihood function does not have a closed-from expression. In this section we show how the EM algorithm may be used to evaluate $p(\mathbf{y}|\bar{\theta})$, the second derivative of the observed-data likelihood function, and hence RDIC for the latent variable models. It is important to point out that we do not need to numerically optimize any function here as in the EM algorithm. Consequently, our method is not subject to the instability problem found in the *M*-step.

Lemma 2.3.2 *For any* θ *and* θ^* *in* Θ*, let* $\mathscr{H}(\theta|\theta^*) = \int \ln p(\mathbf{z}|\mathbf{y},\theta)p(\mathbf{z}|\mathbf{y},\theta^*)d\mathbf{z}$*, the so-called* $\mathscr{H}$ *function in the EM algorithm. It was shown in Dempster et al. (1977) that*

$$\mathscr{L}_o(\mathbf{y},\theta) = \mathscr{Q}\left(\theta|\theta^*\right) - \mathscr{H}\left(\theta|\theta^*\right),$$

where the $\mathscr{Q}$ function is defined in Equation (2.2.2).

Following Lemma 4.3.1, the Bayesian plug-in model fit, $\ln p(\mathbf{y}|\bar{\theta})$, may be obtained as

$$\ln p(\mathbf{y}|\bar{\theta}) = \mathscr{Q}(\bar{\theta}|\bar{\theta}) - \mathscr{H}(\bar{\theta}|\bar{\theta}). \tag{2.3.13}$$

It can be seen that even when $\mathscr{Q}(\bar{\theta}|\bar{\theta})$ is not available in closed form, it is easy to evaluate from the MCMC output because

$$\mathscr{Q}(\bar{\theta}|\bar{\theta}) = \int \ln p(\mathbf{y},\mathbf{z}|\bar{\theta}) p(\mathbf{z}|\mathbf{y},\bar{\theta}) \mathrm{d}\mathbf{z} \approx \frac{1}{M} \sum_{m=1}^{M} \ln p\left(\mathbf{y}, \mathbf{z}^{(m)}|\bar{\theta}\right).$$

where $\{\mathbf{z}^{(m)}, m = 1,2,\cdots,M\}$ are random observations drawn from the posterior distribution $p(z|\mathbf{y},\bar{\theta})$.

For the second term in (2.3.13), if $p(\mathbf{z}|\mathbf{y},\bar{\theta})$ is a standard distribution, $\mathscr{H}(\bar{\theta}|\bar{\theta})$ can be easily evaluated from the MCMC output as

$$\mathscr{H}(\bar{\theta}|\bar{\theta}) = \int \ln p(\mathbf{z}|\mathbf{y},\bar{\theta}) p(\mathbf{z}|\mathbf{y},\bar{\theta}) d\mathbf{z} \approx \frac{1}{M} \sum_{m=1}^{M} \ln p\left(\mathbf{z}^{(m)}|\mathbf{y},\bar{\theta}\right).$$

However, if $p(\mathbf{z}|\mathbf{y},\bar{\theta})$ is not a standard distribution, an alternative approach has to be used, depending on the specific model in consideration. We now consider two situations.

First, if the complete-data $(\mathbf{y}_i,\mathbf{z}_i)$ are independent with $i \neq j$, and $\mathbf{z}_i$ is of low-dimension, say ≤ 5, then a nonparametric approach may be used to approximate the posterior distribution $p(\mathbf{z}|\mathbf{y},\theta)$. Note that

$$\mathscr{H}(\theta|\theta) = \int \ln p(\mathbf{z}|\mathbf{y},\theta)\pi(\mathbf{z}|\mathbf{y},\theta) d\mathbf{z} = \sum_{i=1}^{n} \int \ln p(\mathbf{z}_i|\mathbf{y}_i,\theta)\pi(\mathbf{z}_i|\mathbf{y},\theta) d\mathbf{z}_i = \sum_{i=1}^{n} \mathscr{H}_i(\theta|\theta).$$

The computation of $\mathscr{H}_i(\theta|\theta)$ requires an analytic approximation to $p(\mathbf{z}_i|\mathbf{y}_i,\theta)$ which can be constructed using a nonparametric method. In particular, MCMC allows one to draw some effective samples from $p(\mathbf{z}_i|\mathbf{y}_i,\theta)$. Using these random samples, one can then use nonparametric techniques such as the kernel-based methods to approxi-

mate $p(\mathbf{z}_i|\mathbf{y}_i,\theta)$. In a recent study, Ibrahim et al. (2008) suggested using a truncated Hermite expansion to approximate $p(\mathbf{z}_i|\mathbf{y}_i,\theta)$.

As a simple illustration, we apply this method to the Clark model. When the Gaussian kernel method is used, we get $\ln p(\mathbf{y}|\bar{\theta}) = -1448.97$, RDIC$=$ 2901.46 for Model 1 and $\ln p(\mathbf{y}|\bar{\theta}) = -1449.41$, RDIC$=$ 2902.42 for Model 2. These two sets of numbers are nearly identical. However, if the latent variable models are regarded as parameters, we get $\text{DIC}_7 = 2884.37$ for Model 1 and $\text{DIC}_7 = 2852.85$ for Model 2. The highly distinctive difference between them suggests that DIC_7 is not a reliable model selection criterion for the model. Note that DIC_1 is not really feasible to compute in this case.

Second, for some latent variable models, the latent variables $\mathbf{z}$ follow a multivariate normal distribution and the observed variables $\mathbf{y}$ are independent conditional on $\mathbf{z}$. This class of models is referred to as the Gaussian latent variable models in the literature. In economics and finance, many latent variable models belong to this class of models, including dynamic linear models, dynamic factor models, various forms of stochastic volatility models and credit risk models. In these models, the observed-data likelihood is non-Gaussian but has a Gaussian flavor in the sense that the posterior distribution, $p(\mathbf{z}|\mathbf{y},\theta)$, may be expressed as,

$$p(\mathbf{z}|\mathbf{y},\theta) \propto \exp\left(-\frac{1}{2}\mathbf{z}'V(\theta)\mathbf{z} + \sum_{i=1}^{n} \ln p(\mathbf{y}_i|\mathbf{z}_i,\theta)\right).$$

Rue et al. (2004) and Rue et al. (2009) showed that this type of posterior distribution can be well approximated by a Gaussian distribution that matches the mode and the curvature at the mode. The resulting approximation is known as the Laplace approximation and can be expressed as,

$$p(\mathbf{z}|\mathbf{y},\theta) \propto \exp\left(-\frac{1}{2}\mathbf{z}'(V(\theta) + diag(\mathbf{c}))\mathbf{z}\right),$$

where $\mathbf{c}$ comes from the second order term in the Taylor expansion of $\sum_{i=1}^{n} \ln p(\mathbf{y}_i|\mathbf{z}_i)$ at the mode of $p(\mathbf{z}|\mathbf{y},\theta)$. The Laplace approximation may be employed to compute

$\mathscr{H}(\bar{\theta}|\bar{\theta})$. After $p(\mathbf{y}|\bar{\theta})$ is obtained, it is easy to obtain $D(\bar{\theta})$. It is important to point out that the numerical evaluation of $p(\mathbf{y}|\bar{\theta})$ is needed only once, i.e., at the posterior mean.

To compute P_D^*, we have to calculate the second derivative of the observed-data likelihood function in (2.3.12). The following two lemmas show how to compute the second derivatives.

Lemma 2.3.3 *Under the mild regularity conditions, the observed-data information matrix may be expressed as:*

$$\mathbf{I}(\theta) = -\frac{\partial^2 \mathscr{L}_o(\mathbf{y}|\theta)}{\partial\theta\partial\theta'} = \left\{ -\frac{\partial^2 \mathscr{Q}(\theta|\theta^*)}{\partial\theta\partial\theta'} - \frac{\partial^2 \mathscr{Q}(\theta|\theta^*)}{\partial\theta\partial\theta^{*'}} \right\}_{\theta^*=\theta}. \tag{2.3.14}$$

Lemma 2.3.4 *Let $S(\mathbf{x}|\theta) = \partial\mathscr{L}_c(\mathbf{x}|\theta)/\partial\theta$. Under the mild regularity condition, the observed-data information matrix has an equivalent form:*

$$\begin{aligned} \mathbf{I}(\theta) &= -\frac{\partial^2 \mathscr{L}_o(\mathbf{y}|\theta)}{\partial\theta\partial\theta'} = E_{\mathbf{z}|\mathbf{y},\theta}\left\{ -\frac{\partial^2 \mathscr{L}_c(\mathbf{x}|\theta)}{\partial\theta\partial\theta'} \right\} - Var_{\mathbf{z}|\mathbf{y},\theta}\left\{S(\mathbf{x}|\theta)\right\} \quad (2.3.15) \\ &= E_{\mathbf{z}|\mathbf{y},\theta}\left\{ -\frac{\partial^2 \mathscr{L}_c(\mathbf{x}|\theta)}{\partial\theta\partial\theta'} - S(\mathbf{x}|\theta)S(\mathbf{x}|\theta)' \right\} + E_{\mathbf{z}|\mathbf{y},\theta}\{S(\mathbf{x}|\theta)\}E_{\mathbf{z}|\mathbf{y},\theta}\{S(\mathbf{x}|\theta)\}', \end{aligned}$$

where all the expectations are taken with respect to the conditional distribution of $\mathbf{z}$ *given* $\mathbf{y}$ *and* θ.

Remark 2.3.23 *Lemma 2.3.3 and Lemma 2.3.4 were developed in Oakes (1999) and Louis (1982), respectively, for finding the standard error in the EM algorithm. If the $\mathscr{Q}$ function is available, we can use Lemma 2.3.3 to evaluate the second derivatives. If the $\mathscr{Q}$ function does not have an analytic form, we may use Lemma 2.3.4 to evaluate the second derivatives as follows,*

$$\begin{aligned} & E_{\mathbf{z}|\mathbf{y},\theta}\left\{ -\frac{\partial^2 \mathscr{L}_c(\mathbf{x}|\theta)}{\partial\theta\partial\theta'} - S(\mathbf{x}|\theta)S(\mathbf{x}|\theta)' \right\}, \\ \approx\ & -\frac{1}{M}\sum_{m=1}^{M}\left\{ \frac{\partial^2 \mathscr{L}_c(\mathbf{y},\mathbf{z}^{(m)}|\theta)}{\partial\theta\partial\theta'} + S(\mathbf{y},\mathbf{z}^{(m)}|\theta)S(\mathbf{y},\mathbf{z}^{(m)}|\theta)' \right\}, \\ & E_{\mathbf{z}|\mathbf{y},\theta}\{S(\mathbf{x}|\theta)\} \approx \frac{1}{M}\sum_{m=1}^{M} S(\mathbf{y},\mathbf{z}^{(m)}|\theta), \end{aligned}$$

where $\{\mathbf{z}^{(m)}, m = 1, 2, \cdots, M\}$ *are random observations drawn from the posterior distribution* $p(z|\mathbf{y}, \theta)$.

2.4 Examples

We now illustrate the proposed method in three applications, covering some popular models in economics and finance. In the first example, both $\mathscr{Q}(\theta|\theta)$ and $\mathscr{H}(\theta|\theta)$ are available in closed-form and hence RDIC is trivial to compute. In this example, we pay attention to implications of different distributional representations. In the second example, while $p(\mathbf{y}|\bar{\theta})$ is not available in closed-form, Kalman filter provides a recursive algorithm to evaluate it. Hence, $\mathscr{Q}(\theta|\theta)$ and $\mathscr{H}(\theta|\theta)$ can be calculated in the same manner, facilitating the computation of RDIC. In the third example, $p(\mathbf{y}|\bar{\theta})$ is not available in closed-form and Kalman filter cannot be applied. To compute RDIC, we use the Laplace approximation and the technique suggested in Lemma 2.3.4.

2.4.1 Comparing asset pricing models

Asset pricing theory is fundamentally important in modern finance. A basic assumption required by much asset pricing theory is that the return distribution is normal. Unfortunately, there has been overwhelming empirical evidence against normality for asset returns, which have led researchers to investigate asset pricing models with heavy-tailed distributions, including the family of elliptical distributions discussed in Zhou (1993). Kan and Zhou (2003) suggested to use the multivariate t distribution to replace the multivariate normal distribution. In addition, under the mean-variance efficiency, the asset excess premium should not be statistically different

from zero. In this section, we compare the following six asset pricing models:

$$
\begin{aligned}
&Model\ 1: R_t = \beta' F_t + \varepsilon_t, \varepsilon_t \sim N[0,\Sigma];\\
&Model\ 2: R_t = \alpha + \beta' F_t + \varepsilon_t, \varepsilon_t \sim N[0,\Sigma];\\
&Model\ 3: R_t = \beta' F_t + \varepsilon_t, \varepsilon_t \sim t[0,\Sigma,\nu];\\
&Model\ 4: R_t = \beta' F_t + \varepsilon_t, \varepsilon_t \sim N(0,\Sigma/\omega_t), \omega_t \sim \Gamma\left(\frac{\nu}{2},\frac{\nu}{2}\right);\\
&Model\ 5: R_t = \alpha + \beta' F_t + \varepsilon_t, \varepsilon_t \sim t[0,\Sigma,\nu];\\
&Model\ 6: R_t = \alpha + \beta' F_t + \varepsilon_t, \varepsilon_t \sim N(0,\Sigma/\omega_t), \omega_t \sim \Gamma\left(\frac{\nu}{2},\frac{\nu}{2}\right),
\end{aligned}
$$

where R_t is the excess return of portfolio at period t with $N \times 1$ dimension, F_t a $K \times 1$ vector of factor portfolio excess returns, α a $N \times 1$ vector of intercepts, β a $N \times K$ vector of scaled covariances, ε_t the random error, $t = 1, 2, \cdots, n$. For convenience, we restrict Σ to be a diagonal matrix and ν to be a known constant. Note that Model 4 is the distributional representation of Model 3, and Model 5 is the distributional representation of Model 6. This is especially true if ω_t is not the quantity of interest.

Monthly returns of 25 portfolios, constructed at the end of each June, are the intersections of 5 portfolios formed on size (market equity, ME) and 5 portfolios formed on the ratio of book equity to market equity (BE/ME). The Fama/French's three factors, market excess return, SMB (Small Minus Big), HML (High Minus Low) are used as the explanatory factors (Fama and French (1993)). The sample period is from July 1926 to July 2011, so that $N = 25$, $n = 1021$. The data are freely available from the data library of Kenneth French.[3]

Bayesian analysis of the asset pricing models has attracted a considerable amount of attentions in the empirical asset pricing literature.[4] Here we apply DIC_7 and RDIC to compare Models 1-6. Based on the result of Li and Yu (2012), in the empirical study, we simply set $\nu = 3$. Some vague conjugate prior distributions are

[3] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

[4] Avramov and Zhou (2010) provided an excellent review of the literature on Bayesian portfolio analysis.

used to represent the prior ignorance, namely,

$$\alpha_i \sim N[0,100], \beta_{ij} \sim N[0,100], \phi_{ii}^{-1} \sim \Gamma[0.001, 0.001].$$

The use of uninformative priors implies that P_D^* should be close to the actual number of the parameters, P, if the posterior distribution is well approximated by the normal distribution.

Under these prior specifications, we use WinBUGS to implement Bayesian analysis and to calculate DIC_7. An introduction to WinBUGS can be found in Spiegelhalter et al. (2003). To calculate RDIC, we use R2WinBUGS, a R package that calls WinBUGS and exports the results into R (Sturtz et al. (2005)).[5] Since both $\mathscr{Q}(\theta|\theta)$ and $\mathscr{H}(\theta|\theta)$ are available in closed-form, RDIC is trivial to compute.

We sample 100,000 random observations from the posterior distributions in each model, the first 40,000 of which form the burn-in period. The convergence of the next 60,000 iterations is checked using the Raftery-Lewis diagnostic test statistic (Raftery and Lewis (1992)) with every 3th observation collected. Hence, 20,000 effective observations are used for computing the information criteria. The value of DIC_7 is automatically calculated by WinBUGS. Based on the observed log-likelihood given in formula (.3.1) in Appendix D, we can compute DIC and RDIC for Model 3 and 5. Table I reports DIC_7, RDIC, P_D, and P_D^* for all six models. Note that when there is no latent variable DIC_7 is reduced into DIC_1.

From Table I, we see that P_D is almost identical to P_D^* in each of Models 1, 2, 3 and 5. Not surprisingly, DIC_7 and RDIC are almost the same in each of these models. As expected, DIC_7 in Model 3 is quite different from that in Model 4 although these two models are the same. The main reason for this distinctive difference is that in Model 4, the scale-mixture specification is used and, hence, a sequence of latent variables, $\{\omega_t\}$, is introduced artificially. In DIC_7 the latent variables, $\{\omega_t\}$, are treated as parameters. There is no latent variable for Model 3, however. For the same reason, DIC_7 in Model 5 is quite different from that in Model 6. As argued

[5] R code may be requested from the authors of the present paper.

Table 2.1: Model selection results for Fama-French three factor models

Model	M_1	M_2	M_3	M_4	M_5	M_6
Number of Parameters	100	125	100	100	125	125
P_D	100	125	100	1021	125	1046
DIC_7	-119842	-119880	-133088	-134777	-133202	-134897
P_D^*	100	125	100	100	126	126
RDIC	-119842	-119880	-133087	-133087	-133201	-133201

earlier, this conceptual difficulty is due to the lack of the likelihood principle and is consistent with what has been documented in the literature (Spiegelhalter et al., 2002 and Berg et al., 2004). The most important finding from Table I is that RDIC does not suffer from the same difficulty as DIC_7. RDIC and P_D^* for Model 3 (and Model 5) are nearly identical to those for Model 4 (and Model 6). In terms of the computational cost, for Model 3, after the effective random observations are collected, RDIC takes about 3 minutes in a laptop with Inter Core i5-540M (2.53GHz). On the other hand, DIC_1 involves $\int \ln p(\mathbf{y}|\theta)p(\theta|\mathbf{y})\text{d}\theta$ when computing P_D, which is approximated by $\frac{1}{J}\sum_{j=1}^{J} \ln p\left(\mathbf{y}|\theta^{(j)}\right)$. This quantity is much more expensive to compute because it requires numerical evaluation of $\ln p\left(\mathbf{y}|\theta^{(j)}\right)$ for J times. For Model 3, based on the 20,000 posterior random observations, one has to evaluate $\ln p\left(\mathbf{y}|\theta^{(j)}\right)$ 20,000 times. It requires 11 hours and 4 minutes to compute DIC_1 using the same laptop. The computational relative efficiency of RDIC over DIC_1 is obvious and increases as the number of effective observations increases.

It is important to emphasize that, although our method is motivated from the case of objective priors, informative priors can be also used in our method. In a recent study, Tu and Zhou (2010) explored a general approach to forming informative priors based on economic objectives and found that the proposed informative priors outperform significantly the objective priors in terms of investment performance. RDIC can be used in conjunction with the informative prior specifications. In this case, P_D^* can be quite different from P.

2.4.2 Comparing high dimensional dynamic factor models

For many countries, there exists a rich array of macroeconomic time series and financial time series. To reduce the dimensionality and to extract the information from the large number of time series, factor analysis has been widely used in the empirical macroeconomic literature and in the empirical finance literature. For example, by extending the static factor models previously developed for cross-sectional data, Geweke (1977) proposed the dynamic factor model for time series data. Many empirical studies, such as Sargent and Sims (1977), Giannone et al. (2004), have reported evidence that a large fraction of the variance of many macroeconomic series can be explained by a small number of dynamic factors. Stock and Watson (1999) and Stock and Watson (2002) showed that dynamic factors extracted from a large number of predictors can be used to lead to improvement in predicting macroeconomic variables. Not surprisingly, high dimensional dynamic factor models have become a popular tool under a data rich environment for macroeconomists and policy makers. An excellent review on the dynamic factor models is given by Stock and Watson (2010).

Following Bernanke et al. (2005) (BBE hereafter), the present paper considers the following fundamental dynamic factor model:

$$\begin{aligned} Y_t &= F_t L' + \varepsilon_t', \\ F_t &= F_{t-1}\Phi' + \eta_t, \end{aligned}$$

where Y_t is a $1 \times N$ vector of time series variables, F_t a $1 \times K$ vector of unobserved latent factors which contains the information extracted from all the N time series variables, L an $N \times K$ factor loading matrix, Φ the $K \times K$ autoregressive parameter matrix of unobserved latent factors. It is assumed that $\varepsilon_t \sim N(0, \Sigma)$ and $\eta_t \sim N(0, Q)$. For the purpose of identification, Σ is assume to be diagonal and ε_t and η_t are assumed to be independent with each other. Following BBE (2005), we set the first $K \times K$ block in the loading matrix L to be the identity matrix.

In this dynamic factor model, the observed variable Y_t consists of a balanced panel of 120 monthly macroeconomic time series. These series are initially transformed to induce stationarity. The description of the series and the transformation is provided in BBE (2005). The sample period is from January 1959 to August 2001. Because the data are of high dimension, the analysis of the dynamic factor models via a frequentist method is not trivial; see the discussion in Stock and Watson (2011). In the literature, Bayesian inference via the MCMC techniques has been popular for analyzing the dynamic factor models; see Otrok and Whiteman (1998), Kose et al. (2003), Kose et al. (2008), BBE (2005).

Following BBE (2005), we specify the following prior distribution:

$$\begin{aligned} \Sigma_{ii} &\sim Inverse-\Gamma\left(3,0.001\right), L_i \sim N\left(0,\Sigma_{ii}M_0^{-1}\right), \\ vec\left(\Phi\right)|Q &\sim N\left(0,Q\otimes\Omega_0\right), Q \sim Inverse-\Gamma\left(Q_0,K+2\right), \end{aligned}$$

where M_0 is a $K\times K$ identity matrix, L_i the ith $(i>K)$ column of L. The diagonal elements of Q_0 are set to be the residual variances of the corresponding one lag univariate autoregressions, $\widehat{\sigma}_i^2$. The diagonal elements of Ω_0 are constructed so that the prior variance of parameter on the jth variable in the ith equation equals $\widehat{\sigma}_i^2/\widehat{\sigma}_j^2$.

In this example, we aim to determine the number of factors in the dynamic factor models using model selection criteria. In BBE (2005) model comparison is achieved by graphic methods. Our approach can be regarded as a formal statistical alternative to the graphic methods. It is well documented that the determination of number of factors in the setting of the dynamic factor models is important; see Stock and Watson (1999). As in the previous example, we use DIC_7 and RDIC to compare models with different numbers of factors, namely $K=1$, 2 and 3, which are denoted by M_1, M_2, M_3 respectively. Using the Gibbs sampler, we sample 22,000 random observations from the corresponding posterior distributions. We discard the first 2,000 observations and keep the following 20,000 as the effective samples from the posterior distribution of the parameters.

Table 2.2: Model selection results for dynamic factor models

Model	M_1	M_2	M_3
Number of Parameters	752	1385	2019
P_D	350	965	1391
DIC_7	-135480	-149010	-155060
Number of Parameters	241	363	486
P_D^*	87	20	326
RDIC	-22452	-34868	-40420

Based on the 20,000 samples, we compute DIC_7, RDIC, P_D, P_D^* for all three models. The technique in Lemma 4.3.1 is used to approximate the observed-data likelihood at the posterior mean**.** Table II reports the simple count of the number of parameters (including the latent variables), DIC_7, the P_D component of DIC_7, (i.e. when the data augmentation technique is used), the simple count of the number of parameters (excluding the latent variables), RDIC, and the P_D^* component of RDIC (i.e. when the data augmentation technique is not used). Several conclusions may be drawn from Table II. First, both DIC_7 and RDIC suggest that M_3 is the best model. Second, since some very informative priors have been used, neither P_D nor P_D^* is close to the actual number of parameters. While it is cheap to compute RDIC, it is much harder to compute DIC_1. This is because the observed-data likelihood $p(\mathbf{y}|\theta)$ is not available in closed-form and Kalman filter is used to numerically calculate $p(\mathbf{y}|\theta)$ which involves the computation of $\frac{1}{J}\sum_{j=1}^{J} \ln p(\mathbf{y}|\theta^{(j)})$, for $J = 20,000$. We have to run Kalman filter 20,000 times, which takes more than 4 hours to compute in Matlab.[6] In a sharp contrast, it only took less than 80 seconds to compute RDIC. Obviously, the discrepancy in CPU time increases with J.

2.4.3 Comparing stochastic volatility models

Stochastic volatility (SV) models have been found very useful for pricing derivative securities. In the discrete time log-normal SV models, the logarithmic volatility

[6]Numeircally more efficient algorithms, such as the one proposed by Chan and Jeliazkov (2009), may be used to evaluate $\ln p(\mathbf{y}|\theta^{(j)})$.

is the state variable which is often assumed to follow an AR(1) model. The basic log-normal SV model is of the form:

$$y_t = \alpha + \exp(h_t/2)u_t,\ u_t \sim N(0,1),$$
$$h_t = \mu + \phi(h_{t-1} - \mu) + v_t,\ v_t \sim N(0,\tau^2),$$

where $t = 1,2,\cdots,n$, y_t is the continuously compounded return, h_t the unobserved log-volatility, $h_0 = \mu$, and (u_t, v_t) independently normal variables for all t. In this paper, we denote this model by M_1.

To carry out Bayesian analysis of M_1, following Meyer and Yu (2000), the prior distributions are specified as follows:

$$\begin{aligned} \alpha &\sim N(0,100),\ \ \mu \sim N(0,100), \\ \phi &\sim Beta(1,1),\ \ 1/\tau^2 \sim \Gamma(0.001, 0.001). \end{aligned}$$

An alternative specification of M_1 is given by:

$$\begin{aligned} y_t &= \alpha + \sigma_t u_t,\ u_t \sim N(0,1), \\ \ln\sigma_t^2 &= \mu + \phi\left(\ln\sigma_{t-1}^2 - \mu\right) + v_t,\ v_t \sim N(0,\tau^2), \end{aligned}$$

which is denoted by M_2. Obviously, the only difference between M_2 and M_1 is that the latent variable in M_2 is the exponential transformation of that in M_1. If the same priors are used for the model parameters, $\theta = (\alpha, \mu, \phi, \tau)$, the two models are identical to each other. Our goal here is to compare the two models using DIC_7 and RDIC. In both models, $p(\mathbf{y}|\theta)$ is not available in closed-form. Since the models are of a nonlinear non-Gaussian form, Kalman filter cannot be applied and DIC_1 is infeasible to compute.

The dataset consists of 1,822 daily returns of the Standard & Poor (S&P) 500 index, covering the period between January 3, 2005 and March 28, 2012. For M_1 and M_2, after a burn-in period of 10,000 iterations we save the next 20,000 iterations.

Table 2.3: Model selection results for stochastic volatility models

Model	M_1	M_2
P_D	102.94	89.67
DIC_7	5200.56	5183.12
P_D^*	3.62	3.78
RDIC	5296.20	5296.55

Table III reports DIC_7, RDIC, P_D, P_D^* for both models. To calculate RDIC and P_D^*, since the $\mathscr{Q}$ function does not have a closed-form expression, we employ the technique in Lemma 2.3.3 to compute the second order derivative of the observed-data likelihood. To compute RDIC, we use the Laplace approximation of Rue, Martino and Chopin (2009). The technique in Lemma 4.3.1 is used to approximate the observed-data likelihood at the posterior mean.

The following findings can be obtained from Table III. First, P_D in M_1 is 13 points more than that in M_2. Similarly, DIC_7 in M_1 is nearly 20 points more than that in M_2. These differences are very large and indicate that M_2 is a much better model than M_1 although the two modes are actually the same. Second, P_D^* in M_1 is nearly identical to that in M_2, which is about the same as $P = 4$, the actual number of parameters. Similarly, RDIC in M_1 is nearly identical to that in M_2. Given that M_1 and M_2 are two equivalent representations to each other, the empirical results from RDIC are more reasonable than those from DIC_7.

2.5 Conclusion

This paper introduces a robust deviance information criteria (RDIC) for comparing models with latent variables. Although latent variable models can be conveniently estimated in the Bayesian framework via MCMC if the data augmentation technique is used, we argue that data augmentation cannot be used in connection to DIC. This is because that the justification of DIC rests on the validity of the standard Bayesian asymptotic theory. With data augmentation, the number of parameters increases with the number of observations, making the likelihood nonregular. As a

consequence, the standard Bayesian asymptotic theory does not hold. In addition, the use of the data augmentation makes DIC is very sensitive to transformations and distributional representations.

While in principle one can use the standard DIC (i.e. DIC_1) without resorting to the data augmentation technique, in practice this standard DIC is very difficult to use because the observed-data likelihood is not available in closed-form for many latent variable models and because the standard DIC_1 has to numerically evaluate the observed-data likelihood at each MCMC iteration. These two observations make the implementation of DIC_1 practically non-operational.

The problem is overcome by RDIC. RDIC is defined without augmenting the parameter space and hence can be justified by the standard Bayesian asymptotic theory. We then show that how the EM algorithm can facilitate the computation of RDIC in different contexts. Since the latent variables are not counted as parameters in our approach, RDIC is robust to nonlinear transformations of the latent variables and distributional representations of the model specification. Asymptotic justification, computational tractability and robustness to transformation and specification are the three main advantages of the proposed approach. These advantages are illustrated using several popular models in economics and finance.

Chapter 3 A New Approach to Bayesian Hypothesis Testing

3.1 Introduction

Hypothesis testing plays a fundamental role in making statistical inference about the model specification. After models are estimated, empirical researchers would often like to test a relevant hypothesis to look for evidence to support or to be against a particular theory. An important class of hypotheses involve a single parameter value in the null.

In this paper we are concerned about testing a single point hypothesis under Bayesian paradigm. So far Bayes factor (BF) is the dominant statistic for Bayesian hypothesis testing (Kass and Raftery (1995); Geweke (2007)). The wide range of applicability of BF comes with no surprise. BF computes the posterior odds of the null hypothesis and hence provides a general and intuitive way to evaluate the evidence in favor of the null hypothesis.

In the meantime, unfortunately, BF also suffers from several theoretical and practical difficulties. First, when improper prior distributions are used, BF contains undefined constants and takes arbitrary values. This is known as Bartlett's paradox (Kass and Raftery (1995)). Second, when a proper but vague prior distribution with a large spread is used to represent prior ignorance, BF tends to favor the null hypothesis. The problem may persist even when the sample size is large. This is known as Jeffreys-Lindley's paradox (Kass and Raftery (1995) and Poirier (1995)). Third, the calculation of BF generally requires the evaluation of marginal likelihoods. In

many models, the marginal likelihoods may be difficult to compute.

Several approaches have been proposed in the literature to deal with Bartlett's paradox and Jeffreys-Lindley's paradox. One simple approach is to split the data into two parts, one as a training set, the other for statistical analysis. The non-informative prior is then updated by the training data, which produces a new proper informative prior distribution for computing BF. This idea is shared by the fractional BF (O'Hagan (1995)), and the intrinsic BF (Berger (1985)). In many practical situations, unfortunately, it is not clear how to split the sample. Moreover, the sample split may have a major impact on statistical inference. Without a need to split the sample, several Bayesian hypothesis testing approaches have been proposed based on the decision theory. Noting that the BF approach to Bayesian hypothesis testing is a decision problem with a simple zero-one loss function, Bernardo and Rueda (2002) (BR hereafter) and Li and Yu (2012) (LY hereafter) suggested extending the zero-one loss function into continuous loss functions, resulting in Bayesian test statistics that is well defined under improper priors.

The test statistics of BR and LY relies on threshold values. While in theory these threshold values may be calibrated from simulated data generated from the null hypothesis, in practice they are computationally expensive to obtain. Following McCulloch (1989), LY proposed to choose the threshold values based on the Bernoulli distribution. Although this choice makes the determination of threshold values convenient, there are obvious drawbacks. Not only is the choice of the Bernoulli distribution arbitrary, but also are the threshold values independent of the data and the candidate models. Moreover, it is not clear if the test statistic of LY can resolve Jeffreys-Lindley's paradox.

The main purpose of this paper is to develop a new Bayesian hypothesis testing approach for the point null hypothesis testing. The test statistic is based on the Bayesian deviance and constructed in a decision theoretical framework. It can be regarded as the Bayesian version of the likelihood ratio test. We show that the statistic appeals in four aspects. First, it does not suffer from Bartlett's paradox and,

hence, can be used under improper priors. Second, it does not suffer from Jeffreys-Lindley's paradox and, hence, can be used under vague priors. Third, it is easy to compute. Finally, the threshold values can be easily determined and are dependent on the data as well as the candidate models.

To show the strength of the proposed method, we apply the test to three real examples in economics and finance. In the first example, we compare the performance of the proposed test with that of the BF in the context of CEO salary determination. It is shown that the new test is much more robust than the BF with respect to the prior. In the second example, we test the validity of the three factor Fama-French model and the new test rejects the well-known specification. Finally, we test the absence of the leverage effect in a stochastic volatility model for exchange rates and the new test suggests that there is no leverage effect in the exchange rate series.

The paper is organized as follows. Section 2 reviews the Bayesian literature on testing the point null hypothesis from the viewpoint of decision theory. Section 3 develops the new Bayesian test statistic and establishes its properties. 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 theoretical results.

3.2 Point Null Hypothesis Testing: A Literature Review

3.2.1 The setup

Denote $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$ the vector of observables. Denote $p(\mathbf{y}|\vartheta)$ the likelihood function of the observed data. Denote $\pi(\vartheta)$ the prior distribution and $p(\vartheta|\mathbf{y})$ the posterior. Suppose that researchers may wish to test a hypothesis, the simplest of which contains only a point which may correspond to the prediction of a theory (Robert, 2001). $\theta \in \Theta$, whose dimension is p, the parameters of interest, and $\psi \in \Psi$, whose dimension is q, the nuisance parameters. So $\vartheta = (\theta, \psi)'$. Assume that the

observed data, $\mathbf{y} \in \mathbf{Y}$, is described a probabilistic model $M \equiv \{p(\mathbf{y}|\theta,\psi)\}$. The point null hypothesis is:

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

From the viewpoint of decision theory, the hypothesis testing may be viewed as a decision problem where the action space has two elements, i.e., to accept H_0 (name it d_0) or to reject H_0 (name it d_1). Denote the null model $M_0 \equiv \{p(\mathbf{y}|\theta_0,\psi), \psi \in \Psi\}$, and $M_1 \equiv M$. Suppose a loss is incurred as a function of the actual value of the parameters (θ,ψ) when one accepts H_0 or rejects H_0. Assume the loss function is given by $\{\mathscr{L}[d_i,(\theta,\psi)], i = 0,1\}$. Naturally, one would like to reject H_0 when the expected posterior loss of accepting H_0 is sufficiently larger than the expected posterior loss of rejecting H_0, i.e.,

$$\mathbf{T}(\mathbf{y},\theta_0) = \int_\Theta \int_\Psi \triangle\mathscr{L}[H_0,(\theta,\psi)]p(\theta,\psi|\mathbf{y})\mathrm{d}\theta\mathrm{d}\psi > C,$$

where 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 which can be used to measure the evidence against H_0 as a function of (θ,ψ).

3.2.2 Bayes factors and the discrete loss function

BF employs the zero-one loss function. In particular, if

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

we can get

$$\mathbf{T}(\mathbf{y},\theta_0) = \int_\Psi (-1)\frac{p(\mathbf{y}|\theta_0,\psi)p(\psi|\theta_0)p(\theta_0)}{p(\mathbf{y})}\mathrm{d}\psi + \int_\Theta\int_\Psi 1\frac{p(\mathbf{y}|\theta,\psi)p(\psi|\theta)p(\theta)}{p(\mathbf{y})}\mathrm{d}\theta\mathrm{d}\psi,$$

where $p(\mathbf{y}) = \int p(\mathbf{y}, \vartheta)\mathrm{d}\vartheta$ is the marginal likelihood. In general, to represent a prior ignorance, an equal probability 0.5 is assigned to H_0 and to H_1. A reasonable prior for θ with a discrete support at θ_0 is formulated as $p(\theta) = 0.5$ when $\theta = \theta_0$ and $p(\theta) = 0.5\pi(\theta)$ when $\theta \neq \theta_0$, where $\pi(\theta)$ is a prior distribution. Hence, when $C = 0$, the decision criterion is given by:

$$\text{Reject } H_0 \text{ iff } -\int_{\Psi} p(\mathbf{y}|\theta_0, \psi)p(\psi|\theta_0)\mathrm{d}\psi + \int_{\Theta}\int_{\Psi} p(\mathbf{y}|\theta, \psi)p(\psi|\theta)\pi(\theta)\mathrm{d}\theta\mathrm{d}\psi > 0.$$

which is equivalent to

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

where BF_{01} is the well-known BF (Kass and Raftery (1995)) and is the ratio of two marginal likelihood values.

When a subjective prior is not available, an objective prior or default prior may be used. Often, $\pi(\theta)$ is taken as non-informative priors, such as the Jeffreys or the reference prior (Jeffreys (1961); Bernardo and Rueda (2002)). These non-informative priors are generally improper, and it follows that $\pi(\theta) = C_0 f(\theta)$, where $f(\theta)$ is a nonintegrable function, and C_0 is an arbitrary positive constant. In this case, the BF is

$$BF_{01} = \frac{\int_{\Psi} p(\mathbf{y}|\theta_0, \psi)p(\psi|\theta_0)\mathrm{d}\psi}{C_0\int_{\Theta}\int_{\Psi} p(\mathbf{y}|\theta, \psi)p(\psi|\theta)f(\theta)\mathrm{d}\theta\mathrm{d}\psi}.$$

Clearly, the BF is not well defined since it depends on the arbitrary constant C_0, giving rise to Bartlett's paradox. In addition, if a proper prior is used but has a large variance, the likelihood function may take low values under the alterative hypothesis. This often leads to a smaller marginal likelihood value for the alterative model. Consequently, BF has a tendency to favor H_0, giving rise to Jeffreys-Lindley's paradox; see Poirier (1995), Robert (2001).

The formulation of BF generally requires a positive probability for $\theta = \theta_0$ to

be assigned. When θ is continuous, the prior concentrates a positive probability mass on the single point θ_0. As pointed out by BR, Jeffreys-Lindley's paradox is the consequence of using this non-regular prior structure.

3.2.3 BR and the KL loss function

Instead of using a zero-one loss function, BR (2002) advocated using a continuous function of θ and θ_0 to formulate the loss function. In particular, they suggested using the KL divergence. For any regular probability functions, $p(x)$ and $q(x)$, the KL divergence is defined as:

$$KL[p(x), q(x)] = \int p(x) \log \frac{p(x)}{q(x)} dx. \tag{3.2.2}$$

It can be shown that $KL \geq 0$ for any p and q, and equal to 0 iff $p(x) = q(x)$. In this case, the decision criterion is:

$$\begin{aligned} \mathbf{T}_{BR}(\mathbf{y}, \theta_0) &= \int_\Theta \int_\Psi \triangle\mathscr{L}[H_0, (\theta, \psi)] p(\theta, \psi|\mathbf{y}) \mathrm{d}\theta \mathrm{d}\psi \\ &= \int_\Theta \int_\Psi \left\{ \int \log \frac{p(\mathbf{y}|\theta, \psi)}{p(\mathbf{y}|\theta_0, \psi)} p(\mathbf{y}|\theta, \psi) \mathrm{d}\mathbf{y} \right\} p(\theta, \psi|\mathbf{y}) \mathrm{d}\theta \mathrm{d}\psi > C. \end{aligned} \tag{3.2.3}$$

To ensure the symmetry, BR suggested using the following net loss function:

$$\triangle\mathscr{L}[H_0, (\theta, \psi)] = \min\{KL[p(\mathbf{y}|\theta, \psi), p(\mathbf{y}|\theta_0, \psi)], KL[p(\mathbf{y}|\theta_0, \psi), p(\mathbf{y}|\theta, \psi)]\}. \tag{3.2.4}$$

Obviously, $\mathbf{T}_{BR}(\mathbf{y}, \theta_0) = 0$ under the null hypothesis but is positive under the alternative hypothesis. According to BR, this loss function can be used under the reference priors to maintain objectiveness, overcoming Bartlett's paradox. Although the statistic of BR is well defined under improper priors and has other desirable properties, it has certain practical difficulties. First, when the KL loss function is not available analytically, the test statistic of BR is infeasible to use. Second, threshold values for C, are needed but difficult to find in general.

3.2.4 LY and the $\mathscr{Q}$ loss function

In the context of latent variable models, the likelihood function and the KL loss are not available analytically and the test statistic of BR is difficult to use. To solve this problem, LY developed a Bayesian test statistic based on the $\mathscr{Q}$ function used in the EM algorithm. Denote $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$ the vector of observables and $\mathbf{z} = (z_1, z_2, \cdots, z_n)'$ the vector of latent variables. Let $\mathbf{x} = (\mathbf{y}, \mathbf{z})'$. The latent variable model is dependent on a set of parameters ϑ. Denote $p(\mathbf{y}|\vartheta)$ and $p(\mathbf{x}|\vartheta)$ the likelihood function of the observed data and the likelihood function of complete data, respectively. The two functions are related to each other by

$$p(\mathbf{y}|\vartheta) = \int p(\mathbf{x}|\vartheta) d\mathbf{z} = \int p(\mathbf{y}, \mathbf{z}|\vartheta) d\mathbf{z}. \tag{3.2.5}$$

When the above integral at the right hand side does not have a closed-form solution, instead of using maximum likelihood (ML) method, it is numerically more tractable to carry out Bayesian analysis based on the MCMC algorithm for estimating the latent variable models; see, for example Geweke et al. (2011).

For latent variable models, the complete-data log-likelihood, $L_c(\mathbf{x}|\vartheta) = \log p(\mathbf{x}|\vartheta)$, is related to the observed data log-likelihood, $L_o(\mathbf{y}|\vartheta) = \log p(\mathbf{y}|\vartheta)$. While $L_c(\mathbf{x}|\vartheta)$ is often simple, but $L_o(\mathbf{y}|\vartheta) = \log p(\mathbf{y}|\vartheta)$ is often complicated because the integral Equation (3.2.5) does not have an analytical solution. The EM algorithm is a way to obtain the ML estimator (Dempster et al. (1977)). A standard EM algorithm consists of two steps: the *expectation* (E) step and the *maximization* (M) step. The E-step evaluates the $\mathscr{Q}$ function which is defined by:

$$\mathscr{Q}\left(\vartheta|\vartheta^{(r)}\right) = E_{\mathbf{z}}\left\{L_c(\mathbf{x}|\vartheta)|\mathbf{y}, \vartheta^{(r)}\right\}, \tag{3.2.6}$$

where the expectation is taken with respect to the conditional distribution of latent variables given $\mathbf{y}$ and $\vartheta^{(r)}$, $p(\mathbf{z}|\mathbf{y}, \vartheta^{(r)})$. The M-step determines a $\vartheta^{(r+1)}$ that maximizes $\mathscr{Q}(\vartheta|\vartheta^{(r)})$.

Let $\vartheta_0 = (\theta_0, \psi)$. LY (2012) introduced a continuous net loss 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:

$$\mathbf{T}_{LY}(\mathbf{y}, \theta_0) = E_{\vartheta|\mathbf{y}}[\triangle\mathscr{L}(\vartheta, \vartheta_0)]. \tag{3.2.7}$$

Like the statistic of BR, the test statistic, $T_{LY}(\mathbf{y}, \theta_0)$, is well defined under improper priors and hence is immune to Bartlett's paradox. Also, it is easy to compute if the MCMC output is available. However, like the statistic of BR, the threshold values are needed in practice. Following McCulloch (1989), LY proposed to base the threshold values on two Bernoulli distributions. Although the use of the threshold values is not new in the Bayesian literature (see, for example, Jeffreys' BF scales), it is awkward that these threshold values are independent of the data and the candidate models. It was remarked in LY that a more natural approach is to obtain threshold values from simulated data in repeated sampling, which is computationally time consuming in general.

3.3 A New Method for Bayesian Hypothesis Testing

3.3.1 The test statistic

BR's approach requires the KL loss function must have a closed-form expression and the threshold values for Bayesian hypothesis testing are difficult to obtain. LY's approach is easy to compute, but the threshold values are independent of the data and the candidate models. To avoid these theoretical and computational difficulties, in this section, we introduce a new Bayesian approach for hypothesis testing.

Denote the net loss function as:

$$\triangle\mathscr{L}[H_0,(\theta,\psi)] = -2\log p(\mathbf{y}|\theta_0,\psi) - (-2\log p(\mathbf{y}|\theta,\psi)) = 2\log p(\mathbf{y}|\theta,\psi) - 2\log p(\mathbf{y}|\theta_0,\psi), \tag{3.3.1}$$

where $-2\log p(\mathbf{y}|\theta,\psi)$ represents the residual information in data $\mathbf{y}$ given θ,ψ in the alternative model. According to Good (1956), $-2\log p(\mathbf{y}|\theta,\psi)$ measures the surprise or uncertainty. Similarly, one can interpret $-2\log p(\mathbf{y}|\theta_0,\psi)$. The net loss function is the difference of the two Bayesian deviances, if the Bayesian deviance is defined in the same way as in Spiegelhalter et al. (2002) (Section 2.5). The new Bayesian test statistic is then defined by:

$$\mathbf{T}(\mathbf{y},\theta_0) = 2\int [\log p(\mathbf{y}|\theta,\psi) - \log p(\mathbf{y}|\theta_0,\psi)]\, p(\theta,\psi|\mathbf{y})\mathrm{d}\theta\mathrm{d}\psi. \tag{3.3.2}$$

Under the null, $\mathbf{T}(\mathbf{y},\theta_0) = 0$, whereas under the alternative, $\mathbf{T}(\mathbf{y},\theta_0) \neq 0$. When the deviance of the null hypothesis is sufficiently smaller than that of the alternative, it is reasonable to believe that we should reject the null hypothesis.

BF essentially compares the relative magnitude of

$$\int_\Psi p(\mathbf{y}|\theta_0,\psi)p(\psi|\theta_0)\mathrm{d}\psi$$

and

$$\int_\Theta\int_\Psi p(\mathbf{y}|\theta,\psi)p(\psi|\theta)\pi(\theta)\mathrm{d}\theta\mathrm{d}\psi,$$

whereas our test statistic compares the relative magnitude of

$$\int \log p(\mathbf{y}|\theta_0,\psi)p(\theta,\psi|\mathbf{y})\mathrm{d}\theta\mathrm{d}\psi = \int \log p(\mathbf{y}|\theta_0,\psi)p(\psi|\mathbf{y})\mathrm{d}\psi$$

and

$$\int_\Theta\int_\Psi \log p(\mathbf{y}|\theta,\psi)p(\theta,\psi|\mathbf{y})\mathrm{d}\theta\mathrm{d}\psi.$$

Clearly there are two major differences between the two approaches. First, the likelihood functions in BF are replaced with the log-likelihood functions in our

test. Second and more importantly, the (log-)likelihood functions are averaged over the prior distributions in BF but over the posterior distributions in our method. The second difference suggests that our statistic is less sensitive to the prior distributions.

The first result in this present paper shows that the Bayes risk of $\mathbf{T}(\mathbf{y},\theta_0)$ is just two times the test statistic proposed by BR.

Theorem 3.3.1 *It can be shown that*

$$E_{\mathbf{y}}\left[\mathbf{T}(\mathbf{y},\theta_0)\right] = \int \mathbf{T}(\mathbf{y},\theta_0)p(\mathbf{y})d\mathbf{y} = 2E_{\mathbf{y}}\left[\mathbf{T}_{BR}(\mathbf{y},\theta_0)\right].$$

Remark 3.3.1 $\mathbf{T}(\mathbf{y},\theta_0)$ *may be explained as the Bayesian version of the likelihood ratio test since it is the likelihood ratio averaged over the posterior distribution under the alternative hypothesis.*

Remark 3.3.2 *To show how the new statistic is immune to Bartlett's paradox, consider general improper priors,* $p(\psi|\theta) = Af(\psi|\theta)$, $p(\theta) = Bf(\theta)$, $p(\psi|\theta_0) = C_0 f(\psi|\theta_0)$ *where* $f(\psi|\theta)$, $f(\theta)$ *and* $f(\psi|\theta_0)$ *are nonintegrable functions, and* A,B,C_0 *are arbitrary positive constants. It can be shown that,*

$$\begin{aligned} p(\psi,\theta|\mathbf{y}) &= \frac{p(\mathbf{y},\psi,\theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y},\psi,\theta)}{\int\int p(\mathbf{y},\psi,\theta)d\psi d\theta} = \frac{p(\mathbf{y}|\psi,\theta)p(\psi,\theta)}{\int\int p(\mathbf{y}|\psi,\theta)p(\psi,\theta)d\psi d\theta} \\ &= \frac{p(\mathbf{y}|\psi,\theta)ABf(\psi,\theta)}{\int\int p(\mathbf{y}|\psi,\theta)ABf(\psi,\theta)d\psi d\theta} = \frac{p(\mathbf{y}|\psi,\theta)f(\psi,\theta)}{\int\int p(\mathbf{y}|\psi,\theta)f(\psi,\theta)d\psi d\theta}. \end{aligned}$$

Hence, $p(\psi,\theta|\mathbf{y})$ *is independent of the arbitrary constants. Similarly, we can show that* $p(\psi|\mathbf{y})$ *is also independent of* C_0. *Consequently,* $\mathbf{T}(\mathbf{y},\theta_0)$ *is well defined under improper priors.*

Remark 3.3.3 *To see how the new statistic can avoid Jeffreys-Lindley's paradox, we consider a well known example in the literature; see, for example, Robert (1993). Let* $y \sim N(\theta,1)$. *Suppose we want to test the simple point null hypothesis* $H_0 : \theta = 0$. *The prior distribution of* θ *can be set as* $N(\mu,\tau^2)$ *with* $\mu = 0$. *Then the posterior*

distribution of θ *is* $N(\mu(y), \omega^2)$ *with*

$$\mu(y) = \frac{\mu + \tau^2 y}{1+\tau^2}, \omega^2 = \frac{\tau^2}{1+\tau^2}.$$

BF is given by:

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

As $\tau^2 \to +\infty$, $BF_{10} \to 0$ *which means that the test always supports the null hypothesis regardless whether or not it holds true, giving rise to Jeffreys-Lindley's paradox. The reason for the paradox is that BF compares* $\int p(y|\theta)p(\theta)d\theta$ *with* $p(y|\theta = 0)$. *When* $p(\theta)$ *has a large variance, even if y is far away from 0, there is a fair chance that* $p(y|\theta = 0)$ *is larger. On the other hand, it is easy to show:*

$$\mathbf{T}(y,0) = 2\left[\int \log p(y|\theta)p(\theta|y)d\theta - \log p(y|\theta = 0)\right] = 2y\mu(y) - \mu^2(y) - \omega^2.$$

As $\tau^2 \to +\infty$, $\mu(y) \to y$, $\omega^2 \to 1$. *In this case, the posterior distribution converges to* $N(y,1)$ *and* $\mathbf{T}(y,0) \to y^2 - 1$ *which is distributed exactly as* $\chi^2(1) - 1$ *when* H_0 *is true. Consequently, our proposed test statistic avoids Jeffreys-Lindley's paradox. Essentially, we compare* $\int \log p(y|\theta)dN(\theta; y, 1)$ *with* $\log p(y|\theta = 0)$. *Since the posterior distribution* $N(\theta; y, 1)$ *puts much more weight in the area near y, when y is far away from zero, the former quantity should take a much larger value than the latter. To illustrate the point, if* $y = 3$ *which is 3 standard deviation away under the null hypothesis, we expect a reasonable test should reject the null hypothesis. Table 1 reports* BF_{01} *and* $\mathbf{T}(y,0)$ *when* $\tau = 1, 100, 1000$. *It can be seen that while our method always rejects the null the BF fails to reject the null when* $\tau = 100, 1,000$.

Remark 3.3.4 *When* $p(\mathbf{y}|\theta, \psi)$ *is available in closed-form and the model under alternative hypothesis is estimated by MCMC, it is straightforward to calculate*

Table 3.1: Using BF and the new test to test $\theta = 0$ when $y = 3$.

τ	1	100	1000
BF_{01}	0.15	1.12	11.13
$\mathbf{T}(\mathbf{y}, \theta_0)$	6.25	8.00	8.00

$\mathbf{T}(\mathbf{y}, \theta_0)$ *by*

$$\frac{1}{M}\sum_{m=1}^{M}\left(\log p\left(\mathbf{y}|\theta^{(m)}, \psi^{(m)}\right) - \log p\left(\mathbf{y}|\theta_0, \psi^{(m)}\right)\right),$$

where $\{\theta^{(m)}, \psi^{(m)}\}$, $m = 1, 2, \cdots, M$, *are the draws, generated by the MCMC technique, from the posterior distribution under the alternative hypothesis.*

3.3.2 Latent variable models

In many cases, $p(\mathbf{y}|\vartheta)$ does not have a closed-form expression. For example, in latent variable models, $p(\mathbf{y}|\vartheta)$ often involves integrals that cannot not be solved analytically. In this section, we show how to approximate $\mathbf{T}(\mathbf{y}, \theta_0)$ with the EM algorithm and the MCMC output. To do so, we first impose the following set of regularity conditions.

Assumption 1: The likelihood of the model considered is regular.

Assumption 2: The data generating process is strictly stationary.

Assumption 3: There exists a finite sample size n^*, so that, for $n > n^*$, there is a local maximum at $\hat{\vartheta}$ such that $L_n^{(1)}(\hat{\vartheta}) = 0$ and $L_n^{(2)}(\hat{\vartheta})$ 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 4: The largest eigenvalue λ_n of $\left[-L_n^{(2)}(\hat{\vartheta})\right]^{-1}$ goes to zero when $n \to \infty$.

Assumption 5: For any $\varepsilon > 0$, there exists an integer N and some $\delta > 0$ such that for any $n > \max\{N, n^*\}$ and $\vartheta \in H(\hat{\vartheta}, \delta) = \{\vartheta : ||\vartheta - \hat{\vartheta}|| \leq \delta\}$, $L_n^{(2)}(\vartheta)$ exists and satisfies

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

where $\mathbf{I}_{p+q}$ is an identity matrix and $A(\varepsilon)$ is a positive semidefinite symmetric matrix whose largest eigenvalue goes to zero as $\varepsilon \to 0$. When $\theta = \theta_0$, this assumption also holds.

Assumption 6: For any $\delta > 0$, as $n \to \infty$,

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

where Ω is the support space of ϑ.

Assumption 7: For any $\delta > 0$, when $\vartheta \in H(\hat{\vartheta},\delta)$, $L_n^{(2)}(\vartheta)/n = O_p(1)$.

Remark 3.3.5 *These assumptions are mild regularity conditions and have been used in the literature to develop Bayesian large sample theory; see, for example, Chen (1985), Kim (1994, 1998), Geweke (2005). Based on these regularity conditions, Li et al. (2012) showed that, conditional on the observed data* $\mathbf{y}$*,*

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

where

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

Theorem 3.3.2 *Let* $\bar{\vartheta} = (\bar{\theta},\bar{\psi})'$ *be the posterior mean of* ϑ *under* H_1*,* $\bar{\vartheta}_* = (\theta_0,\bar{\psi})'$*,* $\bar{\vartheta}_b = (1-b)\bar{\vartheta}_* + b\bar{\vartheta}, b \in [0,1]$*,* $S(\mathbf{x}|\vartheta) = \partial \log p(\mathbf{x}|\vartheta)/\partial\vartheta$*,*

$$D = \int_0^1 \left\{(\bar{\theta} - \theta_0)' \left[E_{\mathbf{z}|\mathbf{y},\bar{\vartheta}_b} S_1(\mathbf{x}|\bar{\vartheta}_b)\right]\right\} db, \tag{3.3.3}$$

where $S_1(\mathbf{x}|\vartheta)$ *is the subvector of* $S(\mathbf{x}|\vartheta)$ *corresponding to* θ*. Let*

$$\begin{aligned}
\mathbf{T}_1(\mathbf{y},\theta_0) \;=\; & 2D + 2\left[\log p(\bar{\theta},\bar{\psi}) - \log p(\bar{\psi}|\theta_0)\right] - 2\left[\int \log p(\theta|\psi)p(\vartheta|\mathbf{y})d\vartheta\right]\\
& - \left[p + q - \mathbf{tr}[-L_{0n}^{(2)}(\bar{\psi})V_{22}(\bar{\vartheta})]\right].
\end{aligned} \tag{3.3.4}$$

where $V_{22}(\bar{\vartheta}) = E[(\psi - \bar{\psi})(\psi - \bar{\psi})'|\mathbf{y}, H_1]$*, which is the submatrix of* $V(\bar{\vartheta})$ *corresponding to* ψ *and*

$$L_{0n}^{(2)}(\psi) = \frac{\partial^2 \log p(\mathbf{y}, \psi|\theta_0)}{\partial\psi\partial\psi'}.$$

Under Assumptions 1-7, it can be shown that

$$\mathbf{T}_1(\mathbf{y}, \theta_0) = \mathbf{T}(\mathbf{y}, \theta_0) + o_p(1). \tag{3.3.5}$$

Remark 3.3.6 *According to (3.3.5) we can approximate* $\mathbf{T}(\mathbf{y}, \theta_0)$ *by* $\mathbf{T}_1(\mathbf{y}, \theta_0)$.

Remark 3.3.7 *In many cases, the analytical form of D is not available. Following Gelman and Meng (1998), if D does not have a closed form expression, we can numerically approximate it using the trapezoidal rule. In particular, we can choose a set of fixed grids* $\{b_{(s)}\}_{s=0}^{S}$ *such that* $b_0 = 0 < b_{(1)} < b_{(2)} < \cdots < b_{(S)} < b_{(S+1)} = 1$*, and then approximate D by*

$$\widehat{D} = \frac{1}{2}\left(\bar{\theta} - \bar{\theta}_0\right)' \sum_{s=0}^{S} \left(b_{(s+1)} - b_{(s)}\right) \left(E_{\mathbf{z}|\mathbf{y}, \bar{\vartheta}_{b_{(s)}}} \left[S_1\left(\mathbf{x}|\bar{\vartheta}_{b_{(s)}}\right)\right] + E_{\mathbf{z}|\mathbf{y}, \bar{\vartheta}_{b_{(s+1)}}} \left[S_1\left(\mathbf{x}|\bar{\vartheta}_{b_{(s+1)}}\right)\right] \right). \tag{3.3.6}$$

To calculate $E_{\mathbf{z}|\mathbf{y}, \bar{\vartheta}_{b_{(s)}}} \left[S_1\left(\mathbf{x}|\bar{\vartheta}_{b_{(s)}}\right)\right]$*, we use*

$$E_{\mathbf{z}|\mathbf{y}, \bar{\vartheta}_{b_{(s)}}} \left[S_1\left(\mathbf{x}|\bar{\vartheta}_{b_{(s)}}\right)\right] = E_{\mathbf{z}|\mathbf{y}, \bar{\vartheta}_{b_{(s)}}} \left[S_1\left(\mathbf{y}, \mathbf{z}|\bar{\vartheta}_{b_{(s)}}\right)\right] \approx M^{-1} \sum_{m=1}^{M} S_1\left(\mathbf{y}, \mathbf{z}^{(m)}|\bar{\vartheta}_{b_{(s)}}\right),$$

where $\{\mathbf{z}^{(m)}, m = 1, 2, \cdots, M\}$ *are efficient random observations simulated from* $p\left(\mathbf{z}|\mathbf{y}, \bar{\vartheta}_{b_{(s)}}\right)$ *with* $\bar{\vartheta}_{b_{(s)}} = (1 - b_{(s)})\bar{\vartheta} + b_{(s)}\bar{\vartheta}_*$ *after discarding some burn-in samples. With D being replaced by* $\widehat{D}$ *in (3.3.6), we can approximate* $\mathbf{T}_1(\mathbf{y}, \theta_0)$ *by* $\hat{\mathbf{T}}_1(\mathbf{y}, \theta_0)$.

Remark 3.3.8 *The test statistic clearly requires the evaluation of the observed information matrix, the second derivative of the observed-data likelihood function. For most latent variable models, the observed-data likelihood function does not have a closed-from expression so that the second derivatives are difficult to evalu-*

ate. It is noted that

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

In the EM algorithm, under the mild regularity conditions, if $\mathscr{Q}(\cdot|\cdot)$ has a closed form expression, Oakes (1999) showed that the observed information matrix can be expressed as:

$$\mathbf{I}(\vartheta) = -\frac{\partial^2 L_o(\mathbf{y}|\vartheta)}{\partial\vartheta\partial\vartheta'} = \left\{ -\frac{\partial^2 \mathscr{Q}(\vartheta|\vartheta^*)}{\partial\vartheta\partial\vartheta'} - \frac{\partial^2 \mathscr{Q}(\vartheta|\vartheta^*)}{\partial\vartheta\partial\vartheta^{*\prime}} \right\}_{\vartheta^*=\vartheta}. \tag{3.3.7}$$

When $\mathscr{Q}(\cdot|\cdot)$ does not have a closed form expression, Louis (1982) derived the observed information matrix as:

$$\begin{aligned} \mathbf{I}(\vartheta) &= E_{(\mathbf{z}|\mathbf{y},\vartheta)} \left\{ -\frac{\partial^2 L_c(\mathbf{x}|\vartheta)}{\partial\vartheta\partial\vartheta'} \right\} - Var_{(\mathbf{z}|\mathbf{y},\vartheta)} \{S(\mathbf{x}|\vartheta)\} \qquad (3.3.8) \\ &= E_{(\mathbf{z}|\mathbf{y},\vartheta)} \left\{ -\frac{\partial^2 L_c(\mathbf{x}|\vartheta)}{\partial\vartheta\partial\vartheta'} - S(\mathbf{x}|\vartheta) S(\mathbf{x}|\vartheta)' \right\} + E_{(\mathbf{z}|\mathbf{y},\vartheta)} \{S(\mathbf{x}|\vartheta)\} E_{(\mathbf{z}|\mathbf{y},\vartheta)} \{S(\mathbf{x}|\vartheta)\}', \end{aligned}$$

where the expectations are taken with respect to the conditional distribution of $\mathbf{z}$ *given* $\mathbf{y}$ *and* ϑ*. Hence, the information matrix can be approximated by:*

$$\begin{aligned} & E_{(\mathbf{z}|\mathbf{y},\vartheta)} \left\{ -\frac{\partial^2 L_c(\mathbf{x}|\vartheta)}{\partial\vartheta\partial\vartheta'} - S(\mathbf{x}|\vartheta) S(\mathbf{x}|\vartheta)' \right\} \\ \approx & -\frac{1}{M} \sum_{m=1}^{M} \left\{ \frac{\partial^2 L_c\left(\mathbf{y}, \mathbf{z}^{(m)}|\vartheta\right)}{\partial\vartheta\partial\vartheta'} + S\left(\mathbf{y}, \mathbf{z}^{(m)}|\theta\right) S\left(\mathbf{y}, \mathbf{z}^{(m)}|\vartheta\right)' \right\}, \\ & E_{(\mathbf{z}|\mathbf{y},\vartheta)} \{S(\mathbf{x}|\vartheta)\} \approx \frac{1}{M} \sum_{m=1}^{M} S\left(\mathbf{y}, \mathbf{z}^{(m)}|\vartheta\right), \end{aligned}$$

where $\{\mathbf{z}^{(m)}, m = 1, 2, \cdots, M\}$ are the efficient random draws from the conditional distribution $p(z|\mathbf{y}, \vartheta)$.

3.3.3 Choosing threshold values

To implement the proposed method, we need to specify a threshold value. We shall use the following decision rule to test the hypothesis:

$$\text{Accept } H_0 \text{ if } \mathbf{T}(\mathbf{y},\theta_0) \le C; \text{ Reject } H_0 \text{ if } \mathbf{T}(\mathbf{y},\theta_0) > C,$$

where C is the threshold value to be specified. The following theorem gives the asymptotic distribution of the test statistic. The threshold value can be then set to be a certain percentile of the asymptotic distribution. This compares favorably with Jeffreys' subjective threshold values for BF (Jeffreys (1961)) and the threshold values used in LY.

Theorem 3.3.3 *When the likelihood information dominates the prior information, under Assumptions 1-7, we have, under the null hypothesis*

$$\mathbf{T}(\mathbf{y},\theta_0) + \left[p+q-\mathbf{tr}[-L_{0n}^{(2)}(\bar{\vartheta})V_{22}(\bar{\vartheta})]\right] \overset{a}{\sim} \varepsilon' \left[\mathbf{IJ}_{11}^{1/2}(\vartheta_0)\mathbf{J}_{11}(\vartheta_0)\mathbf{IJ}_{11}^{1/2}(\vartheta_0)\right]\varepsilon, \quad (3.3.9)$$

$$\mathbf{T}_1(\mathbf{y},\theta_0) + \left[p+q-\mathbf{tr}[-L_{0n}^{(2)}(\bar{\vartheta})V_{22}(\bar{\vartheta})]\right] \overset{a}{\sim} \varepsilon' \left[\mathbf{IJ}_{11}^{1/2}(\vartheta_0)\mathbf{J}_{11}(\vartheta_0)\mathbf{IJ}_{11}^{1/2}(\vartheta_0)\right]\varepsilon, \quad (3.3.10)$$

where ε is a standard multivariate normal variate, $\vartheta_0 = (\theta_0, \psi_0)$ the true value of ϑ, $\mathbf{J}(\vartheta_0)$ the Fisher information matrix given by

$$\mathbf{J}(\vartheta_0) = \frac{1}{n}\int -L_n^{(2)}(\vartheta_0)p(\mathbf{y}|\vartheta_0)d\mathbf{y},$$

$\mathbf{IJ}(\vartheta_0)$ the inverse of $\mathbf{J}(\vartheta_0)$, $\mathbf{J}_{11}(\vartheta_0)$ and $\mathbf{IJ}_{11}(\vartheta_0)$ the submatrices of $\mathbf{J}(\vartheta_0)$ and $\mathbf{IJ}(\vartheta_0)$, respectively, corresponding to θ.

Remark 3.3.9 *In general, the asymptotic distributions of $\mathbf{J}_{11}(\vartheta_0)$ and $\mathbf{IJ}_{11}(\vartheta_0)$ are not known. Fortunately, when the alternative hypothesis is assumed to be the true model, $\mathbf{J}(\vartheta_0)$ and $\mathbf{IJ}(\vartheta_0)$ can be consistently estimated by*

$$\mathbf{J}(\vartheta_0) \approx -\frac{1}{n}L_n^{(2)}(\bar{\vartheta}), \mathbf{IJ}(\vartheta_0) \approx nV(\bar{\vartheta}).$$

This greatly facilitates the calculation of the asymptotic distribution.

Remark 3.3.10 *To obtain the asymptotic distribution and the threshold values, since the middle term in the asymptotic distribution,* $\mathbf{IJ}_{11}^{1/2}(\vartheta_0)\mathbf{J}_{11}(\vartheta_0)\mathbf{IJ}_{11}^{1/2}(\vartheta_0)$*, only depends on the model and the data, one only needs to simulate from the standard multivariate normal.*

In some cases, there is no need to simulate the asymptotic distributions of $\mathbf{T}(\mathbf{y},\theta_0)$ and $\mathbf{T}_1(\mathbf{y},\theta_0)$. The following theorem gives such a situation.

Theorem 3.3.4 *If* θ *and* ψ *are orthogonal,* $\mathbf{tr}\left[\mathbf{J}_{22}(\vartheta_0)\mathbf{IJ}_{22}(\vartheta_0)\right]=q$*,* $\mathbf{IJ}_{11}^{1/2}(\vartheta_0)\mathbf{J}_{11}(\vartheta_0)\mathbf{IJ}_{11}^{1/2}(\vartheta_0)=\mathbf{I}_p$*,* $\mathbf{T}(\mathbf{y},\theta_0)\overset{a}{\sim}\chi^2(p)-p$*, and* $T_1(\mathbf{y},\theta_0)\overset{a}{\sim}\chi^2(p)-p$*, where* $\mathbf{J}_{22}(\vartheta_0)$ *and* $\mathbf{IJ}_{22}(\vartheta_0)$ *are the submatrices of* $\mathbf{J}(\vartheta_0)$ *and* $\mathbf{IJ}(\vartheta_0)$ *corresponding to* ψ*.*

Remark 3.3.11 *Theorem 3.3.4 can be simply derived from Theorem 3.3.3. While the likelihood ratio statistic asymptotically follows* $\chi^2(p)$ *and is always positive, the Bayesian version of the likelihood ratio statistic proposed in the present paper asymptotically follows* $\chi^2(p)-p$*. The mean of the asymptotic distribution is zero and hence it is possible that our statistic takes a negative value.*

3.4 Examples

In this section, we illustrate the proposed theory using three examples in economics and finance. In the first example, we compare the performance of BF and that of $\mathbf{T}(\mathbf{y},\theta_0)$ in the context of simple linear regression model, aiming to explore the presence of Jeffreys-Lindley's paradox in BF and the absence of Jeffreys-Lindley's paradox in the proposed method. In the second example, we check the quality of the approximation of $\mathbf{T}_1(\mathbf{y},\theta_0)$ and $\hat{\mathbf{T}}_1(\mathbf{y},\theta_0)$ to $\mathbf{T}(\mathbf{y},\theta_0)$ in the context of linear asset pricing model. In this case both the observed-data log-likelihood and the complete-data log-likelihood have the analytical form. In the third example, we test the presence of leverage effect in a stochastic volatility (SV) model. Since the observed-data log-likelihood is not available in closed-form for the SV model, only $\hat{\mathbf{T}}_1(\mathbf{y},\theta_0)$ is obtained.

3.4.1 Testing the significance in a simple linear regression model

Consider the following simple linear regression model:

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

Denote $\mathbf{y} = (y_1, y_2, ..., y_n)'$ and $X = (x_1, x_2, ..., x_n)'$. We are interested in knowing whether or not the explanatory variable x_i has an explanatory power for y_i, i.e., we test

$$H_0 : \beta = 0,\; H_1 : \beta \neq 0\,.$$

The prior distributions for β and σ^2 are set at

$$\beta \sim N\left(\mu_\beta, V_\beta\right), \sigma^2 \sim IG(a, b)\,.$$

In this example, $\theta = \beta, \psi = \sigma^2$, and the likelihood function has a closed-form expression. Thus, $\mathbf{T}(\mathbf{y}, \theta_0)$ can be computed. Also note that β is orthogonal to σ^2 and, hence, $\mathbf{T}(\mathbf{y}, \theta_0) \overset{a}{\sim} \chi^2(1) - 1$. The marginal likelihood of data can be expressed, under H_0, as:

$$p_0(\mathbf{y}) = \frac{b^a \Gamma\left(a + \frac{n}{2}\right)}{(2\pi)^{n/2} \Gamma(a)} \left[b + \frac{1}{2}\mathbf{y}'\mathbf{y}\right]^{-(a+n/2)},$$

and under H_1, as:

$$p_1(\mathbf{y}) = \frac{b^a \Gamma\left(a + \frac{n}{2}\right) \sqrt{|V^*|}}{(2\pi)^{n/2} \Gamma(a) \sqrt{|V_\beta|}} \left[b + \frac{1}{2}\left(\mu_\beta' V_\beta^{-1} \mu_\beta + \mathbf{y}'\mathbf{y} - \mu^{*\prime} V_\beta^{*-1} \mu^*\right)\right]^{-(a+n/2)},$$

where

$$\mu^* = V^*\left(V_\beta^{-1}\mu_\beta + X'\mathbf{y}\right)\,,\; V^* = \left(V_\beta^{-1} + X'X\right)^{-1}.$$

Hence, $BF_{01} = p_0(\mathbf{y}) / p_1(\mathbf{y})$ has an analytical expression.

To explore the presence of Jeffreys-Lindley's paradox in BF and the absence of Jeffreys-Lindley's paradox in our proposed test, we consider an example used in Wooldridge (2009) (Page 45). In this example, a linear relationship between CEO

salary and firm sales is established. To compute $\mathbf{T}(\mathbf{y}, \theta_0)$, we apply Gibbs sampler to the model corresponding to the alternative hypothesis to carry out the Bayesian analysis. We set the parameters in the priors at:

$$\mu_\beta = 0, \;\; a = 0.001, \;\; b = 0.001,$$

but leave the value of the prior variance V_β varied for the purpose of examining how V_β influences the decision based on BF_{01} and $\mathbf{T}(\mathbf{y}, \theta_0)$, respectively. For the Bayesian MCMC analysis, 10,000 random draws are sampled from the posterior distribution after 1,000 burn-in periods.

The results are reported in Table 2. From this table, we see that as V_β increases, BF_{01} also increases. When the prior variance V_β is moderate, BF is less than 1 and tends to reject the null hypothesis. However, when V_β is large enough, the BF tends to support the null hypothesis. This clearly demonstrates Jeffreys-Lindley's paradox. On the other hand, $\mathbf{T}(\mathbf{y}, \theta_0)$ takes nearly identical values with different V_β. Consequently, $\mathbf{T}(\mathbf{y}, \theta_0)$ is immune to Jeffreys-Lindley's paradox. To test the hypothesis using the proposed theory, since θ and σ^2 are orthogonal to each other, the asymptotic distribution of $\mathbf{T}(\mathbf{y}, \theta_0)$ is $\chi^2(1) - 1$. The 99%, 95%, 90% percentiles of $\chi^2(1) - 1$ are 5.65, 2.84, 1.71. The test statistic $\mathbf{T}(\mathbf{y}, \theta_0)$ is 40.12, suggesting that the null hypothesis is rejected under the 99%, 95%, 90% probability levels. When the frequentist's approach is used, the OLS estimate of β is 0.26 and the standard error is 0.03. This suggests that the null hypothesis has to be rejected, consistent with the finding from our method.

Table 3.2: Testing the significance in a simple linear regression model

V_β	0.1	100	10^5	10^{22}	10^{25}	10^{35}
BF_{01}	2.95×10^{-10}	2.63×10^{-9}	8.32×10^{-8}	26.3051	831.8407	8.31×10^{7}
$\mathbf{T}(\mathbf{y}, \theta_0)$	40.1209	40.1205	40.1205	40.1205	40.1205	40.1205

3.4.2 Hypothesis tests in asset pricing models with heavy tails

Asset pricing theory is a central focus of modern finance. Many econometric approaches have been developed to test asset pricing models. Most of the tests were developed based on the normality assumption, which is often violated in return data due to the presence of heavy tails. The heavy tails have motivated some researchers to develop asset pricing models with heavy-tailed distributions, see Zhou (1993), **?**, and Li and Yu (2012). In this subsection, we apply the proposed method to check the validity of Fama-French three factor asset pricing model (Fama and French (1993)) with a multivariate t distribution.

This asset pricing model with multivariate t distribution can be simply expressed as:

$$\mathbf{R}_t = \alpha + \beta_1 M_t + \beta_2 SMB_t + \beta_3 HML_t + \varepsilon_t, \varepsilon_t \sim t(0, \Sigma, \nu),$$

where $\mathbf{R}_t$ is the excess return of portfolio at period t with $N \times 1$ dimension, M_t the excess return of the whole stock market, SMB_t and HML_t stands for "small (market capitalization) minus big" and for "high (book-to-market ratio) minus low" which measures the historical excess returns of small caps over big caps and of value stocks over growth stocks, Σ a diagonal matrix, and ν the freedom of degree of t distribution which is assumed to be known for the illustrative purpose and for convenience.

Let $\beta = (\beta_1, \beta_2, \beta_3)'$, $\mathbf{F}_t = (M_t, SMB_t, HML_t)'$. As noted in Kan and Zhou (2006), using the scale mixture representation for t distribution, this model can be equivalently specified as:

$$\mathbf{R}_t = \alpha + \beta F_t + \varepsilon_t, \quad \varepsilon_t \sim N(0 \times 1_N, \Sigma/\omega_t), \quad \omega_t \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right).$$

The mean-variance efficiency suggests that the excess premium α should be zero. Hence, the hypothesis to be tested is given by:

$$H_0 : \alpha = 0 \times 1_N, H_1 : \alpha \neq 0 \times 1_N,$$

where 1_N is an N-dimensional vector with unit elements.

As in the previous example, the likelihood function has a closed-form expression and, hence, both D and $\mathbf{T}(\mathbf{y}, \theta_0)$ can be computed. The purpose of this example is to check the quality of approximation of $\mathbf{T}_1(\mathbf{y}, \theta_0)$ and $\widehat{\mathbf{T}}_1(\mathbf{y}, \theta_0)$ when ω_t is regarded as latent variables.

In this empirical analysis, we consider the monthly returns of 25 portfolios constructed at the end of each June on the basis of the intersections of 5 portfolios formed on size (market equity, ME) and 5 portfolios formed on the ratio of book equity to market equity (BE/ME). This sample period is ranged from July 1926 to July 2011 so that $N = 25$, $T = 1002$. The data are freely available from the data library of Kenneth French.[1]

As noted in Kan and Zhou (2006), it is not easy to make the statistical inference using optimization-based ML methods. Hence, we consider Bayesian statistical inference coupled with MCMC techniques. Following Li and Yu (2012), we assign the vague conjugate prior distributions to represent the prior ignorance as follows:

$$\alpha_i \sim N[0, 100], \beta_i \sim N[0, 100], \Sigma_{ii}^{-1} \sim \Gamma[0.001, 0.001],$$

and set $\nu = 3$.

In this Bayesian analysis, 100,000 random samples are draw from the posterior distribution using Gibbs sampler. The convergence of Gibbs sampler is checked using the Raftery-Lewis diagnostic test statistic (Raftery and Lewis (1992)). The first 50,000 random samples are discarded as burning-in samples. To check the quality of approximation of $\mathbf{T}_1(\mathbf{y}, \theta_0)$ and $\hat{\mathbf{T}}_1(\mathbf{y}, \theta_0)$ to $\mathbf{T}(\mathbf{y}, \theta_0)$, we choose $S = 20$ and set the equal distance between $b_{(s)}$ and $b_{(s+1)}$ for $s = 0, 1, \cdots, 21$.

The results are reported in Table 3. From Table 3, we find that $\widehat{D}$ well approximates D. Not surprisingly, $\hat{\mathbf{T}}_1(\mathbf{y}, \theta_0)$, which is based on $\widehat{D}$, well approximates $\mathbf{T}_1(\mathbf{y}, \theta_0)$ which in turn well approximates $\mathbf{T}(\mathbf{y}, \theta_0)$. All three values are around 141. To obtain the threshold values, we estimate $\mathbf{IJ}_{11}^{1/2}(\vartheta_0)\mathbf{J}_{11}(\vartheta_0)\mathbf{IJ}_{11}^{1/2}(\vartheta_0)$ in

[1] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Table 3.3: Asset pricing testing for Fama-French three factor models

Statistics	D	$\widehat{D}$	$\mathbf{T}_1(\mathbf{y},\theta_0)$	$\hat{\mathbf{T}}_1(\mathbf{y},\theta_0)$	$\mathbf{T}(\mathbf{y},\theta_0)$
Value	82.5833	82.5826	141.1921	141.1914	140.5004

(3.3.9), simulate 1,000 random vectors from the standard multivariate normal variate, and then obtain 1,000 random numbers for $\hat{\mathbf{T}}_1(\mathbf{y},\theta_0)$. From these random numbers, we obtain the following threshold values: $C = 20.2657$ under 99%, $C = 15.5040$ under 95% and $C = 11.3610$ under 90%. Consequently, we reject the null hypothesis under all the probability levels.

3.4.3 Testing the leverage effect in a stochastic volatility model

Stochastic volatility (SV) models have been widely used for pricing options. An important and well documented empirical feature in many financial time series is the financial leverage effect (Black (1976); Christie (1982)). Following Yu (2005), we define the leverage effects SV model as follows:

$$\begin{aligned} y_t|h_t &= \exp(h_t/2)\,u_t, \quad t = 1,...,n, \\ h_{t+1}|h_t,\mu,\phi,\tau^2,\rho &= \mu + \phi\,(h_t-\mu) + \tau v_{t+1}, \quad t = 0,...,n, \end{aligned}$$

with

$$\begin{pmatrix} u_t \\ v_{t+1} \end{pmatrix} \overset{i.i.d}{\sim} N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\},$$

and $h_0 = \mu$, where y_t is the return at time t, h_t the return volatility at period t, ρ the leverage effect parameter. The hypothesis that we test is $H_0 : \rho = 0$.

To carry out Bayesian test of the hypothesis, we use the data that consist of daily returns on Pound/Dollar exchange rates $\{x_t\}$ from 01/10/81 to 28/06/85. The series $\{y_t\}$ is the daily mean-corrected returns. We first estimate the model using

the Bayesian MCMC method. The following vague priors are specified:

$$\mu \sim N[0,100],\ \ \phi \sim Beta[1,1],\ \ \tau^{-2} \sim \Gamma[0.001,0.001],\ \ \rho \sim U[-1,1].$$

The parameter estimates are based on 100,000 iterations after a burn-in of 10,000. To calculate $\widehat{\mathbf{T}}_1(\mathbf{y},\theta_0)$, we take $s=20$, set the equal distances between $b_{(s)}$ and $b_{(s+1)}$ for $s=0,1,...,20$ and find $\widehat{\mathbf{T}}_1(\mathbf{y},\theta_0)=-1.7244$. From simulations, the threshold values are $C=5.1041$ under 99%, $C=2.2291$ under 95% and $C=1.0600$ under 90%. Hence, the null hypothesis cannot be rejected under all three probability levels.

3.5 Conclusion

In this paper, we have proposed a new Bayesian statistic to test a point null hypothesis. The main advantages of the new statistic are fourfold. First, it is immune to Bartlett's paradox. Second, it avoids Jeffreys-Lindley's paradox. Third, it can be easily computed using the MCMC outputs from the posterior distribution. Fourth, the asymptotic distribution can be derived for calibrating the threshold values. The proposed method is illustrated using a simple linear regression model, an asset pricing model and a stochastic volatility model.

Chapter 4 A Bayesian Specification Test for Latent Variable Models

4.1 Introduction

Economic theory has long been used to justify a particular choice of econometric models. It almost always does do by using a set of economic assumptions. When some of these assumptions are invalid, the corresponding econometric models may be misspecified. In a worse scenario, economic theory may not be available and the choice of econometric model can then be more arbitrary and, hence, the model is more vulnerable to the specification errors.

In general misspecification of econometric models can potentially lead to inconsistent estimation, which in turn may have serious implications for statistical inferences such as hypothesis testing and out-of-sample forecasting and serious implications for economic decision making such as policy recommendation and investment decision. Consequently and not surprisingly, considerable amount of efforts have been devoted in econometrics to detect model misspecification.

One strand of the literature on specification tests unifies under the m-test of Newey (1985), Tauchen (1985) and White (1987). These tests include as a special case of the Lagrange multipler (LM) test, the tests of Sargan (1958) and Hansen (1982) , the tests of Cox (1961, 1962), the Hausman (1978) test, the information matrix test of White (1982), the conditional moment test of Newey (1985) , the *IOS* test of Presnell and Boos (2004). These tests typically require the parameters in the null hypothesis are estimated by the maximum likelihood (ML) method, or the

generalized method of moments (GMM).

Another strand of the literature is based on distances between nonparametric and parameter counterparts. The idea originated from the Kolmogorov-Smirnov test or the closely related family such as the Cramer-von Mises and Andersen-Darling tests. Examples in this case include Eubank and Spiegelman (1990), Wooldrige (1992), Fan and Li (1996), Gozalo (1993), Zheng (2000), Ait-Sahalia (1996), and Hong and Li (2005). All the tests in this category require either a nonparametric estimate of a function or a nonparametric estimate of a density (either a marginal density or a conditional density).

For many widely used latent variable models in economics, such as nonlinear non-Gaussian state space models, it is not easy to obtain the ML estimate or construct a nonparametric estimate. Not surprisingly, it is difficult to apply the specification tests in the above mentioned two strands of the literature. On the other hand, there has been increasing interest in Bayesian methods to analyze latent variable models. With the advancement of the Markov chain Monte Carlo (MCMC) algorithms and the rapid growth in computer capability, the estimation of latent variable models has become increasingly easier and easier.

Given the popularity of Bayesian MCMC methods for estimating latent variable models, it is therefore natural to introduce a Bayesian test to assess the goodness-of-fit of the model. Unfortunately, model specification test is a challenge in the Bayesian paradigm. Perhaps the most obvious way to assess the goodness-of-fit of the model in the Bayesian paradigm is to compare the posterior model probability in consideration with the posterior model probability of a competing model. This can be achieved by using, for example, Bayes factors (BFs), although BFs are not free of problems. However, it is often not clear how to specify the alternative model and empirical researchers may simply wish to know if the model she employs is adequate or not without worrying about any alternative model.

The question we ask in the present paper is, after the model is estimated by a Bayesian approach, how we can assess the validity of the model specification. The

main purpose of this paper is to introduce a Bayesian approach to testing model specification without specifying an alternative model. The proposed Bayesian test statistic is the Bayesian version of a m-type test. We show how to compute the test statistic from MCMC output in the context of latent variable models. To implement our method, threshold values are needed. We then show that the threshold values can be obtained using Monte Carlo simulations.

The paper is organized as follows. Section 2 briefly reviews the literature on the misspecification tests. Section 3 proposes the new Bayesian test statistic and show how to compute the statistic and the threshold values in latent variable models. Section 4 illustrates the new method using a real example in finance and a real example macroeconomics. Section 5 concludes the paper. Appendix collects the proof of the theoretical results in the paper and the derivation of the quantities needed to compute the statistic.

4.2 Specification Tests: A Literature Review

To begin, let $\mathbf{Y} = (y_1, ..., y_n)$ denote observations drawn from a probability measure P_0 on the probability space (Ω, F, P_0). Let model P be a collection of candidate models indexed by parameters θ whose dimension is p. Denote P indexed by θ by P_θ. Following White (1987), if there exists θ, such that $P_0 \in P_\theta$, we say the model P is correctly specified. However, if for all θ, $P_0 \notin P_\theta$, we say the model P is misspecified.

One of the earliest specification test is based on the informative matrix equivalence due to White (1982). Let $p(\mathbf{Y}|\theta)$ denote the likelihood function of model P and

$$\mathbf{s}(\mathbf{Y}, \theta) := \partial \log p(\mathbf{Y}|\theta)/\partial\theta, \mathbf{h}(\mathbf{Y}, \theta) := \partial^2 \log p(\mathbf{Y}|\theta)/\partial\theta\partial\theta',$$
$$\mathbf{H}(\theta) := \int \mathbf{h}(\mathbf{Y}, \theta) p(\mathbf{Y}|\theta) \mathrm{d}\mathbf{Y}, \mathbf{J}(\theta) := \int \mathbf{s}(\mathbf{Y}, \theta)\mathbf{s}'(\mathbf{Y}, \theta) p(\mathbf{Y}|\theta) \mathrm{d}\mathbf{Y}.$$

When the model is correctly specified, note that

$$\int p(\mathbf{Y}|\theta)\mathrm{d}\mathbf{Y} = 1.$$

Differentiate the above equation twice and use the fact that $\partial f/\partial\theta = f\partial \log f/\partial\theta$, we get

$$0 = \int \left[\mathbf{h}(\mathbf{Y},\theta) + \mathbf{s}(\mathbf{Y},\theta)\mathbf{s}'(\mathbf{Y},\theta)\right] p(\mathbf{Y}|\theta)\mathrm{d}\mathbf{Y} = \mathbf{H}(\theta) + \mathbf{J}(\theta).$$

Define

$$d(\mathbf{Y},\theta) := vech\left[\mathbf{h}(\mathbf{Y},\theta) + \mathbf{s}(\mathbf{Y},\theta)\mathbf{s}'(\mathbf{Y},\theta)\right],$$

where *vech* is the column-wise vectorization with the upper portion excluded. Hence, $d(\mathbf{Y},\theta) = (d_k(\mathbf{Y},\theta))$ is a $q := p(p+1)/2$ dimensional vector. Denote

$$\hat{\mathbf{H}}(\hat{\theta}) := \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}(y_i,\hat{\theta}), \hat{\mathbf{J}}(\hat{\theta}) := \frac{1}{n}\sum_{i=1}^{n}\mathbf{s}(y_i,\hat{\theta})\mathbf{s}'(y_i,\hat{\theta}),$$

where $\hat{\theta}$ is the maximum likelihood (ML) estimator of θ. The corresponding elements of $\hat{\mathbf{H}}(\hat{\theta}) - \hat{\mathbf{J}}(\hat{\theta})$ is given by

$$D_{nk}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} d_k(y_i,\hat{\theta}), k = 1,\cdots,q.$$

White (1982) proposed the following test

$$IMT = nD_n(\hat{\theta})V_n^{-1}(\hat{\theta})D_n(\hat{\theta}), \tag{4.2.1}$$

where $V_n(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[d(y_i,\hat{\theta}) + \dot{d}(y_i,\hat{\theta})\hat{\mathbf{H}}^{-1}(\hat{\theta})\mathbf{s}(y_i,\hat{\theta})\right]\left[d(y_i,\hat{\theta}) + \dot{d}(y_i,\hat{\theta})\hat{\mathbf{H}}^{-1}(\hat{\theta})\mathbf{s}(y_i,\hat{\theta})\right]'$, $D_n(\hat{\theta}) = (D_{nk}(\hat{\theta}))$ is a q-dimensional vector and $\dot{d}(y_i,\hat{\theta}) = \partial d(y_i,\hat{\theta})/\partial\theta$ is a $q\times p$ matrix.

When $y_1,...,y_n$ are iid, under regularity conditions, White (1982) showed that $IMT \overset{a}{\sim} \chi^2(q)$ under the null hypothesis. White (1987) extended the method to

cover dynamic models. Lancaster (1984) pointed out how the covariance matrix of the information matrix test can be estimated without computing the third derivatives of the density function analytically. Dhaene and Hoorelbeke (2004) suggested using the bootstrap method to estimate the covariance matrix. Moreover, it is well documented that the asymptotic χ^2 distribution can be a poor approximation in finite sample so that the test statistic suffers from a serious size distortion; see Orme (1990), Chesher and Spady (1991), Davidson and Mackinnon (1992), Horowitz (1994). To improve the finite sample performance of IMT, Chesher and Spady (1991) used the high-order Edgeworth expansion to obtain the better critical values while Horowitz (1994) advocated the use of bootstrap method to get critical values.

To deal with the difficulties associated with the information matrix test, Presnell and Boos (2004) proposed an "in-and-out" likelihood ratio test. Let $\hat{\theta}_{(i)}$ be the MLE of θ when the ith observation y_i is deleted from the whole sample. In the iid framework, from the predictive perspective, the single likelihood $f(y_i, \hat{\theta}_i)$ can be regarded as the predictive likelihood by the other observations. Presnell and Boos (2004) defined the "in-and-out" likelihood ratio test as:

$$IOS = \log \frac{\prod_{i=1}^{n} p(y_i, \hat{\theta})}{\prod_{i=1}^{n} p(y_i, \hat{\theta}_{(i)})} = \sum_{i=1}^{n} \left[\log p(y_i|\hat{\theta}) - \log p(y_i|\hat{\theta}_{(i)}) \right].$$

and showed that the asymptotic form of IOS is

$$IOS_a = \mathbf{tr}\left[-\hat{\mathbf{H}}^{-1}(\hat{\theta}) \hat{\mathbf{J}}(\hat{\theta}) \right].$$

Note that IOS_a is the penalty term of the well-known information criterion TIC developed by Takeuchi (1976). As in TIC, to compute IOS_a, one has to calculate the inverse of $\hat{H}(\hat{\theta})$ which is generally difficult when the dimension of θ is high.

One does not necessarily need to base the specification test on the ML estimation. Newey (1985) developed a class of specification tests which are based on a finite set of moment conditions and the GMM estimator. Under some regularity conditions, like the test of White, the test statistic of Newey follow asymptotically

a chi-square distribution. It was shown that his test includes as a special case of the Hausman (1978) test, the Hansen (1982) test, the Hausman and Taylor (1980) test.

Specification of a stationary dynamic model implicitly implies a distributional assumption for the marginal density and that for the conditional density. Not surprisingly, many specification tests check the validity of these distributional assumptions based on the Kolmogorov-Smirnov test or the closely related family such as the Cramer-von Mises and Andersen-Darling tests. Examples include Zheng (2000), Andrews (1997), Corradi and Swanson (2004), Duan (2004), Ait-Sahalia (1996), and Hong and Li (2005). For example, Ait-Sahalia (1996) compares the parametric marginal density implied by the assumed continuous time model to the marginal density estimated nonparametrically. The nonparametric test of Hong and Li (2005) is based on the transition density.

4.3 A New Bayesian Approach for Specification Test

4.3.1 Latent variable models

The tests reviewed in Section 2 are based on frequentist's methods. For many econometric models, such as dynamic latent variable models, frequentist's methods are less widely used than Bayesian methods for a number of reasons. First, to use the ML method, the likelihood function must have a attractable form. For many models, such as the nonlinear or non-Gaussian state space models, the log-likelihood function of the observed variables is not analytically tractable. Second, efficiency of GMM depends on the choice of moment conditions. In practice, unfortunately, it often lacks of guidelines as to which and how many moment conditions to use and, hence, the GMM estimator may be inefficient than the likelihood based methods. See, for example, Jacquier et al. (2004) for the comparison of GMM and the likelihood based methods in the context of stochastic volatility models. Because of these problems in using the frequentist methods, there has been increasing interest in the Bayesian methods to analyze latent variable models. Today latent variable models

has been routinely estimated by MCMC algorithms.

To introduce a latent variable model, let $\mathbf{Y} = (y_1, ..., y_n)$ denote observed variables and $z = (z_1, \cdots, z_n)$ denote latent variables. The model is given by

$$\begin{cases} y_t = F(z_t, u_t) \\ z_t = G(z_{t-1}, v_t) \end{cases}. \tag{4.3.1}$$

The first equation relating y_t to z_t is the observation equation where u_t is the error term whose distribution is given. The second equation determining the dynamic of the latent variable is the state equation where v_t is the error term whose distribution is also given. When the distribution of u_t or v_t is non-Gaussian or the functional form of F or G is nonlinear, the model is often referred to as the nonlinear non-Gaussian state space model in the literature.

Let $p(\mathbf{Y}|\theta)$ denote the likelihood function of the observed data, and $p(\mathbf{Y}, z|\theta)$, the complete data likelihood function. Obviously these two functions are related to each other by

$$p(\mathbf{Y}|\theta, \psi) = \int p(\mathbf{Y}, z|\theta, \psi) dz. \tag{4.3.2}$$

The complete data likelihood function $p(\mathbf{Y}, z|\theta)$ can be expressed as $p(\mathbf{Y}|z, \theta)p(z|\theta)$. Usually analytical expressions for $p(\mathbf{Y}|z, \theta)$ and $p(z|\theta)$ are given by the specification of the model. In particular, the observation equation gives the analytical expression for $p(\mathbf{Y}|z, \theta)$ while the state equation gives the analytical expression $p(z|\theta)$. However, in general the integral in (4.3.2) does not have an analytical expression. Consequently, the statistical inferences, such as estimation and hypothesis testing, are difficult to implement if they are based on the ML approach. For linear Gaussian state space models, the observed data likelihood function, $p(\mathbf{Y}|\theta, \psi)$, can be computed numerically by the Kalman filter.

Fortunately, the latent variables models can be efficiently estimated in the Bayesian framework by using MCMC techniques. Let $p(\theta)$ be the prior distribution of θ, and $p(\theta|\mathbf{Y})$ be the posterior distribution of θ. The goal of the Bayesian inference is to obtain $p(\theta|\mathbf{Y})$. The data augmentation strategy of Tanner and Wong (1987), that

augments the parameter space with the latent variable z, is a Bayesian method that uses a MCMC algorithm to generate random samples from the joint posterior distribution $p(\theta, z|y)$. For further details about Bayesian estimation of latent variable models via MCMC such as algorithms, examples and references, see Geweke et al. (2011).

4.3.2 A new Bayesian specification test

The problem concerned in this paper is to assess the goodness-of-fit of the model given that the model is estimated by MCMC. Before proposing the test, we need to introduce some notations. Let $y^{1:t} := (y_1, ..., y_n)$, and

$$\begin{aligned}
&\mathbf{s}(y^{1:t}, \theta) := \frac{\partial \log p(y^{1:t}|\theta)}{\partial \theta}, \mathbf{h}(y^{1:t}, \theta) := \frac{\partial^2 \log p(y^{1:t}|\theta)}{\partial \theta \partial \theta'}, \\
&\mathbf{s}(y_t, \theta) := \mathbf{s}(y^{1:t}, \theta) - \mathbf{s}(y^{1:t-1}, \theta), \mathbf{h}(y_t, \theta) := \mathbf{h}(y^{1:t}, \theta) - \mathbf{h}(y^{1:t-1}, \theta), \\
&\hat{\mathbf{J}}(\theta) := \frac{1}{n}\sum_{t=1}^{n} \mathbf{s}(y_t, \theta)\mathbf{s}'(y_t, \theta), \hat{\mathbf{I}}(\theta) := \frac{1}{n}\sum_{t=1}^{n} \mathbf{h}(y_t, \theta), \\
&L_n(\theta) := \log p(\theta|\mathbf{Y}), L_n^{(k)}(\theta) := \partial^k \log p(\theta|\mathbf{Y})/\partial \theta^k.
\end{aligned}$$

In this paper, we assume that the following mild regularity conditions are satisfied.

Assumption 1: Let $\hat{\theta}$ is the posterior mode such that $L_n^{(1)}(\hat{\theta}) = 0$. For any $\varepsilon > 0$, there exists an integer N_1 and some $\delta > 0$ such that for when $n > N_1$ and $\theta \in H(\hat{\theta}, \delta) = \{\theta : ||\theta - \hat{\theta}|| \leq \delta\}$, $L_n^{(2)}(\theta)$ is negative definite.

Assumption 2: The largest eigenvalue of $[-L_n^{(2)}(\hat{\theta})]^{-1}$ tends to zero as $n \to \infty$.

Assumption 3: For any $\varepsilon > 0$, there exists an integer N_2 and some $\delta > 0$ such that for any $n > \max\{N_1, N_2\}$ and $\theta \in H(\hat{\theta}, \delta) = \{\theta : ||\theta - \hat{\theta}|| \leq \delta\}$, $L_n^{(2)}(\theta)$ satisfies the following inequality

$$-A(\varepsilon) \leq L_n^{(2)}(\theta) L_n^{-(2)}(\hat{\theta}) - \mathbf{I}_p \leq A(\varepsilon),$$

where $\mathbf{I}_p$ is a p-dimensional identity matrix, $A(\varepsilon)$ is a positive semidefinite sym-

metric matrix whose largest eigenvalue goes to zero as $\varepsilon \to 0$.

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

$$\int_{\Omega - H(\hat{\theta},\delta)} p(\theta|\mathbf{y})d\theta \xrightarrow{p} 0,$$

where Ω is the support space of θ.

Assumption 5: When the model is correctly specified, let θ_0 is the true value of θ and let

$$\mathbf{J}(\theta) = \int \hat{\mathbf{J}}(\theta)p(\mathbf{Y}|\theta_0)\mathrm{d}\mathbf{Y}, \mathbf{I}(\theta) = \int \hat{\mathbf{I}}(\theta)p(\mathbf{Y}|\theta_0)\mathrm{d}\mathbf{Y}.$$

It is assumed that

$$\hat{\mathbf{J}}(\theta_0) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{s}(y_t,\theta_0)\mathbf{s}'(y_t,\theta_0) \xrightarrow{p} \int \hat{\mathbf{J}}(\theta_0)p(\mathbf{Y}|\theta_0)\mathrm{d}\mathbf{Y} = \mathbf{J}(\theta_0) = O(1),$$
$$\hat{\mathbf{I}}(\theta_0) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{h}(y_t,\theta_0) \xrightarrow{p} \int \hat{\mathbf{I}}(\theta_0)p(\mathbf{Y}|\theta_0)\mathrm{d}\mathbf{Y} = \mathbf{I}(\theta_0) = O(1).$$

Remark 4.3.1 *These mild regularity assumptions have been used to develop Bayesian large sample theory; see, for example, Chen (1985), Kim (1994, 1998), Geweke (2005) . Based on them, Li et al. (2012) showed that,*

$$\bar{\vartheta} = E[\vartheta|\mathbf{Y}] = \int \vartheta p(\vartheta|\mathbf{Y})d\vartheta = \hat{\vartheta} + o(n^{-1/2}),$$
$$V(\hat{\vartheta}) = -L_n^{-(2)}(\hat{\vartheta}) + o(n^{-1}),$$

where

$$V(\tilde{\vartheta}) = E\left[(\vartheta - \tilde{\vartheta})(\vartheta - \tilde{\vartheta})'|\mathbf{Y}\right] = \int (\vartheta - \tilde{\vartheta})(\vartheta - \tilde{\vartheta})' p(\vartheta|\mathbf{Y})d\vartheta.$$

The new Bayesian test statistic is defined as:

$$\mathbf{BT} = n\int (\theta - \bar{\theta})'\hat{\mathbf{J}}(\bar{\theta})(\theta - \bar{\theta})p(\theta|\mathbf{Y})\mathrm{d}\theta. \tag{4.3.3}$$

Theorem 4.3.1 *Under Assumptions 1-5, when the likelihood information domi-*

nates the prior information and the model is correctly specified, we have

$$\mathbf{BT} = \mathbf{tr}\left[-\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\right] + o_p(1) = p + o_p(1).$$

Remark 4.3.2 *If the observed data likelihood function has a close-form expression, it is easy to compute the test statistic* **BT** *from the MCMC output.*

Remark 4.3.3 *In the iid case, it can be shown that,*

$$\begin{aligned}
& \mathbf{BT} = n\mathbf{tr}\left\{\hat{\mathbf{J}}(\bar{\theta})E\left[(\theta-\bar{\theta})(\theta-\bar{\theta})'|\mathbf{y}\right]\right\} \\
= \;& n\mathbf{tr}\left\{\hat{\mathbf{J}}(\bar{\theta})E\left[(\theta-\bar{\theta})(\theta-\bar{\theta})'|\mathbf{y}\right]\right\} \\
= \;& n\mathbf{tr}\left\{\left[\hat{\mathbf{J}}(\hat{\theta})+o_p(1)\right]E\left[(\theta-\bar{\theta})(\theta-\bar{\theta})'|\mathbf{y}\right]\right\} \\
= \;& -\mathbf{tr}\left\{\left[\hat{\mathbf{J}}(\hat{\theta})+o_p(1)\right]\left[\hat{\mathbf{I}}^{-1}(\hat{\theta})+o_p(1)\right]\right\} \\
= \;& -\mathbf{tr}\left[\hat{\mathbf{J}}(\hat{\theta})\hat{\mathbf{I}}^{-1}(\hat{\theta})\right] + \mathbf{tr}\left[\hat{\mathbf{J}}(\hat{\theta})o_p(1)\right] + \mathbf{tr}\left[\hat{\mathbf{I}}^{-1}(\hat{\theta})o_p(1)\right] + o_p(1) \\
= \;& \mathbf{tr}\left[-\hat{\mathbf{J}}(\hat{\theta})\hat{\mathbf{I}}^{-1}(\hat{\theta})\right] + o_p(1) = IOS_a + o_p(1).
\end{aligned}$$

Hence, our proposed test statistic is asymptotically equivalent to the IOS test of Presnell and Boos (2004). However, an important advantage of our test statistic over the IOS test is that there is no need to invert $\hat{\mathbf{I}}(\hat{\theta})$*. Inversion of* $\hat{\mathbf{I}}(\hat{\theta})$ *may be difficult when the dimension of* $\hat{\theta}$ *is high.*

Remark 4.3.4 *Our test is not only applicable to the iid case but also to the dependent case. This is another important advantage of our test statistic over the IOS test.*

Remark 4.3.5 *When the model is correctly specified,* $-\mathbf{I}(\theta_0) = \mathbf{J}(\theta_0)$ *and hence* $\mathbf{BT} \approx p$*. However, when the model is misspecified,* $-\mathbf{I}(\theta_0) \neq \mathbf{J}(\theta_0)$*,* **BT** *will be away from p. In practice, to implement our test, we need to compare the difference between* **BT** *and p with some threshold values to determine whether the model is misspecified or not. In the iid case, Presnell and Boos (2004) derived the asymptotic distribution for* IOS_a*. Similarly. we can also try to derive the asymptotic distribution for* **BT**. *However, as argued in Presnell and Boos (2004), the finite sample*

behavior of the asymptotic approximation is typically not good. Consequently, they suggested to use the parameter bootstrap to get the critical values for the IOS test. Although the parameter bootstrap is feasible for the IOS test, running MCMC posterior simulation is computationally demanding for our test in bootstrap. Hence, we suggest a simple approach to getting the threshold values based on the Monte Carlo simulation. Note that

$$\mathbf{BT} = n\mathbf{tr}\left\{\hat{\mathbf{J}}(\bar{\theta})E\left[(\theta-\bar{\theta})(\theta-\bar{\theta})'|\mathbf{Y}\right]\right\} = \mathbf{tr}\left\{-\hat{\mathbf{J}}(\bar{\theta})\mathbf{I}^{-1}(\theta_0)\right\} + o_p(1).$$

If $\mathbf{I}^{-1}(\theta_0)$ *is fixed and* $\hat{\mathbf{J}}(\bar{\theta})$ *is not difficult to compute, we can conveniently get the threshold values based on the simulated random observations. The detailed steps may be summarized as follows:*

Step 1*: Set* $\theta_0 = \bar{\theta}$*. Based on the model considered, we generate* $k \times n$ *random observations and run one MCMC simulation under a noninformative prior to get the posterior covariance matrix* $V(\theta)$ *which is used to approximate* $-\mathbf{I}^{-1}(\theta_0)$*.*

Step 2*: Set* $-\mathbf{I}^{-1}(\theta_0) = nkV(\theta)$*. We generate n random observations from the model considered and compute* $-\mathbf{tr}\left\{\hat{\mathbf{J}}(\bar{\theta})\mathbf{I}^{-1}(\theta_0)\right\}$*. Let it denote* $\mathbf{BT}_1$*.*

Step 3*: Repeat Step 2 for* n_1 *times and get a sequence of* $\mathbf{BT}_1, \cdots, \mathbf{BT}_{n_1}$*. The threshold values are obtained from this sequence under desired probability levels.*

4.3.3 Bayesian test for latent variable models

The requirement of an analytical expression for the observed data likelihood function is too strong for many latent variable models. In this section, we show how to calculate the proposed test and the threshold values for latent variable models with the aid of the EM algorithm.

The EM algorithm is a powerful tool to deal with latent variable models. Instead of maximizing the observed data likelihood function, the EM algorithm maximizes

the so-called $\mathscr{Q}$ function given by

$$\mathscr{Q}(\theta|\theta^{(r)}) = E_{\theta^{(r)}}\{\mathscr{L}_c(\mathbf{Y},z|\theta)|\mathbf{y},\theta^{(r)}\}, \tag{4.3.4}$$

where $\mathscr{L}_c(\mathbf{Y},z|\theta) := p(\mathbf{Y},z|\theta)$ is the complete-data likelihood function. The $\mathscr{Q}$-function is the conditional expectation of $\mathscr{L}_c(\mathbf{Y},z|\theta)$ with respect to the conditional distribution $p(z|\mathbf{Y},\theta^{(r)})$ where $\theta^{(r)}$ is a current fit of the parameter. The EM algorithm consists of two steps: the *expectation* (E) step and the *maximization* (M) step. The E-step evaluates $\mathscr{Q}(\theta|\theta^{(r)})$. The M-step determines a $\theta^{(r)}$ that maximizes $\mathscr{Q}(\theta|\theta^{(r)})$. Under some mild regularity conditions, for large enough r, $\{\theta^{(r)}\}$ obtained from the EM algorithm is the MLE $\widehat{\theta}$. For more details about the EM algorithm, see Dempster et al. (1977).

Although the EM algorithm is a good approach to dealing with latent variable models, the numerical optimization in the M-step is often unstable. Not surprisingly, in recent years, the EM algorithm has been less popular to analyze latent variables models compared with the MCMC techniques. However, we will show that, without using the numerical optimization in the M-step, the theoretical properties of the EM algorithm can facilitate the computation of the proposed test for latent variable models.

4.3.4 Computing BT by the EM algorithm

The proposed test statistic **BT** involves $\hat{\mathbf{J}}(\bar{\theta})$ which is based on the first derivative of the observed data log-likelihood function, i.e, $\mathbf{s}(\mathbf{Y},\theta)$. After $\hat{\mathbf{J}}(\bar{\theta})$ is obtained, **BT** can be computed from the MCMC output following the ergodic theorem. In latent variable models, since $p(\mathbf{Y}|\theta)$ and hence $\mathbf{s}(\mathbf{Y},\theta)$ are not analytically available, we propose to use the EM algorithm to compute $\mathbf{s}(\mathbf{Y},\theta)$, as shown in the following lemma.

Lemma 4.3.1 *For any θ and θ^* in Θ, it was shown in Dempster et al. (1977) that*

$$\begin{aligned}\mathbf{s}(\mathbf{Y},\theta) &= \frac{\partial \mathscr{L}_o(\mathbf{Y},\theta)}{\partial \theta} = \frac{\partial \mathscr{Q}(\theta|\theta^*)}{\partial \theta}|_{\theta=\theta^*} = E_{(z|\mathbf{Y},\theta)}\left\{\frac{\partial \mathscr{L}_c(\mathbf{Y},z,\theta)}{\partial \theta}\right\} \\ &= \int \frac{\partial \mathscr{L}_c(\mathbf{Y},\omega,\theta)}{\partial \theta} p(z|\mathbf{Y},\theta)dz,\end{aligned}$$

where $\mathscr{L}_o(\mathbf{Y},\theta) := p(\mathbf{Y}|\theta)$ the observed data likelihood.

Remark 4.3.6 *If the analytical form of the $\mathscr{Q}$-function is available, we can replace the first derivative of the likelihood function $\log p(\mathbf{Y}|\theta)$ with the first derivative of the $\mathscr{Q}$-function.*

Remark 4.3.7 *In Gaussian (perhaps nonlinear) latent variable models, the latent variable z is assumed to follow a multivariate normal distribution and the observed variable $\mathbf{Y}$ are independent conditional on z.*[1] *Rue et al. (2004), Rue et al. (2009) showed that the posterior distribution $p(z|\mathbf{Y},\theta)$ can be well approximated by a Gaussian distribution that matches the mode and the curvature at the mode, i.e.,*

$$p(z|\mathbf{Y},\theta) \propto \exp\left(-\frac{1}{2}z^\top V(\theta)z + \sum_{i=1}^{n}\log p(y_i|\mathbf{z}_i,\theta)\right).$$

Hence, the Laplace approximation for this posterior distribution is

$$p(z|\mathbf{Y},\theta) \propto \exp\left(-\frac{1}{2}z^\top (V(\theta) + diag(\mathbf{c}))z\right),$$

where μ is the mode of $p(z|\mathbf{Y},\theta)$ and $\mathbf{c}$ comes from the second order term in the Taylor expansion of $\sum_{t=1}^{T}\log p(\mathbf{y}_i|\mathbf{z}_i)$ at the mode. When the analytical form of the $\mathscr{Q}$-function is not available, $\hat{\mathbf{J}}(\bar{\theta})$ may be computed based on the Laplace approximation.

Remark 4.3.8 *A more general approach to evaluating the $\mathscr{Q}$-function is to use the*

[1] Many popular latent variable models in economics and finance belong to this class of models. Examples include dynamic linear models, dynamic factor models, some stochastic volatility models.

following formula based on the MCMC output:

$$\mathbf{s}(\mathbf{Y},\theta) \approx \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{\partial \log p(\mathbf{Y},z^{(m)}|\theta)}{\partial \theta}\right\},$$

where $\{z^{(m)}, m = 1,2,\cdots,M\}$ *is a random sample simulated from the posterior distribution* $p(z|\mathbf{y},\theta)$. *However, this approach is computationally more demanding.*

4.4 Empirical Examples

We now illustrate the test using two real examples. The first example is the well-known Fama-French model widely used in empirical finance. In this example, the $\mathscr{Q}$-function has an analytical expression and, hence, the test statistic is relative straightforward to compute. The second example is the stochastic general equilibrium (DSGE) model that has been widely used in empirical macroeconomics. In this example, the $\mathscr{Q}$-function does not have an analytical expression and, hence, computing the test statistic is more involved.

4.4.1 Specification test in asset pricing models

In the first example, we assess the validity of an asset pricing model. It is known in the asset pricing literature that when an asset price model is misspecified, the specification error may lead to a serious loss in investment see Kan and Zhou (2006) and Gospodinov et al. (2012a,b). The particular model we test is the Fama-French asset pricing model with the error term following a multivariate t distribution. The Fama-French model has enjoyed a great deal of attention in the empirical asset pricing literature. Given that heavy tails are commonly found in return distributions, following Zhou (1993) and Kan and Zhou (2006), we generalize the classical Fama-French asset pricing model to

$$R_t = \alpha + \beta' F_t + \varepsilon_t, \varepsilon_t \sim N(0,\Sigma/\omega_t), \omega_t \sim \Gamma\left(\frac{\nu}{2},\frac{\nu}{2}\right), \tag{4.4.1}$$

where R_t is the excess return of portfolio at period t with $N \times 1$ dimension, F_t a $K \times 1$ vector of factor portfolio excess returns, α a $N \times 1$ vector of intercepts, β a $N \times K$ vector of scaled covariances, ε_t the random error, $t = 1, 2, \cdots, n$, Σ a diagonal matrix. For the purpose of illustration, we restrict ν to be a known constant.

The dataset we use is from the data library of Kenneth French.[2] We consider the intersections of 5 portfolios formed on size (market equity, ME) and 5 portfolios formed on the ratio of book equity to market equity (BE/ME), constructed at the end of each June ranged from July 1926 to July 2011. The data comprise monthly returns of 25 portfolios, i.e., $N = 25$ and the sample size $n = 1021$. Following Fama and French (1993), we use Fama/French's three factors, market excess return, SMB (Small Minus Big), HML (High Minus Low) as the explanatory risk factors.

In the empirical study, on the basis of the analysis of Li and Yu (2012), we simply set $\nu = 3$. To represent the prior ignorance, we use the following vague conjugate priors,

$$\alpha_i \sim N[0, 100], \beta_{ij} \sim N[0, 100], \Sigma_{ii}^{-1} \sim \Gamma[0.001, 0.001].$$

Using the Gibbs sampler, 70,000 random observations are draw from the posterior distributions with the first 20,000 discarded. The convergence is checked using the Raftery-Lewis diagnostic test statistic (Raftery and Lewis (1992)). Appendix 2 derives all the derivatives of the $\mathscr{Q}$-function that are needed for computing **BT**. Based on the 50,000 random observations, we get **BT** $= 135.1290$.

To detect whether this model is misspecified or not, we need to obtain some threshold values. Since the sample size is large, we choose $k = 1$, $n_1 = 1000$. Using the Monte Carlo method, we can obtain the threshold interval $[113.7810, 126.0118]$ at the 90% probability level, $[112.8312, 126.9880]$ at the 95% probability level, and $[110.1464, 129.0262]$ at the 99% probability level. Obviously this model is misspecified at all three probability levels.

Further, in order to check the test effect of the statistic **BT** and the correspond-

[2]The data is download from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Table 4.1: Misspecification testing results for Fama-French three factor models

ν	10	25	50	100	$+\infty$
BT	127.6588	128.6581	129.0959	129.2091	129.5321
90%	95	100	100	100	100
95%	81	98	100	100	100
99%	8	32	59	66	78

ing threshold values, we run a simulation study with 100 replications. The same Fama-French three factors with empirical data are used for explanatory risk factors. This estimated values of the parameter is adopted as true value. Then, the data is generated from the model whose specification is the same with the model considered in this paper, except different freedom ν. The numbers of rejecting the heavy tail asset pricing model with $\nu = 3$ under different probability level are shown in Table 1. The first line is the averaging value of **BT** under 100 replications. The second, third and fourth lines are used to show the number of correct decisions under different probability level. From this table, we can find that when the model is misspecified more seriously, i.e., the true ν becomes away from three[3], **BT** becomes larger and away from $p = 125$ to provide stronger evidence against the model used. In addition, the number of correct decisions are also increased. Hence, the statistic **BT** and the threshold values under different probability levels can be used to serve for testing the misspecification well.

4.4.2 Specification test in DSGE models

DSGE models are microfounded and optimization-based. They have become very popular in macroeconomics over the last 30 years. Estimation and evaluation of the DSGE models require one to solve them and then to construct a linear or nonlinear state-space approximation. Bayesian time series methods have been widely applied to estimate the DSGE models. For a linear Gaussian approximation, the Kalman filter can be used to compute the likelihood function numerically; see Schorfheide

[3] $\nu = +\infty$ means that the t distribution is reduced as normal distribution

(2001), Lubik and Schorfheide (2006), An and Schorfheide (2007). For a non-linear non-Gaussian approximation, Fernández-Villaverde and Rubio-Ramírez (2005) and Rubio-Ramirez (2005) used the particle filter to calculate the likelihood numerically.

In this example, following An and Schorfheide (2007), we adopt a linear Gaussian approximation. We estimate and test three DSGE models after the log-linearization

$$\widehat{y}_t = E_t\left[\widehat{y}_{t+1}\right] + \widehat{g}_t - E_t\left[\widehat{g}_{t+1}\right] - \frac{1}{\tau}\left(\widehat{R}_t - E_t\left[\widehat{\pi}_{t+1}\right] - E_t\left[\widehat{z}_{t+1}\right]\right),$$

$$\widehat{\pi}_t = \beta E_t\left[\widehat{\pi}_{t+1}\right] + \kappa\left(\widehat{y}_t - \widehat{g}_t\right),$$

$$\widehat{y}_t = \widehat{c}_t + \widehat{g}_t,$$

where $\widehat{y}_t$, $\widehat{\pi}_t$, $\widehat{R}_t$, $\widehat{c}_t$, $\widehat{g}_t$, $\widehat{z}_t$ are the percentage deviations from the steady state for the output, the inflation, the interest rate, the consumption, the government expenditure, and the technology growth rate, respectively.

Monetary policy is described by an interest rate feedback rule of the form

$$\widehat{R}_t = \rho_R \widehat{R}_{t-1} + \left(1-\rho_R\right)\psi_1 \widehat{\pi}_t + \left(1-\rho_R\right)\psi_2\left(\widehat{y}_t - \widehat{g}_t\right) + \varepsilon_{R,t}.$$

The specification allows that the central bank reacts to the inflation and the deviations of the output from the potential output. The exogenous process is as follows

$$\widehat{g}_t = \rho_g \widehat{g}_{t-1} + \varepsilon_{g,t},$$

$$\widehat{z}_t = \rho_z \widehat{z}_{t-1} + \varepsilon_{z,t},$$

where the monetary policy shock $\varepsilon_{R,t}$, the government spending shock $\varepsilon_{g,t}$, the technology shock $\varepsilon_{z,t}$ are assumed to be serially uncorrelated. The three shocks are independent of each other at all leads and lags and are normally distributed with mean zeros and standard deviations σ_z, σ_g and σ_g. Since $\widehat{y}_t$, $\widehat{\pi}_t$, $\widehat{R}_t$, $\widehat{c}_t$, $\widehat{g}_t$, $\widehat{z}_t$ are not observed, the above six equations are all state equations.

We define a set of measurement equations to relate the state variables to a set of observed variables:

$$\begin{aligned} YGR_t &= \gamma^{(Q)} + 100\left(\widehat{y}_t - \widehat{y}_{t-1} + \widehat{z}_t\right), \\ INFL_t &= \pi^{(A)} + 400\widehat{\pi}_t, \\ INT_t &= \pi^{(A)} + r^{(A)} + 4\gamma^{(Q)} + 400\widehat{R}_t, \end{aligned}$$

where YGR_t is the quarter-to-quarter per capita GDP growth rates, $INFL_t$ and INT_t are the annualized quarter-to-quarter inflation rates and the annualized quarter-to-quarter nominal interest rates, respectively. The parameters $\gamma^{(Q)}$, $\pi^{(A)}$ and $r^{(A)}$ are

$$\gamma = 1 + \frac{\gamma^{(Q)}}{100}, \beta = \frac{1}{1 + r^{(A)}/400}, \pi = 1 + \frac{\pi^{(A)}}{400},$$

where γ/β and π are the steady states of $\widehat{R}_t$ and $\widehat{\pi}_t$, respectively.

The six state equations and the three measurement equations constitute a New Keynesian DSGE model, which we denote Model 1. We set Model 2 the same as Model 1, except that we restrict $\kappa = 5$. Model 3 is a restricted version of Model 1 where we set $\psi_2 = 0$. In Model 3, the central bank does not respond to the output gap.

The data are from Lubik and Schorfheide (2006). They include quarterly U.S. series on the GDP growth rates, the inflation rates, and the nominal interest rates. The other two series are annualized. The sample range is from the first quarter of 1983 to the last quarter of 2002. The priors are the same as in An and Schorfheide (2007).

Following Schorfheide (2001), we estimate Model 1 by using the Random Walk Metropolis-Hasting algorithm in the MATLAB-based DYNARE package (Adjemian et al. (2011)). 1,000,000 draws from the posterior are generated with the first 100,000 draws being discarded. All the quantities needed to compute **BT** are derived in Appendix 3.

To compute the threshold values we set $K = 6$, $n_1 = 480$. Using the Monte

Carlo method, we obtain the threshold interval $[9.4777, 18.3866]$ at the 90% probability level, $[8.888, 19.9232]$ at the 95% probability level, and $[7.0366, 22.7872]$ at the 99% probability level. Based on these threshold values, we cannot reject the hypothesis that Model 1 is correctly specified.

Table 4.2: Misspecification testing results for DSGE models

	Model 2	Model 3
BT	7.8809	8.5018
90%	99	84
95%	92	69
99%	14	5

As in the previous example, we run a simulation with 100 replications to check the test effect and threshold values. The same empirical data is used for Model 2 and Model 3 respectively. The estimated values of the parameter are adopted as true value. And the data is generated from Model 2 and Model 3. The numbers of rejecting the Model 1 under different probability level are shown in Table 2. The first line is the averaging value of **BT** under 100 replications. The second to four lines show the numbers of correct decisions under different probability level. From the results, we can see that Model 2 is misspecified more seriously than Model 3. As pointed out by An and Schorfheide (2007), the DSGE model implies that when actual output is closed to the target flexible price output, inflation will also be close to its target value, hence, deviations of output from target coincide with deviations of inflation from its target value, which makes it difficult to identify the policy rule coefficients and imposing an incorrect value for ψ_2 is not particularly costly in terms of fit.

4.5 Conclusion

In this paper, we have proposed a new Bayesian test statistic to assess the adequacy of specification of a model after the model is estimated by Bayesian MCMC methods. The main advantages of the new statistic can be summarized as follows: (1)

The proposed Bayesian test approach is quite general and can be applied into a variety of models, such as, time series, panel data models, latent variable models (2) The test statistic can be easy to compute and its threshold values are also easily obtained using Monte Carlo simulation technique. At last, we illustrate the newly developed approach using asset pricing models and dynamic stochastic general equilibrium models.

Chapter 5 Summary of Conclusions

In Chapter 2, we introduce a robust deviance information criteria (RDIC) for comparing models with latent variables. Although latent variable models can be conveniently estimated in the Bayesian framework via MCMC if the data augmentation technique is used, we argue that data augmentation cannot be used in connection to DIC. This is because that the justification of DIC rests on the validity of the standard Bayesian asymptotic theory. With data augmentation, the number of parameters increases with the number of observations, making the likelihood nonregular. As a consequence, the standard Bayesian asymptotic theory does not hold. In addition, the use of the data augmentation makes DIC is very sensitive to transformations and distributional representations.

While in principle one can use the standard DIC (i.e. DIC_1) without resorting to the data augmentation technique, in practice this standard DIC is very difficult to use because the observed-data likelihood is not available in closed-form for many latent variable models and because the standard DIC_1 has to numerically evaluate the observed-data likelihood at each MCMC iteration. These two observations make the implementation of DIC_1 practically non-operational.

The problem is overcome by RDIC. RDIC is defined without augmenting the parameter space and hence can be justified by the standard Bayesian asymptotic theory. We then show that how the EM algorithm can facilitate the computation of RDIC in different contexts. Since the latent variables are not counted as parameters in our approach, RDIC is robust to nonlinear transformations of the latent variables and distributional representations of the model specification. Asymptotic justification, computational tractability and robustness to transformation and specification

are the three main advantages of the proposed approach. These advantages are illustrated using several popular models in economics and finance.

In Chapter 3, we propose a new Bayesian statistic to test a point null hypothesis. The main advantages of the new statistic are fourfold. First, it is immune to Bartlett's paradox. Second, it avoids Jeffreys-Lindley's paradox. Third, it can be easily computed using the MCMC outputs from the posterior distribution. Fourth, the asymptotic distribution can be derived for calibrating the threshold values. The proposed method is illustrated using a simple linear regression model, an asset pricing model and a stochastic volatility model.

In Chapter 4, we propose a new Bayesian test statistic to assess the adequacy of specification of a model after the model is estimated by Bayesian MCMC methods. The main advantages of the new statistic can be summarized as follows: (1) The proposed Bayesian test approach is quite general and can be applied into a variety of models, such as, time series, panel data models, latent variable models (2) The test statistic can be easy to compute and its threshold values are also easily obtained using Monte Carlo simulation technique. At last, we illustrate the newly developed approach using asset pricing models and dynamic stochastic general equilibrium models.

Bibliography

ADJEMIAN, S., H. BASTANI, M. JUILLARD, F. MIHOUBI, G. PERENDIA, M. RATTO, AND S. VILLEMOT (2011): "Dynare: Reference manual, version 4," *Dynare Working Papers*, 1.

AIT-SAHALIA, Y. (1996): "Testing continuous-time models of the spot interest rate," *Review of Financial studies*, 9, 385–426.

AKAIKE, H. (1973): "Information theory and an extension of the maximum likelihood principle," in *Second international symposium on information theory*, Springer Verlag, vol. 1, 267–281.

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

ANDO, T. (2007): "Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models," *Biometrika*, 94, 443–458.

ANDO, T. AND R. TSAY (2010): "Predictive likelihood for Bayesian model selection and averaging," *International Journal of Forecasting*, 26, 744–763.

ANDREWS, D. (1997): "A conditional Kolmogorov test," *Econometrica*, 65, 1097–1128.

ANDREWS, D. AND C. MALLOWS (1974): "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society Series B*, 36, 99–102.

BERG, A., R. MEYER, AND J. YU (2004): "Deviance information criterion

for comparing stochastic volatility models," *Journal of Business and Economic Statistics*, 22, 107–120.

BERGER, J. O. (1985): "Statistical decision theory and Bayesian analysis," *Springer Series in Statistics, New York: Springer, 1985, 2nd ed.*, 1.

BERNANKE, B., J. BOIVIN, AND P. ELIASZ (2005): "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach," *The Quarterly Journal of Economics*, 120, 387–422.

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

BLACK, F. (1976): "Studies of stock market volatility changes," *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, 177–181.

BROOKS, S. (2002): "Discussion on the paper by Spiegelhalter, Best, Carlin, and van de Linde (2002)," *Journal of the Royal Statistical Society Series B*, 64, 616–618.

BURNHAM, K. AND D. ANDERSON (2002): *Model selection and multi-model inference: a practical information-theoretic approach*, Springer.

CELEUX, G., F. FORBES, C. ROBERT, AND D. TITTERINGTON (2006): "Deviance Information Criteria for Missing Data Models," *Bayesian Analysis*, 1, 651–674.

CHAN, J. C. AND I. JELIAZKOV (2009): "Efficient simulation and integrated likelihood estimation in state space models," *International Journal of Mathematical Modelling and Numerical Optimisation*, 1, 101–120.

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

CHESHER, A. AND R. SPADY (1991): "Asymptotic expansions of the information matrix test statistic," *Econometrica*, 59, 787–815.

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.

CHRISTIE, A. A. (1982): "The stochastic behavior of common stock variances: Value, leverage and interest rate effects," *Journal of Financial Economics*, 10, 407–432.

CLARK, P. (1973): "A subordinated stochastic process model with finite variance for speculative prices," *Econometrica*, 41, 135–155.

CORRADI, V. AND N. R. SWANSON (2004): "Some recent developments in predictive accuracy testing with nested models and (generic) nonlinear alternatives," *International Journal of Forecasting*, 20, 185–199.

COX, D. R. (1961): "TESTS OF SEPARATE FAMILIES OF HYPOTHESES," in *Proceedings: Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Univ of California Press, 105.

——— (1962): "Further results on tests of separate families of hypotheses," *Journal of the Royal Statistical Society. Series B (Methodological)*, 406–424.

DANIELS, M. AND J. HOGAN (2008): *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, vol. 109, Chapman & Hall.

DAVIDSON, R. AND J. MACKINNON (1992): "A new form of the information matrix test," *Econometrica*, 60, 145–157.

DEIORIO, M. AND C. ROBERT (2002): "Discussion on the paper by Spiegelhalter, Best, Carlin, and van de Linde (2002)," *Journal of the Royal Statistical Society Series B*, 64, 629–630.

DEMPSTER, A., N. LAIRD, AND D. RUBIN (1977): "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B*, 39, 1–38.

DHAENE, G. AND D. HOORELBEKE (2004): "The information matrix test with bootstrap-based covariance matrix estimation," *Economics Letters*, 82, 341–347.

DUAN, J.-C. (2004): "A Specification Test for Time Series Models by a Normality," in *Econometric Society 2004 North American Winter Meetings*, Econometric Society, 467.

EUBANK, R. L. AND C. H. SPIEGELMAN (1990): "Testing the goodness of fit of a linear model via nonparametric regression techniques," *Journal of the American Statistical Association*, 85, 387–392.

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

FAN, Y. AND Q. LI (1996): "Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms," *Econometrica*, 64, 865–90.

FERNÁNDEZ-VILLAVERDE, J. AND J. RUBIO-RAMÍREZ (2005): "Estimating dynamic equilibrium economies: linear versus nonlinear likelihood," *Journal of Applied Econometrics*, 20, 891–910.

GELFAND, A. AND M. TREVISANI (2002): "Discussion on the paper by Spiegelhalter, Best, Carlin, and van de Linde (2002)," *Journal of the Royal Statistical Society Series B*, 64, 629–630.

GELMAN, A. (2003): *Bayesian Data Analysis*, CRC Press.

GELMAN, A. AND X.-L. MENG (1998): "Simulating normalizing constants: from importance sampling to bridge sampling to path sampling," *Statistical Science*, 13, 163–185.

GEWEKE, J. (1977): "The dynamic factor analysis of economic time-Series models," in *Latent variables in socio-economic models*, ed. by A. Aigner and A. Goldberger, North-Holland, 365–395.

——— (2005): *Contemporary Bayesian Econometrics and Statistics*, vol. 537, Wiley-Interscience.

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

GEWEKE, J., G. KOOP, AND H. VAN DYK (2011): *Oxford Handbook of Bayesian Econometrics*, Oxford Univ Press.

GHOSH, J. AND R. RAMAMOORTHI (2003): *Bayesian nonparametrics*, Springer Verlag.

GIANNONE, D., L. REICHLIN, AND L. SALA (2004): "Monetary policy in real time," *NBER Macroeconomics Annual*, 161–200.

GOOD, I. (1956): "The surprise index for the multivariate normal distribution," *The Annals of Mathematical Statistics*, 27, 1130–1135.

GOSPODINOV, N., R. KAN, AND C. ROBOTTI (2012a): "Chi-squared tests for evaluation and comparison of asset pricing models," *Forthcoming in Journal of Econometrics*.

——— (2012b): "Robust Inference in Linear Asset Pricing Models," *Working Paper*.

HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

HAUSMAN, J. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

HAUSMAN, J. A. AND W. E. TAYLOR (1980): "Comparing specification tests and classical tests," .

HERBST, E. (2010): "Gradient and Hessian-based MCMC for DSGE Models," .

HONG, Y. AND H. LI (2005): "Nonparametric specification testing for continuous-time models with applications to term structure of interest rates," *Review of Financial Studies*, 18, 37–84.

HOROWITZ, J. (1994): "Bootstrap-based critical values for the information matrix test," *Journal of Econometrics*, 61, 395–411.

HUANG, S. AND J. YU (2010): "Bayesian analysis of structural credit risk models with microstructure noises," *Journal of Economic Dynamics and Control*, 34, 2259–2272.

IBRAHIM, J., H. ZHU, AND N. TANG (2008): "Model selection criteria for missing-data problems using the EM algorithm," *Journal of the American Statistical Association*, 103, 1648–1658.

ISKREV, N. (2008): "Evaluating the information matrix in linearized DSGE models," *Economics Letters*, 99, 607–610.

JACQUIER, E., N. G. POLSON, AND P. E. ROSSI (2004): "Bayesian analysis of stochastic volatility models with fat-tails and correlated errors," *Journal of Econometrics*, 122, 185–212.

JEFFREYS, H. (1961): "Theory of probability," *Clarendon Press Oxford*.

KAN, R. AND G. ZHOU (2003): "Modeling non-normality using multivariate t: Implications for asset pricing," .

——— (2006): "Modelling Non-normality using Multivariate t: Implications to Asset Pricing," *Working Paper*.

KASS, R. AND A. RAFTERY (1995): "Bayes factors," *Journal of the American Statistical Association*, 90, 773–795.

KIM, J. (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*, 66, 359–380.

KOSE, A. M., C. OTROK, AND C. WHITEMAN (2003): "International business cycles: World, region, and country-specific factors," *American Economic Review*, 93, 1216–1239.

——— (2008): "Understanding the evolution of world business cycles," *Journal of International Economics*, 75, 110–130.

LANCASTER, T. (1984): "The Covariance Matrix of the Information Matrix Test," *Econometrica*, 52, 1051–53.

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

LI, Y., T. ZENG, AND J. YU (2012): "Robust Deviation Information Criterion for Latent Variable Models," *Working Paper*.

LOUIS, T. A. (1982): "Finding the Observed Information Matrix when Using the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 44, 226–233.

LUBIK, T. AND F. SCHORFHEIDE (2006): "A Bayesian look at the new open economy macroeconomics," in *NBER Macroeconomics Annual 2005, Volume 20*, MIT Press, 313–382.

MAGNUS, J. AND H. NEUDECKER (1999): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, Chichester.

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

MCLACHLAN, G. AND T. KRISHNAN (2008): *The EM Algorithm and Extensions*, vol. 382, John Wiley and Sons.

MEYER, R. AND J. YU (2000): "BUGS for a Bayesian analysis of stochastic volatility models," *The Econometrics Journal*, 3, 198–215.

NEWEY, W. (1985): "Generalized method of moments specification testing," *Journal of econometrics*, 29, 229–256.

OAKES, D. (1999): "Direct calculation of the information matrix via the EM," *Journal of the Royal Statistical Society Series B*, 61, 479–482.

O'HAGAN, A. (1995): "Fractional Bayes factors for model comparison," *Journal of the Royal Statistical Society. Series B. Methodological*, 57, 99–138.

ORME, C. (1990): "The small-sample performance of the information-matrix test," *Journal of Econometrics*, 46, 309–331.

OTROK, C. AND C. WHITEMAN (1998): "Bayesian leading indicators: measuring and predicting economic conditions in Iowa," *International Economic Review*, 39, 997–1014.

PHILLIPS, P. (1996): "Econometric model determination," *Econometrica*, 64, 763–812.

PHILLIPS, P. AND W. PLOBERGER (1996): "An asymptotic theory of bayesian inference for time series," *Econometrica*, 64, 381–412.

POIRIER, D. J. (1995): "Intermediate Statistics and Econometrics: A Comparative Approach," *The MIT Press*.

PRESNELL, B. AND D. BOOS (2004): "The IOS test for model misspecification," *Journal of the American Statistical Association*, 99, 216–227.

RAFTERY, A. AND S. LEWIS (1992): "How many iterations in the Gibbs sampler," *Bayesian statistics*, 4, 763–773.

ROBERT, C. (2001): *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Springer.

ROBERT, C. P. (1993): "A note on Jeffreys-Lindley paradox," *Statistica Sinica*, 3, 601–608.

RUE, H., S. MARTINO, AND N. CHOPIN (2009): "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society Series B*, 71, 319–392.

RUE, H., I. STEINSLAND, AND S. ERLAND (2004): "Approximating hidden Gaussian Markov random fields," *Journal of the Royal Statistical Society Series B*, 66, 877–892.

SARGAN, J. (1958): "The Estimation of Economic Relationships using Instrumental Variables," *Econometrica*, 26, 393–415.

SARGENT, T. AND C. SIMS (1977): "Business cycle modeling without pretending to have too much a priori economic theory," *New Methods in Business Research, Federal Reserve Bank of Minneapolis, Minneapolis*.

SCHORFHEIDE, F. (2001): "Loss function-based evaluation of DSGE models," *Journal of Applied Econometrics*, 15, 645–670.

SHUMWAY, R. AND D. STOFFER (2006): "Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics)," .

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

SPIEGELHALTER, D., A. THOMAS, N. BEST, AND D. LUNN (2003): "WinBUGS version 1.4 user manual," .

STOCK, J. AND M. WATSON (1999): "Forecasting inflation," *Journal of Monetary Economics*, 44, 293–335.

——— (2002): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20, 147–162.

——— (2010): "Dynamic factor models," *Prepared for the Oxford Handbook of Economic Forecasting*.

STURTZ, S., U. LIGGES, AND A. GELMAN (2005): "R2WinBUGS: A Package for Running WinBUGS from R," *Journal of Statistical Software*, 12, 1–16.

TAKEUCHI, K. (1976): "Distribution of Informational Statistics and a Criterion of Model Fitting," *Suri-Kagaku (Mathematic Sciences)*, 153, 12–18.(in Japanese).

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

TAUCHEN, G. (1985): "Diagnostic testing and evaluation of maximum likelihood models," *Journal of Econometrics*, 30, 415–443.

TU, J. AND G. ZHOU (2010): "Incorporating economic objectives into Bayesian priors: Portfolio choice under parameter uncertainty," *Journal of Financial and Quantitative Analysis*, 45, 959–986.

VAIDA, F. AND S. BLANCHARD (2005): "Conditional Akaike information for mixed-effects models," *Biometrika*, 92, 351–370.

WANG, J., J. CHAN, AND S. CHOY (2011): "Stochastic volatility models with leverage and heavy-tailed distributions: A Bayesian approach using scale mixtures," *Computational Statistics and Data Analysis*, 55, 852–862.

WHITE, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

——— (1987): "Specification testing in dynamic models," in *Advances in econometrics, Fifth world congress*, vol. 1, 1–58.

WOOLDRIDGE, J. M. (2009): *Introductory econometrics: A modern approach*, South-Western Pub.

WU, C. (1983): "On the convergence properties of the EM algorithm," *The Annals of Statistics*, 11, 95–103.

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

ZHENG, J. X. (2000): "A consistent test of conditional parametric distributions," *Econometric Theory*, 16, 667–691.

ZHOU, G. (1993): "Asset-pricing tests under alternative distributions," *Journal of Finance*, 48, 1927–1942.

Appendix

.1 Proofs in Chapter 2

.1.1 Proof of Lemma 2.3.1

Using the Taylor-expansion on the log-posterior probability density function, we can show that

$$\begin{aligned}\ln p(\theta|\mathbf{y}) &= \ln p(\hat{\theta}_m|\mathbf{y}) + L_n^{(1)}(\hat{\theta}_m)'(\theta-\hat{\theta}_m) + \frac{1}{2}(\theta-\hat{\theta}_m)'L_n^{(2)}(\tilde{\theta})(\theta-\hat{\theta}_m) \\ &= \ln p(\hat{\theta}_m|\mathbf{y}) + \frac{1}{2}(\theta-\hat{\theta}_m)'L_n^{(2)}(\tilde{\theta})(\theta-\hat{\theta}_m),\end{aligned}$$

where $\tilde{\theta}$ lies on the segment between θ and $\hat{\theta}_m$. It follows that

$$p(\theta|\mathbf{y}) = p(\hat{\theta}_m|\mathbf{y})\exp\left[\frac{1}{2}(\theta-\hat{\theta}_m)'L_n^{(2)}(\tilde{\theta})(\theta-\hat{\theta}_m)\right].$$

Let $\omega = \sqrt{n}(\theta-\hat{\theta}_m)$, $J(\theta) = -\frac{1}{n}L_n^{(2)}(\theta)$, $c_n^* = \int \exp[-\frac{1}{2}\omega' J(\tilde{\theta})\omega]\mathrm{d}\omega$, $c_n = \int \exp[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega]\mathrm{d}\omega$. It can be shown that

$$p(\omega|\mathbf{y}) \propto \exp\left[\frac{1}{2}(\theta-\hat{\theta}_m)'L_n^{(2)}(\tilde{\theta})(\theta-\hat{\theta}_m)\right] = \exp\left\{-\frac{1}{2}\omega' J(\tilde{\theta})\omega\right\}.$$

Then, we have

$$
\begin{aligned}
& P_n = \int \left| p(\omega|\mathbf{y}) - \frac{1}{c_n} \exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right] \right| \mathrm{d}\omega \\
= & \int \left| \frac{1}{c_n^*} \exp\left[-\frac{1}{2}\omega' J(\tilde{\theta})\omega\right] - \frac{1}{c_n} \exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right] \right| \mathrm{d}\omega \\
= & \frac{1}{c_n} \int \left| \frac{c_n}{c_n^*} \exp\left[-\frac{1}{2}\omega' J(\tilde{\theta})\omega\right] - \exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right] \right| \mathrm{d}\omega \\
= & \frac{1}{c_n} \int \left| \frac{c_n - c_n^*}{c_n^*} \exp\left[-\frac{1}{2}\omega' J(\tilde{\theta})\omega\right] + \exp\left[-\frac{1}{2}\omega' J(\tilde{\theta})\omega\right] - \exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right] \right| \mathrm{d}\omega \\
\leq & \frac{1}{c_n} \left\{ \int \left| \frac{c_n - c_n^*}{c_n^*} \right| \exp\left[-\frac{1}{2}\omega' J(\tilde{\theta})\omega\right] \mathrm{d}\omega + \int \left| \exp\left[-\frac{1}{2}\omega' J(\tilde{\theta})\omega\right] - \exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right] \right| \mathrm{d}\omega \right\} \\
\leq & \frac{|c_n - c_n^*|}{c_n} + \frac{1}{c_n} \int \left| \exp\left[-\frac{1}{2}\omega' J(\tilde{\theta})\omega\right] - \exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right] \right| \mathrm{d}\omega \\
\leq & \frac{2}{c_n} \int \left| \exp\left[-\frac{1}{2}\omega' J(\tilde{\theta})\omega\right] - \exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right] \right| \mathrm{d}\omega \\
\leq & \frac{2}{c_n} \int \left| \exp\left\{-\frac{1}{2}\omega' \left[J(\tilde{\theta}) - J(\hat{\theta}_m)\right]\omega\right\} - 1 \right| \exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right] \mathrm{d}\omega.
\end{aligned}
$$

By Assumption 3, for any $\varepsilon > 0$, there exists some $\delta > 0$ such that when $\Omega = \{\omega : ||\omega|| < \sqrt{n}\delta\}$ we have $\theta \in H(\hat{\theta}_m, \delta)$ and $-A(\varepsilon) \leq [J(\tilde{\theta})J^{-1}(\hat{\theta}_m) - I_P] \leq A(\varepsilon)$. By Hölder inequality, we have

$$
\begin{aligned}
& \lim_{n\to\infty} Q_n = \lim_{n\to\infty} \int \left| \exp\left\{-\frac{1}{2}\omega' \left[J(\tilde{\theta}) - J(\hat{\theta}_m)\right]\omega\right\} - 1 \right| \exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right] \mathrm{d}\omega \\
= & \lim_{n\to\infty} \int_\Omega \left| \exp\left\{-\frac{1}{2}\omega' \left[J(\tilde{\theta}) - J(\hat{\theta}_m)\right]\omega\right\} - 1 \right| \exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right] \mathrm{d}\omega \\
= & \lim_{n\to\infty} \int_\Omega \left| \exp\left\{-\frac{1}{2}\omega' \left[J(\tilde{\theta})J^{-1}(\hat{\theta}_m) - I_P\right] J(\hat{\theta}_m)\omega\right\} - 1 \right| \exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right] \mathrm{d}\omega \\
\leq & \left\{ \lim_{n\to\infty} \int_\Omega \left| \exp\left\{-\frac{1}{2}\omega' \left[J(\tilde{\theta})J^{-1}(\hat{\theta}_m) - I_P\right] J(\hat{\theta}_m)\omega\right\} - 1 \right|^2 \exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right] \mathrm{d}\omega \right\}^{1/2} \\
= & (D_1 - 2D_2 + D_3)^{1/2},
\end{aligned}
$$

where

$$
\begin{aligned}
D_1 &= \lim_{n\to\infty}\int_{\Omega}\exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right]\mathrm{d}\omega,\\
D_2 &= \lim_{n\to\infty}\int_{\Omega}\exp\left\{-\frac{1}{2}\omega'\left[J(\tilde{\theta})J^{-1}(\hat{\theta}_m)-I_P\right]J(\hat{\theta}_m)\omega\right\}\exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right]\mathrm{d}\omega\\
&= \lim_{n\to\infty}\int_{\Omega}\exp\left[-\frac{1}{2}\omega' J(\tilde{\theta})\omega\right]\mathrm{d}\omega,\\
D_3 &= \lim_{n\to\infty}\int_{\Omega}\exp\left\{-\omega'\left[J(\tilde{\theta})J^{-1}(\hat{\theta}_m)-I_P\right]J(\hat{\theta}_m)\omega\right\}\exp\left[-\frac{1}{2}\omega' J(\hat{\theta}_m)\omega\right]\mathrm{d}\omega.
\end{aligned}
$$

It can be shown that $D_1 = (2\pi)^{P/2}|J(\hat{\theta}_m)|^{-1/2}$. Following the proof of the posterior normality in Lemma 2.1 and Theorem 2.1 of Chen (1985), we have $D_2^- \le D_2 \le D_2^+, D_3^- \le D_2 \le D_3^+$ and

$$
\begin{aligned}
D_2^+ &= \left|J(\hat{\theta}_m)\right|^{1/2}|I_P - A(\varepsilon)|^{-1/2}\int_{||Z||<s_n}\exp\left[-\frac{1}{2}Z'Z\right]\mathrm{d}Z,\\
D_2^- &= \left|J(\hat{\theta}_m)\right|^{1/2}|I_P + A(\varepsilon)|^{-1/2}\int_{||Z||<t_n}\exp\left[-\frac{1}{2}Z'Z\right]\mathrm{d}Z,\\
D_3^+ &= \left|J(\hat{\theta}_m)\right|^{1/2}|I_P - 2A(\varepsilon)|^{-1/2}\int_{||Z||<s_n}\exp\left[-\frac{1}{2}Z'Z\right]\mathrm{d}Z,\\
D_3^- &= \left|J(\hat{\theta}_m)\right|^{1/2}|I_P + 2A(\varepsilon)|^{-1/2}\int_{||Z||<t_n}\exp\left[-\frac{1}{2}Z'Z\right]\mathrm{d}Z,
\end{aligned}
$$

where $s_n = \delta(1-e^*(\varepsilon))^{1/2}/\sigma_n^*$ and $t_n = \delta(1+e(\varepsilon))^{1/2}/\sigma_n$, σ_n^2 and σ_n^{*2} is the largest and smallest eigenvalue of $\{nJ(\hat{\theta}_m)\}^{-1}$, $e(\varepsilon)$ and $e^*(\varepsilon)$ is the largest and smallest eigenvalue of $A(\varepsilon)$. Under the regularity conditions, when $n\to\infty$, $s_n\to\infty$ and $t_n\to\infty$, if $\varepsilon\to 0$, we get

$$
\begin{aligned}
&\lim_{n\to\infty}|I_P \pm A(\varepsilon)| = 1, \lim_{n\to\infty}|I_P \pm 2A(\varepsilon)| = 1,\\
&\lim_{n\to\infty}\int_{||Z||<s_n}\exp\left[-\frac{1}{2}Z'Z\right]\mathrm{d}Z = (2\pi)^{P/2},\\
&\lim_{n\to\infty}\int_{||Z||<t_n}\exp\left[-\frac{1}{2}Z'Z\right]\mathrm{d}Z = (2\pi)^{P/2}.
\end{aligned}
$$

Then, we can show that $D_1 = D_2 = D_3 = (2\pi)^{P/2}|J(\hat{\theta}_m)|^{-1/2}$ which implies that $\lim_{n\to\infty} Q_n = 0$ and that $\lim_{n\to\infty} P_n = 0$.

For $i, j = 1, 2, \cdots, P$, it can be shown that

$$
\begin{aligned}
& \int \omega_i \left\{ p(\omega|\mathbf{y}) - \frac{1}{c_n} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \right\} \mathrm{d}\omega \\
\leq & \int |\omega_i| \left| p(\omega|\mathbf{y}) - \frac{1}{c_n} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \right| \mathrm{d}\omega \\
\leq & \frac{2}{c_n} \int |\omega_i| \left| \exp\left\{ -\frac{1}{2} \omega' \left[J(\tilde{\theta}) - J(\hat{\theta}_m) \right] \omega \right\} - 1 \right| \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \mathrm{d}\omega \\
\leq & \frac{2}{c_n} \left\{ \lim_{n\to\infty} \int_\Omega |\omega_i|^2 \left| \exp\left\{ -\frac{\omega' \left[J(\tilde{\theta}) J^{-1}(\hat{\theta}_m) - I_P \right] J(\hat{\theta}_m) \omega}{2} \right\} - 1 \right|^2 \exp\left[-\frac{\omega' J(\hat{\theta}_m) \omega}{2} \right] \mathrm{d}\omega \right\}^{\frac{1}{2}} \\
= & \frac{2}{c_n} (ED_1 - 2ED_2 + ED_3)^{1/2},
\end{aligned}
$$

$$
\begin{aligned}
& \int \omega_i \omega_j \left\{ p(\omega|\mathbf{y}) - \frac{1}{c_n} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \right\} \mathrm{d}\omega \\
\leq & \int |\omega_i \omega_j| \left| p(\omega|\mathbf{y}) - \frac{1}{c_n} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \right| \mathrm{d}\omega \\
\leq & \frac{2}{c_n} \int |\omega_i \omega_j| \left| \exp\left\{ -\frac{1}{2} \omega' \left[J(\tilde{\theta}) - J(\hat{\theta}_m) \right] \omega \right\} - 1 \right| \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \mathrm{d}\omega \\
= & \frac{2}{c_n} (VD_1 - 2VD_2 + VD_3)^{1/2},
\end{aligned}
$$

where

$$
\begin{aligned}
ED_1 &= \lim_{n\to\infty} \int_\Omega \omega_i^2 \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \mathrm{d}\omega, \\
ED_2 &= \lim_{n\to\infty} \int_\Omega \omega_i^2 \exp\left\{ -\frac{1}{2} \omega' \left[J(\tilde{\theta}) J^{-1}(\hat{\theta}_m) - I_P \right] J(\hat{\theta}_m) \omega \right\} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \mathrm{d}\omega, \\
&= \lim_{n\to\infty} \int_\Omega \omega_i^2 \exp\left[-\frac{1}{2} \omega' J(\tilde{\theta}) \omega \right] \mathrm{d}\omega, \\
ED_3 &= \lim_{n\to\infty} \int_\Omega \omega_i^2 \exp\left\{ -\omega' \left[J(\tilde{\theta}) J^{-1}(\hat{\theta}_m) - I_P \right] J(\hat{\theta}_m) \omega \right\} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \mathrm{d}\omega, \\
VD_1 &= \lim_{n\to\infty} \int_\Omega \omega_i^2 \omega_j^2 \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \mathrm{d}\omega, \\
VD_2 &= \lim_{n\to\infty} \int_\Omega \omega_i^2 \omega_j^2 \exp\left\{ -\frac{1}{2} \omega' \left[J(\tilde{\theta}) J^{-1}(\hat{\theta}_m) - I_P \right] J(\hat{\theta}_m) \omega \right\} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \mathrm{d}\omega, \\
&= \lim_{n\to\infty} \int_\Omega \omega_i^2 \omega_j^2 \exp\left[-\frac{1}{2} \omega' J(\tilde{\theta}) \omega \right] \mathrm{d}\omega, \\
VD_3 &= \lim_{n\to\infty} \int_\Omega \omega_i^2 \omega_j^2 \exp\left\{ -\omega' \left[J(\tilde{\theta}) J^{-1}(\hat{\theta}_m) - I_P \right] J(\hat{\theta}_m) \omega \right\} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \mathrm{d}\omega.
\end{aligned}
$$

For the same argument, we can prove that $ED_1 = ED_2 = ED_3$ and $VD_1 = VD_2 = VD_3$. Hence, we have

$$\begin{aligned}
& \int \omega_i \left\{ p(\omega|\mathbf{y}) - \frac{1}{c_n} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \right\} \mathrm{d}\omega \\
\leq & \int |\omega_i| \left| p(\omega|\mathbf{y}) - \frac{1}{c_n} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \right| \mathrm{d}\omega \to 0, \\
& \int \omega_i \omega_j \left\{ p(\omega|\mathbf{y}) - \frac{1}{c_n} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \right\} \mathrm{d}\omega \\
\leq & \int |\omega_i \omega_j| \left| p(\omega|\mathbf{y}) - \frac{1}{c_n} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \right| \mathrm{d}\omega \to 0.
\end{aligned}$$

Note that

$$\begin{aligned}
& \int \omega_i \left\{ \frac{1}{c_n} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \right\} \mathrm{d}\omega = 0, \\
& \int \omega_i \omega_j \left\{ \frac{1}{c_n} \exp\left[-\frac{1}{2} \omega' J(\hat{\theta}_m) \omega \right] \right\} \mathrm{d}\omega = J_{ij}^{-1}(\hat{\theta}_m),
\end{aligned}$$

where $J_{ij}^{-1}(\hat{\theta}_m)$ is the $(i,j)^{th}$ element of $J^{-1}(\hat{\theta}_m)$. Hence, we have $E(\omega|\mathbf{y}) = 0 + o(1)$ and $E(\omega\omega'|\mathbf{y}) = J^{-1}(\hat{\theta}_m) + o(1)$ which imply that

$$E[(\theta - \hat{\theta}_m)|\mathbf{y}] = o(n^{-1/2}), E[(\theta - \hat{\theta}_m)(\theta - \hat{\theta}_m)'|\mathbf{y}] = -L_n^{-(2)}(\hat{\theta}_m) + o(n^{-1}).$$

.1.2 Proof of Theorem 2.3.1

Under Assumption 6, when $n \to \infty$, we have

$$\frac{\partial \ln p(\mathbf{y}|\theta)}{\partial \theta} = L_n^{(1)}(\theta), -I(\theta) = \frac{\partial^2 \ln p(\mathbf{y}|\theta)}{\partial \theta \theta'} = L_n^{(2)}(\theta),$$

and the ML estimator $\hat{\theta}$ is asymptotically equivalent to the posterior mode $\hat{\theta}_m$. According to Lemma 2.3.1, we can show that $\bar{\theta} = E(\theta|\mathbf{y}) = \hat{\theta}_m + o(n^{-1/2})$. Hence, there exists an integer N, when $n > N$, $\bar{\theta} \in H(\hat{\theta}, \delta)$. We can then find some δ_1 with $0 < \delta_1 < ||\hat{\theta} - \bar{\theta}||$ so that $H(\bar{\theta}, \delta_1) \subset H(\hat{\theta}, \delta)$.

Applying the Taylor expansion to the log-likelihood function, we get

$$\ln p(\mathbf{y}|\theta) = \ln p(\mathbf{y}|\bar{\theta}) + L_n^{(1)}(\bar{\theta})(\theta - \bar{\theta}) + \frac{1}{2}(\theta - \bar{\theta})' L_n^{(2)}(\tilde{\theta})(\theta - \bar{\theta}),$$

where $\tilde{\theta}$ is some θ lying on the segment between θ and $\bar{\theta}$. When $n \to \infty$, $H(\bar{\theta}, \delta_1) \subset H(\hat{\theta}, \delta)$ and $\tilde{\theta} \in H(\bar{\theta}, \delta_1) \subset H(\hat{\theta}, \delta)$. Hence, for any $\varepsilon > 0$, there exists an integer N such that for any $n > N$, $L_n^{(2)}(\tilde{\theta})$ satisfies

$$[I_P - A(\varepsilon)]\,[-L_n^{(2)}(\hat{\theta})] \leq -L_n^{(2)}(\tilde{\theta}) = \left[L_n^{(2)}(\tilde{\theta}) L_n^{-(2)}(\hat{\theta})\right]\left[-L_n^{(2)}(\hat{\theta})\right] \leq [I_P + A(\varepsilon)]\left[-L_n^{(2)}(\hat{\theta})\right].$$

That is,

$$[I_P - A(\varepsilon)]\, I(\hat{\theta}) \leq I(\tilde{\theta}) = \left[I(\tilde{\theta}) I^{-1}(\hat{\theta})\right] I(\hat{\theta}) \leq [I_P + A(\varepsilon)]\, I(\hat{\theta}).$$

Hence, under the regularity conditions, when $n \to \infty$, we have

$$\begin{aligned}
P_D &= -2\int_\Theta \left[\ln p(\mathbf{y}|\theta) - \ln p(\mathbf{y}|\bar{\theta})\right] p(\theta|\mathbf{y})\mathrm{d}\theta \\
&= -2\int_\Theta \left[L_n^{(1)}(\bar{\theta})(\theta - \bar{\theta}) + \frac{1}{2}(\theta - \bar{\theta})' L_n^{(2)}(\tilde{\theta})(\theta - \bar{\theta})\right] p(\theta|\mathbf{y})\mathrm{d}\theta \\
&= \int_\Theta -(\theta - \bar{\theta})' L_n^{(2)}(\tilde{\theta})(\theta - \bar{\theta}) p(\theta|\mathbf{y})\mathrm{d}\theta \\
&= \int_{H(\hat{\theta},\delta)} (\theta - \bar{\theta})' I(\tilde{\theta})(\theta - \bar{\theta}) p(\theta|\mathbf{y})\mathrm{d}\theta \\
&= \int_{H(\hat{\theta},\delta)} (\theta - \bar{\theta})' I(\tilde{\theta}) I^{-1}(\hat{\theta}) I(\hat{\theta})(\theta - \bar{\theta}) p(\theta|\mathbf{y})\mathrm{d}\theta,
\end{aligned}$$

which is bounded above by

$$P_D^+ = \int_{H(\hat{\theta},\delta)} (\theta - \bar{\theta})' [I_P + A(\varepsilon)] I(\hat{\theta})(\theta - \bar{\theta}) p(\theta|\mathbf{y})\mathrm{d}\theta = \mathbf{tr}\left\{[I_P + A(\varepsilon)] I(\hat{\theta}) V(\bar{\theta})\right\},$$

and below by

$$P_D^- = \int_{H(\hat{\theta},\delta)} (\theta - \bar{\theta})' [I_P - A(\varepsilon)] I(\hat{\theta})(\theta - \bar{\theta}) p(\theta|\mathbf{y})\mathrm{d}\theta = \mathbf{tr}\left\{[I_P - A(\varepsilon)] I(\hat{\theta}) V(\bar{\theta})\right\}.$$

Under the regularity conditions, for $\varepsilon \to 0$, we have $\lim_{n\to\infty} P_D = \mathbf{tr}\{-L_n^{(2)}(\hat{\theta})V(\bar{\theta})\}$ or $P_D = \mathbf{tr}\{I(\hat{\theta})V(\bar{\theta})\} + o(1)$.

Conditional on the observed data $\mathbf{y}$, note that $L_n^{(2)}(\bar{\theta})/n = O(1)$, $L_n^{(2)}(\hat{\theta})/n = O(1)$, we get $L_n^{(2)}(\bar{\theta})/n = L_n^{(2)}(\hat{\theta})/n + o(1)$. According to Lemma 2.3.1, we have $nV(\hat{\theta}) = n[V(\bar{\theta}) + (\hat{\theta}-\bar{\theta})(\hat{\theta}-\bar{\theta})'] = nV(\bar{\theta}) + no(n^{-1}) = nV(\bar{\theta}) + o(1)$ and $nV(\hat{\theta}) = [-L_n^{(2)}(\hat{\theta})/n]^{-1} + o(1) = O(1)$ so that $nV(\bar{\theta}) = O(1)$. Thus, we have

$$\begin{aligned} & P_D = \mathbf{tr}\left\{I(\hat{\theta})V(\bar{\theta})\right\} + o(1) = \mathbf{tr}\left\{[I(\hat{\theta})/n][nV(\bar{\theta})]\right\} + o(1) \\ = \;& \mathbf{tr}\left\{[I(\bar{\theta})/n][nV(\bar{\theta})]\right\} + o(1)O(1) + o(1) \\ = \;& \mathbf{tr}\left\{[I(\bar{\theta})/n][nV(\bar{\theta})]\right\} + o(1) = \mathbf{tr}\left\{I(\bar{\theta})V(\bar{\theta})\right\} + o(1) = P_D^* + o(1). \end{aligned}$$

Similarly, DIC_1 =RDIC+$o(1)$ and the theorem is proved.

.1.3 Proof of Theorem 2.3.2

When $n \to \infty$, there exists δ_1 such that $H(\bar{\theta}, \delta_1) \subset H(\hat{\theta}, \delta)$. Under Assumption 7, we have

$$p(\mathbf{y}_{rep}|\theta) = p(\mathbf{y}_{rep}|\bar{\theta}) + \frac{\partial p(\mathbf{y}_{rep}|\bar{\theta})}{\partial \theta}(\theta-\bar{\theta}) + \frac{1}{2}(\theta-\bar{\theta})' \frac{\partial^2 p(\mathbf{y}_{rep}|\bar{\theta})}{\partial \theta \partial \theta'}(\theta-\bar{\theta}) + o_p(1),$$

and

$$\begin{aligned} & p(\mathbf{y}_{rep}|\mathbf{y}) = \int p(\mathbf{y}_{rep}|\theta)p(\theta|\mathbf{y})\mathrm{d}\theta \\ = \;& p(\mathbf{y}_{rep}|\bar{\theta}) + \frac{\partial p(\mathbf{y}_{rep}|\bar{\theta})}{\partial \theta}\int(\theta-\bar{\theta})\mathrm{d}\theta + + \frac{1}{2}\int\left[(\theta-\bar{\theta})' \frac{\partial^2 p(\mathbf{y}_{rep}|\bar{\theta})}{\partial \theta \partial \theta'}(\theta-\bar{\theta})\right]\mathrm{d}\theta + o_p(1) \\ = \;& p(\mathbf{y}_{rep}|\bar{\theta}) + \frac{1}{2}\int\left[(\theta-\bar{\theta})' \frac{\partial^2 p(\mathbf{y}_{rep}|\bar{\theta})}{\partial \theta \partial \theta'}(\theta-\bar{\theta})\right]\mathrm{d}\theta + o_p(1) \\ = \;& p(\mathbf{y}_{rep}|\bar{\theta})\left\{1 + \frac{1}{2}\int\left[(\theta-\bar{\theta})'\left[\frac{1}{p(\mathbf{y}_{rep}|\bar{\theta})}\frac{\partial^2 p(\mathbf{y}_{rep}|\bar{\theta})}{\partial \theta \partial \theta'}\right](\theta-\bar{\theta})\right]\mathrm{d}\theta\right\} + o_p(1) \\ = \;& p(\mathbf{y}_{rep}|\bar{\theta})\left\{1 + \frac{1}{2}\left[\frac{1}{p(\mathbf{y}_{rep}|\bar{\theta})}\frac{\partial^2 p(\mathbf{y}_{rep}|\bar{\theta})}{\partial \theta \partial \theta'}\right]V(\bar{\theta})\right\} + o_p(1). \end{aligned}$$

Noting that $\int p(\mathbf{y}_{rep}|\theta)\mathrm{d}\mathbf{y}_{rep} = 1$, we get

$$\int \left[\frac{\partial^2 \log p(\mathbf{y}_{rep}|\theta)}{\partial\theta\partial\theta'} + \frac{\partial \log p(\mathbf{y}_{rep}|\theta)}{\partial\theta} \frac{\partial \log p(\mathbf{y}_{rep}|\theta)}{\partial\theta'} \right] p(\mathbf{y}_{rep}|\theta)\mathrm{d}\mathbf{y}_{rep} = 0.$$

When θ is set at the true value θ_0, using the central limit theorem, we get

$$\frac{1}{n}\left[\frac{\partial^2 \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta\partial\theta'} + \frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta} \frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta'} \right] = o_p(1).$$

Using Lemma 3.1 and the asymptotic theory for maximum likelihood, we have $\bar{\theta} = \hat{\theta} + o_p(n^{-1/2})$ and $\hat{\theta} = \theta_0 + O_p(n^{-1/2})$. Hence,

$$\begin{aligned} & \frac{1}{n}\left[\frac{\partial^2 \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta\partial\theta'} + \frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta} \frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta'} \right] \\ = & \frac{1}{n}\left[\frac{\partial^2 \log p(\mathbf{y}_{rep}|\bar{\theta})}{\partial\theta\partial\theta'} + \frac{\partial \log p(\mathbf{y}_{rep}|\bar{\theta})}{\partial\theta} \frac{\partial \log p(\mathbf{y}_{rep}|\bar{\theta})}{\partial\theta'} \right] + o_p(1) \\ = & \frac{1}{n}\frac{1}{p(\mathbf{y}_{rep}|\bar{\theta})} \frac{\partial^2 p(\mathbf{y}_{rep}|\bar{\theta})}{\partial\theta\partial\theta'} + o_p(1). \end{aligned}$$

Based on Lemma 3.1, we get $V(\bar{\theta}) = O_p(n^{-1})$ and

$$\left[\frac{1}{p(\mathbf{y}_{rep}|\bar{\theta})} \frac{\partial^2 p(\mathbf{y}_{rep}|\bar{\theta})}{\partial\theta\partial\theta'} \right] V(\bar{\theta}) = o_p(n) O_p(n^{-1}) = o_p(1).$$

Therefore, we have

$$\begin{aligned} & -2\log p(\mathbf{y}_{rep}|\mathbf{y}) = -2\log p(\mathbf{y}_{rep}|\bar{\theta}) - 2\log\left\{ 1 + \frac{1}{2}\left[\frac{1}{p(\mathbf{y}_{rep}|\bar{\theta})} \frac{\partial^2 p(\mathbf{y}_{rep}|\bar{\theta})}{\partial\theta\partial\theta'} \right] V(\bar{\theta}) \right\} + o_p(1) \\ = & -2\log p(\mathbf{y}_{rep}|\bar{\theta}) - \left\{ \left[\frac{1}{p(\mathbf{y}_{rep}|\bar{\theta})} \frac{\partial^2 p(\mathbf{y}_{rep}|\bar{\theta})}{\partial\theta\partial\theta'} \right] V(\bar{\theta}) \right\} + o_p(1) \\ = & -2\log p(\mathbf{y}_{rep}|\bar{\theta}) + o_p(1). \end{aligned}$$

Using the Taylor expansion, we have

$$\log p(\mathbf{y}_{rep}|\bar{\theta}) = \log p(\mathbf{y}_{rep}|\hat{\theta}) + \frac{\partial \log p(\mathbf{y}_{rep}|\hat{\theta})}{\partial\theta}(\bar{\theta} - \hat{\theta}) + \frac{1}{2}(\bar{\theta} - \hat{\theta})' \frac{\partial^2 \log p(\mathbf{y}_{rep}|\tilde{\theta})}{\partial\theta\partial\theta'}(\bar{\theta} - \hat{\theta})$$

where $\tilde{\theta}_1$ lies on the segment between $\bar{\theta}$ and $\hat{\theta}$. Noting that $\bar{\theta} = \hat{\theta} + o_p(n^{-1/2})$ and $\hat{\theta} = \theta_0 + O_p(n^{-1/2})$, we get $\tilde{\theta}_1 = \theta_0 + O_p(n^{-1/2})$ and

$$\begin{aligned}
& \frac{1}{n}\left[\frac{\partial^2 \log p(\mathbf{y}_{rep}|\tilde{\theta}_1)}{\partial\theta\partial\theta'}\right] = \frac{1}{n}\left[\frac{\partial^2 \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta\partial\theta'}\right] + o_p(1) \\
= & \int \frac{1}{n}\left[\frac{\partial^2 \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta\partial\theta'}\right] p(\mathbf{y}_{rep})\mathrm{d}\mathbf{y}_{rep} + o_p(1) \\
= & \int \frac{1}{n}\left[\frac{\partial^2 \log p(\mathbf{y}|\theta_0)}{\partial\theta\partial\theta'}\right] p(\mathbf{y})\mathrm{d}\mathbf{y} + o_p(1) \\
= & \frac{1}{n}\left[\frac{\partial^2 \log p(\mathbf{y}|\hat{\theta})}{\partial\theta\partial\theta'}\right] + o_p(1) = O_p(1).
\end{aligned}$$

Furthermore, we get

$$\frac{\partial \log p(\mathbf{y}_{rep}|\hat{\theta})}{\partial\theta} = \frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta} + \frac{\partial^2 \log p(\mathbf{y}_{rep}|\tilde{\theta}_2)}{\partial\theta\partial\theta'}(\bar{\theta} - \theta_0),$$

where $\tilde{\theta}_2$ lies on the segment between $\hat{\theta}$ and θ_0. It is also noted that $\tilde{\theta}_2 = \theta_0 + O_p(n^{-1/2})$ and, hence,

$$\frac{\partial \log p(\mathbf{y}_{rep}|\hat{\theta})}{\partial\theta} = \frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta} + O_p(n)O_p(n^{-1/2}) = \frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta} + O_p(n^{1/2})$$

Furthermore,

$$\begin{aligned}
& E_{\mathbf{y}_{rep}}\left[\frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta}\right] = \int \frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta} p(\mathbf{y}_{rep})\mathrm{d}\mathbf{y}_{rep} \\
= & \int \frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta} p(\mathbf{y}_{rep}|\theta_0)\mathrm{d}\mathbf{y}_{rep} \\
= & \int \frac{\partial p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta}\mathrm{d}\mathbf{y}_{rep} = \frac{\partial \int p(\mathbf{y}_{rep}|\theta_0)\mathrm{d}\mathbf{y}_{rep}}{\partial\theta} = 0
\end{aligned}$$

Based on the derivation of AIC shown in Burnham and Anderson (2002) and given the observable data $\mathbf{y}$, we have

$$
\begin{aligned}
& E_{\mathbf{y}}E_{\mathbf{y}_{rep}}[\mathscr{L}(\mathbf{y}_{rep},\mathbf{y})] = E_{\mathbf{y}}E_{\mathbf{y}_{rep}}[-2\log p(\mathbf{y}_{rep}|\mathbf{y})] \\
= \;& E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\log p(\mathbf{y}_{rep}|\bar{\theta})\right] + o(1) \\
= \;& -2E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\log p(\mathbf{y}_{rep}|\hat{\theta}) + \frac{\partial \log p(\mathbf{y}_{rep}|\bar{\theta})}{\partial\theta}(\bar{\theta}-\hat{\theta}) + \frac{1}{2}(\bar{\theta}-\hat{\theta})'\frac{\partial^2 \log p(\mathbf{y}_{rep}|\tilde{\theta})}{\partial\theta\partial\theta'}(\bar{\theta}-\hat{\theta})\right] + o(1) \\
= \;& -2E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\log p(\mathbf{y}_{rep}|\hat{\theta}) + \frac{\partial \log p(\mathbf{y}_{rep}|\bar{\theta})}{\partial\theta}(\bar{\theta}-\hat{\theta}) + o(n^{-1/2})O(n)o(n^{-1/2})\right] + o(1) \\
= \;& -2E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\log p(\mathbf{y}_{rep}|\hat{\theta}) + \frac{\partial \log p(\mathbf{y}_{rep}|\bar{\theta})}{\partial\theta}(\bar{\theta}-\hat{\theta})\right] + o(1) \\
= \;& -2E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\log p(\mathbf{y}_{rep}|\hat{\theta}) + \frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta}(\bar{\theta}-\hat{\theta}) + O(n^{1/2})o(n^{-1/2})\right] + o(1) \\
= \;& -2E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\log p(\mathbf{y}_{rep}|\hat{\theta}) + \frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta}(\bar{\theta}-\hat{\theta})\right] + o(1) \\
= \;& -2E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\log p(\mathbf{y}_{rep}|\hat{\theta})\right] - 2E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta}(\bar{\theta}-\hat{\theta})\right] + o(1). \\
= \;& -2E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\log p(\mathbf{y}_{rep}|\hat{\theta})\right] - 2\left[E_{\mathbf{y}_{rep}}\frac{\partial \log p(\mathbf{y}_{rep}|\theta_0)}{\partial\theta}\right]\left[E_{\mathbf{y}}(\bar{\theta}-\hat{\theta})\right] + o(1) \\
= \;& -2E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\log p(\mathbf{y}_{rep}|\hat{\theta})\right] + 0 + o(1) \\
= \;& E_{\mathbf{y}}(AIC) + o(1) == E_{\mathbf{y}}(AIC) + o(1) = E_{\mathbf{y}}(RDIC) + o(1)
\end{aligned}
$$

.1.4 The derivation of RDIC for the asset pricing models

It has been noted in Kan and Zhou (2003) that under the multivariate t specification, a direct numerical optimization of the observed data likelihood function is very difficult. By using normal-gamma scale-mixture distribution to replace the t distribution, the powerful EM algorithm can be used to obtain the $\mathscr{Q}$ function. Since Models 1-5 are nested by Model 6, we only need to derive the first and second derivatives for Model 6.

Let $\mathbf{R} = \{\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_n\}$, $\mathbf{F} = \{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_n\}$, $\omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$, $\theta = (\alpha, \beta, \Sigma)$. The density function of the multivariate t is given by

$$
f(\varepsilon_t) = \frac{\Gamma(\frac{\nu+N}{2})}{(\pi\nu)^{\frac{2}{N}}\Gamma(\frac{\nu}{2})|\Sigma|^{\frac{1}{2}}}\left\{1 + \frac{\varepsilon_t'\Sigma^{-1}\varepsilon_t}{\nu}\right\}^{-\frac{\nu+N}{2}}.
$$

Hence, the observed data log-likelihood function, $L_o(\mathbf{R}|\theta)$, is:

$$L_o(\mathbf{R}|\theta) = C(\nu) - \frac{n}{2}\ln|\Sigma| - \frac{\nu+N}{2}\sum_{t=1}^{n}\log\left[\nu+\varphi\left(R_t,F_t,\theta\right)\right], \qquad (.1.1)$$

where

$$C(\nu) = -\frac{nN}{2}\log(\pi\nu) + n\left[\ln\Gamma\left(\frac{\nu+N}{2}\right) - \ln\Gamma\left(\frac{\nu}{2}\right)\right] + \frac{n(\nu+N)\ln\nu}{2},$$
$$\varphi\left(R_t,F_t,\theta\right) = \left(R_t-\alpha-\beta F_t\right)' \Sigma^{-1}\left(R_t-\alpha-\beta F_t\right).$$

Based on the normal-gamma mixture representation for the multivariate t distribution, the complete log-likelihood, $L_c(\mathbf{R},\omega|\theta)$, can be expressed as

$$-\frac{1}{2}nN\ln(2\pi) + \frac{N}{2}\sum_{t=1}^{n}\ln\omega_t - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{t=1}^{n}\omega_t\varphi\left(R_t,F_t,\theta\right)$$
$$-n\ln\Gamma\left(\frac{\nu}{2}\right) + \frac{n\nu}{2}\ln\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\sum_{t=1}^{n}(\ln\omega_t-\omega_t) - \sum_{t=1}^{n}\ln\omega_t.$$

Thus, the posterior expectation of ω_t is

$$\omega_t|\mathbf{y} \sim \Gamma\left[\frac{\nu+N}{2}, \frac{\nu+\varphi\left(\mathbf{R}_t,\mathbf{F}_t\theta\right)}{2}\right].$$

According to McLachlan and Krishnan (2008), it can be shown that

$$E\left(\omega_t|\theta,\mathbf{R}_t\right) = \frac{\nu+N}{\nu+\varphi\left(R_t,F_t,\theta\right)},$$
$$E\left(\ln\omega_t|\theta,\mathbf{R}_t\right) = \ln E\left(\omega_t|\theta,\mathbf{R}_t\right) + \psi\left(\frac{\nu+N}{2}\right) - \ln\left(\frac{\nu+N}{2}\right),$$

where $\psi(x)$ is the Digamma function, $\partial\Gamma(x)/\partial x/\Gamma(x)$. Hence, we get

$$\begin{aligned}
&\mathscr{Q}(\theta|\theta^*) = \int L_c(\mathbf{R},\omega|\theta)p(\omega|\mathbf{R},\theta^*)\mathrm{d}\omega \\
= \;& -\frac{1}{2}nK\ln(2\pi) + \frac{N}{2}\sum_{t=1}^{n}E(\ln\omega_t|\mathbf{R}_t,\theta^*) - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{t=1}^{n}E(\omega_t|\mathbf{R}_t,\theta^*)\varphi(R_t,F_t,\theta) \\
& -n\ln\Gamma\left(\frac{\nu}{2}\right) + \frac{n\nu}{2}\ln\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\sum_{t=1}^{n}E(\ln\omega_t-\omega_t|\mathbf{R}_t,\theta^*) - \sum_{t=1}^{n}E(\ln\omega_t|\mathbf{R}_t,\theta^*).
\end{aligned}$$

For the asset price models considered in this paper, we obtain the second derivatives:

$$\begin{aligned}
\frac{\partial \mathscr{Q}(\theta|\theta^*)}{\partial\theta\partial\theta^{'}} &= \frac{\partial^2(-\frac{n}{2}\ln|\Sigma|)}{\partial\theta\partial\theta^{'}} - \frac{1}{2}\sum_{t=1}^{n} E(\omega_t|\mathbf{R}_t,\theta^*)\frac{\partial^2\varphi(R_t,F_t,\theta)}{\partial\theta\partial\theta^{'}} \\
\frac{\partial \mathscr{Q}(\theta|\theta^*)}{\partial\theta\partial\theta^{*'}} &= -\frac{1}{2}\sum_{t=1}^{n}\frac{\partial\varphi(R_t,F_t,\theta)}{\partial\theta}\frac{\partial E(\omega_t|\mathbf{R}_t,\theta^*)}{\partial\theta^{*'}} \\
= \quad \frac{1}{2}\sum_{t=1}^{n}&\frac{1}{\nu+\varphi(R_t,F_t,\theta^*)}E(\omega_t|\mathbf{R}_t,\theta^*)\frac{\partial\varphi(R_t,F_t,\theta)}{\partial\theta}\frac{\partial\varphi(R_t,F_t,\theta^*)}{\partial\theta^{*'}}
\end{aligned}$$

For $i,j=1,2,\cdots,N$, letting $\phi_i=\sigma_{ii}^{-1}$, we get

$$\begin{aligned}
&\frac{\partial^2(-\frac{n}{2}\ln|\Sigma|)}{\partial\alpha\partial\alpha^{'}}=0, \frac{\partial^2(-\frac{n}{2}\ln|\Sigma|)}{\partial\alpha\partial\beta^{'}}=0, \frac{\partial^2(-\frac{n}{2}\ln|\Sigma|)}{\partial\alpha\partial\phi_i}=0, \\
&\frac{\partial^2(-\frac{n}{2}\ln|\Sigma|)}{\partial\beta\partial\beta^{'}}=0, \frac{\partial^2(-\frac{n}{2}\ln|\Sigma|)}{\partial\beta\partial\phi_i}=0, \frac{\partial^2(-\frac{n}{2}\ln|\Sigma|)}{\partial\phi_i^2}=-\frac{n}{2\phi_i^2}, \\
&\frac{\partial\varphi(R_t,F_t,\theta)}{\partial\alpha_i}=-2\phi_i(R_{it}-\alpha_i-\beta_i^{'}\mathbf{F}_t), \\
&\frac{\partial\varphi(R_t,F_t,\theta)}{\partial\beta_i}=-2\phi_i(R_{it}-\alpha_i-\beta_i^{'}\mathbf{F}_t)\mathbf{F}_t, \\
&\frac{\partial\varphi(R_t,F_t,\theta)}{\partial\phi_i}=(R_{it}-\alpha_i-\beta_i^{'}\mathbf{F}_t)^2, \\
&\frac{\partial^2\varphi(R_t,F_t,\theta)}{\partial\alpha_i^2}=2\phi_i, \frac{\partial\varphi^2(R_t,F_t,\theta)}{\partial\alpha_i\partial\alpha_j}=0, i\neq j, \\
&\frac{\partial^2\varphi(R_t,F_t,\theta)}{\partial\alpha_i\partial\beta_i}=2\phi_i F_t, \frac{\partial\varphi^2(R_t,F_t,\theta)}{\partial\alpha_i\partial\beta_j}=0, i\neq j, \\
&\frac{\partial^2\varphi(R_t,F_t,\theta)}{\partial\alpha_i\partial\phi_i}=-2(R_{it}-\alpha_i-\beta_i^{'}F_t), \frac{\partial\varphi^2(R_t,F_t,\theta)}{\partial\alpha_i\partial\phi_j}=0, i\neq j, \\
&\frac{\partial^2\varphi(R_t,F_t,\theta)}{\partial\beta_i\partial\beta_i^{'}}=2\phi_i F_t F_t^{'}, \frac{\partial\varphi^2(R_t,F_t,\theta)}{\partial\beta_i\partial\beta_j}=0, i\neq j, \\
&\frac{\partial^2\varphi(R_t,F_t,\theta)}{\partial\beta_i\partial\phi_i}=-2(R_{it}-\alpha_i-\beta_i^{'}F_t)F_t, \frac{\partial^2\varphi(R_t,F_t,\theta)}{\partial\beta_i\partial\phi_j}=0, i\neq j, \\
&\frac{\partial^2\varphi(R_t,F_t,\theta)}{\partial\phi_i^2}=0, \frac{\partial\varphi^2(R_t,F_t,\theta)}{\partial\phi_i\partial\phi_j}=0, i\neq j.
\end{aligned}$$

.1.5 The derivation of RDIC for the dynamic factor models

The complete-data log-likelihood function is:

$$\begin{aligned}\ln f(Y,F|L,\Sigma,\Phi,Q) &= -\frac{(K+N)T-K}{2}\ln 2\pi-\frac{T}{2}\ln|\Sigma|-\frac{1}{2}\mathbf{tr}\left[\Sigma^{-1}\left(Y-FL'\right)'\left(Y-FL'\right)\right]\\ &\quad -\frac{T-1}{2}\ln|Q|-\frac{1}{2}\mathbf{tr}\left[Q^{-1}\left(F_{+1}-F_{-1}\Phi'\right)'\left(F_{+1}-F_{-1}\Phi'\right)\right],\end{aligned}$$

where $Y=[Y_1',Y_2',...,Y_T']'$, $F=[F_1',F_2',...,F_T']'$, $F_{+1}=\left[F_2',F_3',...,F_T'\right]'$, $F_{-1}=\left[F_1',F_2',...,F_{T-1}'\right]'$. Denote this function by $\varphi(L,\Sigma,\Phi,Q)$, In this appendix, we derive the first and second derivative of the complete-data log-likelihood function. The matrix differentiation used here follows the rules discussed in Magnus and Neudecker (1999).

The first order derivatives of $\varphi(L,\Sigma,\Phi,Q)$:

Whenever there is no confusion, we denote $\varphi(L,\Sigma,\Phi,Q)$ simply by φ. The differential of $\varphi(L,\Sigma,\Phi,Q)$ with respect to L is

$$\begin{aligned}d_L(\varphi) &= d\left(-\frac{1}{2}\mathbf{tr}\left[\Sigma^{-1}\left(Y-FL'\right)'\left(Y-FL'\right)\right]\right)\\ &= -\frac{1}{2}\mathbf{tr}\left\{-\Sigma^{-1}(dL)F'\left(Y-FL'\right)+\Sigma^{-1}\left(Y-FL'\right)'\left(-F(dL)'\right)\right\}\\ &= \frac{1}{2}\mathbf{tr}\left\{\Sigma^{-1}dLF'\left(Y-FL'\right)+\Sigma^{-1}\left(Y-FL'\right)'F(dL)'\right\}\\ &= \frac{1}{2}\mathbf{tr}\left\{F'\left(Y-FL'\right)\Sigma^{-1}dL+dLF'\left(Y-FL'\right)\left(\Sigma^{-1}\right)'\right\}\\ &= \frac{1}{2}\mathbf{tr}\left\{F'\left(Y-FL'\right)\left(\left(\Sigma^{-1}\right)'+\Sigma^{-1}\right)dL\right\}\\ &= \mathbf{tr}\left(\widetilde{c}dL\right),\end{aligned}$$

where

$$\widetilde{c}=\frac{1}{2}F'\left(Y-FL'\right)\left(\left(\Sigma^{-1}\right)'+\Sigma^{-1}\right).$$

Taking *vec* both sides, we get

$$d\left(vec\left(-\frac{1}{2}\mathbf{tr}\left[\Sigma^{-1}(Y-FL')'(Y-FL)\right]\right)\right)=d(vec(\varphi))=\left(vec(\widetilde{c}')\right)'d(vec(L)).$$

The first derivative of $\varphi(L,\Sigma,\Phi,Q)$ is

$$D_L(\varphi)=\left(vec\left(\left[\frac{1}{2}F'\left(Y-FL'\right)\left(\left(\Sigma^{-1}\right)'+\Sigma^{-1}\right)\right]'\right)\right)'.$$

Similarly, we have

$$D_\Sigma(\varphi)=\left(vec\left(-\frac{T}{2}\Sigma^{-1}+\frac{1}{2}\Sigma^{-1}\left(Y-FL'\right)'\left(Y-FL'\right)\Sigma^{-1}\right)'\right)',$$

$$D_\Phi(\varphi)=\left(vec\left(\left[\frac{1}{2}F'_{-1}\left(F_{+1}-F_{-1}\Phi'\right)\left(\left(Q^{-1}\right)'+Q^{-1}\right)\right]'\right)\right)',$$

$$D_Q(\varphi)=\left(vec\left(-\frac{T-1}{2}Q^{-1}+\frac{1}{2}Q^{-1}\left(F_{+1}-F_{-1}\Phi'\right)'\left(F_{+1}-F_{-1}\Phi'\right)Q^{-1}\right)'\right)'.$$

The second order derivatives of $\varphi(L,\Sigma,\Phi,Q)$:

The first order derivative of $\widetilde{c}$ is

$$d\widetilde{c}=d\left(\frac{1}{2}F'\left(Y-FL'\right)\left(\left(\Sigma^{-1}\right)'+\Sigma^{-1}\right)\right)=-\frac{1}{2}F'F\left(dL\right)'\left(\left(\Sigma^{-1}\right)'+\Sigma^{-1}\right).$$

And the second order derivative is

$$\begin{aligned}d_L^2\varphi &= \mathbf{tr}\left(d\widetilde{c}*dL\right)\\ &= \mathbf{tr}\left(-\frac{1}{2}F'F\left(dL\right)'\left(\left(\Sigma^{-1}\right)'+\Sigma^{-1}\right)dL\right).\end{aligned}$$

Then, we have,

$$D_{L,L}(\varphi)=-\frac{1}{2}\left(F'F\otimes\left(\left(\Sigma^{-1}\right)'+\Sigma^{-1}\right)\right),$$

$$\begin{aligned}H &= G(T)=T',\quad T=S(\Sigma)=\frac{1}{2}F'\left(Y-FL'\right)\left(\left(\Sigma^{-1}\right)'+\Sigma^{-1}\right),\\ D(G(T)) &= K_{KN},\\ D(S(\Sigma)) &= I_N\otimes\left(F'\left(Y-FL'\right)\right)\cdot\left(-\frac{1}{2}\left(K_{NN}+I_{NN}\right)\right)\cdot\left(\left(\Sigma^{-1}\right)'\otimes\Sigma^{-1}\right),\\ DH(\Sigma) &= \left(DG(T)\right)\left(DS(\Sigma)\right),\end{aligned}$$

where K_{KN} is the commutation matrix for a matrix with K rows and N columns. Thus, we have

$$\begin{aligned} D_{L,\Sigma}(\varphi) &= \frac{\partial D_L(\varphi)}{(\partial vec\Sigma)'} = (DG(T))(DS(\Sigma)) \\ &= K_{KN}\cdot I_N \otimes \left(F'\left(Y-FL'\right)\right)\cdot\left(-\frac{1}{2}\left(K_{NN}+I_{NN}\right)\right)\cdot\left(\left(\Sigma^{-1}\right)'\otimes\Sigma^{-1}\right), \end{aligned}$$

$$\begin{aligned} D_{L,\Phi}(\varphi) &= 0, \\ D_{L,Q}(\varphi) &= 0, \end{aligned}$$

$$D_{\Sigma,\Sigma}(\varphi) = K_{NN}\cdot\left(\begin{array}{c} \frac{T}{2}\cdot\frac{1}{2}\left(\left(\Sigma^{-1}\right)'\otimes\Sigma^{-1}+\left(\Sigma^{-1}\right)'\otimes\Sigma^{-1}\right) \\ -\frac{1}{2}\left(\begin{array}{c}\left(\Sigma^{-1}\left(Y-FL'\right)'\left(Y-FL'\right)\Sigma^{-1}\right)'\otimes\Sigma^{-1} \\ +\left(\Sigma^{-1}\right)'\otimes\left(\Sigma^{-1}\left(Y-FL'\right)'\left(Y-FL'\right)\Sigma^{-1}\right)\end{array}\right)\end{array}\right),$$

$$\begin{aligned} D_{\Sigma,\Phi}(\varphi) &= 0, \\ D_{\Sigma,Q}(\varphi) &= 0, \end{aligned}$$

$$\begin{aligned} & D_{\Phi,Q}(\varphi) \\ = \; & K_{KK}\cdot\left(I_K\otimes F'_{-1}\left(F_{+1}-F_{-1}\Phi'\right)\right)\cdot\left(-\frac{1}{2}\left(K_{KK}+I_{KK}\right)\right)\cdot\left(\left(Q^{-1}\right)'\otimes Q^{-1}\right), \end{aligned}$$

$$D_{\Phi,\Phi}(\varphi) = -\frac{1}{2}\left(F'_{-1}F_{-1}\otimes\left(\left(Q^{-1}\right)'+Q^{-1}\right)\right),$$

$$D_{Q,Q}(\varphi) = K_{KK}\cdot\left(\begin{array}{c} \frac{T-1}{2}\cdot\frac{1}{2}\left(\left(Q^{-1}\right)'\otimes Q^{-1}+\left(Q^{-1}\right)'\otimes Q^{-1}\right) \\ -\frac{1}{2}\left(\begin{array}{c}\left(Q^{-1}\left(F_{+1}-F_{-1}\Phi'\right)'\left(F_{+1}-F_{-1}\Phi'\right)Q^{-1}\right)'\otimes Q^{-1} \\ +\left(Q^{-1}\right)'\otimes\left(\Sigma^{-1}\left(F_{+1}-F_{-1}\Phi'\right)'\left(F_{+1}-F_{-1}\Phi'\right)Q^{-1}\right)\end{array}\right)\end{array}\right).$$

The special structure of parameter matrix:

Let L,Σ,Φ,Q have some special structures. In particular, let

$$L^* = vec\left(\overline{L}\right),$$

where $\overline{L}$ is the last $(N-K)\times K$ block of L, and

$$\Sigma^{*}=diag\left(\Sigma\right),\Phi^{*}=vec\left(\Phi\right),Q^{*}=vech\left(Q\right).$$

The first order derivatives are as follows:

$$\begin{aligned}
D_{L^{*}}\left(\varphi\right) &= D_{L}\left(\varphi\right)\cdot D_{L^{*}}\left(L\left(L^{*}\right)\right)=D_{L}\left(\varphi\right)\cdot\dot{I}_{L^{*}},\\
D_{\Sigma^{*}}\left(\varphi\right) &= D_{\Sigma}\left(\varphi\right)\cdot D_{\Sigma^{*}}\left(\Sigma\left(\Sigma^{*}\right)\right)=D_{\Sigma}\left(\varphi\right)\cdot\dot{I}_{\Sigma^{*}},\\
D_{\Phi^{*}}\left(\varphi\right) &= D_{\Phi}\left(\varphi\right)\cdot\dot{I}_{\Phi^{*}},\\
D_{Q^{*}}\left(\varphi\right) &= D_{Q}\left(\varphi\right)\cdot\dot{I}_{Q^{*}}.
\end{aligned}$$

The second order derivatives are as follows:

$$\begin{aligned}
D_{L^{*},L^{*}}\left(\varphi\right) &= D_{L^{*}}\left(D_{L^{*}}\left(\varphi\right)\right)=D_{L^{*}}\left(D_{L}\left(\varphi\right)\cdot\dot{I}_{L^{*}}\right)\\
&= \left(\dot{I}_{L^{*}}^{'}\otimes I_{1}\right)\cdot D_{L^{*}}\left(D_{L}\left(\varphi\right)\right)\\
&= \left(\dot{I}_{L^{*}}^{'}\otimes I_{1}\right)\cdot D_{L,L}\left(\varphi\right)\cdot\dot{I}_{L^{*}},\\
D_{L^{*},\Sigma^{*}}\left(\varphi\right) &= D_{\Sigma^{*}}\left(D_{L^{*}}\left(\varphi\right)\right)=D_{\Sigma^{*}}\left(D_{L}\left(\varphi\right)\cdot\dot{I}_{L^{*}}\right)\\
&= \left(\dot{I}_{L^{*}}^{'}\otimes I_{1}\right)\cdot D_{\Sigma^{*}}\left(D_{L}\left(\varphi\right)\right)\\
&= \left(\dot{I}_{L^{*}}^{'}\otimes I_{1}\right)\cdot D_{\Sigma}\left(D_{L}\left(\varphi\right)\right)\cdot D_{\Sigma^{*}}\left(\Sigma\left(\Sigma^{*}\right)\right)\\
&= \dot{I}_{L^{*}}^{'}\cdot D_{L,\Sigma}\left(\varphi\right)\cdot\dot{I}_{\Sigma^{*}},\\
D_{L^{*},\Phi^{*}}\left(\varphi\right) &= 0,\\
D_{L^{*},Q^{*}}\left(\varphi\right) &= 0,
\end{aligned}$$

$$\begin{aligned}
D_{\Sigma^{*},\Sigma^{*}}\left(\varphi\right) &= D_{\Sigma^{*}}\left(D_{\Sigma^{*}}\left(\varphi\right)\right)=D_{\Sigma^{*}}\left(D_{\Sigma}\left(\varphi\right)\cdot\dot{I}_{\Sigma^{*}}\right)\\
&= \dot{I}_{\Sigma^{*}}^{'}\otimes I_{1}\cdot D_{\Sigma^{*}}\left(D_{\Sigma}\left(\varphi\right)\right)\\
&= \dot{I}_{\Sigma^{*}}^{'}\cdot D_{\Sigma}\left(D_{\Sigma}\left(\varphi\right)\right)\cdot\dot{I}_{\Sigma^{*}},\\
D_{\Sigma^{*},\Phi^{*}}\left(\varphi\right) &= 0,\\
D_{\Sigma^{*},Q^{*}}\left(\varphi\right) &= 0.
\end{aligned}$$

$$
\begin{aligned}
D_{\Phi^*,\Phi^*}(\varphi) &= \dot{I}'_{\Phi^*} \cdot (D_{\Phi,\Phi}(\varphi)) \cdot \dot{I}_{\Phi^*}, \\
D_{\Phi^*,Q^*}(\varphi) &= \dot{I}'_{\Phi^*} \cdot \left(D_{\Phi,Q}(\varphi)\right) \cdot \dot{I}_{Q^*}, \\
D_{Q^*,Q^*}(\varphi) &= \dot{I}'_{Q^*} \cdot D_{Q,Q}(\varphi) \cdot \dot{I}_{Q^*},
\end{aligned}
$$

where $D_{L^*}(L(L^*)) = \dot{I}_{L^*}$, $D_{\Sigma^*}(\Sigma(\Sigma^*)) = \dot{I}_{\Sigma^*}$.

For $\dot{I}_{L^*}$ which is a block diagonal matrix, we have

$$
\dot{I}_{L^*} = diag(P_1, P_2, ..., P_K),
$$

where

$$
P_i = \begin{bmatrix} 0_{K\times(N-K)} \\ I_{N-K} \end{bmatrix}.
$$

And for $\dot{I}_{\Sigma^*}$, which is an $N^2 \times N$ matrix whose n^{th} column has 1 in the $((n-1)\times N + n)^{th}$ row and other elements are all zeros. For $\dot{I}_{\Phi^*}$, we have

$$
\dot{I}_{\Phi^*} = I_{K*K}.
$$

For $\dot{I}_{Q^*}$, we have

$$
\dot{I}_{Q^*} = diag(R_1, R_2, ...R_k, ..., R_K).
$$

where

$$
R_k = \begin{bmatrix} 0_{(k-1)\times(K-k+1)} \\ I_{K-k+1} \end{bmatrix}_{K\times(K-k+1)},
$$

since Q is a symmetric matrix.

The first order derivative matrix of the complete-data likelihood with respect to $L^*, \Sigma^*, \Phi^*, Q^*$ is:

$$
vec\left(\begin{bmatrix} D_{L^*}(\varphi) & D_{\Sigma^*}(\varphi) & D_{\Phi^*}(\varphi) & D_{Q^*}(\varphi) \end{bmatrix}\right).
$$

The second order derivative matrix of the complete-data likelihood with respect to

$L^*, \Sigma^*, \Phi^*, Q^*$ should be:

$$\begin{bmatrix} D_{L^*,L^*}(\varphi) & D_{L^*,\Sigma^*}(\varphi) & 0 & 0 \\ D_{\Sigma^*,L^*}(\varphi) & D_{\Sigma^*,\Sigma^*}(\varphi) & 0 & 0 \\ 0 & 0 & D_{\Phi^*,\Phi^*}(\varphi) & D_{\Phi^*,Q^*}(\varphi) \\ 0 & 0 & D_{Q^*,\Phi^*}(\varphi) & D_{Q^*,Q^*}(\varphi) \end{bmatrix}.$$

.1.6 The derivation of RDIC for the stochastic volatility model

The derivatives of the complete-data log-likelihood for M_1

The complete-data log-likelihood function

$$\begin{aligned} \ln p(\mathbf{y},\mathbf{h}|\theta) &= -n\ln 2\pi + \frac{n}{2}\ln \nu - \frac{1}{2}\sum_{t=1}^{n} h_t - \frac{1}{2}\sum_{t=1}^{n}\frac{(y_t-\alpha)^2}{\exp(h_t)} \\ &\quad -\frac{1}{2}\nu\left[\sum_{t=1}^{n}(h_t-\mu-\phi(h_{t-1}-\mu))^2\right], \end{aligned}$$

where $\mathbf{y} = (y_1, y_2, \ldots y_n)'$, $\mathbf{h} = (h_1, h_2, \ldots h_n)'$, $\nu = 1/\tau^2$.

The first order derivatives

$$\frac{\partial \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \alpha} = \sum_{t=1}^{n}\frac{(y_t-\alpha)}{\exp(h_t)},$$

$$\begin{aligned} \frac{\partial \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \mu} &= -\frac{1}{2}\nu\left[-2\sum_{t=1}^{n}(h_t-\mu-\phi(h_{t-1}-\mu))(1-\phi)\right] \\ &= \nu\left[(1-\phi)\sum_{t=1}^{n}(h_t-\mu-\phi(h_{t-1}-\mu))\right], \end{aligned}$$

$$\begin{aligned}\frac{\partial \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \phi} &= -\frac{1}{2}\nu\left[-2\sum_{t=1}^{n}(h_t-\mu-\phi(h_{t-1}-\mu))(h_{t-1}-\mu)\right] \\ &= \nu\left[\sum_{t=1}^{n}(h_t-\mu-\phi(h_{t-1}-\mu))(h_{t-1}-\mu)\right],\end{aligned}$$

$$\frac{\partial \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \nu} = \frac{n}{2}\frac{1}{\nu} - \frac{1}{2}\left[\sum_{t=1}^{n}(h_t-\mu-\phi(h_{t-1}-\mu))^2\right].$$

The second order derivatives

$$\frac{\partial^2 \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \alpha \partial \alpha} = -\sum_{t=1}^{n}\frac{1}{\exp(h_t)} = -\sum_{t=1}^{n}\exp(-h_t),$$

$$\frac{\partial^2 \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \alpha \partial \mu} = \frac{\partial^2 \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \alpha \partial \phi} = \frac{\partial^2 \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \alpha \partial \nu} = 0,$$

$$\begin{aligned}\frac{\partial^2 \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \mu \partial \mu} &= \nu\left[-\left(1-\phi^2\right)-(1-\phi)\sum_{t=1}^{n}(1-\phi)\right] \\ &= -\nu\left[n(1-\phi)^2\right],\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \mu \partial \phi} &= \nu\left[-\sum_{t=1}^{n}(h_t-\mu-\phi(h_{t-1}-\mu))-(1-\phi)\sum_{t=1}^{n}(h_{t-1}-\mu)\right] \\ &= -\nu\left[\sum_{t=1}^{n}(h_t-\mu-\phi(h_{t-1}-\mu))+(1-\phi)\sum_{t=1}^{n}(h_{t-1}-\mu)\right],\end{aligned}$$

$$\frac{\partial^2 \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \mu \partial \nu} = (1-\phi)\sum_{t=1}^{n}(h_t-\mu-\phi(h_{t-1}-\mu)),$$

$$\frac{\partial^2 \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \phi \partial \phi} = \nu\left[-\sum_{t=1}^{n}(h_{t-1}-\mu)^2\right],$$

$$\frac{\partial^2 \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \phi \partial \nu} = \sum_{t=1}^{n}(h_t-\mu-\phi(h_{t-1}-\mu))(h_{t-1}-\mu),$$

$$\frac{\partial^2 \ln p(\mathbf{y},\mathbf{h}|\theta)}{\partial \nu \partial \nu} = -\frac{n}{2\nu^2}.$$

The derivatives of the complete-data log-likelihood for M_2

The complete-data log-likelihood function

$$\begin{aligned}\ln p\left(\mathbf{y},\sigma^2|\theta\right) &= \sum_{t=1}^{n}\ln\sigma_t^2-\frac{n}{2}\ln 2\pi+\frac{n}{2}\ln\nu-\frac{1}{2}\nu\left[\sum_{t=1}^{n}\left(\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)^2\right]\\ &\quad -\frac{1}{2}\sum_{t=1}^{n}\frac{\left(y_t-\alpha\right)^2}{\sigma_t^2}-\frac{n}{2}\ln 2\pi-\frac{1}{2}\sum_{t=1}^{n}\ln\sigma_t^2,\end{aligned}$$

where $\boldsymbol{\sigma^2}=\left(\sigma_1^2,\sigma_2^2,...\sigma_n^2\right)'$.

The first order derivatives

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\theta\right)}{\partial\alpha}=\sum_{t=1}^{n}\frac{y_t-\alpha}{\sigma_t^2},$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\theta\right)}{\partial\mu} = \nu\left[\left(1-\phi\right)\sum_{t=1}^{n}\left(\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)\right],$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\theta\right)}{\partial\phi} = \nu\left[\sum_{t=1}^{n}\left(\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)\left(\ln\sigma_{t-1}^2-\mu\right)\right],$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\theta\right)}{\partial\nu}=\frac{n}{2\nu}-\frac{1}{2}\left[\sum_{t=1}^{n}\left(\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)^2\right].$$

The second order derivatives

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\theta\right)}{\partial\alpha\partial\alpha}=-\sum_{t=1}^{n}\frac{1}{\sigma_t^2},$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\theta\right)}{\partial\alpha\partial\mu}=\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\theta\right)}{\partial\alpha\partial\phi}=\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\theta\right)}{\partial\alpha\partial\nu}=0,$$

$$\begin{aligned}\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\theta\right)}{\partial\mu\partial\mu} &= \nu\left[-\left(1-\phi^2\right)-\left(1-\phi\right)\sum_{t=1}^{n}\left(1-\phi\right)\right]\\ &= -\nu\left[n\left(1-\phi\right)^2\right],\end{aligned}$$

$$\begin{aligned}\frac{\partial \ln p\left(\mathbf{y},\boldsymbol{\sigma}^{\mathbf{2}}|\theta\right)}{\partial\mu\partial\phi} &= \nu\left[-\sum_{t=1}^{n}\left(\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)-(1-\phi)\sum_{t=1}^{n}\left(\ln\sigma_{t-1}^2-\mu\right)\right]\\ &= -\nu\left[\sum_{t=1}^{n}\left(\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)+(1-\phi)\sum_{t=1}^{n}\left(\ln\sigma_{t-1}^2-\mu\right)\right],\end{aligned}$$

$$\frac{\partial \ln p\left(\mathbf{y},\boldsymbol{\sigma}^{\mathbf{2}}|\theta\right)}{\partial\mu\partial\nu} = (1-\phi)\sum_{t=1}^{n}\left(\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right),$$

$$\frac{\partial \ln p\left(\mathbf{y},\boldsymbol{\sigma}^{\mathbf{2}}|\theta\right)}{\partial\phi\partial\phi} = \nu\left[-\sum_{t=1}^{n}\left(\ln\sigma_{t-1}^2-\mu\right)^2\right],$$

$$\frac{\partial \ln p\left(\mathbf{y},\boldsymbol{\sigma}^{\mathbf{2}}|\theta\right)}{\partial\phi\partial\nu} = \sum_{t=1}^{n}\left(\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)\left(\ln\sigma_{t-1}^2-\mu\right),$$

$$\frac{\partial^2 \ln p\left(\mathbf{y},\boldsymbol{\sigma}^{\mathbf{2}}|\theta\right)}{\partial\nu\partial\nu} = -\frac{n}{2\nu^2}.$$

Gaussian Approximation

The complete-data log-likelihood function of M_1 can be also expressed as follows:

$$\begin{aligned}\ln\left(p\left(\mathbf{y},\mathbf{h}|\theta\right)\right) &= -\frac{n}{2}\ln\left(2\pi\right)-\frac{n}{2}\ln\left(\tau^2\right)-\frac{1}{2}\left(\mathbf{h}-\mu\right)'Q\left(\mathbf{h}-\mu\right)\\ &\quad -\frac{n}{2}\ln\left(2\pi\right)-\frac{1}{2}\sum_{t=1}^{n}h_t-\sum_{t=1}^{n}\frac{\left(y_t-\alpha\right)^2}{2}\exp\left(-h_t\right),\end{aligned}$$

where $\mathbf{h}=(h_1,h_2,...,h_n)$, $\mu=\mu\mathbf{e}$, $\mathbf{e}'=(1,\ldots,1)_n$, Q is a tri-diagonal precision matrix, $Q=Q^*/\tau^2$, Q^* is defined as follows:

$$Q^*=\begin{pmatrix}\phi^2 & -\phi & & & & & \\ -\phi & 1+\phi^2 & -\phi & & & & \\ & & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & & \\ & & & -\phi & 1+\phi^2 & -\phi & \\ & & & & -\phi & 1 & \end{pmatrix}.$$

The posterior density of $\mathbf{h}$ is

$$\begin{aligned} p(\mathbf{h}|\mathbf{y},\theta) &\propto \exp\left[-\frac{1}{2}(\mathbf{h}-\mu)'Q(\mathbf{h}-\mu)-\sum_{t=1}^{n}\left(\frac{1}{2}h_t+\frac{(y_t-\alpha)^2}{2}\exp(-h_t)\right)\right] \\ &= \exp(f(\mathbf{h}))\approx\exp\left(-\frac{1}{2}\mathbf{h}'c\mathbf{h}+b\mathbf{h}+cons\right). \end{aligned}$$

In order to obtain the parameters c and b of the canonical form, we use the first and second order derivatives:

$$\begin{aligned} \dot{f}(\mathbf{h}) &= -\mathbf{h}'Q+\mu'Q-\frac{1}{2}\mathbf{e}'+\frac{1}{2}\left(\mathbf{y}^{*2}\right)'\odot\exp(-\mathbf{h})' \\ \ddot{f}(\mathbf{h}) &= -Q-diag\left(\frac{1}{2}(\mathbf{y}^{*})^2\odot\exp(-\mathbf{h})\right), \end{aligned}$$

where $\mathbf{y}^{*}=\mathbf{y}-\alpha$ and $\alpha=\alpha\mathbf{e}$, $\mathbf{e}'=(1,\ldots,1)_n$, $\mathbf{y}^{*2}=(y_1^{*2},\ldots,y_n^{*2})'$ and $\exp(-\mathbf{h})=(\exp(-h_1),\ldots,\exp(-h_n))'$.

Denoting the mode of f by $\mathbf{m}$, we apply the Taylor expansion to $f(x)$:

$$\begin{aligned} f(\mathbf{h}) &\approx (\mathbf{h}-\mathbf{m})'\frac{\ddot{f}(\mathbf{m})}{2}(\mathbf{h}-\mathbf{m})+\dot{f}(\mathbf{m})(\mathbf{h}-\mathbf{m})+cons \\ &= -\frac{1}{2}\mathbf{h}'\left(-\ddot{f}(\mathbf{m})\right)\mathbf{h}-\mathbf{m}'\ddot{f}(\mathbf{m})\mathbf{h}+\dot{f}(\mathbf{m})\mathbf{h}+cons \\ &= -\frac{1}{2}\mathbf{h}'\mathbf{c}\mathbf{h}+\mathbf{b}\mathbf{h}+cons. \end{aligned}$$

Now, we obtain $\mathbf{c}$ and $\mathbf{b}$ as

$$\mathbf{c}=-\ddot{f}(\mathbf{m})=\mathbf{Q}+diag\left(\frac{1}{2}\mathbf{y}^{*2}\odot\exp(-\mathbf{m})\right),$$

$$\begin{aligned} \mathbf{b} &= -\mathbf{m}'\ddot{f}(\mathbf{m})+\dot{f}(\mathbf{m}) \\ &= \mathbf{m}'\mathbf{Q}+\mathbf{m}'diag\left(\frac{1}{2}\mathbf{y}^{*2}\odot\exp(-\mathbf{m})\right) \\ &\quad -\mathbf{m}'\mathbf{Q}+\mu'\mathbf{Q}-\frac{1}{2}\mathbf{e}'+\frac{1}{2}\left(\mathbf{y}^{*2}\right)'\odot\exp(-\mathbf{m})' \\ &= \mathbf{m}'diag\left(\frac{1}{2}\mathbf{y}^{*2}\odot\exp(-\mathbf{m})\right)+\frac{1}{2}\left(\mathbf{y}^{*2}\right)'\odot\exp(-\mathbf{m})'+\mu'\mathbf{Q}-\frac{1}{2}\mathbf{e}'. \end{aligned}$$

Using

$$-\frac{1}{2}\mathbf{h}'\mathbf{c}\mathbf{h}+\mathbf{b}\mathbf{h}+cons=-\frac{1}{2}\left(\mathbf{h}-\mathbf{m}^*\right)'\mathbf{Q}^*\left(\mathbf{h}-\mathbf{m}^*\right),$$

we obtain

$$\begin{aligned}\mathbf{Q}^* &= \mathbf{c}=\mathbf{Q}+diag\left(\frac{1}{2}\mathbf{y}^{*2}\odot\exp\left(-\mathbf{m}\right)\right),\\ \mathbf{m}^* &= \mathbf{Q}^{*-1}\mathbf{b}'.\end{aligned}$$

In order to obtain the optimal mode of $\mathbf{Q}^*$ and $\mathbf{m}^*$, we run the above procedure recursively until convergence.

.2 Proofs in Chapter 3

.2.1 Proof of Theorem 3.3.1

It can be shown that

$$\begin{aligned}
& E_{\mathbf{y}}\left\{\int\left[\int \log p(\mathbf{y}|\vartheta)p(\mathbf{y}|\vartheta)\mathrm{d}\mathbf{y}\right]p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\right\} \\
= & \int\int\int \log p(\mathbf{y}|\vartheta)p(\mathbf{y}|\vartheta)\mathrm{d}\mathbf{y}p(\vartheta|\mathbf{y})\mathrm{d}\vartheta p(\mathbf{y})\mathrm{d}\mathbf{y} \\
= & \int\int\int \log p(\mathbf{y}|\vartheta)p(\mathbf{y}|\vartheta)\mathrm{d}\mathbf{y}p(\mathbf{y},\vartheta)\mathrm{d}\mathbf{y}\mathrm{d}\vartheta \\
= & \int\left\{\int\left[\int \log p(\mathbf{y}|\vartheta)p(\mathbf{y}|\vartheta)\mathrm{d}\mathbf{y}\right]p(\mathbf{y}|\vartheta)\mathrm{d}\mathbf{y}\right\}p(\vartheta)\mathrm{d}\vartheta \\
= & \int\left\{\left[\int \log p(\mathbf{y}|\vartheta)p(\mathbf{y}|\vartheta)\mathrm{d}\mathbf{y}\right]\left[\int p(\mathbf{y}|\vartheta)\mathrm{d}\mathbf{y}\right]\right\}p(\vartheta)\mathrm{d}\vartheta \\
= & \int\left[\int \log p(\mathbf{y}|\vartheta)p(\mathbf{y}|\vartheta)\mathrm{d}\mathbf{y}\right]p(\vartheta)\mathrm{d}\vartheta \\
= & \int\left[\int \log p(\mathbf{y}|\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\right]p(\mathbf{y})\mathrm{d}\mathbf{y}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
& E_{\mathbf{y}}\left\{\int\left[\int \log p(\mathbf{y}|\vartheta_0)p(\mathbf{y}|\vartheta)\mathrm{d}\mathbf{y}\right]p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\right\} \\
= & \int\left[\int \log p(\mathbf{y}|\vartheta_0)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\right]p(\mathbf{y})\mathrm{d}\mathbf{y}.
\end{aligned}$$

Thus,

$$\begin{aligned}
& E_{\mathbf{y}}\left[T_{BR}(\mathbf{y},\theta_0)\right] = E_{\mathbf{y}}\left\{\int KL\left[p(\mathbf{y}|\theta),p(\mathbf{y}|\theta_0)\right]p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\right\} \\
= & E_{\mathbf{y}}\left\{\int\left[\int \log p(\mathbf{y}|\vartheta)p(\mathbf{y}|\vartheta)\mathrm{d}\mathbf{y}\right]p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\right\} - E_{\mathbf{y}}\left\{\int\left[\int \log p(\mathbf{y}|\vartheta_0)p(\mathbf{y}|\vartheta)\mathrm{d}\mathbf{y}\right]p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\right\} \\
= & \int\int \log p(\mathbf{y}|\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta p(\mathbf{y})\mathrm{d}\mathbf{y} - \int\int \log p(\mathbf{y}|\vartheta_0)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta p(\mathbf{y})\mathrm{d}\mathbf{y} \\
= & E_{\mathbf{y}}\int\left[\log p(\mathbf{y}|\vartheta) - \log p(\mathbf{y}|\vartheta_0)\right]p(\vartheta|\mathbf{y})\mathrm{d}\vartheta.
\end{aligned}$$

Theorem 4.3.1 is proven.

.2.2 Proof of Theorem 3.3.2

Applying the Taylor expansion on the logarithm of the posterior density, we get

$$\begin{aligned} & \log p(\vartheta|\mathbf{y}) = \log p(\hat{\vartheta}|\mathbf{y}) + L_n^{(1)}(\hat{\vartheta})'(\vartheta-\hat{\vartheta}) + \frac{1}{2}(\vartheta-\hat{\vartheta})' L_n^{(2)}(\tilde{\vartheta})(\vartheta-\hat{\vartheta}) \\ = \;& \log p(\hat{\vartheta}|\mathbf{y}) + \frac{1}{2}(\vartheta-\hat{\vartheta})' L_n^{(2)}(\tilde{\vartheta})(\vartheta-\hat{\vartheta}), \end{aligned}$$

where $\tilde{\vartheta}$ lies on the segment between ϑ and $\hat{\vartheta}$. Note that

$$p(\vartheta|\mathbf{y}) = \frac{p(\mathbf{y},\vartheta)}{p(\mathbf{y})}.$$

Hence,

$$\begin{aligned} & \log p(\vartheta|\mathbf{y}) - \log p(\hat{\vartheta}|\mathbf{y}) = \log p(\mathbf{y},\vartheta) - \log p(\mathbf{y}) - \log p(\mathbf{y},\hat{\vartheta}) + \log p(\mathbf{y}) \\ = \;& \log p(\mathbf{y},\vartheta) - \log p(\mathbf{y},\hat{\vartheta}) = \frac{1}{2}(\vartheta-\hat{\vartheta})' L_n^{(2)}(\tilde{\vartheta})(\vartheta-\hat{\vartheta}). \end{aligned}$$

Then, for any $\varepsilon > 0$, there exists an integer N_2 such that for any $n > N_2$, $L_n^{(2)}(\tilde{\vartheta})$ satisfies

$$\left[I_{p+q} - A(\varepsilon)\right]\left[-L_n^{(2)}(\hat{\vartheta})\right] \leq -L_n^{(2)}(\tilde{\vartheta}) = \left[L_n^{(2)}(\tilde{\vartheta}) L_n^{-(2)}(\hat{\vartheta})\right]\left[-L_n^{(2)}(\hat{\vartheta})\right] \leq \left[I_{p+q} + A(\varepsilon)\right]\left[-L_n^{(2)}(\hat{\vartheta})\right].$$

Following the proof of Theorem 3.2 in Li et al. (2012), under Assumptions 1-7, we note that there exists N, when $n > N$, we have

$$\begin{aligned} & \int (\vartheta-\hat{\vartheta})'\left[-L_n^{(2)}(\tilde{\vartheta})\right](\vartheta-\hat{\vartheta}) p(\vartheta|\mathbf{y})\mathrm{d}\vartheta \\ = \;& \int_{H(\hat{\vartheta},\delta)} [(\vartheta-\hat{\vartheta})'\left[-L_n^{(2)}(\tilde{\vartheta})\right](\vartheta-\hat{\vartheta})] p(\vartheta|\mathbf{y})\mathrm{d}\vartheta \\ = \;& \int_{H(\hat{\vartheta},\delta)} [(\vartheta-\hat{\vartheta})'\left[L_n^{(2)}(\tilde{\vartheta}) L_n^{-(2)}(\hat{\vartheta})\right]\left[-L_n^{(2)}(\hat{\vartheta})\right](\vartheta-\hat{\vartheta}) p(\vartheta|\mathbf{y})\mathrm{d}\vartheta, \end{aligned}$$

which is bounded above by

$$\begin{aligned}&\int_{H(\hat{\vartheta},\delta)}[(\vartheta-\hat{\vartheta})'\left[I_{p+q}+A(\varepsilon)\right]\left[-L_n^{(2)}(\hat{\vartheta})\right](\vartheta-\hat{\vartheta})]p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\\&\quad=\mathbf{tr}\left\{\left[I_{p+q}+A(\varepsilon)\right]\left[-L_n^{(2)}(\hat{\vartheta})\right]V(\hat{\vartheta})\right\},\end{aligned}$$

and below by

$$\begin{aligned}&\int_{H(\hat{\vartheta},\delta)}[(\vartheta-\hat{\vartheta})'\left[I_{p+q}-A(\varepsilon)\right]\left[-L_n^{(2)}(\hat{\vartheta})\right](\vartheta-\hat{\vartheta})]p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\\&\quad=\mathbf{tr}\left\{\left[I_{p+q}-A(\varepsilon)\right]\left[-L_n^{(2)}(\hat{\vartheta})\right]V(\hat{\vartheta})\right\}.\end{aligned}$$

Hence, under the regularity conditions, for $\varepsilon\to 0$, we have

$$\lim_{n\longrightarrow\infty}\int(\vartheta-\hat{\vartheta})'\left[-L_n^{(2)}(\tilde{\vartheta})\right](\vartheta-\hat{\vartheta})p(\vartheta|\mathbf{y})\mathrm{d}\vartheta=\mathbf{tr}\left\{\left[-L_n^{(2)}(\hat{\vartheta})\right]V(\hat{\vartheta})\right\}.$$

Furthermore, it can be shown that

$$\mathbf{tr}\left\{\left[-L_n^{(2)}(\hat{\vartheta})\right]V(\hat{\theta})\right\}=\mathbf{tr}\left\{\left[-L_n^{(2)}(\hat{\vartheta})\right]\left[-L_n^{(2)}(\hat{\vartheta})\right]^{-1}\right\}+o(1)=p+q+o(1).$$

Hence, conditional on the observed data $\mathbf{y}$, we get

$$\begin{aligned}&\int\log p(\mathbf{y}|\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta=\int\left[\log p(\mathbf{y},\vartheta)-\log p(\vartheta)\right]p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\\=\;&\int\log p(\mathbf{y},\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta-\int\log p(\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\\=\;&\int\left[\frac{1}{2}(\vartheta-\hat{\vartheta})'L_n^{(2)}(\tilde{\vartheta})(\vartheta-\hat{\vartheta})\right]p(\vartheta|\mathbf{y})\mathrm{d}\vartheta+\log p(\mathbf{y},\hat{\vartheta})-\int\log p(\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\\=\;&-\frac{1}{2}\mathbf{tr}\left\{\left[-L_n^{(2)}(\hat{\vartheta})\right]V(\hat{\theta})\right\}+o(1)+\log p(\mathbf{y},\hat{\vartheta})-\int\log p(\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\\=\;&\log p(\mathbf{y},\hat{\vartheta})-\int\log p(\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta-\frac{1}{2}\mathbf{tr}\left\{\left[-L_n^{(2)}(\hat{\vartheta})\right]V(\hat{\vartheta})\right\}+o(1)\\=\;&\log p(\mathbf{y},\hat{\vartheta})-\int\log p(\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta-\frac{1}{2}(p+q)+o(1).\end{aligned}$$

Furthermore, it is noted that

$$\log p(\mathbf{y}, \bar{\vartheta}) = \log p(\mathbf{y}, \hat{\vartheta}) + \frac{1}{2}(\bar{\vartheta} - \hat{\vartheta})' L_n^{(2)}(\tilde{\theta})(\bar{\vartheta} - \hat{\vartheta}),$$

where $\tilde{\theta}$ lies on the segment between $\bar{\theta}$ and $\hat{\theta}$. Using Assumption 7, we can show that $\log p(\mathbf{y}, \bar{\vartheta}) = \log p(\mathbf{y}, \hat{\vartheta}) + o_p(1)$.

Similarly, under the null hypothesis, it can be shown that

$$\begin{aligned}
&\log p(\mathbf{y}, \psi|\theta_0) = \log p(\mathbf{y}, \bar{\psi}) + \frac{\log p(\mathbf{y}, \psi|\theta_0)}{\partial \psi}|_{\psi=\bar{\psi}}(\psi - \bar{\psi}) \\
&+\frac{1}{2}(\psi - \bar{\psi})' \left[\frac{\partial^2 \log p(\mathbf{y}, \psi|\theta_0)}{\partial \psi \partial \psi'}|_{\psi=\tilde{\psi}^*} \right] (\psi - \bar{\psi}),
\end{aligned}$$

where $\tilde{\psi}^*$ lies on the segment between ψ and $\bar{\psi}$. When $n \to \infty$, we have $H(\bar{\psi}, \delta_1) \subset H(\hat{\psi}, \delta)$ and $\tilde{\psi}^* \in H(\bar{\psi}, \delta_1) \subset H(\hat{\psi}, \delta)$. Then,

$$\begin{aligned}
&\int (\psi - \bar{\psi})' \left[-L_{0n}^{(2)}(\tilde{\psi}^*) \right] (\psi - \bar{\psi}) p(\vartheta|\mathbf{y}) \mathrm{d}\vartheta \\
= \;& \mathbf{tr} \left\{ \left[-L_{0n}^{(2)}(\tilde{\psi}^*) \right] \left[\int (\psi - \bar{\psi})(\psi - \bar{\psi})' p(\vartheta|\mathbf{y}) \mathrm{d}\vartheta \right] \right\} \\
= \;& \mathbf{tr} \left\{ \left[-L_{0n}^{(2)}(\hat{\psi}) \right] E[(\psi - \bar{\psi})(\psi - \bar{\psi})'|\mathbf{y}, H_1] \right\} + o_p(1).
\end{aligned}$$

Moreover, we get

$$\int \left[\frac{\log p(\mathbf{y}, \psi|\theta_0)}{\partial \psi}|_{\psi=\bar{\psi}} \right] (\psi - \bar{\psi}) p(\vartheta|\mathbf{y}) \mathrm{d}\vartheta = \left[\frac{\log p(\mathbf{y}, \psi|\theta_0)}{\partial \psi}|_{\psi=\bar{\psi}} \right] (\bar{\psi} - \bar{\psi}) = 0.$$

and

$$\begin{aligned}
&\int \log p(\mathbf{y}, \psi|\theta_0) p(\vartheta|\mathbf{y}) \mathrm{d}\vartheta \\
= \;& \log p(\mathbf{y}, \bar{\psi}|\theta_0) - \frac{1}{2} \mathbf{tr} \left\{ \left[-L_{0n}^{(2)}(\hat{\psi}) \right] E[(\psi - \bar{\psi})(\psi - \bar{\psi})'|\mathbf{y}, H_1] \right\} + o_p(1).
\end{aligned}$$

Hence,

$$
\begin{aligned}
& E[(\vartheta-\bar{\vartheta})(\vartheta-\bar{\vartheta})'|\mathbf{y},H_1] \\
= \;& E[(\vartheta-\hat{\vartheta})(\vartheta-\hat{\vartheta})'|\mathbf{y},H_1]+2E[(\vartheta-\hat{\vartheta})|\mathbf{y},H_1](\hat{\vartheta}-\bar{\vartheta})'+(\hat{\vartheta}-\bar{\vartheta})(\hat{\vartheta}-\bar{\vartheta})' \\
= \;& E[(\vartheta-\hat{\vartheta})(\vartheta-\hat{\vartheta})'|\mathbf{y},H_1]+o_p(n^{-1/2})o_p(n^{-1/2}) \\
= \;& E[(\vartheta-\hat{\vartheta})(\vartheta-\hat{\vartheta})'|\mathbf{y},H_1]+o_p(n^{-1}) = -L_n^{-(2)}(\hat{\vartheta})+o_p(n^{-1}) \\
= \;& \frac{1}{n}\left[\frac{1}{n}L_n^{(2)}(\hat{\vartheta})\right]^{-1}+o_p(n^{-1}) = \frac{1}{n}O_p(1)+o_p(n^{-1}) = O_p(n^{-1}),
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathbf{tr}\left\{\left[-L_{0n}^{(2)}(\hat{\psi})\right]E[(\psi-\bar{\psi})(\psi-\bar{\psi})'|\mathbf{y},H_1]\right\} \\
= \;& \mathbf{tr}\left\{\left[-\frac{1}{n}L_{0n}^{(2)}(\hat{\psi})\right]nE[(\psi-\bar{\psi})(\psi-\bar{\psi})'|\mathbf{y},H_1]\right\} \\
= \;& \mathbf{tr}\left\{\left[-\frac{1}{n}L_{0n}^{(2)}(\bar{\psi})+o_p(1)\right]nE[(\psi-\bar{\psi})(\psi-\bar{\psi})'|\mathbf{y},H_1]\right\} \\
= \;& \mathbf{tr}\left\{\left[-L_{0n}^{(2)}(\bar{\psi})\right]E[(\psi-\bar{\psi})(\psi-\bar{\psi})'|\mathbf{y},H_1]\right\}+o_p(1).
\end{aligned}
$$

We can further show that

$$
\begin{aligned}
& T(\mathbf{y},\theta_0) = 2\left[\int \log p(\mathbf{y}|\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta - \int \log p(\mathbf{y}|\vartheta_0)p(\psi|\mathbf{y})\mathrm{d}\psi\right] \\
= \;& 2\log p(\mathbf{y},\bar{\vartheta}) - 2\int \log p(\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta - (p+q) \\
& -2\log p(\mathbf{y},\bar{\psi}|\theta_0)+2\int \log p(\psi)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta+\mathbf{tr}[-L_{0n}^{(2)}(\bar{\psi})V_{22}(\bar{\vartheta})]+o_p(1) \\
= \;& 2[\log p(\mathbf{y},\bar{\vartheta})-\log p(\mathbf{y},\bar{\psi}|\theta_0)]-2\left[\int \log p(\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta - \int \log p(\psi)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\right] \\
& -\left[p+q-\mathbf{tr}[-L_{0n}^{(2)}(\bar{\vartheta}_0)V_{22}(\bar{\vartheta})]\right]+o_p(1) \\
= \;& 2[\log p(\mathbf{y}|\bar{\vartheta})-\log p(\mathbf{y}|\theta_0,\bar{\psi}]+2\left[\log p(\bar{\theta},\bar{\psi})-\log p(\bar{\psi}|\theta_0)\right]-2\left[\int \log p(\theta|\psi)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta\right] \\
& -\left[p+q-\mathbf{tr}[-L_{0n}^{(2)}(\bar{\psi})V_{22}(\bar{\vartheta})]\right]+o_p(1).
\end{aligned}
$$

For latent variable models, $p(\mathbf{y}|\vartheta)$ generally does not have an analytical form.

Using the path sampling technique of Gelman and Meng (1998), we get:

$$p(\mathbf{z}|\mathbf{y},\bar{\vartheta}_b) = \frac{p(\mathbf{z},\mathbf{y}|\bar{\vartheta}_b)}{p(\mathbf{y}|\bar{\vartheta}_b)} = \frac{p(\mathbf{z},\mathbf{y}|\bar{\vartheta}_b)}{f(b)},$$

where $f(b) = p(\mathbf{y}|\bar{\vartheta}_b)$ such that $f(1) = p(\mathbf{y}|\bar{\vartheta})$ and $f(0) = p(\mathbf{y}|\bar{\vartheta}_*)$. Then,

$$\begin{aligned}
& \frac{\partial \log f(b)}{\partial b} = \frac{f^{'}(b)}{f(b)} = \frac{1}{f(b)} \int \frac{\partial p(\mathbf{y},\mathbf{z}|\bar{\vartheta}_b)}{\partial b} \mathrm{d}\mathbf{z} = \frac{1}{f(b)} \int \frac{\partial \log p(\mathbf{y},\mathbf{z}|\bar{\vartheta}_b)}{\partial b} p(\mathbf{y},\mathbf{z}|\bar{\vartheta}_b) \mathrm{d}\mathbf{z} \\
= & \int \frac{\partial \log p(\mathbf{y},\mathbf{z}|\bar{\vartheta}_b)}{\partial b} \frac{p(\mathbf{y},\mathbf{z}|\bar{\vartheta}_b)}{f(b)} \mathrm{d}\mathbf{z} = \int \frac{\partial \log p(\mathbf{y},\mathbf{z}|\bar{\vartheta}_b)}{\partial b} p(\mathbf{z}|\mathbf{y},\bar{\vartheta}_b) \mathrm{d}\mathbf{z} \\
= & E_{\mathbf{z}|\mathbf{y},\bar{\vartheta}_b} \left[\frac{\partial \log p(\mathbf{y},\mathbf{z}|\bar{\vartheta}_b)}{\partial b} \right] = E_{\mathbf{z}|\mathbf{y},\bar{\vartheta}_b} \left[\frac{\partial \bar{\vartheta}_b}{\partial b} \frac{\partial \log p(\mathbf{y},\mathbf{z}|\bar{\vartheta}_b)}{\partial \bar{\vartheta}_b} \right].
\end{aligned}$$

Hence, we get

$$\begin{aligned}
& \log p(\mathbf{y}|\bar{\vartheta}) - \log p(\mathbf{y}|\bar{\vartheta}_*) = \log \frac{f(1)}{f(0)} = \int_0^1 \frac{\partial \log f(b)}{\partial b} \mathrm{d}b \\
= & \int_0^1 \left\{ (\bar{\vartheta} - \bar{\vartheta}_*)' E_{\mathbf{z}|\mathbf{y},\bar{\vartheta}_b} \left[\frac{\partial \log p(\mathbf{y},\mathbf{z}|\vartheta)}{\partial \vartheta} |_{\vartheta=\bar{\vartheta}_b} \right] \right\} \mathrm{d}b \\
= & \int_0^1 \left\{ (\bar{\theta} - \theta_0)' E_{\mathbf{z}|\mathbf{y},\bar{\vartheta}_b} \left[\frac{\partial \log p(\mathbf{y},\mathbf{z}|\vartheta)}{\partial \theta} |_{\theta=\bar{\theta}_b} \right] \right\} \mathrm{d}b \\
& + \int_0^1 \left\{ (\bar{\psi} - \bar{\psi})' E_{\mathbf{z}|\mathbf{y},\bar{\vartheta}_b} \left[\frac{\partial \log p(\mathbf{y},\mathbf{z}|\vartheta)}{\partial \psi} |_{\psi=\bar{\psi}_b} \right] \right\} \mathrm{d}b \\
= & \int_0^1 \left\{ (\bar{\theta} - \theta_0)' E_{\mathbf{z}|\mathbf{y},\bar{\vartheta}_b} [S_1(\mathbf{x}|\bar{\vartheta}_b)] \right\} \mathrm{d}b.
\end{aligned}$$

Theorem 3.3.2 is proven.

.2.3 Proof of Theorem 3.3.3

When $n \to \infty$, the prior information is negligible. Hence, we have

$$\frac{\partial \log p(\mathbf{y}|\theta)}{\partial \theta} = L_n^{(1)}(\theta), \frac{\partial^2 \log p(\mathbf{y}|\theta)}{\partial \theta \partial \theta'} = L_n^{(2)}(\theta),$$

and the ML estimator is asymptotically equivalent to the posterior mode $\hat{\theta}$. Furthermore, according to Theorem 3.2, it can be shown that

$$\begin{aligned} & T(\mathbf{y},\theta_0) = 2\left[\int \log p(\mathbf{y}|\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta - \int \log p(\mathbf{y}|\psi,\theta_0)p(\psi|\mathbf{y})\mathrm{d}\psi\right] \\ = \;& 2\left[\log p(\mathbf{y}|\bar{\vartheta}) - \log p(\mathbf{y}|\theta_0,\bar{\psi})\right] - \left[p+q-\mathbf{tr}[-L_{0n}^{(2)}(\bar{\psi})V_{22}(\bar{\vartheta})]\right] + o_p(1). \end{aligned}$$

In Theorem 3.2, it is shown that $\log p(\mathbf{y}|\bar{\vartheta}) = \log p(\mathbf{y}|\hat{\vartheta}) + o_p(1)$. Similarly, when H_0 is true, let $\bar{\vartheta}_* = (\theta_0,\bar{\psi})$, we can show that

$$\begin{aligned} & \log p(\mathbf{y}|\theta_0,\bar{\psi}) = \log p(\mathbf{y}|\hat{\theta},\hat{\psi}) + L_n^{(1)}(\hat{\vartheta})(\bar{\vartheta}_* - \hat{\vartheta}) + \frac{1}{2}(\bar{\vartheta}_* - \hat{\vartheta})' L_n^{(2)}(\hat{\vartheta})(\bar{\vartheta}_* - \hat{\vartheta}) + o_p(1) \\ = \;& \log p(\mathbf{y}|\hat{\theta},\hat{\psi}) + \frac{1}{2}(\bar{\vartheta}_* - \hat{\vartheta})' L_n^{(2)}(\hat{\vartheta})(\bar{\vartheta}_* - \hat{\vartheta}) + o_p(1). \end{aligned}$$

Furthermore, under the null hypothesis, it is noted that $\bar{\psi} = \hat{\psi} + o_p(n^{-\frac{1}{2}})$, $\frac{1}{n}L_n^{(2)}(\hat{\vartheta}) = O_p(1)$ and

$$\begin{aligned} & -\frac{1}{n}L_n^{(2)}(\hat{\vartheta}) = -\frac{1}{n}L_n^{(2)}(\vartheta_0) + o_p(1) = \mathbf{J}(\vartheta_0) + o_p(1), \\ & \left[-\frac{1}{n}L_n^{(2)}(\hat{\vartheta}_0)\right]^{-1} = -\left[\frac{1}{n}L_n^{(2)}(\vartheta_0)\right]^{-1} + o_p(1) = \mathbf{J}^{-1}(\vartheta_0) + o_p(1) = \mathbf{IJ}(\vartheta_0) + o_p(1). \end{aligned}$$

Thus, we have

$$\begin{aligned} & (\bar{\vartheta}_* - \hat{\vartheta})' L_n^{(2)}(\hat{\vartheta})(\bar{\vartheta}_* - \hat{\vartheta}) \\ = \;& (\theta_0 - \hat{\theta})' L_{n,11}^{(2)}(\hat{\vartheta})(\theta_0 - \hat{\theta}) + 2(\theta_0 - \hat{\theta})' L_{n,12}^{(2)}(\hat{\vartheta})(\bar{\psi} - \hat{\psi}) + (\bar{\psi} - \hat{\psi})' L_{n,22}^{(2)}(\hat{\vartheta})(\bar{\psi} - \hat{\psi}) \\ = \;& (\theta_0 - \hat{\theta})' L_{n,11}^{(2)}(\hat{\vartheta})(\theta_0 - \hat{\theta}) + 2O_p(n^{-1/2})O_p(n)o_p(n^{-1/2}) + o_p(n^{-1/2})O_p(n)o_p(n^{-1/2}) \\ = \;& (\theta_0 - \hat{\theta})' L_{n,11}^{(2)}(\hat{\vartheta})(\theta_0 - \hat{\theta}) + o_p(1) \\ = \;& \sqrt{n}(\theta_0 - \hat{\theta})' \left[\frac{1}{n}L_{n,11}^{(2)}(\hat{\vartheta})\right] \sqrt{n}(\theta_0 - \hat{\theta}) + o_p(1) \\ = \;& -\sqrt{n}(\theta_0 - \hat{\theta})' \left[\mathbf{J}_{11}(\vartheta_0) + o_p(1)\right] \sqrt{n}(\theta_0 - \hat{\theta}) + o_p(1) \\ = \;& -\sqrt{n}(\theta_0 - \hat{\theta})' \left[\mathbf{J}_{11}(\vartheta_0)\right] \sqrt{n}(\theta_0 - \hat{\theta}) + o_p(1)O_p(1)O_p(1) + o_p(1) \\ = \;& -\sqrt{n}(\theta_0 - \hat{\theta})' \left[\mathbf{J}_{11}(\vartheta_0)\right] \sqrt{n}(\theta_0 - \hat{\theta}) + o_p(1). \end{aligned}$$

According to the ML theory, we know that $\sqrt{n}(\hat{\theta}-\theta_0) \sim N[\mathbf{0}, \mathbf{I}\mathbf{J}_{11}(\vartheta_0)]$ so that $\varepsilon = \sqrt{n}\mathbf{I}\mathbf{J}_{11}^{-1/2}(\vartheta_0)(\hat{\theta}-\theta_0) \sim N[\mathbf{0}, \mathbf{I}_q]$. Hence, we have

$$
\begin{aligned}
& 2\left[\log p(\mathbf{y}|\hat{\theta},\hat{\psi}) - \log p(\mathbf{y}|\theta_0,\bar{\psi})\right] = (\bar{\vartheta}_* - \hat{\vartheta})' \left[-L_n^{(2)}(\hat{\vartheta})\right](\bar{\vartheta}_* - \hat{\vartheta}) + o_p(1) \\
= \;& \sqrt{n}(\theta_0 - \hat{\theta})'[\mathbf{J}_{11}(\vartheta_0)]\sqrt{n}(\theta_0 - \hat{\theta}) + o_p(1) \\
= \;& \sqrt{n}(\theta_0 - \hat{\theta})'\mathbf{I}\mathbf{J}_{11}^{-1/2}(\vartheta_0)\left[\mathbf{I}\mathbf{J}_{11}^{1/2}(\vartheta_0)\mathbf{J}_{11}(\vartheta_0)\mathbf{I}\mathbf{J}_{11}^{1/2}(\vartheta_0)\right]\mathbf{I}\mathbf{J}_{11}^{-1/2}\sqrt{n}(\theta_0 - \hat{\theta}) + o_p(1) \\
= \;& \varepsilon'\left[\mathbf{I}\mathbf{J}_{11}^{1/2}(\vartheta_0)\mathbf{J}_{11}(\vartheta_0)\mathbf{I}\mathbf{J}_{11}^{1/2}(\vartheta_0)\right]\varepsilon + o_p(1).
\end{aligned}
$$

Further, when the null hypothesis is true, we can get that

$$
\begin{aligned}
& T(\mathbf{y},\theta_0) + \left[p+q-\mathbf{tr}[-L_{0n}^{(2)}(\bar{\vartheta})V_{22}(\bar{\vartheta})]\right] \\
= \;& 2\left[\int \log p(\mathbf{y}|\vartheta)p(\vartheta|\mathbf{y})\mathrm{d}\vartheta - \int \log p(\mathbf{y}|\psi,\theta_0)p(\psi|\mathbf{y})\mathrm{d}\psi\right] + \left[p+q-\mathbf{tr}[-L_{0n}^{(2)}(\bar{\vartheta})V_{22}(\bar{\vartheta})]\right] \\
= \;& T_1(\mathbf{y},\theta_0) + \left[p+q-\mathbf{tr}[-L_{0n}^{(2)}(\bar{\vartheta})V_{22}(\bar{\vartheta})]\right] + o_p(1) \\
= \;& 2[\log p(\mathbf{y}|\bar{\vartheta}) - \log p(\mathbf{y}|\theta_0,\bar{\psi})] + o_p(1) \\
= \;& 2\left[\log p(\mathbf{y}|\hat{\theta},\hat{\psi}) - \log p(\mathbf{y}|\theta_0,\bar{\psi})\right] + o_p(1) \\
\sim \;& \varepsilon'\left[\mathbf{I}\mathbf{J}_{11}^{1/2}(\vartheta_0)\mathbf{J}_{11}(\vartheta_0)\mathbf{I}\mathbf{J}_{11}^{1/2}(\vartheta_0)\right]\varepsilon.
\end{aligned}
$$

.3 Proofs in Chapter 4

.3.1 Proof of Theorem 4.3.1

When the likelihood information dominates the prior information, we have

$$\frac{1}{n}L_n^{(2)}(\theta)=\frac{1}{n}\frac{\partial^2\log p(\mathbf{y}|\theta)}{\partial\theta\partial\theta'}+\frac{1}{n}\frac{\partial^2\log p(\theta)}{\partial\theta\partial\theta'}=\frac{1}{n}\frac{\partial^2\log p(\mathbf{y}|\theta)}{\partial\theta\partial\theta'}+o_p(1)=\hat{\mathbf{I}}(\theta)+o_p(1),$$

and the posterior mode θ is equivalent to the ML estimator. When the model is specified correctly, using the regularity conditions $\mathbf{J}(\theta_0)=O(1)$, we get

$$\begin{aligned}
&\hat{\mathbf{J}}(\theta_0)=\frac{1}{n}\sum_{i=1}^{n}s(\mathbf{y}_i,\theta_0)s'(\mathbf{y}_i,\theta_0)\\
=&\int\left[\frac{1}{n}\sum_{i=1}^{n}s(\mathbf{y}_i,\theta_0)s'(\mathbf{y}_i,\theta_0)\right]p(\mathbf{y}|\theta_0)\mathrm{d}\mathbf{y}+o_p(1)=\mathbf{J}(\theta_0)+o_p(1)\\
=&O(1)+o_p(1)=O_p(1).
\end{aligned}$$

From the standard ML theory, we get $\theta_0=\hat{\theta}+O_p(n^{-1/2})$ and $\bar{\theta}=\hat{\theta}+o_p(n^{-1/2})$ so that $\bar{\theta}=\theta_0+O_p(n^{-1/2})=\theta_0+o_p(1)$. Then, we have $\hat{\mathbf{J}}(\theta_0)=\hat{\mathbf{J}}(\bar{\theta})+o_p(1)$.

We can further show that

$$\begin{aligned}
&\mathbf{BT}=n\int(\theta-\bar{\theta})'\hat{\mathbf{J}}(\bar{\theta})(\theta-\bar{\theta})p(\theta|\mathbf{y})\mathrm{d}\theta\\
=&n\int(\theta-\bar{\theta})'\left[\hat{\mathbf{J}}(\theta_0)+o_p(1)\right](\theta-\bar{\theta})p(\theta|\mathbf{y})\mathrm{d}\theta\\
=&n\int(\theta-\bar{\theta})'\hat{\mathbf{J}}(\theta_0)(\theta-\bar{\theta})p(\theta|\mathbf{y})\mathrm{d}\theta+n\int(\theta-\bar{\theta})'o_p(1)(\theta-\bar{\theta})p(\theta|\mathbf{y})\mathrm{d}\theta\\
=&n\int(\theta-\bar{\theta})'\hat{\mathbf{J}}(\theta_0)(\theta-\bar{\theta})p(\theta|\mathbf{y})\mathrm{d}\theta+o_p(1)\\
=&n\int(\theta-\bar{\theta})'\left[\mathbf{J}(\theta_0)+o_p(1)\right](\theta-\bar{\theta})p(\theta|\mathbf{y})\mathrm{d}\theta+o_p(1)\\
=&n\int(\theta-\bar{\theta})'\mathbf{J}(\theta_0)(\theta-\bar{\theta})p(\theta|\mathbf{y})\mathrm{d}\theta+n\int(\theta-\bar{\theta})'o_p(1)(\theta-\bar{\theta})p(\theta|\mathbf{y})\mathrm{d}\theta+o_p(1)\\
=&n\int(\theta-\bar{\theta})'\mathbf{J}(\theta_0)(\theta-\bar{\theta})p(\theta|\mathbf{y})\mathrm{d}\theta+o_p(1)\\
=&n\mathbf{tr}\left\{\mathbf{J}(\theta_0)E\left[(\theta-\bar{\theta})(\theta-\bar{\theta})'|\mathbf{y}\right]\right\}.
\end{aligned}$$

Using the regularity condition, we have

$$\hat{\mathbf{I}}(\theta_0) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{h}(\mathbf{y}_t, \theta_0) \xrightarrow{p} \int \left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{h}(\mathbf{y}_t, \theta_0)\right] p(\mathbf{y}|\theta_0)\mathrm{d}\mathbf{y} = \mathbf{I}(\theta_0) = O(1).$$

Then, we get $\hat{\mathbf{I}}(\theta_0) = \mathbf{I}(\theta_0) + o_p(1) = O_p(1) + o_p(1) = O_p(1)$, $\hat{\mathbf{I}}(\theta_0) = \hat{\mathbf{I}}(\hat{\theta}) + o_p(1)$ and

$$\begin{aligned}
& E\left[(\theta-\bar{\theta})(\theta-\bar{\theta})'|\mathbf{y}\right] = E\left[(\theta-\hat{\theta}+\hat{\theta}-\bar{\theta})(\theta-\hat{\theta}+\hat{\theta}-\bar{\theta})'|\mathbf{y}\right] \\
= \;& E\left[(\theta-\hat{\theta})(\theta-\hat{\theta})'|\mathbf{y}\right] + 2E\left[(\theta-\hat{\theta})|\mathbf{y}\right](\hat{\theta}-\bar{\theta}) + (\hat{\theta}-\bar{\theta})(\hat{\theta}-\bar{\theta})' \\
= \;& E\left[(\theta-\hat{\theta})(\theta-\hat{\theta})'|\mathbf{y}\right] + 2(\bar{\theta}-\hat{\theta})(\hat{\theta}-\bar{\theta}) + (\hat{\theta}-\bar{\theta})(\hat{\theta}-\bar{\theta})' \\
= \;& E\left[(\theta-\hat{\theta})(\theta-\hat{\theta})'|\mathbf{y}\right] - (\hat{\theta}-\bar{\theta})(\hat{\theta}-\bar{\theta})' \\
= \;& -L_n^{-(2)}(\hat{\theta}) + o_p(n^{-1}) + o_p(n^{-\frac{1}{2}})o_p(n^{-\frac{1}{2}}) \\
= \;& -\frac{1}{n}\left[\hat{\mathbf{I}}(\theta_0) + o_p(1)\right]^{-1} + o_p(n^{-1}) \\
= \;& -\frac{1}{n}\hat{\mathbf{I}}^{-1}(\theta_0) + o_p(1)\frac{1}{n} + o_p(n^{-1}) \\
= \;& -\frac{1}{n}\left[\mathbf{I}(\theta_0) + o_p(1)\right]^{-1} + o_p(n^{-1}) \\
= \;& -\frac{1}{n}\mathbf{I}^{-1}(\theta_0) + o_p(1)\frac{1}{n} + o_p(n^{-1}) \\
= \;& -\frac{1}{n}\mathbf{I}^{-1}(\theta_0) + o_p(n^{-1}).
\end{aligned}$$

When H_0 is true, we have $\mathbf{J}(\theta_0) = -\mathbf{I}(\theta_0)$. Therefore, we get

$$\begin{aligned}
& \mathbf{BT} = n\mathbf{tr}\left\{\mathbf{J}(\theta_0)E\left[(\theta-\bar{\theta})(\theta-\bar{\theta})'|\mathbf{y}\right]\right\} \\
= \;& -n\mathbf{tr}\left\{\mathbf{J}(\theta_0)\left[\frac{1}{n}\mathbf{I}^{-1}(\theta_0) + o_p(n^{-1})\right]\right\} \\
= \;& -\mathbf{tr}\left\{\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right\} + n\mathbf{tr}\left\{\mathbf{J}(\theta_0)o_p(n^{-1})\right\} \\
= \;& -\mathbf{tr}\left\{\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right\} + n\mathbf{tr}\left\{O_p(1)o_p(n^{-1})\right\} \\
= \;& \mathbf{tr}\left[-\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o_p(1) = p + o_p(1).
\end{aligned}$$

Theorem 4.3.1 is proven.

.3.2 The derivation of BT for the asset pricing models

Let $\mathbf{R}=\{\mathbf{R}_1,\mathbf{R}_2,\cdots,\mathbf{R}_n\}$, $\mathbf{F}=\{\mathbf{F}_1,\mathbf{F}_2,\cdots,\mathbf{F}_n\}$, $\omega=\{\omega_1,\omega_2,\cdots,\omega_n\}$, $\theta=(\alpha,\beta,\Sigma)$. The observed data log-likelihood function, $\mathscr{L}_o(\mathbf{R}|\theta)$, is:

$$\mathscr{L}_o(\mathbf{R}|\theta)=C(\nu)-\frac{n}{2}\ln|\Sigma|-\frac{\nu+N}{2}\sum_{t=1}^{n}\log\left[\nu+\varphi\left(R_t,F_t,\theta\right)\right], \tag{.3.1}$$

where

$$C(\nu)=-\frac{nN}{2}\log(\pi\nu)+n\left[\ln\Gamma\left(\frac{\nu+N}{2}\right)-\ln\Gamma\left(\frac{\nu}{2}\right)\right]+\frac{n(\nu+N)\ln\nu}{2},$$
$$\varphi\left(R_t,F_t,\theta\right)=\left(R_t-\alpha-\beta F_t\right)^{'}\Sigma^{-1}\left(R_t-\alpha-\beta F_t\right).$$

It is noted in Li et al. (2012), using the normal-gamma mixture representation for the multivariate t distribution, the complete log-likelihood, $\mathscr{L}_c(\mathbf{R},\omega|\theta)$, is given by

$$-\frac{1}{2}nN\ln(2\pi)+\frac{N}{2}\sum_{t=1}^{n}\ln\omega_t-\frac{n}{2}\ln|\Sigma|-\frac{1}{2}\sum_{t=1}^{n}\omega_t\varphi\left(R_t,F_t,\theta\right)$$
$$-n\ln\Gamma\left(\frac{\nu}{2}\right)+\frac{n\nu}{2}\ln\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\sum_{t=1}^{n}(\ln\omega_t-\omega_t)-\sum_{t=1}^{n}\ln\omega_t.$$

Hence, for any θ and θ^*, we have

$$\begin{aligned}
&\mathscr{Q}(\theta|\theta^*)=\int L_c(\mathbf{R},\omega|\theta)p(\omega|\mathbf{R},\theta^*)\mathrm{d}\omega\\
=\;&-\frac{1}{2}nK\ln(2\pi)+\frac{N}{2}\sum_{t=1}^{n}E(\ln\omega_t|\mathbf{R}_t,\theta^*)-\frac{n}{2}\ln|\Sigma|-\frac{1}{2}\sum_{t=1}^{n}E(\omega_t|\mathbf{R}_t,\theta^*)\varphi(R_t,F_t,\theta)\\
&-n\ln\Gamma\left(\frac{\nu}{2}\right)+\frac{n\nu}{2}\ln\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\sum_{t=1}^{n}E(\ln\omega_t-\omega_t|\mathbf{R}_t,\theta^*)-\sum_{t=1}^{n}E(\ln\omega_t|\mathbf{R}_t,\theta^*).
\end{aligned}$$

For $i = 1, 2, \cdots, N$, letting $\phi_i = \Sigma_{ii}^{-1}$, we have

$$\begin{aligned}
\frac{\mathscr{Q}(\theta|\theta^*)}{\partial \alpha_i}|_{\theta^*=\theta} &= \sum_{t=1}^{n} E\left(\omega_t|\theta, \mathbf{R}_t\right) \phi_i (R_{it} - \alpha_i - \beta_i^{'} \mathbf{F}_t), \\
\frac{\mathscr{Q}(\theta|\theta^*)}{\partial \beta_i}|_{\theta^*=\theta} &= \sum_{t=1}^{n} E\left(\omega_t|\theta, \mathbf{R}_t\right) \phi_i (R_{it} - \alpha_i - \beta_i^{'} \mathbf{F}_t) \mathbf{F}_t, \\
\frac{\mathscr{Q}(\theta|\theta^*)}{\partial \phi_i}|_{\theta^*=\theta} &= \sum_{t=1}^{n} \left[\frac{1}{\phi_i} - \frac{1}{2} E\left(\omega_t|\theta, \mathbf{R}_t\right) (R_{it} - \alpha_i - \beta_i^{'} \mathbf{F}_t)^2 \right],
\end{aligned}$$

where

$$E\left(\omega_t|\theta, \mathbf{R}_t\right) = \frac{\nu + N}{\nu + \varphi\left(R_t, F_t, \theta\right)}.$$

Since the data are conditionally independent on $\mathbf{F}$, one can simply delete the summation in the above $\mathscr{Q}$ function to get the first derivative for each observation.

.3.3 The derivation of BT for the DSGE model

The equilibrium object for a DSGE model is a collection of the nonlinear equations defining optimality conditions, markets clearing conditions, et cetera. We follow the standard practice and linearize these conditions around a steady state. Then, the model can be written as linear expectation system,

$$\Gamma_0(\theta) x_t = \Gamma_1(\theta) E_t\left[x_{t+1}\right] + \Gamma_2(\theta) x_{t-1} + \Gamma_3(\theta) \varepsilon_t, \tag{.3.2}$$

where x_t are the state variables, ε_t the exogenous shocks, θ the structural parameters of interest, and $\{\Gamma_1\}$ matrix functions that map the equilibrium conditions of the model, where $\Gamma_0(\theta), \Gamma_1(\theta), \Gamma_2(\theta)$ are $n_s \times n_s$, $\Gamma_3(\theta)$ is $n_s \times n_e$. The solution to the system takes the form of a $VAR(1)$,

$$x_t = T(\theta) x_{t-1} + R(\theta) \varepsilon_t, \tag{.3.3}$$

The mapping from θ to T and R must be solved numerically for all models of interest, where T is $n_s \times n_s$, R is $n_s \times n_e$. The model variables x_t are linked to observed y_t via a state space system:

$$
\begin{aligned}
x_t &= T(\theta)\, x_{t-1} + R(\theta)\, \varepsilon_t, \\
y_t &= D(\theta) + Z(\theta)\, x_t + \xi_t,
\end{aligned}
$$

where y_t, D are $n_y \times 1$, Z is $n_y \times n_s$, θ is $n_q \times 1$.

Consider the state space system

$$
\begin{aligned}
x_t &= T x_{t-1} + R\varepsilon_t, \\
y_t &= D + Z x_t + \xi_t,
\end{aligned}
$$

where $\varepsilon_t \sim N(0, Q)$, $\xi_t \sim N(0, H)$.

Let $Y_s = (y_1, y_2 ..., y_s)$, we can define

$$
\begin{aligned}
x_t^s &= E(x_t | Y_s), \\
P_t^s &= E\{(x_t - x_t^s)(x_t - x_t^s)' | Y_s\}.
\end{aligned}
$$

Then for the linear Gaussian state-space model specified in above equation, with initial condition x_0^0 and P_0^0, for $t = 1, 2 ... n$, the Kalman Filter algorithm is as follows (Shumway and Stoffer (2006)):

$$
\begin{aligned}
x_t^{t-1} &= T x_{t-1}^{t-1}, \\
P_t^{t-1} &= T P_{t-1}^{t-1} T' + RQR',
\end{aligned}
$$

with

$$
\begin{aligned}
x_t^t &= x_t^{t-1} + K_t \left(y_t - D - Z x_t^{t-1}\right), \\
P_t^t &= [I_{n_s} - K_t Z] P_t^{t-1},
\end{aligned}
$$

where

$$K_t = P_t^{t-1} Z' \left[Z P_t^{t-1} Z' + H \right]^{-1}.$$

From the Kalman Filter, the likelihood of the data is as follows:

$$\begin{aligned} \log \ell &= -\sum_{t=1}^{n} \left[\frac{n_y}{2} \log 2\pi + \frac{1}{2} \log |F_t| + \frac{1}{2} \left(y_t - D - Z x_t^{t-1} \right)' F_t^{-1} \left(y_t - D - Z x_t^{t-1} \right) \right] \\ &= -\sum_{t=1}^{n} \left[\frac{n_y}{2} \log 2\pi + \frac{1}{2} \log |F_t| + \frac{1}{2} \omega_t' F_t^{-1} \omega_t \right], \end{aligned}$$

where

$$\begin{aligned} F_t &= Z P_t^{t-1} Z' + H, \\ \omega_t &= y_t - D - Z(\theta) x_t^{t-1}. \end{aligned}$$

Before we compute the derivatives of the model, we will first introduce some notations from Magnus and Neudecker (1999) about the matrix derivative.

Definition .3.1 *Let $F = (f_{st})$ be an $m \times p$ matrix function of an $n \times q$ matrix of variables $X = \left(x_{ij}\right)$. Any $mp \times nq$ matrix A containing all the partial derivatives such that each row contains the partial derivatives of one function with respect to all variables, and each column contains the partial derivatives of all functions with respect to one variable x_{ij}, is called a derivative of F. We define the α-derivative as:*

$$DF(X) = \frac{\partial vecF(X)}{\partial (vecX)'}.$$

In our case, $\partial (vec\theta)' = \partial \theta'$ since θ is a vector.

Definition .3.2 *Let A be an $m \times n$ matrix. There exists a unique $mn \times mn$ permutation matrix K_{mn} which is defined as:*

$$K_{mn} \cdot vecA = vec\left(A'\right).$$

Since K_{mn} is a permutation matrix it is orthogonal, $K_{mn}^{-1} = K_{mn}^{'}$.

In order to compute the first order derivative of the likelihood, we have the following

$$\frac{\partial vec(\omega_t)}{\partial \theta'} = -\frac{\partial vec(D)}{\partial \theta'} - \left(x_t^{t-1\prime} \otimes I_{n_y}\right) \frac{\partial vec(Z)}{\partial \theta'} - (I_1 \otimes Z) \frac{\partial vec\left(x_t^{t-1}\right)}{\partial \theta'},$$

$$\begin{aligned} \frac{\partial vec(F_t)}{\partial \theta'} &= \left(\left(P_t^{t-1}Z'\right)' \otimes I_{n_y} + \left(I_{n_y} \otimes \left(ZP_t^{t-1}\right)\right) K_{n_y n_s}\right) \frac{\partial vec(Z)}{\partial \theta'} \\ &\quad + (Z \otimes Z) \frac{\partial vec\left(P_t^{t-1}\right)}{\partial \theta'} + \frac{\partial vecH}{\partial \theta'}, \end{aligned}$$

$$\begin{aligned} \frac{\partial vec\left(F_t^{-1}\right)}{\partial \theta'} &= -\left(\left(F_t^{-1}\right)' \otimes F_t^{-1}\right) \frac{\partial vec(F_t)}{\partial \theta'}, \\ \frac{\partial vec(\log|F_t|)}{\partial \theta'} &= \left(vec\left[\left(F_t^{-1}\right)'\right]\right)' \frac{\partial vec(F_t)}{\partial \theta'}, \end{aligned}$$

$$\begin{aligned} \frac{\partial vec\left(\omega_t' F_t^{-1} \omega_t\right)}{\partial \theta'} &= \left[\left(F_t^{-1}\omega_t\right)' \otimes I_1\right] K_{n_y 1} \frac{\partial vec(\omega_t)}{\partial \theta'} + \left(\omega_t' \otimes \omega_t'\right) \frac{\partial vec\left(F_t^{-1}\right)}{\partial \theta'} \\ &\quad + \left[I_1 \otimes \left(\omega_t' F_t^{-1}\right)\right] \frac{\partial vec(\omega_t)}{\partial \theta'}. \end{aligned}$$

In the above equations, the first order derivatives of the matrix D, Z, Q, H are easy to get, and according to Iskrev (2008) and Herbst (2010), we can get the first order derivatives of matrix T and R, substitute (.3.3) into (.3.2), we have

$$\Gamma_0(\theta) x_t = \Gamma_1(\theta) T(\theta) x_t + \Gamma_2(\theta) x_{t-1} + \Gamma_3(\theta) \varepsilon_t.$$

Furthermore

$$(\Gamma_0(\theta) - \Gamma_1(\theta) T(\theta)) x_t = \Gamma_2(\theta) x_{t-1} + \Gamma_3(\theta) \varepsilon_t. \tag{.3.4}$$

From (.3.4)

$$(\Gamma_0(\theta)-\Gamma_1(\theta)T(\theta))x_t=(\Gamma_0(\theta)-\Gamma_1(\theta)T(\theta))T(\theta)x_{t-1}+(\Gamma_0(\theta)-\Gamma_1(\theta)T(\theta))R(\theta)\varepsilon_t. \tag{.3.5}$$

Comparing (.3.4) and (.3.5), we have

$$(\Gamma_0(\theta)-\Gamma_1(\theta)T(\theta))T(\theta)-\Gamma_2(\theta) = 0. \tag{.3.6}$$

$$(\Gamma_0(\theta)-\Gamma_1(\theta)T(\theta))R(\theta)-\Gamma_3(\theta) = 0. \tag{.3.7}$$

Consider the Eq.(.3.6), we can get the derivatives of matrix T by solving the following equation

$$\begin{aligned}\left[(I_{n_s}\otimes\Gamma_0)-(I_{n_s}\otimes\Gamma_1 T)-\left(T'\otimes\Gamma_1\right)\right]\frac{\partial vec(T)}{\partial\theta'} &- \left(T'^2\otimes I_{n_s}\right)\frac{\partial vec(\Gamma_1)}{\partial\theta'}\\ &+ \left(T'\otimes I_{n_s}\right)\frac{\partial vec(\Gamma_0)}{\partial\theta'}-\frac{\partial vec(\Gamma_2)}{\partial\theta'}=0.\end{aligned}$$

From (.3.7), the first order derivatives of matrix R is as follows:

$$\frac{\partial vec(R)}{\partial\theta'}=-\left(\Gamma_3'\otimes I_{n_s}\right)\left(W^{'-1}\otimes W^{-1}\right)\frac{\partial vec(W)}{\partial\theta'}+\left(I_{n_e}\otimes W^{-1}\right)\frac{\partial vec(\Gamma_3)}{\partial\theta'}.$$

From Herbst (2010), where

$$W=\Gamma_0-\Gamma_1 T,$$

$$\frac{\partial vec(W)}{\partial\theta'}=\frac{\partial vec(\Gamma_0)}{\partial\theta'}-\left(T'\otimes I_{n_s}\right)\frac{\partial vec(\Gamma_1)}{\partial\theta'}-(I_{n_s}\otimes\Gamma_1)\frac{\partial vec(T)}{\partial\theta'}.$$

Given the initial conditions P_0^0 and x_0^0, we have the following recursive equations

$$\frac{\partial vec\left(x_t^{t-1}\right)}{\partial\theta'}=(I_1\otimes T)\frac{\partial vec\left(x_{t-1}^{t-1}\right)}{\partial\theta'}+\left(x_{t-1}^{t-1\prime}\otimes I_{n_s}\right)\frac{\partial vec(T)}{\partial\theta'},$$

$$\begin{aligned}\frac{\partial vec\left(P_t^{t-1}\right)}{\partial\theta'} &= \left(\left(P_{t-1}^{t-1}T'\right)'\otimes I_{n_s}\right)\frac{\partial vec(T)}{\partial\theta'}+(T\otimes T)\frac{\partial vec\left(P_{t-1}^{t-1}\right)}{\partial\theta'}\\ &\quad+\left(I_{n_s}\otimes TP_{t-1}^{t-1}\right)K_{n_sn_s}\frac{\partial vec(T)}{\partial\theta'}+\frac{\partial vec(RQR')}{\partial\theta'},\end{aligned}$$

$$\begin{aligned}\frac{\partial vec\left(x_t^t\right)}{\partial \theta^{*\prime}} &= \frac{\partial vec\left(x_t^{t-1}\right)}{\partial \theta^{*\prime}} + \left[\left(y_t - D - Zx_t^{t-1}\right)' \otimes I_{n_s}\right] \frac{\partial vec\left(K_t\right)}{\partial \theta^{*\prime}} \\ &\quad - \left(I_1 \otimes K_t\right) \frac{\partial vec\left(D\right)}{\partial \theta^{*\prime}} - \left(x_t^{t-1\prime} \otimes K_t\right) \frac{\partial vec\left(Z\right)}{\partial \theta^{*\prime}} - \left(I_1 \otimes K_t Z\right) \frac{\partial vec\left(x_t^{t-1}\right)}{\partial \theta^{*\prime}},\end{aligned}$$

$$\begin{aligned}\frac{\partial vec\left(P_t^t\right)}{\partial \theta^{*\prime}} &= -\left(\left(ZP_t^{t-1}\right)' \otimes I_{n_s}\right) \frac{\partial vec\left(K_t\right)}{\partial \theta^{*\prime}} - \left(P_t^{t-1\prime} \otimes K_t\right) \frac{\partial vec\left(Z\right)}{\partial \theta^{*\prime}} \\ &\quad + \left(I_{n_s} \otimes \left(I_{n_s} - K_t Z\right)\right) \frac{\partial vec\left(P_t^{t-1}\right)}{\partial \theta^{*\prime}},\end{aligned}$$

where

$$\begin{aligned}\frac{\partial vec\left(K_t\right)}{\partial \theta^{*\prime}} &= \left[\left(Z'F_t^{-1}\right)' \otimes I_{n_s}\right] \frac{\partial vec\left(P_t^{t-1}\right)}{\partial \theta^{*\prime}} + \left[\left(F_t^{-1}\right)' \otimes P_t^{t-1}\right] K_{n_y n_s} \frac{\partial vec\left(Z\right)}{\partial \theta^{*\prime}} \\ &\quad + \left[I_{n_y} \otimes P_t^{t-1} Z'\right] \frac{\partial vec\left(F_t^{-1}\right)}{\partial \theta^{*\prime}},\end{aligned}$$

and

$$\frac{\partial vec\left(RQR'\right)}{\partial \theta'} = \left[\left(RQ' \otimes I_{n_s}\right) + \left(I_{n_s} \otimes RQ\right) K_{n_s n_e}\right] \frac{\partial vecR}{\partial \theta'} + \left(R \otimes R\right) \frac{\partial vecQ}{\partial \theta'}.$$

The initial condition is given as

$$\begin{aligned}x_0^0 &= 0, \\ P_0^0 &= TP_0^0 T' + RQR'.\end{aligned}$$

From the above, we have

$$vec\left(P_0^0\right) = \left(I_{n_s^2} - T \otimes T\right)^{-1} vec\left(RQR'\right).$$

Then

$$\frac{\partial vec\left(P_0^0\right)}{\partial \theta'} = \left[\left(TP_0^0 \otimes I_{n_s}\right) + \left(I_{n_s} \otimes TP_0^0\right) K_{n_s n_s}\right] \frac{\partial vec\left(T\right)}{\partial \theta'} + \left(T \otimes T\right) \frac{\partial vec\left(P_0^0\right)}{\partial \theta'} + \frac{\partial vec\left(RQR'\right)}{\partial \theta'}.$$