Chapter 4

Hypothesis testing, specification testing, and model selection based on the MCMC output using R[☆]

Yong Li[a], Jun Yu[b] and Tao Zeng[c,*]

[a]*Hanqing Advanced Institute of Economics and Finance, Renmin University of China, Beijing, China*

[b]*School of Economics and Lee Kong Chian School of Business, Singapore Management University, Singapore, Singapore*

[c]*School of Economics, Academy of Financial Research, and Institute for Fiscal Big-Data & Policy of Zhejiang University, Zhejiang University, Zhejiang, China*

[*]*Corresponding author: e-mail: ztzt6512@gmail.com*

Abstract

This chapter overviews several MCMC-based test statistics for hypothesis testing and specification testing and MCMC-based model selection criteria developed in recent years. The statistics for hypothesis testing can be viewed as the MCMC version of the "trinity" of test statistics based in maximum likelihood (ML), namely, the likelihood ratio (LR) test, the Lagrange multiplier (LM) test, and the Wald test. The model selection criteria correspond to two predictive distributions. One of them can be viewed as the MCMC version of widely used information criterion, AIC. The asymptotic distributions of the test statistics and model selection criteria are discussed. The test statistics and model selection criteria are applied to several popular models using real data, one of which involves latent variables. The implementation is illustrated in R with the MCMC output obtained by R2WinBUGS.

JEL classification: C11, C12

Keywords: AIC, DIC, Information matrix, LR test, LM test, Markov chain Monte Carlo, Latent variable, Wald test

[☆]Li gratefully acknowledges the financial support of the Chinese Natural Science Fund (No. 71773130). Yu would like to acknowledge the financial support from Singapore Ministry of Education Academic Research Fund Tier 3 under the grant number MOE2013-T3-1-017. R code that implement our methods can be found at http://www.mysmu.edu/faculty/yujun/Handbook_Rcode.zip.

Handbook of Statistics, Vol. 41. https://doi.org/10.1016/bs.host.2018.12.003

1 Introduction

In economics and finance, statistical models with increasing complexity have been used more and more often. Typically empirical analysis of statistical models involves calculating and maximizing the log-likelihood function, leading to the maximum likelihood (ML) estimator. The ML estimator (MLE) has desirable asymptotic properties of consistency, normality, and efficiency under broad conditions, facilitating hypothesis testing, specification testing, and model selection. The asymptotic normality and efficiency of MLE make the well-known trinity of tests in ML popular in practice, i.e., the likelihood ratio (LR) test, the Wald test, and the Lagrange Multiplier (LM) test. In addition, some specification tests, such as the information matrix based tests, are based on MLE. Furthermore, some widely used information criteria for model selection, such as AIC, BIC, and HQ, are based on MLE.

Unfortunately, many statistical models face with a great deal of difficulties empirically in the sense that they cannot be easily estimated by ML. Examples include but not are restricted to latent variable models, continuous time models, models with complicated parameter restrictions, models in which the log-likelihood is not available in closed-form or is unbounded, models in which parameters are not point identified, high dimensional models for which numerical optimization is difficult to use, models with multiple local optimum in the log-likelihood function.

While for some of these models, alternative estimation methods, such as GMM, can be used. These alternative methods are generally less efficient than ML. With rapidly enhanced power in computing technology, the MCMC method has been used more and more frequently to provide the full likelihood analysis of models. MCMC is typically regarded as a Bayesian approach as it samples from the posterior distribution and the posterior mean is often chosen to be the Bayesian parameter estimate.

After the MCMC output is obtained, a few questions naturally arise. The first question is how to conduct hypothesis testing as one typically does after MLE is used to estimate a model. The second question is how to perform the specification test of the estimated model. The third question is how to compare alternative models that are not necessarily nested by each other. Hypothesis testing, specification testing and model selection are of fundamental importance in empirical studies. Therefore, MCMC-based answers to these questions become critically in practice. The traditional Bayesian answer to these questions is to use the gold standard, the Bayes factors (BFs), or it variants. The BFs basically compare the posterior model probabilities of candidate models, conditional on the data. Despite its appeal in the statistical interpretation, BFs suffer a few serious theoretical and computational difficulties. For example, it is not well-defined under improper priors. It subjects to Jeffreys-Lindley's paradox, that is, it tends to reject the null hypothesis even when the null is correct. For many models, BFs are difficult to compute.

The aim of this chapter is to overview the literature on MCMC-based statistical inference. However, we focus on test statistics and model selection criteria which can be justified in a frequentist set up, in the same way as how the ML-based methods are justified. Since MCMC was introduced initially as a Bayesian tool, it is not immediately obvious how to make statistical inference based on the MCMC output in the frequentist framework. The essence of the literature is to treat MCMC as a sampling method and resort to the frequentist framework to obtain the asymptotic theory of various statistics based on the MCMC output in repeated sampling.

The statistics for hypothesis testing developed in the literature can be viewed as the MCMC version of the "trinity" of the tests in ML. The statistics for specification testing can be viewed as the MCMC version of the information matrix based test. One of the model selection criteria can be viewed as the MCMC version of AIC. Their asymptotic properties of these statistics are reviewed. The methods are illustrated using some important models widely used in economics and finance in a real data setting. The implementation is illustrated in R with the MCMC output obtained by R2WinBUGS.

MCMC can be used to sample from distributions other than the posterior. In a seminar paper, Chernozhukov and Hong (2003) proposed to use MCMC to sample from quasi-posterior. Moreover, the MCMC output may be used for other types of statistical inference. One example is to construct the confidence sets for identified sets of parameters in econometric models defined through a likelihood or a vector of moments; see Chen et al. (2016). Review of these studies are beyond of the scope of this chapter.

The chapter is organized as follows. Section 2 reviews the MCMC technique and introduces the implementation of MCMC using the R package. We also briefly explain the inferencial approach typically adopted in the Bayesian literature. Section 3 overviews several statistics for hypothesis testing based on the MCMC output. Section 4 overviews the MCMC-based test statistics for specification. Section 5 reviews DIC, an MCMC version of AIC, and other related information criteria. Section 6 gives the empirical illustrations. Section 7 concludes the chapter. R code that implement our methods can be found at http://www.mysmu.edu/faculty/yujun/Handbook_Rcode.zip.

2 MCMC and its implementation in R

Without loss of generality, we take the latent variable models as an example, to explain why ML is difficult to use and to describe how to obtain the MCMC output. Let $\mathbf{y}=(y_1, \ldots, y_n)$ denote the data generated from a probability measure P_0 on the probability space $(\Omega, \mathrm{F}, P_0)$. Let $\mathbf{z}=(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)'$ be the latent variables. The latent variable model is indexed by the some P-dimensional parameter vector, $\boldsymbol{\theta}$. Furthermore, $p(\mathbf{y}|\boldsymbol{\theta})$ is used to denote the observed-data likelihood function, and $p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$ is denoted as the complete-data likelihood function. The relationship between these two likelihood functions is given by

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z}. \tag{1}$$

In many latent variable modes, especially dynamic latent variable models, the latent variable $\mathbf{z}$ is often dependent on the sample size and its dimension is the same as or larger than the number of the sample size. When the sample size is large, the integral is high-dimensional. Often the integral does not have a closed-form solution and cannot be reduced into lower dimension integrals. In this case, it will be very difficult to accurately approximate the integral numerically. Consequently, ML is difficult to implement.

Now, we review the basic idea of MCMC. Let $p(\boldsymbol{\theta})$ be prior distribution assigned for parameter $\boldsymbol{\theta}$. Since the observed likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ is intractable, it is very difficult to draw the random observations from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ directly. To deal with this difficulty, the data-augmentation strategy (Tanner and Wong, 1987) can be applied to augment the parameter space from $\boldsymbol{\theta}$ to $(\boldsymbol{\theta}, \mathbf{z})$. As a result, the likelihood function becomes $p(\mathbf{y}|\boldsymbol{\theta},\mathbf{z})$ which typically is available in closed-form. The MCMC technique, such as Gibbs sampler, draws random samples from the joint posterior distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$. More concretely, we start with an initial value $[\boldsymbol{\theta}^{(0)}, \mathbf{z}^{(0)}]$, and then at the jth iteration, conditional on the current values $[\boldsymbol{\theta}^{(j)}, \mathbf{z}^{(j)}]$,

(a) generate $\boldsymbol{\theta}^{(j+1)}$ from $p(\boldsymbol{\theta}|\mathbf{z}^{(j)}, \mathbf{y})$;
(b) generate $\mathbf{z}^{(j+1)}$ from $p(\mathbf{z}|\boldsymbol{\theta}^{(j+1)}, \mathbf{z})$.

To get rid of the effect of the initial value, some random observations are discarded as the burn-in observations. After that, the simulated random samples can be regarded as efficient random draws (though correlated in general) from the joint posterior distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$. These correlated random samples are the MCMC output.

Based on the MCMC output, the parameter estimate can be obtained. For example, Bayesian estimates of $\boldsymbol{\theta}$ can be easily obtained as the sample mean of the generated random samples. Specifically, let $\{\boldsymbol{\theta}^{(j)}, j=1, 2, \ldots, J\}$ be effective random observations generated form the joint posterior distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$. Then Bayesian estimates of $\boldsymbol{\theta}$ is

$$\overline{\theta}=\frac{1}{J}\sum_{i=1}^{J}\theta^{(j)}.$$

This estimate is justified when the loss function is quadratic.

Under some regularity conditions, it is well documented in the literature (see, for example, Gelman et al., 2013) that the posterior distribution has a limiting normal distribution given by

$$\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}|\mathbf{y}\overset{a}{\sim}N\left(0,\ \left[-\frac{1}{n}\frac{\partial^2\ \ln p\left(\hat{\boldsymbol{\theta}}|\ \mathbf{y}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]^{-1}\right), \tag{2}$$

where $\hat{\boldsymbol{\theta}}$ is the posterior mode $\left(\text{i.e., } \hat{\boldsymbol{\theta}} = \arg\max \ \ln p(\boldsymbol{\theta}|\,\mathbf{y})\right)$ and

$$p(\boldsymbol{\theta}|\,\mathbf{y}) = \frac{p(\mathbf{y}|\,\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{y}|\,\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}.$$

Furthermore, under extra regularity conditions, when $p(\boldsymbol{\theta}) = O_p(1)$, Li et al. (2017a) showed that the relationship between the posterior mean $\overline{\boldsymbol{\theta}}$ and the posterior mode $\hat{\boldsymbol{\theta}}$ can be expressed as

$$\overline{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + O_p\left(n^{-1}\right), \tag{3}$$

$$\widehat{Var}(\boldsymbol{\theta}|\,\mathbf{y}) = \left[-\frac{\partial^2 \ln p\left(\mathbf{y}|\,\hat{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]^{-1} + O_p\left(n^{-2}\right). \tag{4}$$

The large sample properties in (2), (3) and (4) provide the fountainhead from which all the methods reviewed in this chapter springs.

In practice, however, MCMC procedures are not easy to implement using nonconventional software that is not widely available among researchers and practitioners. Therefore, it is practically important to find efficient software packages which can free the researchers from tedious programming and debugging. For this purpose, under the R language environment, Sturtz et al. (2005) introduced a so-called R2WinBUGS package combined with a free software WinBUGS1.4 to obtain the MCMC output. R is an extremely powerful language and environment for statistical computation and graphics which is available free of charge. WinBUGS is a user-friendly software package that implements the Gibbs sampler. It does sampling-based posterior computations for a variety of statistical models such as random effects, generalized linear, proportional hazards, latent variable, and frailty models. The latest version of WinBUGS is Win-BUGS1.4 which was developed by the medical Research Council Biostatistics Unit and the department of Epidemiology and Public Health of the Imperial College School of Medicine at St Mary's Hospital. It is available free of charge at http://www.mrc-bsu.cam.ac.uk/bugs/ An introduction to this software can be found in Spiegelhalter et al. (2003).

In this chapter, using the R language, we implement R2WinBUGS to get the MCMC outputs and then use R to compute the test statistics and the information criteria discussed below. The R code can be downloaded online where the detailed explanation for R commands is provided line by line in the R scripts by us. For more details about R2WinBUGS and WinBUGS1.4, one can refer to Sturtz et al. (2005) and Spiegelhalter et al. (2003). Special tailored R packages to obtain the MCMC output to fit particular statistical models are also available. For example, the R package named MCMC-Pack was developed by Martin and Quinn (2005). Our R code to compute the test statistics and the information criteria discussed below may be also applied to the MCMC output generated by MCMCPack.

3 Hypothesis testing based on the MCMC output

3.1 Hypothesis testing under decision theory

Assume that a statistical model $M \equiv \{p(\mathbf{y}|\boldsymbol{\theta})\}$ is used to fit the data. The P-dimensional parameter vector $\boldsymbol{\theta}$ can be divided into two parts $\boldsymbol{\theta} = (\vartheta', \boldsymbol{\psi}')'$ where $\boldsymbol{\vartheta} \in \boldsymbol{\Theta}$ denote a vector of p-dimensional parameter of interest and $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ a vector of q-dimensional nuisance parameter. We are interested in knowing whether or not $\boldsymbol{\vartheta}$ is equal to some value to verify a particular theory. Hence, the point null hypothesis problem can be written as

$$\begin{cases} H_0 : \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0 \\ H_1 : \boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}_0 \end{cases}. \tag{5}$$

In this section, we discuss the hypothesis testing problem from a decision viewpoint.

Consider a decision problem whose decision space has two statistical decisions, to accept H_0 (name it d_0) or to reject H_0 (name it d_1). We may specify a loss function denoted by $\{\mathcal{L}[d_i, (\boldsymbol{\theta}, \boldsymbol{\psi})], i = 0, 1\}$ to measure the consequence of the statistical decision d_i. Let $p(\boldsymbol{\vartheta}, \boldsymbol{\psi}|\mathbf{y})$ be the posterior distribution with some given prior $p(\boldsymbol{\vartheta}, \boldsymbol{\psi})$, and $\mathbf{T}(\mathbf{y}, \boldsymbol{\vartheta}_0)$ be a test statistic for hypothesis testing which is a function of the data $\mathbf{y}$. When the expected posterior loss of accepting H_0 is sufficiently larger than the expected posterior loss of rejecting H_0, i.e.,

$$\mathbf{T}(\mathbf{y}\boldsymbol{\vartheta}_0) = \int_{\Theta}\int_{\Psi} \{\mathcal{L}[d_0(\boldsymbol{\vartheta}\boldsymbol{\psi})] - \mathcal{L}[d_1(\boldsymbol{\vartheta}\boldsymbol{\psi})]\} p(\boldsymbol{\vartheta}, \boldsymbol{\psi}|\,\mathbf{y}) \mathrm{d}\boldsymbol{\vartheta} \mathrm{d}\boldsymbol{\psi} > c,$$

we can say that the statistical decision of accepting H_0 might be inappropriate with some confidence so that the statistical decision to reject H_0 can be done naturally. For more details about hypothesis testing under decision theory, one can refer to Bernardo and Rueda (2002) and Bernardo and Smith (2006).

In practice, it is enough to specify the net loss function denoted by $\Delta\mathcal{L}[H_0, (\boldsymbol{\vartheta}, \boldsymbol{\psi})] = \mathcal{L}[d_0, (\boldsymbol{\vartheta}, \boldsymbol{\psi})] - \mathcal{L}[d_1, (\boldsymbol{\vartheta}, \boldsymbol{\psi})]$. Hence, the test statistic can be rewritten as

$$\mathbf{T}(\mathbf{y}, \boldsymbol{\vartheta}_0) = \int_{\Theta}\int_{\Psi} \Delta\mathcal{L}[H_0, (\boldsymbol{\vartheta}, \boldsymbol{\psi})] p(\boldsymbol{\vartheta}, \boldsymbol{\psi}|\,\mathbf{y}) \mathrm{d}\boldsymbol{\vartheta} \mathrm{d}\boldsymbol{\psi} = E_{\boldsymbol{\theta}|\mathbf{y}}(\Delta\mathcal{L}[H_0, (\boldsymbol{\vartheta}, \boldsymbol{\psi})]).$$

3.2 The choice of loss function for hypothesis testing

In the subsection, we review the loss functions for the purpose of constructing hypothesis test statistics. We show that the BFs correspond to the discrete loss function that takes values of 0 and 1. To overcome the shortcomings of BFs, alterative continuous loss functions have been proposed in the literature to construct new test statistics based on the MCMC output. There is a more fundamental difference between these new test statistics and the BFs. The new test statistics are justified in a frequentist setup, that is, by assuming that **y** comes out of the data generating process in a repeated experiment whereas BFs is justified in a Bayesian setup, that is, the decision is made conditional on **y**.

3.2.1 BFs and 0–1 loss function

If the 0–1 loss function is used, that is,

$$\mathcal{L}[d_0,(\boldsymbol{\vartheta},\boldsymbol{\psi})]=\begin{cases}0 & \text{if } \boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0\\ 1 & \text{if } \boldsymbol{\vartheta}\neq\boldsymbol{\vartheta}_0\end{cases},\quad \mathcal{L}[d_1,(\boldsymbol{\vartheta},\boldsymbol{\psi})]=\begin{cases}1 & \text{if } \boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0\\ 0 & \text{if } \boldsymbol{\vartheta}\neq\boldsymbol{\vartheta}_0\end{cases},$$

the net loss function $\Delta\mathcal{L}[H_0,(\boldsymbol{\vartheta},\boldsymbol{\psi})]$ is given by

$$\Delta\mathcal{L}[H_0,(\boldsymbol{\vartheta},\boldsymbol{\psi})]=\begin{cases}-1 & \text{if } \boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0\\ 1 & \text{if } \boldsymbol{\vartheta}\neq\boldsymbol{\vartheta}_0\end{cases}.$$

Hence, the test statistic based on this discrete loss function is given by

$$\begin{aligned}\mathbf{T}(\mathbf{y},\boldsymbol{\vartheta}_0)&=\int_\Theta\int_\Psi \Delta\mathcal{L}[H_0,(\boldsymbol{\vartheta},\boldsymbol{\psi})]p(\boldsymbol{\vartheta},\boldsymbol{\psi}|\,\mathbf{y})\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi}\\ &=\int_\Theta\int_\Psi \Delta\mathcal{L}[H_0,(\boldsymbol{\vartheta},\boldsymbol{\psi})]\frac{p(\mathbf{y}|\,\boldsymbol{\vartheta},\boldsymbol{\psi})p(\boldsymbol{\vartheta},\boldsymbol{\psi})}{p(\mathbf{y})}\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi},\end{aligned}$$

where $p(\mathbf{y})=\int_\Theta\int_\Psi p(\mathbf{y}\,|\,\boldsymbol{\vartheta},\boldsymbol{\psi})p(\boldsymbol{\vartheta},\boldsymbol{\psi})\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi}$ is the marginal likelihood.

In general, a positive probability w is assigned to the event $\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0$, such that a reasonable prior for $\boldsymbol{\vartheta}$ with a discrete support at $\boldsymbol{\vartheta}_0$ can be given by

$$p(\boldsymbol{\vartheta})=\begin{cases}w & \text{if}\,\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0\\ (1-w)\pi(\boldsymbol{\vartheta}) & \text{if}\,\boldsymbol{\vartheta}\neq\boldsymbol{\vartheta}_0\end{cases}.$$

where $\pi(\boldsymbol{\vartheta})$ is a prior distribution. Hence, the test statistic under this discrete prior distribution can be expressed as

$$\begin{aligned}\mathbf{T}(\mathbf{y},\boldsymbol{\vartheta}_0)\ &=\int_\Theta\int_\Psi \Delta\mathcal{L}[H_0,(\boldsymbol{\vartheta},\boldsymbol{\psi})]\frac{p(\mathbf{y}|\,\boldsymbol{\vartheta},\boldsymbol{\psi})p(\boldsymbol{\vartheta},\boldsymbol{\psi})}{p(\mathbf{y})}\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi}\\ &=-\int_\Psi\frac{p(\mathbf{y}|\,\boldsymbol{\vartheta}_0,\boldsymbol{\psi})p(\boldsymbol{\vartheta}_0,\boldsymbol{\psi})}{p(\mathbf{y})}\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi}+\int_\Theta\int_\Psi\frac{p(\mathbf{y}|\,\boldsymbol{\vartheta},\boldsymbol{\psi})p(\boldsymbol{\vartheta},\boldsymbol{\psi})}{p(\mathbf{y})}\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi}\\ &=-\int_\Psi\frac{p(\mathbf{y}|\,\boldsymbol{\vartheta}_0,\boldsymbol{\psi})p(\boldsymbol{\psi}|\,\boldsymbol{\vartheta}_0)p(\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0)}{p(\mathbf{y})}\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi}+\int_\Theta\int_\Psi\frac{p(\mathbf{y}|\,\boldsymbol{\vartheta},\boldsymbol{\psi})p(\boldsymbol{\psi}|\,\boldsymbol{\vartheta})p(\boldsymbol{\vartheta})}{p(\mathbf{y})}\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi}\\ &=-\int_\Psi\frac{p(\mathbf{y}|\,\boldsymbol{\vartheta}_0,\boldsymbol{\psi})p(\boldsymbol{\psi}|\,\boldsymbol{\vartheta}_0)w}{p(\mathbf{y})}\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi}+\int_\Theta\int_\Psi\frac{p(\mathbf{y}|\,\boldsymbol{\vartheta},\boldsymbol{\psi})p(\boldsymbol{\psi}|\,\boldsymbol{\vartheta})(1-w)\pi(\boldsymbol{\vartheta})}{p(\mathbf{y})}\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi},\end{aligned}$$

where $p(\boldsymbol{\psi}\,|\,\boldsymbol{\vartheta})$ is the conditional prior distribution.

From this formula, we can see that the decision criterion can be made as

$$\text{Reject } H_0 \text{iff} \quad \int_{\Psi} p(\mathbf{y}|\,\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0,\boldsymbol{\psi})\omega p(\boldsymbol{\psi}|\,\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0)\mathrm{d}\boldsymbol{\psi} < \int_{\Theta}\int_{\Psi} p(\mathbf{y}|\,\boldsymbol{\vartheta},\boldsymbol{\psi})p(\boldsymbol{\psi}|\,\boldsymbol{\vartheta})(1-w)p(\boldsymbol{\vartheta})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\psi}$$

To represent the prior ignorance, in practice, the probability w is set to 1/2 and the criterion becomes:

$$\text{Reject } H_0 \text{iff} B_{01} = \frac{\int_{\Psi} p(\mathbf{y}|\,\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0,\boldsymbol{\psi})p(\boldsymbol{\psi}|\,\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0)\mathrm{d}\boldsymbol{\psi}}{\int_{\Theta}\int_{\Psi} p(\mathbf{y}|\,\boldsymbol{\vartheta},\boldsymbol{\psi})p(\boldsymbol{\psi}|\,\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi}} = \frac{m_0}{m_1} < 1,$$

where $\{m_k,\ k=0,\ 1\}$ are marginal likelihoods. B_{01} is the well-known BF defined as the ratio of the marginal likelihoods (Kass and Raftery, 1995).

Although BF is intuitively appealing and has a strong probabilistic interpretation, it is known to suffer from some theoretical and computational difficulties. First, when a subjective prior $\pi(\boldsymbol{\vartheta})$ is not available, Jeffreys' prior or reference prior (Bernardo and Smith, 2006; Jeffreys, 1961) are often used to reflect the lack of prior information. Jeffreys' prior and reference prior are generally improper. It follows that $\pi(\boldsymbol{\vartheta})=C\,f\,(\boldsymbol{\vartheta})$, where $f\,(\boldsymbol{\vartheta})$ is a nonintegrable function, and C is an arbitrary positive constant. In this case, the BF can be expressed as

$$B_{01} = \frac{1}{C}\frac{\int_{\Psi} p(\mathbf{y}|\,\boldsymbol{\psi},\boldsymbol{\vartheta}_0)p(\boldsymbol{\psi}|\,\boldsymbol{\vartheta}_0)\mathrm{d}\boldsymbol{\psi}}{\int_{\Theta}\int_{\Psi} p(\mathbf{y}|\,\boldsymbol{\vartheta},\boldsymbol{\psi})p(\boldsymbol{\psi}|\,\boldsymbol{\vartheta})f(\boldsymbol{\vartheta})\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi}}.$$

Clearly, the BF is ill-defined since it depends on the arbitrary constant, C.

Second, to address the ill-defined problem of BF under the improper prior, a proper prior $\pi(\boldsymbol{\vartheta})$ with a large variance (that is a vague prior) has been proposed to represent the prior ignorance. While in this case the BF is well-defined, it has a tendency to favor the null hypothesis even when the null hypothesis is correct, giving rise to the notorious Jeffreys-Lindley's paradox; see Poirier (1995), Robert (1993, 2001). Jeffreys-Lindley's paradox leads to researchers to find variations to the BF. Examples include *partial Bayes factor* (O'Hagan, 1991), the *intrinsic Bayes factor* (Berger and Perrichi, 1996), and the *fractional Bayes factor* (O'Hagan, 1995). These variants basically split the data $\mathbf{y}$ into a training sample and a testing sample. The training sample is used to update an uninformative prior to obtain an informative prior. Unfortunately, they suffer from more or less arbitrary choices of training samples, weights for averaging training samples, and fractions, respectively.

Last but not least, for the latent variable model and many other models, calculation of the marginal likelihood M_k, $k=0$, 1 often involves intractable high-dimensional integrals, and, as a result, BFs are generally very difficult to calculate; see Han and Carlin (2001) for an excellent review of methods for calculating the BFs from the MCMC output.

3.2.2 Bernardo and Rueda (2002) and the KL loss function

Bernardo and Rueda (2002, BR hereafter) pointed out that if $\boldsymbol{\vartheta}$ is a continuous parameter, hypothesis testing forces the use of a nonregular (not absolutely continuous) "sharp" prior concentrating a positive probability mass so that the null hypothesis H_0 must have a strictly positive prior probability. This nonregular prior structure leads to the theoretical difficulties of BFs. To overcome these difficulties, Bernardo and Rueda (2002) suggested using a continuous loss function based on the Kullback–Leibler0 (KL) divergence to replace the discrete loss function, i.e.,

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

where $p(x)$ and $q(x)$ are any two regular probability density functions. Then, the corresponding hypothesis test statistic can be given by:

$$\mathbf{T}_{BR}(\mathbf{y}, \boldsymbol{\vartheta}_0) = E_{\boldsymbol{\theta}|\mathbf{y}}(\min\{KL[p(\mathbf{y}|\,\boldsymbol{\vartheta}, \boldsymbol{\psi}), p(\mathbf{y}|\,\boldsymbol{\vartheta}_0, \boldsymbol{\psi})], KL[p(\mathbf{y}|\,\boldsymbol{\vartheta}_0, \boldsymbol{\psi}), p(\mathbf{y}|\,\boldsymbol{\vartheta}, \boldsymbol{\psi})]\}).$$

While $\mathbf{T}_{BR}$ $(\mathbf{y}, \boldsymbol{\vartheta}_0)$ is well-defined under improper priors, since the KL divergence function often does not have a closed-form expression, $\mathbf{T}_{BR}$ $(\mathbf{y}, \boldsymbol{\vartheta}_0)$ is difficult to compute for the latent variable model. Moreover, BR suggested choosing threshold values based on the normal distribution to implement the test. The rationale for basing threshold values on the normal distribution conceivably comes from the fact that many test statistics are asymptotically normally distributed. Therefore, BR's approach is not Bayesian as the sampling distribution of the test statistic is used and it is based on the idea of repeated sampling, not conditional on $\mathbf{y}$.

3.2.3 Li and Yu (2012) and the $\mathcal{Q}$ loss function

To address the computational problem in $\mathbf{T}_{BR}$ $(\mathbf{y}, \boldsymbol{\theta}_0)$, Li and Yu (2012, LY hereafter) proposed a loss function based on the $\mathcal{Q}$function used in the EM algorithm (Dempster et al., 1977) to replace the KL divergence function. For any two points such as $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ defined in the parameter space, the $\mathcal{Q}$ function can be expressed as

$$\mathcal{Q}(\boldsymbol{\theta}_1|\,\boldsymbol{\theta}_2) = E_{\mathbf{Z}|\mathbf{y},\boldsymbol{\theta}_2}[\,\ln p(\mathbf{y}, \mathbf{z}|\,\boldsymbol{\theta}_1)].$$

Compared with the observed data likelihood function $p(\mathbf{y}\,|\,\boldsymbol{\theta})$, the $\mathcal{Q}$ function is easier to evaluate for the latent variable model. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\vartheta}_0, \boldsymbol{\psi})$, Li and Yu (2012) defined a new continuous net loss function as:

$$\Delta\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\theta}_0) = \{\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}) - \mathcal{Q}(\boldsymbol{\theta}_0,\boldsymbol{\theta})\} + \{\mathcal{Q}(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0) - \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\},$$

and proposed a MCMC-based test statistic as:

$$\mathrm{T}_{LY}(\mathbf{y},\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}|\mathbf{y}}[\Delta\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\vartheta}_0)].$$

While T_{LY} $(\mathbf{y}, \boldsymbol{\theta}_0)$ is well-defined under improper priors and easy to compute for the latent variable model, one still needs to specify some threshold values. Again, threshold values lack of rigorous statistical justifications. Importantly, the need to specify some threshold values suggests that LY's approach is not Bayesian.

3.2.4 Li et al. (2014) and LR-type loss function

To address the problem in choosing threshold values, Li et al. (2014, LZY hereafter) introduced another net continuous loss function based on the deviance function (Spiegelhalter et al., 2002) given by

$$\Delta\mathcal{L}[H_0,(\boldsymbol{\vartheta},\boldsymbol{\psi})] = 2\ \ln p(\mathbf{y}|\,\boldsymbol{\vartheta},\boldsymbol{\psi}) - 2\ \ln p(\mathbf{y}|\,\boldsymbol{\vartheta}_0,\boldsymbol{\psi}).$$

The corresponding test statistic is

$$\mathbf{T}_{LZY}(\mathbf{y},\boldsymbol{\vartheta}_0) = 2\int[\ \ln p(\mathbf{y}|\,\boldsymbol{\vartheta},\boldsymbol{\psi}) -\ \ln p(\mathbf{y}|\,\boldsymbol{\vartheta}_0,\boldsymbol{\psi})]p(\boldsymbol{\vartheta},\boldsymbol{\psi}|\,\mathbf{y})\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\psi}. \tag{6}$$

Since the likelihood function $p(\mathbf{y}\,|\,\boldsymbol{\vartheta}, \boldsymbol{\psi})$ is often intractable for the latent variable model, to achieve computational tractability, under some regularity conditions, Li et al. (2014) developed an asymptotically equivalent form for $\mathbf{T}_{LZY}$ $(\mathbf{y}, \boldsymbol{\vartheta}_0)$, i.e.,

$$\mathbf{T}^{*}_{LZY}(\mathbf{y},\boldsymbol{\vartheta}_0) = 2D + 2\left[\ \ln p\left(\overline{\boldsymbol{\vartheta}},\overline{\boldsymbol{\psi}}\right) -\ \ln p(\overline{\boldsymbol{\psi}}|\,\boldsymbol{\vartheta}_0)\right] - 2\int\ \ln p(\boldsymbol{\vartheta}|\,\boldsymbol{\psi})p(\boldsymbol{\theta}|\,\mathbf{y})\mathrm{d}\boldsymbol{\theta} \\ - \left[p + q - \mathbf{tr}\left[-L^{(2)}_{0n}(\overline{\boldsymbol{\psi}})V_{22}\left(\overline{\boldsymbol{\theta}}\right)\right]\right], \tag{7}$$

where $\overline{\boldsymbol{\theta}} = \left(\overline{\boldsymbol{\vartheta}},\overline{\boldsymbol{\psi}}\right)'$ is the posterior mean of $\boldsymbol{\theta}$ under H_1, and

$$D = \int_0^1\left\{\left(\overline{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\right)'\left[E_{\mathbf{z}|\mathbf{y},\overline{\boldsymbol{\theta}}_b}\left(S_1\left(\mathbf{y},\mathbf{z}|\,\overline{\boldsymbol{\theta}}_b\right)\right)\right]\right\}\mathrm{d}b,$$

with $\overline{\boldsymbol{\theta}}_b = (1-b)\overline{\boldsymbol{\theta}}_* + b\overline{\boldsymbol{\theta}},$ for $b \in [0,\ 1]$, $\overline{\boldsymbol{\theta}}_* = (\boldsymbol{\vartheta}_0,\overline{\boldsymbol{\psi}})'$, $S(\mathbf{y},\ \mathbf{z}\,|\,\boldsymbol{\theta}) = \partial$ ln $p(\mathbf{y},\ \mathbf{z}\,|\,\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, $S_1(\cdot)$ being the subvector of $S(\mathbf{y},\ \mathbf{z}\,|\,\boldsymbol{\theta})$ corresponding to $\boldsymbol{\vartheta}$, $V_{22}\left(\overline{\boldsymbol{\theta}}\right) = E\left[(\boldsymbol{\psi}-\overline{\boldsymbol{\psi}})(\boldsymbol{\psi}-\overline{\boldsymbol{\psi}})'\,|\,\mathbf{y},H_1\right]$, the submatrix of $V\left(\overline{\boldsymbol{\theta}}\right)$ corresponding to $\boldsymbol{\psi}$, and $L^{(2)}_{0n}(\boldsymbol{\psi}) = \partial^2$ ln $p(\mathbf{y},\ \boldsymbol{\psi}\,|\,\boldsymbol{\vartheta}_0)/\partial\boldsymbol{\psi}\,\partial\boldsymbol{\psi}'$.

To compute $\mathbf{T}^*_{LZY}(\mathbf{y}, \boldsymbol{\vartheta}_0)$, one mainly needs to evaluate the second derivative of ln $p(\mathbf{y}\,|\,\boldsymbol{\theta})$. The well-known Louis formula by Louis (1982) suggests

$$\begin{aligned}\frac{\partial^2 \ln p(\mathbf{y}|\,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} &= E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\left\{\frac{\partial^2 \ln(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\} + Var_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\{S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})\}\\ &= E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\left\{\frac{\partial^2 \ln(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})'\right\}\\ &\quad - E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\{S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})\}E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\{S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})\}',\end{aligned}$$

where all the expectations are taken with respect to the conditional distribution of $\mathbf{z}$ given $\mathbf{y}$ and $\boldsymbol{\theta}$. Hence, we can use the following formula to calculate the second derivative of the observed-data likelihood function,

$$\begin{aligned}&E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\left\{\frac{\partial^2 \ln(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})'\right\}\\ &\approx \frac{1}{J}\sum_{i=1}^{J}\left\{\frac{\partial^2 \ln\left(\mathbf{y},\mathbf{z}^{(j)}|\,\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + S\left(\mathbf{y},\mathbf{z}^{(j)}|\,\boldsymbol{\theta}\right)S\left(\mathbf{y},\mathbf{z}^{(j)}|\,\boldsymbol{\theta}\right)'\right\},\\ &E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\{S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})\} \approx \frac{1}{J}\sum_{i=1}^{J} S\left(\mathbf{y},\mathbf{z}^{(m)}|\,\boldsymbol{\theta}\right) = \frac{1}{J}\sum_{i=1}^{J}\frac{\partial \ln p\left(\mathbf{y},\mathbf{z}^{(j)}|\,\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}},\end{aligned}$$

where $\{\mathbf{z}^{(j)}, j = 1, 2, \ldots, J\}$ are the MCMC samples of $\mathbf{z}$.

Since $\mathbf{T}_{LZY}$ is the posterior mean of the difference in deviance, $\mathbf{T}_{LZY}$ and $\mathbf{T}^*_{LZY}$ can be understood as the MCMC version of LR test. Li et al. (2014) pointed out that the proposed test statistic appeals in four aspects. First, they are well-defined under improper priors. Second, they do not suffer from Jeffreys-Lindley's paradox and, hence, can be used under non-informative vague priors. Third, at least, $\mathbf{T}^*_{LZY}$ is not difficult to compute. For the latent variable model, $\mathbf{T}^*_{LZY}(\mathbf{y}, \boldsymbol{\vartheta}_0)$ only involves the second derivative which is not very difficult to evaluate from the MCMC output.

Finally, under some mild regularity conditions, when the likelihood information dominates the prior information, Li et al. (2014) proved that under the null hypothesis

$$\begin{aligned}\mathbf{T}_{LZY}(\mathbf{y}, \boldsymbol{\vartheta}_0) &\overset{a}{\sim} \boldsymbol{\epsilon}'\left[\mathbf{IJ}_{11}^{1/2}(\boldsymbol{\theta}_0)\mathbf{J}_{11}(\boldsymbol{\theta}_0)\mathrm{IJ}_{11}^{1/2}(\boldsymbol{\theta}_0)\right]\boldsymbol{\epsilon}\\ &\quad - \left[p + q - \mathbf{tr}[-L_{0n}^{(2)}(\overline{\boldsymbol{\theta}})V_{22}(\overline{\boldsymbol{\theta}})]\right],\end{aligned} \tag{8}$$

$$\begin{aligned}\mathbf{T}^*_{LZY}(\mathbf{y}, \boldsymbol{\vartheta}_0) &\overset{a}{\sim} \boldsymbol{\epsilon}'\left[\mathbf{IJ}_{11}^{1/2}(\boldsymbol{\theta}_0)\mathbf{J}_{11}(\boldsymbol{\theta}_0)\mathrm{IJ}_{11}^{1/2}(\boldsymbol{\theta}_0)\right]\boldsymbol{\epsilon}\\ &\quad - \left[p + q - \mathbf{tr}[-L_{0n}^{(2)}(\overline{\boldsymbol{\theta}})V_{22}(\overline{\boldsymbol{\theta}})]\right],\end{aligned} \tag{9}$$

where $\boldsymbol{\epsilon}$ is a standard multivariate normal variate, $\boldsymbol{\theta_0} = (\boldsymbol{\vartheta_0}, \boldsymbol{\psi}0)$ the true value of $\boldsymbol{\theta}$, $\mathbf{J}(\boldsymbol{\theta_0})$ the Fisher information matrix given by

$$\mathbf{J}(\boldsymbol{\theta}_0) = \frac{1}{n}\int -L_n^{(2)}(\boldsymbol{\theta}_0)p(\mathbf{y}|\,\boldsymbol{\theta}_0)\mathrm{d}\mathbf{y},$$

$\mathbf{IJ}(\boldsymbol{\theta_0})$ the inverse of $\mathbf{J}(\boldsymbol{\theta_0})$, $\mathbf{J}_{11}(\boldsymbol{\theta_0})$, and $\mathbf{IJ}_{11}(\boldsymbol{\theta_0})$ the submatrices of $\mathbf{J}(\boldsymbol{\theta_0})$ and $\mathbf{IJ}(\boldsymbol{\theta_0})$, respectively, corresponding to $\boldsymbol{\vartheta}$. The asymptotic distributions given in (8) and (9) are obtained under the assumptions of repeated sampling and the diverged sample size. Clearly, the set up is also in the frequentist domain. A drawback of the test is that it is not asymptotically pivotal because the asymptotic distribution depends on some unknown population parameters.

3.2.5 Li et al. (2015) and LM-type loss function

To address the nonpivotal problem in the test statistic of Li et al. (2014, 2015) proposed to use a quadratic loss function given by

$$\Delta\mathcal{L}[H_0, \boldsymbol{\theta}] = \left(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}}\right)' C_{\vartheta\vartheta}\left(\overline{\boldsymbol{\theta}}_0\right)\left(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}}\right), \tag{10}$$

where

$$C(\boldsymbol{\theta}) = s(\boldsymbol{\theta})s(\boldsymbol{\theta})', \mathbf{s}(\boldsymbol{\theta}) = \frac{\partial\, \ln p(\mathbf{y}|\,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}},$$

and $s(\boldsymbol{\theta})$ the score function of $\boldsymbol{\theta}$, $C_{\vartheta\vartheta}$ $(\boldsymbol{\theta})$ is the submatrix of $C(\boldsymbol{\theta})$ corresponding to $\boldsymbol{\vartheta}$ and is semipositive definite, $\overline{\boldsymbol{\theta}}_0 = (\boldsymbol{\vartheta}_0, \overline{\boldsymbol{\psi}}_0)$ is the posterior mean of $\boldsymbol{\vartheta}$ under H_0, $\overline{\boldsymbol{\theta}}$ is the posterior mean of $\boldsymbol{\theta}$ under H_1. Based on this quadratic loss, naturally, the test statistic is given by

$$\mathbf{T}_{LLY}(\mathbf{y}, \boldsymbol{\vartheta}_0) = \int \Delta\mathcal{L}[H_0, \boldsymbol{\theta}]p(\boldsymbol{\theta}|\,\mathbf{y})\mathrm{d}\boldsymbol{\theta} = \int \left(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}}\right)' C_{\vartheta\vartheta}\left(\overline{\boldsymbol{\theta}}_0\right)\left(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}}\right)p(\boldsymbol{\theta}|\,\mathbf{y})\mathrm{d}\boldsymbol{\theta}, \tag{11}$$

where $p(\boldsymbol{\theta}\,|\,\mathbf{y})$ is the posterior distribution of $\boldsymbol{\theta}$ under H_1.

To compute $\mathbf{T}_{LLY}$ $(\mathbf{y}, \boldsymbol{\vartheta}_0)$, one mainly needs to evaluate the first derivative of ln $p(\mathbf{y}\,|\,\boldsymbol{\theta})$. For the latent variable model, ln $p(\mathbf{y}\,|\,\boldsymbol{\theta})$ is often intractable. Under the EM algorithm (Dempster et al., 1977), it can be shown that

$$\frac{\partial\, \ln p(\mathbf{y}|\,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\{S(\mathbf{y}, \mathbf{z}|\,\boldsymbol{\theta})\} \approx \frac{1}{J}\sum_{i=1}^{J} S\left(\mathbf{y}, \mathbf{z}^{(j)}|\,\boldsymbol{\theta}\right) = \frac{1}{J}\sum_{i=1}^{J}\frac{\partial\, \ln p\left(\mathbf{y}, \mathbf{z}^{(j)}|\,\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}$$

where $\{\mathbf{z}^{(j)}, j = 1, 2, \ldots, J\}$ are the MCMC samples of $\mathbf{z}$.

The proposed test can be viewed as the MCMC version of LM test. To see the link, let the LM statistic (Breusch and Pagan, 1980) be

$$\mathbf{LM} = s_{\vartheta}\left(\hat{\boldsymbol{\theta}}_0\right)\left[-\mathbf{IL}_{\vartheta\vartheta}^{(2)}\left(\hat{\boldsymbol{\theta}}\right)\right]s_{\vartheta}\left(\hat{\boldsymbol{\theta}}_0\right),$$

where $\hat{\boldsymbol{\theta}}_0 = (\boldsymbol{\vartheta}_0, \hat{\boldsymbol{\psi}}_0)$ is the MLE of $\boldsymbol{\theta}$ under the null hypothesis, $s_{\vartheta}(\boldsymbol{\theta})$ is subvector of $s(\boldsymbol{\theta})$ corresponding to $\boldsymbol{\vartheta}$, $\mathbf{IL}_{\vartheta\vartheta}(\boldsymbol{\theta})$ is the submatrix of $\mathbf{IL}(\boldsymbol{\theta})$ corresponding to $\boldsymbol{\vartheta}$, $\mathbf{IL}^{(2)}(\boldsymbol{\theta})$ is the inverse matrix of $\mathbf{L}^{(2)}(\boldsymbol{\theta}) := \partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$. Under some regularity assumptions, when the null hypothesis is true and the likelihood dominates the prior, Li et al. (2015) showed that

$$\mathbf{T}_{LLY}(\mathbf{y}, \boldsymbol{\vartheta}_0) = \mathbf{LM} + o_p(1) \xrightarrow{d} \chi^2(p).$$

The test statistic $\mathbf{T}_{LLY}(\mathbf{y}, \boldsymbol{\vartheta}_0)$ has a few nice properties. For example, it is well-defined under an improper prior and immune to Jeffreys-Lindley's paradox. In addition, for the latent variable model it is not difficult to compute with the EM algorithm. Finally, it follows a pivotal χ^2_p asymptotically, and hence, it is easy to obtain threshold values.

3.2.6 Li et al. (2019) and Wald-type loss function

Although the test statistic proposed by Li et al. (2015) is convenient to calculate and has some good properties, it requires the MCMC output to be obtained twice, one under H_0 and the other under H_1. Based on another quadratic loss function, Li et al. (2019) proposed a test statistic which is only by-product of the MCMC output under H_1, and hence, is easier to compute.

Let the posterior covariance matrix under the alterative hypothesis be

$$\mathbf{V}(\overline{\boldsymbol{\theta}}) = E\left[(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}})(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}})' | \mathbf{y}, H_1\right] = \int (\boldsymbol{\theta} - \overline{\boldsymbol{\theta}})(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}})' p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta},$$

where $\overline{\boldsymbol{\theta}}$ is the posterior mean of $\boldsymbol{\theta}$ under the alternative hypothesis H_1. Li et al. (2019) proposed the following net loss function for hypothesis testing

$$\Delta\mathcal{L}[H_0, \boldsymbol{\theta}] = (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)' \left[\mathbf{V}_{\vartheta\vartheta}(\overline{\boldsymbol{\theta}})\right]^{-1} (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0),$$

where $\mathbf{V}_{\vartheta\vartheta}(\overline{\boldsymbol{\theta}})$ is the submatrix of $\mathbf{V}(\overline{\boldsymbol{\theta}})$ corresponding to $\boldsymbol{\vartheta}$, $\left[\mathbf{V}_{\vartheta\vartheta}(\overline{\boldsymbol{\theta}})\right]^{-1}$ is the inverse matrix of $\mathbf{V}_{\vartheta\vartheta}(\overline{\boldsymbol{\theta}})$. Then, the test statistic can be established as follows:

$$\mathbf{T}_{LLYZ}(\mathbf{y}, \boldsymbol{\vartheta}_0) = \int (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)' \left[\mathbf{V}_{\vartheta\vartheta}(\overline{\boldsymbol{\theta}})\right]^{-1} (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0) p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta}. \qquad (12)$$

To see the link between $\mathbf{T}_{LLYZ}(\mathbf{y}, \boldsymbol{\vartheta}_0)$ and the Wald statistic, define the Wald statistic by (Engle, 1984)

$$\mathbf{Wald} = \left(\hat{\boldsymbol{\vartheta}}_{ML} - \boldsymbol{\vartheta}_0\right)' \left[-\mathbf{IL}^{(2)}_{\vartheta\vartheta}\left(\hat{\boldsymbol{\theta}}_{ML}\right)\right]^{-1} \left(\hat{\boldsymbol{\vartheta}}_{ML} - \boldsymbol{\vartheta}_0\right)',$$

where $\hat{\boldsymbol{\theta}}_{ML} := \left(\hat{\boldsymbol{\vartheta}}_{ML}, \hat{\boldsymbol{\psi}}_{ML}\right)$ is the ML estimate of $\boldsymbol{\theta}$. Under some regularity assumptions, when the null hypothesis is true and the likelihood dominates the prior, Li et al. (2019) showed that

$$\mathbf{T}_{LLYZ}(\mathbf{y}, \boldsymbol{\vartheta}_0) = \mathbf{Wald} + o_p(1) \xrightarrow{d} \chi^2(p).$$

This is why $\mathbf{T}_{LLYZ}(\mathbf{y}, \boldsymbol{\vartheta}_0)$ may be viewed as a MCMC version of the Wald test.

It can be seen that $\mathbf{T}_{LLYZ}$ ($\mathbf{y}$, $\boldsymbol{\vartheta}_0$) shared some nice properties with the test of Li et al. (2015). First, it is well-defined under improper prior distributions and avoids Jeffreys-Lindley's paradox. Second, the asymptotic distribution is pivotal so that the threshold values can be easily obtained from the $\chi^2(p)$ distribution. Most importantly, it is only by-product of the posterior output under H_1, and hence, is easier to compute.

Table 1 summarize the MCMC-based trinity of the tests and their key properties. It is important to emphasize that although they are constructed from the MCMC output which contains random draw from the Bayesian posterior distribution, the statistical inference made by the three tests is not conditional on the data. Instead, the justification of the three tests is done in a frequentist framework, requiring repeated sampling from the DGP and an asymptotic argument.

4 Specification testing based on the MCMC output

Detection of specification problems in economics has been a major concern. After ML is applied to estimate the model, several specification tests may be used, including the information matrix test of White (1982), the IOS and IOS_A tests of Presnell and Boos (2004). Recently, Li et al. (2018) proposed a specification test based on the MCMC output which can assess the validity of the model specification and can tell the source of model misspecification if the null model is rejected.

Let model P be a collection of candidate models indexed by parameters $\boldsymbol{\theta}$ whose dimension is q. Let $P_{\boldsymbol{\theta}}$ denote P indexed by $\boldsymbol{\theta}$. We say the model P is correctly specified if there exists $\boldsymbol{\theta}$, such that $P_0 \in P_{\boldsymbol{\theta}}$.

Arguably the best known specification test is based on the information matrix proposed by White (1982). For *i.i.d.* case, let $p(\mathbf{y}|\boldsymbol{\theta})$ denote the likelihood function of Model $P\boldsymbol{\theta}$ and

$$\begin{aligned}\mathbf{s}(\mathbf{y},\boldsymbol{\theta}) &:= \partial \ln p(\mathbf{y}|\,\boldsymbol{\theta})/\partial\boldsymbol{\theta}, \mathrm{h}(\mathbf{y},\boldsymbol{\theta}) &&:= \partial^2 \ln p(\mathbf{y}|\,\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}' \\ \mathbf{H}(\boldsymbol{\theta}) &:= \int \mathbf{h}(\mathbf{y},\boldsymbol{\theta})p(\mathbf{y}|\,\boldsymbol{\theta})\mathrm{d}\mathbf{y}, \mathrm{J}(\boldsymbol{\theta}) &&:= \int \mathbf{s}(\mathbf{y},\boldsymbol{\theta})\mathbf{s}'(\mathbf{y},\boldsymbol{\theta})p(\mathbf{y}|\,\boldsymbol{\theta})\mathrm{d}\mathbf{y}\end{aligned}$$

Let $d(\mathbf{y},\theta) := vech[\mathbf{h}(\mathbf{y},\theta)+\mathbf{s}(\mathbf{y},\theta)\mathbf{s}'(\mathbf{y},\theta)]$, where *vech* is the columnwise vectorization with the upper portion excluded. Let the ML-based sample counterparts of $\mathbf{H}(\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta})$ be

$$\hat{\mathbf{H}}_n\left(\hat{\boldsymbol{\theta}}_{ML}\right) := \frac{1}{n}\sum_{t=1}^{n}\mathbf{h}\left(y_t,\hat{\boldsymbol{\theta}}_{ML}\right), \hat{\mathbf{J}}_n\left(\hat{\boldsymbol{\theta}}_{ML}\right) := \frac{1}{n}\sum_{t=1}^{n}\mathbf{s}\left(y_{t,}\hat{\boldsymbol{\theta}}_{ML}\right)\mathbf{s}'\left(y_t,\hat{\boldsymbol{\theta}}_{ML}\right).$$

Let $D_n\left(\hat{\boldsymbol{\theta}}_{ML}\right) = \frac{1}{n}\sum_{t=n}^{n} d\left(y_t,\hat{\boldsymbol{\theta}}_{ML}\right)$ and $\dot{D}_n\left(\hat{\boldsymbol{\theta}}_{ML}\right) = \partial D_n\left(\hat{\boldsymbol{\theta}}_{ML}\right)/\partial\boldsymbol{\theta}$. If the model is correctly specified, then $\mathbf{H}(\boldsymbol{\theta})+\mathbf{J}(\boldsymbol{\theta})=0$. White (1982) proposed the following information matrix test

$$\text{IMT} = nD_n\left(\hat{\boldsymbol{\theta}}_{ML}\right)V_n^{-1}\left(\hat{\boldsymbol{\theta}}_{ML}\right)D_n\left(\hat{\boldsymbol{\theta}}_{ML}\right), \tag{13}$$

TABLE 1 Summary of MCMC-based trinity of tests

	T_{LZY}	T_{LLY}	T_{LLYZ}
Expression	$2\left[\ \ln p(\overline{\boldsymbol{\vartheta}},\overline{\boldsymbol{\psi}})-\ \ln p(\overline{\boldsymbol{\psi}}\mid\boldsymbol{\vartheta}_0)\right]$ $-2\int\ \ln p(\boldsymbol{\vartheta}\mid\boldsymbol{\psi})p(\boldsymbol{\theta}\mid\mathbf{y})\mathrm{d}\boldsymbol{\theta}+2D$ $-\left[p+1-\mathbf{tr}[-L_{0n}^{(2)}(\overline{\boldsymbol{\psi}})V_{22}(\overline{\boldsymbol{\theta}})\right]$	$\int(\boldsymbol{\vartheta}-\overline{\boldsymbol{\vartheta}})'\mathrm{C}_{\vartheta\vartheta}(\overline{\boldsymbol{\vartheta}}_0)$ $(\boldsymbol{\vartheta}-\overline{\boldsymbol{\vartheta}})p(\boldsymbol{\vartheta}\mid\mathbf{y})\mathrm{d}\boldsymbol{\vartheta}$	$\int(\boldsymbol{\vartheta}-\overline{\boldsymbol{\vartheta}}_0)'[\mathbf{V}_{\vartheta\vartheta}\overline{\boldsymbol{\theta}}]^{-1}$ $(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_0)p(\boldsymbol{\theta}\mid\mathbf{y})\mathrm{d}\boldsymbol{\theta}$
Prior	Improper or proper	Improper or proper	Improper or proper
Jeffreys-Lindley's Paradox	No	No	No
Asymptotic theory	$\boldsymbol{\epsilon}'\left[\mathbf{IJ}_{11}^{1/2}(\boldsymbol{\vartheta}_0)\mathbf{J}_{11}(\boldsymbol{\vartheta}_0)\mathbf{IJ}_{11}^{1/2}(\boldsymbol{\vartheta}_0)\right]\boldsymbol{\epsilon}$ $-\left[p+q-\mathbf{tr}\left[-L_{0n}^{(2)}(\overline{\boldsymbol{\theta}})V_{22}(\overline{\boldsymbol{\theta}})\right]\right]$	$\chi^2(p)$	$\chi^2(p)$
Asymptotic pivotal	No	Yes	Yes

where

$$V_n\left(\hat{\boldsymbol{\theta}}_{ML}\right) = \frac{1}{n}\sum_{t=1}^{n} v_t\left(\hat{\boldsymbol{\theta}}_{ML}\right) v_t\left(\hat{\boldsymbol{\theta}}_{ML}\right)',$$

$$v_t\left(\hat{\boldsymbol{\theta}}_{ML}\right) = d\left(y_t, \hat{\boldsymbol{\theta}}_{ML}\right) - \dot{D}_n\left(\hat{\boldsymbol{\theta}}_{ML}\right)\hat{\mathbf{H}}_n^{-1}\left(\hat{\boldsymbol{\theta}}_{ML}\right)\mathbf{s}\left(y_t, \hat{\boldsymbol{\theta}}_{ML}\right).$$

He then showed that IMT $\xrightarrow{d}\chi^2$ as $n \to \infty$ under the null hypothesis.

Presnell and Boos (2004) proposed an alternative test—the "in-and-out" likelihood ratio (IOS) test for models with *i.i.d.* observations,

$$\text{IOS} = \ \ln \frac{\prod_{t=1}^{n} p\left(y_t, \hat{\boldsymbol{\theta}}_{ML}\right)}{\prod_{t=1}^{n} p\left(y_t, \hat{\boldsymbol{\theta}}_{ML}^{(t)}\right)} = \sum_{t=1}^{n}\left[\ \ln p\left(y_t \mid \hat{\boldsymbol{\theta}}_{ML}\right) - \ \ln p\left(y_t, \hat{\boldsymbol{\theta}}_{ML}^{(t)}\right)\right],$$

where $\hat{\boldsymbol{\theta}}_{ML}^{(t)}$ be the MLE of $\boldsymbol{\theta}$ when the t-th observation, y_t, is deleted from the whole sample. They showed that the asymptotic form of IOS is

$$\text{IOS}_A = \mathbf{tr}\left[-\hat{\mathbf{H}}_n^{-1}\left(\hat{\boldsymbol{\theta}}_{ML}\right)\hat{\mathbf{J}}_n\left(\hat{\boldsymbol{\theta}}_{ML}\right)\right], \tag{14}$$

and IOS − $\text{IOS}_A = o_p\ (n^{-1/2})$. Like IMT, IOS_A also compares $\hat{H}_n\left(\hat{\boldsymbol{\theta}}_{ML}\right)$ with $\hat{J}_n\left(\hat{\boldsymbol{\theta}}_{ML}\right)$, but in a ratio form instead of an additive form. Under the null hypothesis, $\text{IOS}_A \xrightarrow{p} q$ and $n^{1/2}$ ($\text{IOS}_A - q$) converges to a normal distribution with zero mean and finite variance. It is well documented in the literature that the asymptotic distributions poorly approximate their finite sample counterparts for IMT, IOS, and IOS_A. As a result, they all suffer from serious bias distortions if the critical values for testing are based on the asymptotic distributions. The poor finite sample performance of these tests is not surprising as the asymptotic theory is derived based on the convergence of the sample high order moments, whose speed is slow. To reduce the size distortion of these tests, bootstrap methods have been proposed to obtain the critical values. Unfortunately, bootstrap methods are computationally demanding.

For weakly dependent data, let $\mathbf{y}^t := (y_1, \ldots, y_t)$ and

$$\begin{aligned}
\mathbf{s}(\mathbf{y}^t, \boldsymbol{\theta}) &:= \frac{\partial \ \ln p(\mathbf{y}^t \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, & \mathbf{h}(\mathbf{y}^t, \boldsymbol{\theta}) &:= \frac{\partial^2 \ \ln p(\mathbf{y}^t \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}, \\
\mathbf{s}_t(\boldsymbol{\theta}) &:= \mathbf{s}(\mathbf{y}^t, \boldsymbol{\theta}) - \mathbf{s}(\mathbf{y}^{t-1}, \boldsymbol{\theta}), & \mathbf{h}_t(\boldsymbol{\theta}) &:= \mathbf{h}(\mathbf{y}^t, \boldsymbol{\theta}) - \mathbf{h}(\mathbf{y}^{t-1}, \boldsymbol{\theta}), \\
\hat{\mathbf{J}}_n(\boldsymbol{\theta}) &:= \frac{1}{n}\sum_{t=1}^{n} \mathbf{s}_t(\boldsymbol{\theta})\mathbf{s}_t'(\boldsymbol{\theta}), & \hat{\mathbf{H}}_n(\boldsymbol{\theta}) &:= \frac{1}{n}\sum_{t=1}^{n} \mathbf{h}_t(\boldsymbol{\theta}).
\end{aligned}$$

and $V(\overline{\boldsymbol{\theta}}) = \int (\boldsymbol{\theta} - \overline{\boldsymbol{\theta}})(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}})' p(\boldsymbol{\theta} \mid \mathbf{y}) \mathrm{d}\boldsymbol{\theta}$, a natural MCMC-based informative matrix test statistic can be defined as:

$$\text{BIMT} = \mathbf{tr}\left[nV(\overline{\boldsymbol{\theta}})\hat{\mathbf{J}}_n(\overline{\boldsymbol{\theta}})\right] = n\int (\boldsymbol{\theta} - \overline{\boldsymbol{\theta}})'\hat{\mathbf{J}}_n(\overline{\boldsymbol{\theta}})(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}) p(\boldsymbol{\theta} \mid \mathbf{y}) \mathrm{d}\boldsymbol{\theta}, \tag{15}$$

Under some mild regularity conditions, Li et al. (2018) showed that under the null hypothesis, $n^{1/2}$ (BIMT/$q-$ 1) has the same asymptotic distribution as $n^{1/2}$ (IOS$_A$/$q-$ 1). Hence, BIMT may be regarded as the MCMC-based version of IOS$_A$. Unfortunately but not surprisingly, BIMT inherits the size distortion problem of IOS$_A$ and bootstrap methods must be used.

Due to this size distortion problem, Li et al. used a technique of Fan et al. (2015) to construct a new specification test statistic. In particular, they propose to expand $p(\mathbf{y}|\boldsymbol{\theta})$, the model in concern, to a larger model denoted by $p(\mathbf{y}|\boldsymbol{\theta}_L)$ where $\boldsymbol{\theta}_L=(\boldsymbol{\theta}',\boldsymbol{\theta}_E')'$ with $\boldsymbol{\theta}_E$ being a q_E-dimensional vector. So the expanded model $p(\mathbf{y}|\boldsymbol{\theta}_L)$ nests the original model $p(\mathbf{y}|\boldsymbol{\theta})$.

It is assumed that if the specification $p(\mathbf{y}|\boldsymbol{\theta})$ is correct, then the true value of $\boldsymbol{\theta}_E$ is zero. The final specification test statistic of Li et al. (2018) has the form of

$$\text{BMT}=\mathbf{tr}\left\{C_E\left(\mathbf{y},\left(\overline{\boldsymbol{\theta}},\boldsymbol{\theta}_E=0\right)\right)V_E\left(\overline{\boldsymbol{\theta}}_L\right)\right\}+\sqrt{n}(\text{BIMT}/q-1)^2, \qquad (16)$$

where $C_E\left(\mathbf{y},\left(\overline{\boldsymbol{\theta}},\boldsymbol{\theta}_E=0\right)\right)$ is the submatrix of C ($\mathbf{y}$, $\boldsymbol{\theta}_L$) corresponding to $\boldsymbol{\theta}_E$ evaluated at $\left(\overline{\boldsymbol{\theta}},\boldsymbol{\theta}_E=0\right)$ and $V_E\left(\overline{\boldsymbol{\theta}}_L\right)$ is the submatrix of $V_E(\boldsymbol{\theta}_L)$ corresponding to $\boldsymbol{\theta}_E$ evaluated at, $\overline{\boldsymbol{\theta}}_L$ and

$$s(\mathbf{y},\boldsymbol{\theta}_L)=\frac{\partial\ \ln p(\mathbf{y}|\,\boldsymbol{\theta}_L)}{\partial\boldsymbol{\theta}_L},C(\mathbf{y},\boldsymbol{\theta}_L)=s(\mathbf{y},\boldsymbol{\theta}_L)s(\mathbf{y},\boldsymbol{\theta}_L)',$$

$$V\left(\overline{\boldsymbol{\theta}}_L\right)=E\left[\left(\boldsymbol{\theta}_L-\overline{\boldsymbol{\theta}}_L\right)\left(\boldsymbol{\theta}_L-\overline{\boldsymbol{\theta}}_L\right)'|\,\mathbf{y}\right]=\int\left(\boldsymbol{\theta}_L-\overline{\boldsymbol{\theta}}_L\right)\left(\boldsymbol{\theta}_L-\overline{\boldsymbol{\theta}}_L\right)'p(\boldsymbol{\theta}_L|\mathbf{y})\mathrm{d}\boldsymbol{\theta}_L,$$

with $\overline{\boldsymbol{\theta}}_L$ being the posterior mean of $\boldsymbol{\theta}_L$ in the expanded model. It can be seen that BIMT is used as the power enhancement function.

Under a set of regularity conditions, Li et al. showed that if the model is correctly specified, $BMT\xrightarrow{d}\chi^2(q_E)$; but if the model is misspecified with $q^*\neq q$, then

$$\mathbf{tr}\left\{C_E\left(\mathbf{y},\left(\overline{\boldsymbol{\theta}},\boldsymbol{\theta}_E=0\right)\right)V_E\left(\overline{\boldsymbol{\theta}}_L\right)\right\}=\sqrt{n}(q^*/q-1)^2+O_p\left(\sqrt{n}\right),\text{BMT}\sim O_p\left(\sqrt{n}\right),$$

where $q^*=\mathbf{tr}[-\mathbf{H}\,(\boldsymbol{\theta}^*)^{-1}\,\mathbf{J}\,(\boldsymbol{\theta}^*)]$ with $\boldsymbol{\theta}^*$ being the pseudo true value of $\boldsymbol{\theta}$, where

$$\mathbf{H}(\boldsymbol{\theta}^*)\quad :=\lim_{n\to\infty}\mathbf{H}_n(\boldsymbol{\theta}^*)\text{and}\mathbf{J}(\boldsymbol{\theta}^*):=\lim_{n\to\infty}\mathbf{J}_n(\boldsymbol{\theta}^*),$$

$$\mathbf{J}_n(\boldsymbol{\theta})\quad :=\int\hat{\mathbf{J}}_n(\boldsymbol{\theta})p(\mathbf{y})d\mathbf{y},\mathbf{H}_n(\boldsymbol{\theta}):=\int\hat{\mathbf{H}}_n(\boldsymbol{\theta})p(\mathbf{y})d\mathbf{y},$$

BMT has several nice properties. First, compared with IM, IOS, and IOS$_A$, BMT is based on the MCMC output. When the likelihood function is difficult to optimize but the MCMC draws from the posterior distribution are available, BMT is easier to compute than IM, IOS, and IOS$_A$. Second, when $\sqrt{n}(\text{BIMT}/q-1)^2$ does not have the size distortion problem, it is most likely that BMT will not suffer from size distortion. As a result, no bootstrap method is needed and intensive computational effort is avoided.

5 Model selection based on the MCMC output

Model selection is a very important statistical decision in practice. Many important and widely used information criteria have been proposed to select from candidate models in the literature. Examples include AIC, BIC, and HQ. Most of them require that MLE is available. The most well-known model selection criterion based on the MCMC output is DIC of Spiegelhalter et al. (2002). 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 from the MCMC output when the likelihood function is available in closed-form. Second, DIC is applicable to a wide range of statistical models. Third, unlike BFs, DIC is not subject to Jeffreys-Lindley's paradox and can be defined under improper priors. In this section, we first review the DIC for models when the asymptotic theory for ML is applicable, paying particular attention to the asymptotic justification of DIC. We also discuss how to obtain DICs when there are latent variables. In both cases, the loss function is the plug-in predictive loss. We also discuss the information criteria when the loss function is the Bayesian predictive loss.

5.1 DIC for regular models

We first review DIC for regular models, that is, when the asymptotic theory given by (2), (3) and (4) holds true. Spiegelhalter et al. (2002) proposed the DIC for Bayesian model comparison. The criterion is based on the deviance

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

and takes the form of

$$\text{DIC} = D\left(\overline{\boldsymbol{\theta}}\right) + 2P_D, \tag{17}$$

where P_D, 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:

$$P_D = \overline{D(\boldsymbol{\theta})} - D\left(\overline{\boldsymbol{\theta}}\right) = -2\int [\ln p(\mathbf{y}|\boldsymbol{\theta}) - \ln p(\mathbf{y}|\boldsymbol{\theta})] p(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{18}$$

with $\overline{\boldsymbol{\theta}}$ being the posterior mean of $\boldsymbol{\theta}$.

Under some regularity conditions, Li et al. (2017a) gives a rigorous decision-theoretic justification. Let $g(\mathbf{y})$ be the data generating process of $\mathbf{y}$, $\mathbf{y}_{rep} = (y_{1,rep}, \ldots, y_{n,rep})'$ denote the future replicate data with $\mathbf{y}$. Hence, the plug-in predictive distribution based on replicate data is $-2 \ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}(\mathbf{y})\right)$

where $\overline{\boldsymbol{\theta}}(\mathbf{y})$ is the posterior mean under the data $\mathbf{y}$. Consider the plug-in predictive distribution $p\ \mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}(\mathbf{y})$ in the following KL divergence

$$KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\,\overline{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right] = E_{\mathbf{y}_{rep}}\left[\ln\frac{g\left(\mathbf{y}_{rep}\right)}{p\left(\mathbf{y}_{rep}|\,\overline{\boldsymbol{\theta}}_n(\mathbf{y})\right)}\right]$$

$$= E_{\mathbf{y}_{rep}}\left[\ln g\left(\mathbf{y}_{rep}\right)\right] + E_{\mathbf{y}_{rep}}\left[-\ln p\left(\mathbf{y}_{rep}|\,\overline{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right].$$

The smaller this KL divergence, the better the candidate model in predicting $g(\mathbf{y}_{rep})$. Since $g(\mathbf{y}_{rep})$ is the true DGP and $E_{\mathbf{y}_{rep}}$ ln $g(\mathbf{y}_{rep})$ is independent with candidate models, it is dropped from the above equation. Li et al. (2017a) showed that DIC is an unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ \ln p\left(\mathbf{y}_{rep}|\,\overline{\boldsymbol{\theta}}(\mathbf{y})\right)\right]$ asymptotically, i.e., $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ \ln p\left(\mathbf{y}_{rep}|\,\overline{\boldsymbol{\theta}}\right)\right] = E_{\mathbf{y}}(\text{DIC}) + o(1)$. The key assumptions to obtain the asymptotic unbiasedness include that the candidate models are good approximation to the true DGP, the consistency and asymptotic normality of MLE, and the expression for the asymptotic variance of MLE. For details, see Li et al. (2017a).

The above decision-theoretic justification to DIC is that DIC selects a model that asymptotically minimizes the risk, which is the expected KL divergence between the DGP and the plug-in predictive distribution $p\left(\mathbf{y}_{rep}|\,\overline{\boldsymbol{\theta}}(\mathbf{y})\right)$ where the expectation is taken with respect to the DGP. A key difference between AIC and DIC is that the plug-in predictive distribution is based on different estimators. In AIC, the ML estimate, $\hat{\boldsymbol{\theta}}_{ML}(\mathbf{y})$, is used while in DIC the Bayesian posterior mean, $\overline{\boldsymbol{\theta}}(\mathbf{y})$, is used.

When ln $p(\mathbf{y}\,|\,\boldsymbol{\theta})$ has a closed-form expression, it can be seen that DIC is trivial to compute from the MCMC output. DIC has been incorporated into a Bayesian software, WinBUGS. This explains why DIC has been widely used in practice for model selection.

5.2 Bayesian predictive distribution as the loss function

Unfortunately, the plug-in predictive distribution is not invariant to parameterization. As a result, DIC is sensitive to parameterization. Alternatively, we may use the Bayesian predictive distribution as a loss function. The Bayesian predictive distribution is not only a full proper predictive distribution, but also invariant to reparameterization.

Let $p(\mathbf{y}_{rep}\,|\,\mathbf{y})$ be the Bayesian predictive distribution, that is,

$$p\left(\mathbf{y}_{rep}|\,\mathbf{y}\right) = \int p\left(\mathbf{y}_{rep}|\,\boldsymbol{\theta}\right)p(\boldsymbol{\theta}|\,\mathbf{y})d\boldsymbol{\theta}.$$

The KL divergence based on the Bayesian predictive distribution is given by

$$KL\left[g\left(\mathbf{y}_{rep}\right),p\left(\mathbf{y}_{rep}|\,\mathbf{y}\right)\right]=E_{\mathbf{y}_{rep}}\left(\ln g\left(\mathbf{y}_{rep}\right)\right)-E_{\mathbf{y}_{rep}}\left(\ln p\left(\mathbf{y}_{rep}|\,\mathbf{y}\right)\right). \tag{19}$$

Li et al. (2017a) obtained the information criterion based on the Bayesian predictive distribution as

$$\mathrm{DIC}^{BP}=D\left(\overline{\boldsymbol{\theta}}\right)+(1+\ln 2)P_D. \tag{20}$$

Under some regularity assumptions, Li et al. showed that DIC^{BP} is an unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}[-2\ln p(\mathbf{y}_{rep}|\mathbf{y})]$ asymptotically, i.e., $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}[-2\ln p(\mathbf{y}_{rep}|\mathbf{y})]=E_y(DIC^{BP})+o(1)$. Clearly, DIC^{BP} is as easy to compute as DIC. Since DIC is monitored in WinBUGS, no additional effort is needed for calculating DIC^{BP}.

5.3 Integrated DIC for latent variable models

Unfortunately, not all models are regular. A well-known nonregular model in economics is a class of models with incidental parameters which leads to the incidental parameter problem. In this class of models, the information about the incidental parameters stops accumulating after a finite number of observations have been taken; see Neyman and Scott (1948) and Lancaster (2000) for details about the incidental parameter problem.

As shown in Gelman et al. (2013), the incidental parameter problem can lead that the ML estimator is inconsistent and Bayesian large sample theory becomes invalid. When this is the case, the asymptotic justification of DIC does not hold because of the failure of these standard asymptotic theory.

In general, the latent variable model given in (1) does not have incidental parameters and hence the incidental parameter problem is not applicable. As explained earlier, for many latent variable models, the likelihood function is very difficult to be accurately approximate, rendering ML difficult to implement. To facilitate the posterior analysis, the data-augmentation strategy of Tanner and Wong (1987) is often used to augment the parameter space to $(\boldsymbol{\theta}, \mathbf{z})$, changing the likelihood function to $p(\mathbf{y}\,|\,\boldsymbol{\theta}, \mathbf{z})$ which typically has a closed-form expression. Denote the sample mean of $\mathbf{z}$, $\boldsymbol{\theta}$ by $\overline{\mathbf{z}}$, $\overline{\boldsymbol{\theta}}$, obtained from the MCMC output. Applying DIC developed earlier to the data-augmented MCMC output leads to

$$\mathrm{DIC}^{DA}=D\left(\overline{\mathbf{z}},\overline{\boldsymbol{\theta}}\right)+2P_D^{DA}, \tag{21}$$

$$\begin{aligned} P_D^{DA}&=\overline{D(\mathbf{z},\boldsymbol{\theta})}-D\left(\overline{\mathbf{z}},\overline{\boldsymbol{\theta}}\right)\\ &=-2\int\left[\ln p(\mathbf{y}|\,\mathbf{z},\boldsymbol{\theta})-\ln p\left(\mathbf{y}|\,\overline{z},\overline{\boldsymbol{\theta}}\right)\right]p(\mathbf{z},\boldsymbol{\theta}|\,\mathbf{y})\mathrm{d}\mathbf{z}\mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{22}$$

where $D(\mathbf{z},\boldsymbol{\theta})=-2\ln p(\mathbf{y}\,|\,\mathbf{z},\theta)$ which is typically available in closed-form. This way of calculating DIC is monitored and implemented in Win-BUGS, following the suggestion of Spiegelhalter et al. (2002). Clearly the use of data augmentation not only facilitates MCMC sampling, but also makes DIC easier to calculate from the MCMC output.

Unfortunately, the data augmentation technique introduces incidental parameters to the model which lead to the incidental parameter problem. This is because, as discussed before, in many latent variable models, the latent variable **z** is often dependent on the sample size and its dimension is the same as or larger than the number of the sample size. As a result, the model becomes non-regular after the parameter space is expanded to $(\boldsymbol{\theta}, \mathbf{z})$. In particular, the ML estimator of **z** is typically inconsistent and the Bayesian large sample theory is invalid for **z**. Although data augmentation makes DIC easy to calculate, it invalidates the asymptotic justification of DIC. DIC based on the data augmentation technique, as calculated in (21) and (22), is no longer asymptotically unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ \ln p\left(\mathbf{y}_{rep}|\,\overline{\boldsymbol{\theta}}(\mathbf{y})\right)\right]$. As a result, for the latent variable model, DIC, as how it is currently monitored and implemented in Win-BUGS, should not be used.

To address this problem, Li et al. (2017b) introduced an integrated DIC (IDIC) which integrates the latent variable out of the deviance and the penalty term. IDIC is given by

$$\text{IDIC} = D\left(\overline{\boldsymbol{\theta}}\right) + 2P_D^I, \tag{23}$$

where

$$P_D^I = \mathbf{tr}\left\{\mathbf{I}\left(\overline{\boldsymbol{\theta}}\right)V\left(\overline{\boldsymbol{\theta}}\right)\right\}, \tag{24}$$

and

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

Li et al. (2017b) showed that under regularity conditions, IDIC is an asymptotically unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ \ln p\left(\mathbf{y}_{rep}|\,\overline{\boldsymbol{\theta}}(\mathbf{y})\right)\right]$.

Similarly, if the loss function the Bayesian predictive distribution, one may obtain an alternative information criterion, which is $IDIC^{BP}$ by Li et al. (2017b) and is defined as

$$\text{IDIC}^{BP} = D\left(\overline{\boldsymbol{\theta}}\right) + (1 + \ \ln 2)P_D^I, \tag{25}$$

As shown in Li et al. (2017a), $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(-2\ln p(\mathbf{y}_{rep}\,|\,\mathbf{y})) = E_{\mathbf{y}}[\text{IDIC}^{BP}] + o(1)$.

5.4 Computing IDIC for latent variable models

For the latent variable model, ln $p(\mathbf{y}\,|\,\boldsymbol{\theta})$ generally does not have an analytical expression. As a result, computing ln $p\left(\mathbf{y}|\,\overline{\boldsymbol{\theta}}\right)$ and P_D^I is not trivial, in sharp contrast to the quantities in (21) and (22). Li et al. (2017b) introduced a very general approach to computing IDIC.

Let

$$
\begin{aligned}
&p\left(\mathbf{y},\mathbf{z}|\,\bar{\boldsymbol{\theta}},b\right)=p\left(\mathbf{y}|\,\mathbf{z},\bar{\boldsymbol{\theta}}\right)^{b}p\left(\mathbf{z}|\,\bar{\boldsymbol{\theta}}\right)\\
&p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}},b\right)=\int p\left(\mathbf{y},\mathbf{z}|\,\bar{\boldsymbol{\theta}},b\right)\mathrm{d}\mathbf{z}=\int p\left(\mathbf{y}|\,\mathbf{z},\bar{\boldsymbol{\theta}}\right)^{b}p\left(\mathbf{z}|\,\bar{\boldsymbol{\theta}}\right)\mathrm{d}\mathbf{z},\\
&p\left(\mathbf{z}|\,\mathbf{y},\bar{\boldsymbol{\theta}},b\right)=\frac{p\left(\mathbf{y},\mathbf{z}|\,\bar{\boldsymbol{\theta}},b\right)}{p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}},b\right)}=\frac{p\left(\mathbf{y}|\,\mathbf{z},\bar{\boldsymbol{\theta}}\right)^{b}p\left(\mathbf{z}|\,\bar{\boldsymbol{\theta}}\right)}{p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}},b\right)},
\end{aligned}
$$

so that

$$
\begin{aligned}
&p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}},1\right)=\int p\left(\mathbf{y}|\,\mathbf{z},\bar{\boldsymbol{\theta}}\right)p\left(\mathbf{z}|\,\bar{\boldsymbol{\theta}}\right)\mathrm{d}\mathbf{z}=\int p\left(\mathbf{y},\mathbf{z}|\,\bar{\boldsymbol{\theta}}\right)\mathrm{d}\mathbf{z}=p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}}\right),\\
&p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}},0\right)=\int p\left(\mathbf{y}|\,\mathbf{z},\bar{\boldsymbol{\theta}}\right)^{0}p\left(\mathbf{z}|\,\bar{\boldsymbol{\theta}}\right)\mathrm{d}\mathbf{z}=\int p\left(\mathbf{z}|\,\bar{\boldsymbol{\theta}}\right)\mathrm{d}\mathbf{z}=1
\end{aligned}
$$

$$
\begin{aligned}
&p\left(\mathbf{z}|\,\mathbf{y},\bar{\boldsymbol{\theta}},1\right)=\frac{p\left(\mathbf{z},\mathbf{y}|\,\bar{\boldsymbol{\theta}},1\right)}{p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}},1\right)}=\frac{p\left(\mathbf{y}|\,\mathbf{z},\bar{\boldsymbol{\theta}}\right)p\left(\mathbf{z}|\,\bar{\boldsymbol{\theta}}\right)}{p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}},1\right)}=\frac{p\left(\mathbf{y}|\,\mathbf{z},\bar{\boldsymbol{\theta}}\right)p\left(\mathbf{z}|\,\bar{\boldsymbol{\theta}}\right)}{p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}}\right)}=p\left(\mathbf{z}|\,\mathbf{y},\bar{\boldsymbol{\theta}}\right),\\
&p\left(\mathbf{z}|\,\mathbf{y},\bar{\boldsymbol{\theta}},0\right)=\frac{p\left(\mathbf{z},\mathbf{y}|\,\bar{\boldsymbol{\theta}},0\right)}{p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}},0\right)}=\frac{p\left(\mathbf{y}|\,\mathbf{z},\bar{\boldsymbol{\theta}}\right)^{0}p\left(\mathbf{z}|\,\bar{\boldsymbol{\theta}}\right)}{p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}},0\right)}=\frac{p\left(\mathbf{z}|\,\bar{\boldsymbol{\theta}}\right)}{1}=p\left(\mathbf{z}|\,\bar{\boldsymbol{\theta}}\right).
\end{aligned}
$$

Using the path sampling technique of Gelman and Meng (1998), Li et al. showed that

$$
\begin{aligned}
\ln p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}}\right)-\ln 1&=\ln\frac{f(1)}{f(0)}=\int_0^1\frac{\partial\ln f(b)}{\partial b}\mathrm{d}b\\
&=\int_0^1 E_{\mathbf{z}|\mathbf{y},\bar{\boldsymbol{\theta}},b}\left[\ln p\left(\mathbf{y}|\,\mathbf{z},\bar{\boldsymbol{\theta}}\right)\right]\mathrm{d}b:=\int_0^1 u(b)\mathrm{d}b,
\end{aligned}
\tag{26}
$$

where $f(b)=p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}},b\right)$ such that $f(1)=p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}}\right)$ and $f\,(0)=1$.

In many cases, $\int_0^1 u(b)\mathrm{d}b$ in (26) does not have an analytical solution. Following Gelman and Meng (1998), we can numerically approximate it using the trapezoidal rule. In particular, we can choose a set of fixed grids $\left\{b_{(s)}=\frac{s}{S}\right\}_{s=0}^{S}$ such that $b_{(0)}=0<b_{(1)}<b_{(2)}<\ldots<b_{(S)}=1$, and then approximate the integral by

$$
\ln p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}}\right)\approx\frac{1}{S}\left(\frac{u(0)}{2}+\sum_{s=1}^{s-1}u(b_s)+\frac{u(1)}{2}\right).
$$

Since $\ln\,p\left(\mathbf{y}|\,\mathbf{z},\bar{\boldsymbol{\theta}}\right)$ often has an analytical expression, $\ln\,p\left(\mathbf{y}|\,\bar{\boldsymbol{\theta}}\right)$ can be conveniently obtained using the above formula.

To compute P_D^I, it mainly needs to evaluate the second derivative of ln $p(\mathbf{y}\,|\,\boldsymbol{\theta})$. Again, the well-known Louis formula suggests that

$$\begin{aligned}\frac{\partial^2 \ln p(\mathbf{y}|\,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} &= E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\left\{\frac{\partial^2 \ln(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\} + Var_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\{S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})\} \\ &= E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\left\{\frac{\partial^2 \ln(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})'\right\} \\ &\quad - E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\{S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})\}E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\{S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})\}'.\end{aligned}$$

Hence, we can use the following formula to calculate the second derivative of the observed-data likelihood function,

$$\begin{aligned}&E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\left\{\frac{\partial^2 \ln(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})'\right\} \\ &\approx \frac{1}{J}\sum_{j=1}^{J}\left\{\frac{\partial^2 \ln\left(\mathbf{y},\mathbf{z}^{(j)}|\,\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + S\left(\mathbf{y},\mathbf{z}^{(j)}|\,\boldsymbol{\theta}\right)S\left(\mathbf{y},\mathbf{z}^{(m)}|\,\boldsymbol{\theta}\right)'\right\}, \\ &E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\{S(\mathbf{y},\mathbf{z}|\,\boldsymbol{\theta})\} \approx \frac{1}{J}\sum_{j=1}^{J} S\left(\mathbf{y},\mathbf{z}^{(j)}|\,\boldsymbol{\theta}\right),\end{aligned}$$

where $\{\mathbf{z}^{(j)}, j=1, 2, \ldots, J\}$ are the MCMC samples.

The main difference between DIC, given in (17) and (18), and IDIC, given in (23) and (24), lies in P_D and P_D^I. To compute P_D, we need to evaluate $E_{\boldsymbol{\theta}|\mathbf{y}}[\ \ln p(\mathbf{y}|\,\boldsymbol{\theta}))] \approx \frac{1}{J}\sum_{j=1}^{J} \ln p\left(\mathbf{y}|\,\boldsymbol{\theta}^{(j)}\right)$. For the latent variable models, without knowing the analytical form of ln $p(\mathbf{y}\,|\,\boldsymbol{\theta})$, computing $\frac{1}{J}\sum_{j=1}^{J} \ln p\left(\mathbf{y}|\,\boldsymbol{\theta}^{(j)}\right)$ is very expensive since one has to evaluate ln p $(\mathbf{y}\,|\,\boldsymbol{\theta}^{(j)})$ for J times with J being large. To compute P_D^I in IDIC, one only needs to compute the second derivative once.

Two well-known classes of latent variable models are the linear Gaussian state space model and the nonlinear non-Gaussian state space model. For these two classes of models, some recursive algorithms, such as the Kalman filter and particle filter algorithms, can be used to facilitate the computation of IDIC. There are existing R packages to implement the Kalman filter and particle filter algorithms; see Tusell (2011). Hence, the proposed method here can be combined with these R packages.

6 Empirical illustrations

In this section, we illustrate the proposed test statistics and model selection criteria using three popular examples in economics and finance. The first example contains asset pricing models with a t error distributions. The likelihood functions of these models not only have the analytical form, but also can be rewritten as in the latent variable form. These two alternative ways of rewritting the models allow us to check the problem in DIC with data augmentation. The second example contains stochastic volatility models, where the volatility is latent. In the second example, the analytical expression of the observed data likelihood does not exist.

6.1 Statistical inference in asset pricing models

Asset pricing models are one of important models in modern finance. There models generally assume 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. Zhou (1993) and Kan and Zhou (2017) suggested to use the multivariate t distribution to replace the multivariate normal distribution. Moreover, on the basis of the efficient market theory, the asset excess premium should not be statistically different from zero. At last, the multivariate t distribution can be rewritten as scale-mixture framework to become a latent variable model. Hence, we consider the following six asset pricing models:

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

where R_t is the excess return of portfolio at period t with $N \times 1$ dimension, $\boldsymbol{F}_t$ a $K \times 1$ vector of factor portfolio excess returns, $\boldsymbol{\alpha}$ a $N \times 1$ vector of intercepts, $\boldsymbol{\beta}$ a $N \times K$ vector of scaled covariances, ϵ_t the random error, $t = 1, 2, \ldots, n$. For convenience, we restrict $\boldsymbol{\Sigma}$ to be a diagonal matrix and ν to be a known constant as $\nu = 3$. It is noted that Model 4 is the scale-mixture distributional representation of Model 3, and Model 5 is the scale mixture distributional representation of Model 6.

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.[a]

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

Making inference for the asset pricing models has attracted a considerable amount of attentions in the empirical asset pricing literature. Avramov and Zhou (2010) provided an excellent review of the literature on Bayesian portfolio analysis. As to Bayesian inference, we need specify the prior distributions for parameters. Here, to represent the prior ignorance, we assign some vague conjugate prior distributions, that is,

$$\alpha_i \sim N[0,100], \beta_{ij} \sim N[0,100], \phi_{ii}^{-1} \sim \Gamma[0.01, 0.01].$$

Based on the R language, we use R2WinBUGS to get the MCMC outputs, and draw 100,000 random observations from the posterior distributions in each model where the first 40,000 is used as the burn-in sample, and the next 60,000 iterations is collected with every 3th observation as effective observations. Hence, these are 20,000 effective observations.

6.1.1 Hypothesis testing for asset pricing models

In asset pricing theory, the efficient market theory suggests that the excess premium $\boldsymbol{\alpha}$ should be zero. Hence, we can write this problem as a hypothesis to be tested as:

$$H_0 : \boldsymbol{\alpha} = 0 \times \mathbf{1}_N, H_1 : \boldsymbol{\alpha} \neq 0 \times \mathbf{1}_N,$$

where $\mathbf{1}_N$ is an N-dimensional vector with unit elements. Model 6 is the most general model which can nest other models, hence, based on this model, we discuss the asset pricing testing problem above.

In Section 4, among of those approaches, we have shown that the threshold values by Bernardo and Rueda (2002) and Li and Yu (2012) are difficult to calibrate. Hence, here, we only consider the statistics respectively developed by (Li et al., 2014, 2015, 2019). Based on 20,000 MCMC samples, we calculate the three test statistics, $\mathbf{T}_{LZY}$ ($\mathbf{y}$, $\boldsymbol{\vartheta}_0$), $\mathbf{T}_{LLY}$ ($\mathbf{y}$, $\boldsymbol{\vartheta}_0$) and $\mathbf{T}_{LLYZ}$ ($\mathbf{y}$, $\boldsymbol{\vartheta}_0$). We report the results in Table 2.

Obviously, from these results, according to the critical values from $\chi^2(25)$, under 5% significant level, all the test statistics reject the null hypothesis. Hence, we can conclude that the mean–variance efficiency does not held in practice. As to these test statistics, more details, one can refer to Li et al. (2014, 2015, 2019). At last, according to the Savage-Dickey Density Ratio approach by Verdinelli and Wasserman (1995), it can be shown that.

$\hat{BF} = 1.069$ which provide mild evidence to support H_0 which is contractive to the results from the hypothesis testing statistics. This reason lies that in this section, we use the vague prior to do the hypothesis testing so that BFs suffer from the Jeffreys-Lindley's paradox. It should be very suggested to use BFs for doing hypothesis testing when the prior information is not available. More details about the Jeffreys-Lindley's paradox, see the discussion by Li et al. (2015).

TABLE 2 Asset pricing testing in M_6

Hypothesis	$\alpha = 0$
$\mathbf{T}_{LZY}(\mathbf{y}, \vartheta_0)$	140.5191
$\mathbf{T}_{LLY}(\mathbf{y}, \vartheta_0)$	153.5680
$\mathbf{T}_{LLZY}(\mathbf{y}, \vartheta_0)$	184.4315

6.1.2 Specification testing for asset pricing models

In this subsection, we take the standard Fama–French three-factor asset pricing model (Fama and French, 1993) that is, model 2 as an example for illustrating the proposed approach. The standard asset pricing model is given by

$$Model2 : R_t = \alpha + \beta_1 R_{mt} + \beta_2 SMB_t + \beta_3 HML_t + \epsilon_t, \epsilon_t \sim N[\mathbf{0}, \boldsymbol{\Sigma}]$$

where R_m is the excess market return, SMB stands for "Small [market capitalization] Minus Big" and $H\,ML$ for "High [book-to-market ratio] Minus Low"; they measure the historic excess returns of small caps over big caps and of value stocks over growth stocks.

Here, for checking the model misspecification, the expanded model can be specified as

$$Model2E : R_t = \boldsymbol{\alpha} + \boldsymbol{\beta}_1 R_{mt} + \boldsymbol{\beta}_{1E} R_{mt}^2 + \boldsymbol{\beta}_2 SMB_t + \boldsymbol{\beta}_3 HML_t + \epsilon_t, \epsilon_t \sim N[\mathbf{0}, \boldsymbol{\Sigma}]$$

Hence, according to Section 4, we can write this model misspecification problem as a hypothesis to be tested as:

$$H_0 : \beta_{1E} = 0, H_1 : \beta_{1E} \neq 0$$

Following Section 4, the proposed test statistic can be given by

$$\mathrm{BMT} = \mathbf{tr}\left\{C_E[\mathbf{y}, ((\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\Sigma}), \boldsymbol{\beta}_{1E} = 0)] V_E\left(\overline{\boldsymbol{\theta}}_L\right)\right\} + \sqrt{1021}(\mathrm{BIMT}/125 - 1)^2$$

Hence, based on 20,000 effective observation drawn from the posterior distribution, we can compute the corresponding statistics which are reported in Table 3. It is noted that if the model is correctly specified, *BMT* converges to $\chi^2(25)$ distribution. Given this χ^2 distribution, under 0.05 significant level, the critical value is 37.65. Hence, according to the table, we can conclude that *BMT* strongly reject the null hypothesis which means that the asset price model is misspecified (Table 3).

6.1.3 Model comparison for asset pricing models

We make a model comparison of these asset pricing models. Based on 20,000 effective observations, we calculate DICs, and BFs. Table 4 reports P_D, P_D^{DA},

TABLE 3 Results of specification test for model 2

Item	Value
BIMT	610
$\mathbf{tr}\left\{C_E[\mathrm{y}, ((\alpha, \beta_1, \beta_2, \beta_3, \Sigma), \beta_{1E} = 0)] V_E\left(\bar{\boldsymbol{\theta}}_{\mathrm{L}}\right)\right.$	444
$\sqrt{1021}(\mathrm{BIMT}/125 - 1)^2$	481
BMT	925

TABLE 4 Model selection results for Fama–French three factor models

Model	M_1	M_2	M_3	M_4	M_5	M_6
# of Parameters	100	125	100	100	125	125
P_D	100	125	100	100	125	125
DIC	−119,842	−119,880	−133,088	−133,088	−133,202	−133,202
DIC^{BP}	−119,872	−119,918	−133,118	−133,118	−133,240	−133,240
P_D^{DA}	—	—	—	1021	—	1046
DIC^{DA}	—	—	—	−134,777	—	−134,897
P_D^I	100	125	100	100	126	126
IDIC	−119,842	−119,880	−133,087	−133,087	−133,201	−133,201
IDIC^{BP}	−119,873	−119,918	−133,118	−133,118	−133,240	−133,240

P_D^I, DIC, DIC^{BP}, DIC^{DA}, IDIC, and IDIC^{BP} for all six models. Note that only M_4 and M_6 has the latent variable so that P_D^{DA} and DIC^{DA} are only reported for these two models. Furthermore, M_3 and M_4 are the same model with different distribution expression, M_5 and M_6 are the same model with different distribution expression. Hence, as to the same model with different distribution expression, P_D, P_D^I, DIC, DIC^{BP}, IDIC, and IDIC^{BP} are equal for the same model.

From Table 4, we can get some interesting finding. First, as expected, DIC^{DA} in Model 3 is quite different from that in Model 4 although these two models are the same, but only have different distribution expression. The main reason is that in Model 4, the scale-mixture specification is used and, hence, a sequence of latent variables, $\{\omega_t\}$ are treated as parameters. For the same reason, DIC^{DA} in Model 5 is quite different from that in Model 6. As argued earlier, this conceptual difficulty is due to lack of the theoretical foundation. Second, DIC, DIC^{BP}, IDIC, and IDIC^{BP} do not suffer from the same difficulty as DIC^{DA}. For Model 3 (and Model 5), they are identical to those for Model 4 (and Model 6). Third, the theoretical results show that P_D and P_D^I should be close to the actual number of the parameters, P, if the posterior distribution is well approximated by the normal distribution and the use of uninformative priors is used. The results can be confirmed from this table. Most importantly, we see that P_D is almost identical to P_D^I in all models. Not surprisingly, DIC and IDIC are almost the same in all models and DIC^{BP} and IDIC^{BP} are almost the same. This confirm the theoretical result that P_D and P_D^I can be well approximated. In addition, all DICs provide the evidence to support M_6 is the best model for prediction among these six models.

In addition, as to P_D and P_D^I, we need point out that in terms of the computational cost, for Models 3 and 5, P_D^I can require less efforts than P_D. The reason is that P_D involves $\int \ln p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$, which is approximated by $\frac{1}{J}\sum_{j=1}^{J} \ln p\left(\mathbf{y}|\boldsymbol{\theta}^{(j)}\right)$. This quantity is much more expensive to compute because it requires numerical evaluation of $\ln p(\mathbf{y}|\boldsymbol{\theta}^{(j)})$ for J times. For example, here, based on the 20,000 posterior random observations, one has to evaluate $\ln p(\mathbf{y}|\boldsymbol{\theta}^{(j)})$ 20,000 times. Fortunately, as to asset pricing models, $\ln p(\mathbf{y}|\boldsymbol{\theta}^{(j)})$ has closed-form. However, as to other models such that $\ln p(\mathbf{y}|\boldsymbol{\theta})$ does not have analytical form, obviously, IDIC is more advantageous than DIC.

At last, in order to check the reliability of the general computation approach by Section 5.4, we take model 6 as an example. Since the likelihood function $\ln p(\mathbf{y}|\boldsymbol{\theta})$ has analytical form, we can easily get that $D(\overline{\boldsymbol{\theta}}) = -133452$. Using the approximation approach in Section 5.4, we give the approximated value of $D(\overline{\boldsymbol{\theta}})$, that is, $\hat{D}(\overline{\boldsymbol{\theta}})$ under different grids and report the results in the Table 5. From this table, it can be observed that with the increasing grid S, the proposed approach can approximate $D(\overline{\boldsymbol{\theta}})$ very well.

TABLE 5 The approximated value of $D(\overline{\boldsymbol{\theta}})$ based on Section 5.4

Hypothesis	$\hat{D}(\overline{\boldsymbol{\theta}})$
$S = 200$	−133,436
$S = 400$	−133,437
$S = 800$	−133,451
$S = 900$	−133,452

6.2 Statistical inference in stochastic volatility models

Stochastic volatility (SV) models are one of the important models to model the time-varying volatility in financial econometrics. The basic SV model is composed of two equations, one is measurement equation, the other is state equation where the logarithmic volatility is the state variable which is often assumed to follow an AR(1) model. The basic form can be written as

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

where $t=1, 2, \ldots, n$, y_t is the continuously compounded return, h_t the unobserved log-volatility, $h_0=\mu$, u_t, and v_t are independent for all t. In this chapter, we denote this model by M_1.

An important and well documented empirical feature in many financial time series is the leverage effect (Black, 1976). Hence, following Yu (2005), a fundamental extension of the basic SV model is to incorporate the leverage effect. The leverage effect SV model can be defined as:

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

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 $h_0=\mu$. In this model, ρ captures the leverage effect if $\rho < 0$. In the empirical literature, there is a negative relationship between the expected future volatility and the current return. We denote this model as M_2.

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

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

This type prior can be regarded as a noninformative prior to represent the prior ignorance.

The dataset consists of 945 daily mean-corrected returns on Pound/−Dollar exchange rates, covering the period between 01/10/81 and 28/06/85. Here, using R language, we use R2WinBUGS to run MCMC to get the outputs. After a burn-in period of 10,000 iterations, we save every 20th value for the next 100,000 iterations to get 5000 effective draws. The same dataset was used in Kim et al. (1998) and Meyer and Yu (2000). The posterior mean and standard error of parameters in the two competing model are reported in Table 6. Note that the in M_2, the posterior mean of ρ is very close to zero, relative to its posterior standard error.

TABLE 6 Posterior mean and standard error of parameters in M_1 and M_2

Parameter	M_1		M_2	
	Mean	*SE*	*Mean*	*SE*
μ	−0.6733	0.3282	−0.6485	0.3377
φ	0.9733	0.0127	0.9802	0.0138
ρ	—	—	−0.0575	0.1570
τ	0.1698	0.0378	0.1661	0.0391

6.2.1 Hypothesis testing for stochastic volatility models

In this chapter, the hypothesis that we are concerned can be expressed as:

$$H_0 : \rho = 0,\ H_1 : \rho \neq 0$$

Here, ρ is the interest parameter, the nuisance parameter is denoted by $\boldsymbol{\psi} = (\mu, \varphi, \tau^{-2})$, $\boldsymbol{\theta} = (\rho, \boldsymbol{\psi}) = \rho, (\mu, \varphi, \tau^2)$. Again, based on 20,000 effective observation, we calculate the three test statistic, that is, $\mathbf{T}_{LZY}$ ($\mathbf{y}$, $\boldsymbol{\vartheta}_0$), $\mathbf{T}_{LLY}$ ($\mathbf{y}$, $\boldsymbol{\vartheta}_0$), and $\mathbf{T}_{LLYZ}$ ($\mathbf{y}$, $\boldsymbol{\vartheta}_0$). We report all the results in Table 7.

From this table, according to the critical values calibrated from their asymptotic distribution, under 5% significant level, all three test statistics fail to reject the null hypothesis. The result is correspond with estimation result, that is, $\rho = -0.0575$. Furthermore, this provide enough evidence to support that leverage effect in this exchange data is not obvious.

6.2.2 Specification testing for SV models

The dataset used here contains the daily returns on AUD/USD exchange rates from January 2005 to December 2012. Following a suggestion of a referee,

TABLE 7 Hypothesis hypothesis results for the leverage effect

Hypothesis	$\rho = 0$
$\mathbf{T}^*_{LZY}(\mathbf{y}, \boldsymbol{\vartheta}_0)$	−0.6870
$\mathbf{T}_{LLY}$ ($\mathbf{y}$, $\boldsymbol{\vartheta}_0$)	0.1659
$\mathbf{T}_{LLZY}$ ($\mathbf{y}$, $\boldsymbol{\vartheta}_0$)	1.7050

before we apply BMT to the SV model, we first test the *i.i.d.* normal model with constant mean and constant variance given by

$$y_t = \alpha + \varepsilon_t, \varepsilon_t \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right) \tag{27}$$

An AR(1) model is used as the expanded model

$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t, \varepsilon_t \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right). \tag{28}$$

The Bayesian MCMC method is implemented to estimate the parameters with the following vague prior

$$\alpha \sim N\left(0, 100\sigma^2\right), \beta \sim N\left(0, 100\sigma^2\right), \sigma^{-2} \sim \Gamma(0.001, 0.001).$$

For the above two models, we draw 20,000 MCMC samples from the posterior distribution and compute BMT.

The critical value of $\chi^2(1)$ is 6.63 at the 1% significance level. BMT is 251.52, rejecting the *i.i.d.* normal model. This conclusion is not surprising as the volatility of stock returns is stochastic. However, J_1 is 0.2858 (i.e., $J_0 = 251.23$) which is less than the critical value of $\chi^2(1)$. Using J_1 alone only suggests that we cannot reject $\beta = 0$ in Model (28). This conclusion is also not surprising as the weekly returns have very weak serial correlations.

Next, we change the null model to the following basic SV model,

$$\begin{aligned} y_t &= \alpha + \exp(h_t/2) u_t, u_t \overset{i.i.d.}{\sim} N(0, 1), \\ h_t &= \mu + \phi(h_{t-1} - \mu) + \tau v_t, v_t \overset{i.i.d.}{\sim} N(0, 1). \end{aligned} \tag{29}$$

The expanded model is as follows:

$$\begin{aligned} y_t &= \alpha + \beta_1 y_{t-1} + \exp(h_t/2) u_t, u_t \overset{i.i.d.}{\sim} N(0, 1). \\ h_t &= \mu + \phi(h_{t-1} - \mu) + \tau v_t, v_t \overset{i.i.d.}{\sim} N(0, 1). \end{aligned} \tag{30}$$

The following vague priors are used

$$\begin{aligned} &\alpha \sim N(0, 100), \ \ \phi \sim Beta(1, 1), \\ &\tau^{-2} \sim \Gamma(0.001, 0.001), \ \ \beta_1 \sim N(0.5, 100). \end{aligned}$$

To obtain BMT, we draw 110,000 MCMC samples from the posterior distribution and discard the first 10,000 as burning-in observations, and store the remaining samples as effective observations in both models. In this case, BMT = 0.4279 which is less than the critical value of $\chi^2(1)$, suggesting that the basic SV model is not misspecified.

6.2.3 *Model comparison of SV models*

Hence, we consider the model comparison of these two models. Since the models are of a nonlinear non-Gaussian form and both $p(\mathbf{y}|\boldsymbol{\theta})$ are not available in closed-form, the approach provided in Section 5 is implemented to compute

TABLE 8 Model selection results for M_1 and M_2

Model	M_1	M_2
P_D^{DA}	53.60	31.33
$D(\mathbf{z},\overline{\boldsymbol{\theta}})$	1695.40	1693.36
DIC^{DA}	1802.52	1756.21
P_D^I	2.32	3.24
$D(\overline{\boldsymbol{\theta}})$	1837.81	1837.78
IDIC	1842.50	1844.30
IDIC^{BP}	1841.80	1843.30
BF_{21}	0.2174	

DICs, and the Savage Dickey density ratio (Verdinelli and Wasserman, 1995) is implemented to calculate BFs. Hence, DIC requires tedious computational efforts. Here, we only report the results of DIC^{DA}, IDIC, P_D^{DA}, P_D^I, and BFs in Table 8.

From this table, we can get the following findings. First, DIC^{DA} and IDIC suggest different rankings of the competing models where DIC^D suggests that M_2 is better that M_1, IDIC and IDIC^{BP} both suggest M_1. According to DIC^{DA}, it can be observed that M_1 and M_2 perform nearly the same judged by the model fit term, $D(\mathbf{z},\overline{\boldsymbol{\theta}})$. However, M_2 reduces P_D^7 by 22.3 over M_1. This reduction of the model complexity is the reason why DIC^{DA} prefers M_2. This result is surprising as the posterior mean of the leverage effect is nearly zero as reported in Table 8 and not accord with the hypothesis testing results. Obviously, as to SV models, when the latent variable is regarded as parameters, the number of parameters exceeds the number of observations, say $n+3$ in M_1 and $n+4$ in M_1. Hence, an important season to lead the surprising results lie that DIC^{DA} is lack of rigorously theoretical foundation and should be cautious to be used in practice although its computation is simple.

Second, IDIC and IDIC^{BP} both suggest that M_1 is slightly better that M_2 although the difference is not large. In IDIC, P_D^I is 2.32 in M_1 and 3.24 in M_2. These values are very close to the actual numbers of parameters in the two models. It is noted that M_2 has one extra parameter so that this difference is reasonable. Moreover, M_1 and M_2 perform nearly the same judged by $D(\overline{\boldsymbol{\theta}})$. These findings give the reason why M_1 is slightly better that M_2. Third, BFs suggest that M_1 is the better model, consistent with the ranking of IDIC. This empirical example clearly demonstrates that IDIC is a more reliable model selection criterion that DIC^{DA}. In addition, although IDIC and IDIC^{BP} both select the basic SV model, they imply that different predictive distribution should be used. From the theoretical analysis, as to predictive problem, the model selection results suggest that the basic SV model with Bayesian predictive distribution should be used because this decision can yield smallest risk asymptotically when M_1, M_2, plug-in predictive distribution and Bayesian predictive distribution are candidate use.

7 Concluding remarks

In this chapter, instead of making refinements for BFs, we overviews some alterative approaches developed in the recent literature for hypothesis testing and model selection methods. The approaches are established after the MCMC output is available. We show that these approaches not only have good theoretical properties, but also, do not require tedious additional computational efforts. Hence, with the advance of MCMC techniques and expanding computing facility, these approaches can be applied into a variety of complex models, especially latent variable models.

As to the hypothesis testing, we overviews several statistics for hypothesis testing which can be regarded as the MCMC version of the "trinity" of test statistics widely used in the frequentist domain, namely, LR test, LM test, and Wald test. Their asymptotic distributions are discussed based on a set of regular conditions. Furthermore, we overview the well-known DIC and its extensions. The asymptotic property of DICs are also discussed compared with AIC. At last, we illustrate the methods using econometric models with real data, some of which involve latent variables. The implementation is illustrated by R code with the MCMC output obtained by R2WinBUGS.

References

Avramov, D., Zhou, G., 2010. Bayesian portfolio analysis. Annu. Rev. Financ. Econ. 2, 25–47.

Berger, J.O., Perrichi, L.R., 1996. The intrinsic Bayes factor for model selection and prediction. J. Am. Stat. Assoc. 91, 109–122.

Bernardo, J.M., Rueda, R., 2002. Bayesian hypothesis testing: a reference approach. Int. Stat. Rev. 70, 351–372.

Bernardo, J.M., Smith, A.F.M., 2006. Bayesian Theory, second ed. John Wiley & Sons Canada, Limited, Chichester.

Black, F., 1976. Studies of stock market volatility changes. Proc. Am. Stat. Assoc. Bus. Econ. Stat. Sec. 177–181.

Breusch, T.S., Pagan, A.R., 1980. The Lagrange multiplier test and its applications to model specification in econometrics. Rev. Econ. Stud. 47, 239–253.

Chen, X., Christensen, T.M., O'Hara, K., Tamer, E., 2016. MCMC Confidence Sets for Identified Sets. Working Paper, Yale University.

Chernozhukov, V., Hong, H., 2003. An MCMC approach to classical estimation. J. Econ. 115 (2), 293–346.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39 (1), 1–38.

Engle, R.F., 1984. Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. Handb. Econ. 2, 775–826.

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33, 3–56.

Fan, J., Liao, Y., Yao, J., 2015. Power enhancement in high-dimensional cross-sectional tests. Econometrica 83 (4), 1497–1541.

Gelman, A., Meng, X.-L., 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Stat. Sci. 13, 163–185.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2013. Bayesian Data Analysis, third ed. Chapman and Hall/CRC.

Han, C., Carlin, B.P., 2001. Markov chain Monte Carlo methods for computing Bayes factor: a comparative review. J. Am. Stat. Assoc. 96, 1122–1132.

Jeffreys, H., 1961. Theory of Probability, third ed. Oxford University Press, Oxford.

Kan, R., Zhou, G., 2017. Modeling non-normality using multivariate t: implications for asset pricing. China Financ. Rev. Inter. 7, 2–32.

Kass, R.E., Raftery, A.E., 1995. Bayes factor. J. Am. Stat. Assoc. 90, 773–795.

Kim, S., Shephard, N., Chib, S., 1998. Stochastic volatility: Likelihood inferenceand comparison with ARCH models. Rev. Econ. Stud. 65, 361–393.

Lancaster, T., 2000. The incidental parameter problem since 1948. J. Econ. 95 (2), 391–413.

Li, Y., Yu, J., 2012. Bayesian hypothesis testing in latent variable models. J. Econ. 166, 237–246.

Li, Y., Zeng, T., Yu, J., 2014. A new approach to Bayesian hypothesis testing. J. Econ. 178, 602–612.

Li, Y., Liu, X.B., Yu, J., 2015. A Bayesian chi-squared test for hypothesis testing. J. Econ. 189, 54–69.

Li, Y., Yu, J., Zeng, T., 2017a. Deviation Information Criterion: Justification and Variation. Working paper. Singapore Management University.

Li, Y., Yu, J., Zeng, T., 2017b. Integrated Deviation Information Criterion for Latent Variable Models. Working paper. Singapore Management University.

Li, Y., Yu, J., Zeng, T., 2018. Specification tests based on MCMC output. J. Econometrics 207, 237–260.

Li, Y., Liu, X.B., Yu, J., Zeng, T., 2019. A posterior-based Wald-type statistic for hypothesis testing. Working paper. School of Economics, Singapore Management University.

Louis, T.A., 1982. Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. Ser. B 44, 226–233.

O'Hagan, A., 1991. Discussion of posterior Bayes factors by M. Aitkin. J. R. Stat. Soc. Ser. B 53, 136.

O'Hagan, A., 1995. Fractional Bayes factors for model comparison, (with discussion). J. R. Stat. Soc. Ser. B 57, 99–138.

Martin, A.D., Quinn, K.M., 2005. MCMCpack 0.6-6. http://mcmcpack.wustl.edu/.

Meyer, R., Yu, J., 2000. BUGS for a Bayesian analysis of stochastic volatility models. Econ. J. 3, 198–215.

Neyman, J., Scott, E.L., 1948. Consistent estimates based on partially consistent observations. Econometrica 16, 1–32.

Poirier, D.J., 1995. Intermediate Statistics and Econometrics: A Comparative Approach. The MIT Press.

Presnell, B., Boos, D.D., 2004. The IOS test for model misspecification. J. Am. Stat. Assoc. 99 (465), 216–227.

Robert, C., 1993. A note on Jeffreys-Lindley paradox. Stat. Sin. 3, 601–608.

Robert, C., 2001. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, second ed. Springer Texts in Statistics.

Spiegelhalter, D., Best, N.G., Carlin, B., van der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B 64, 583–639.

Spiegelhalter, D., Thomas, A., Best, N.G., Lunn, D., 2003. WinBUGS User Manual. Version 1.4. MRC Biostatistics Unit, Cambridge, England.

Sturtz, S., Ligges, U., Gelman, A., 2005. R2WinBUGS: a package for running WinBUGS from R. J. Stat. Softw. 39 (3), 1–16.

Tanner, T.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. 82, 528–540.

Tusell, F., 2011. Kalman filtering in R. J. Stat. Softw. 12 (2), 1–27.

Verdinelli, I., Wasserman, L., 1995. Computing Bayes factors using a generalization of the Savage-Dickey density. J. Am. Stat. Assoc. 90 (430), 614–618.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–25.

Yu, J., 2005. On leverage in a stochastic volatility models. J. Econ. 127, 165–178.

Zhou, G., 1993. Asset-pricing tests under alternative distributions. J. Financ. 48, 1927–1942.

Further reading

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory, vol. 1. Springer Verlag, pp. 267–281.

Chen, C.F., 1985. On asymptotic normality of limiting density function with Bayesian implications. J. R. Stat. Soc. Ser. B 47, 540–546.

Chib, S., 1995. Marginal likelihood from the Gibbs output. J. Am. Stat. Assoc. 90, 1313–1321.

Chib, S., Jeliazkov, I., 2001. Marginal likelihood from the Metropolis-Hastings output. J. Am. Stat. Assoc. 96, 270–281.

Creal, D., 2012. A survey of sequential Monte Carlo methods for economics and finance. Econ. Rev. 31, 245–296.

Doucet, A., Johansen, A.M., 2011. A tutorial on particle filtering and smoothing: fifteen years later. In: Crisan, D., Boris, R. (Eds.), The Oxford Handbook of Nonlinear Filtering. Oxford University Press.

Doucet, A., Shephard, N., 2012. Robust Inference on Parameters via Particle Filters and Sandwich Covariance Matrices. Working Paper, Harvard University.

Geweke, J., 2007. Bayesian model comparison and validation. Am. Econ. Rev. 97, 60–64.

Kadane, J.B., Lazar, N.A., 2004. Methods and criteria for model selection. J. Am. Stat. Assoc. 99 (465), 279–290.

Shephard, N., 2005. Stochastic Volatility: Selective Readings. Oxford University Press.

Spiegelhalter, D., Best, N., Carlin, B., van der Linde, A., 2014. The deviance information criterion: 12 years on. J. R. Stat. Soc. Ser. B 76, 485–493.

Vehtari, A., Ojanen, J., 2012. A survey of Bayesian predictive methods for model assessment, selection and comparison. Stat. Surv. 6, 142–228.