Journal of Econometrics 189 (2015) 54–69

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

A Bayesian chi-squared test for hypothesis testing[☆]

Yong Li[a,*], Xiao-Bin Liu[b], Jun Yu[b,c]

[a] Hanqing Advanced Institute of Economics and Finance, Renmin University of China, Beijing, 100872, PR China
[b] School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903, Singapore
[c] Lee Kong Chian School of Business, Singapore Management University, 90 Stamford Road, Singapore 178903, Singapore

ARTICLE INFO

Article history:
Received 2 June 2014
Received in revised form
24 January 2015
Accepted 21 June 2015
Available online 7 July 2015

JEL classification:
C11
C12

Keywords:
Bayes factor
Decision theory
EM algorithm
Lagrange multiplier
Markov chain Monte Carlo
Latent variable models

ABSTRACT

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

1. Introduction

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

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

In the first class, refinements are made to the BF to overcome the theoretical and computational difficulties. For example, to reduce the influence of the prior on the BF, one may split the data into two parts, a training sample and a sample for statistical analysis. The training sample is used to update the non-informative prior and to obtain a new proper informative prior, as in the fractional BF (O'Hagan, 1995). In practice, however, this strategy is not always satisfactory because it relies on an arbitrary division of the data. To alleviate this difficulty, Berger and Perrichi (1996) proposed the so-called intrinsic BF which is based on the minimal training

☆ Li gratefully acknowledges the financial support of the Chinese Natural Science Fund (No. 71271221), Program for New Century Excellent Talents in University, Beijing Higher Education Young Elite Teacher Project and Beijing Social Sciences Federation Young Teacher Project (No. 2014SKL020). Yu would like to acknowledge the financial support from Singapore Ministry of Education Academic Research Fund Tier 2 under the grant number MOE2011-T2-2-096 and Tier 3 under the grant number MOE2013-T3-1-017. We would like to acknowledge helpful comments from the co-editor, an associate editor, two referees, Tao Zeng and the seminar and conference participants at Peking University, Princeton University, the inaugural China Meeting of the Econometric Society.

* Corresponding author.
E-mail addresses: gibbsli@ruc.edu.cn (Y. Li), yujun@smu.edu.sg (J. Yu).
URL: http://www.mysmu.edu/faculty/yujun/ (J. Yu).

http://dx.doi.org/10.1016/j.jeconom.2015.06.021
0304-4076/

sample that results in proper posteriors. In general, the minimal training sample is not unique. Hence, the intrinsic BF is obtained by averaging the partial BFs calculated from all possible minimal training samples. Unfortunately, the intrinsic BF is computationally demanding, especially for latent variable models. O'Hagan (1995) discussed properties of the fractional and the intrinsic BFs.

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

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

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

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

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

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

2. Bayesian hypothesis testing under decision theory

2.1. Testing a point null hypothesis

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

$$\begin{cases} H_0: & \boldsymbol{\theta} = \boldsymbol{\theta}_0 \\ H_1: & \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \end{cases}. \tag{1}$$

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

$$\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0) = \int_\Theta \int_\Psi \{\mathcal{L}[d_0, (\boldsymbol{\theta}, \boldsymbol{\psi})] - \mathcal{L}[d_1, (\boldsymbol{\theta}, \boldsymbol{\psi})]\}\, p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\psi} > c \geq 0,$$

where $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ is a Bayesian test statistic; $p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y})$ is the posterior distribution with some given prior $p(\boldsymbol{\theta}, \boldsymbol{\psi})$; c is a threshold value. Let $\triangle\mathcal{L}[H_0, (\boldsymbol{\theta}, \boldsymbol{\psi})] = \mathcal{L}[d_0, (\boldsymbol{\theta}, \boldsymbol{\psi})] - \mathcal{L}[d_1, (\boldsymbol{\theta}, \boldsymbol{\psi})]$ be the net loss difference function which can generally be used to measure the evidence against H_0 as a function of $(\boldsymbol{\theta}, \boldsymbol{\psi})$. Hence, the Bayesian test statistic can be rewritten as

$$\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0) = E_{\boldsymbol{\vartheta}|\mathbf{y}} \left(\triangle\mathcal{L}[H_0, (\boldsymbol{\theta}, \boldsymbol{\psi})]\right).$$

2.2. A literature review

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

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

where $M_0 := \{p(\mathbf{y}|\boldsymbol{\theta}_0, \boldsymbol{\psi}), \boldsymbol{\psi} \in \boldsymbol{\Psi}\}$ is the model under the null; $M_1 := M$ is the model under the alternative. The two marginal likelihood functions are defined as

$$p(\mathbf{y}|M_0) = \int_{\Psi} p(\mathbf{y}|\boldsymbol{\theta}_0, \boldsymbol{\psi}) p(\boldsymbol{\psi}|\boldsymbol{\theta}_0) \mathrm{d}\boldsymbol{\psi},$$

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

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

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

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

$$\text{reject } H_0 \quad \text{iff} \quad BF_{01} = \frac{\int_{\Psi} p(\mathbf{y}|\boldsymbol{\theta}_0, \boldsymbol{\psi}) p(\boldsymbol{\psi}|\boldsymbol{\theta}_0) \mathrm{d}\boldsymbol{\psi}}{\int_{\Theta} \int_{\Psi} p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\psi}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\psi}} < 1.$$

Remark 2.1. The BF has several disadvantages. If the Jeffreys or the reference prior (Jeffreys, 1961) is used to reflect the objectiveness, the BF is not well-defined since it depends on an arbitrary constant (Bernardo and Rueda, 2002). In addition, if a proper prior with a large spread is used to represent the prior ignorance, the BF has a tendency to favor the null hypothesis, giving rise to Jeffreys–Lindley's paradox; see Poirier (1995) and Robert (1993, 2001). Moreover, for many models in economics, such as latent variable models and the DSGE models, the marginal likelihood and, hence, the BF are very difficult to evaluate; see Han an Carlin (2001) for a good review of methods for calculating the BF from the MCMC output.

Bernardo and Rueda (2002) suggested using a continuous loss function based on the KL divergence,

$$KL[p(x), q(x)] = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x, \tag{2}$$

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

$$\begin{aligned} \mathbf{T}_{BR}(\mathbf{y}, \boldsymbol{\theta}_0) = E_{\boldsymbol{\vartheta}|\mathbf{y}} (\min \{ & KL[p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi}), p(\mathbf{y}|\boldsymbol{\theta}_0, \boldsymbol{\psi})], \\ & KL[p(\mathbf{y}|\boldsymbol{\theta}_0, \boldsymbol{\psi}), p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi})]\}). \end{aligned} \tag{3}$$

Remark 2.2. It is shown in Bernardo and Rueda (2002) that $\mathbf{T}_{BR}(\mathbf{y}, \boldsymbol{\theta}_0)$ is well-defined under improper distributions. This is an important advantage over the BF. However, the BR test is not without its problems. First, the KL divergence function often does not have a closed-form expression. Consequently, $\mathbf{T}_{BR}(\mathbf{y}, \boldsymbol{\theta}_0)$ may be difficult to compute. Second, BR suggested choosing threshold values based on the normal distribution to implement the test. Unfortunately, the choice of the normal distribution and, hence, the threshold values are subjective and lack rigorous statistical justifications. A different distribution will lead to different threshold values.

To alleviate the computational problems of $\mathbf{T}_{BR}(\mathbf{y}, \boldsymbol{\theta}_0)$ in the context of latent variable models, Li and Yu (2012) proposed a new loss difference function, based on the $\mathcal{Q}$ function used in the EM algorithm (Dempster et al., 1977). Let $\mathbf{z} = (z_1, z_2, \ldots, z_n)'$ denote the latent variables and $\mathbf{x} = (\mathbf{y}', \mathbf{z}')'$. Let $p(\mathbf{y}|\boldsymbol{\vartheta})$ and $p(\mathbf{x}|\boldsymbol{\vartheta})$ $(:= p(\mathbf{y}, \mathbf{z}|\boldsymbol{\vartheta}))$ be the observed data likelihood function and the complete data likelihood function, respectively. The relationship between these two likelihood functions is

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

For any $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$, the $\mathcal{Q}$ function is:

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

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

$$\mathcal{Q}(\boldsymbol{\vartheta}_1|\boldsymbol{\vartheta}_2) \approx \frac{1}{G} \sum_{g=1}^{G} \log p\left(\mathbf{y}, \mathbf{z}^{(g)}|\boldsymbol{\vartheta}_1\right),$$

where $\{\mathbf{z}^{(g)}, g = 1, 2, \ldots, G\}$ are the effective MCMC draws from the posterior distribution $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\vartheta}_2)$. Let $\boldsymbol{\vartheta}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\psi})$. Li and Yu (2012) defined a new continuous net loss difference function as:

$$\triangle\mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}_0) = \{\mathcal{Q}(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}) - \mathcal{Q}(\boldsymbol{\vartheta}_0, \boldsymbol{\vartheta})\} + \{\mathcal{Q}(\boldsymbol{\vartheta}_0, \boldsymbol{\vartheta}_0) - \mathcal{Q}(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}_0)\},$$

and proposed a Bayesian test statistic as:

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

Remark 2.3. It is shown in Li and Yu (2012) that the test statistic, $T_{LY}(\mathbf{y}, \boldsymbol{\theta}_0)$, is well-defined under improper priors and also easy to compute. However, this test statistic has some practical disadvantages. First, like the test statistic of BR, some threshold values have to be specified. Following the idea of McCulloch (1989), Li and Yu (2012) proposed to choose threshold values based on the Bernoulli distribution. Unfortunately, the choice of the Bernoulli distribution is arbitrary. If another distribution is used, the threshold values will be different. Second, it is not clear whether this test statistic is immune to Jeffreys–Lindley's paradox.

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

$$\triangle\mathcal{L}[H_0, (\boldsymbol{\theta}, \boldsymbol{\psi})] = 2 \log p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi}) - 2 \log p(\mathbf{y}|\boldsymbol{\theta}_0, \boldsymbol{\psi}),$$

$$\begin{aligned} \mathbf{T}_{LZY}(\mathbf{y}, \boldsymbol{\theta}_0) &= 2 \int [\log p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi}) - \log p(\mathbf{y}|\boldsymbol{\theta}_0, \boldsymbol{\psi})] \\ &\quad \times p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\psi}. \end{aligned} \tag{4}$$

$\mathbf{T}_{LZY}$ can be understood as the Bayesian version of the likelihood ratio test. However, for latent variable models, the likelihood function $p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi})$ generally is not available in closed-form. To achieve computational tractability, under some regularity conditions, Li et al. (2014) gave an asymptotically equivalent form for $\mathbf{T}_{LZY}(\mathbf{y}, \boldsymbol{\theta}_0)$, i.e.,

$$\begin{aligned} \mathbf{T}^*_{LZY}(\mathbf{y}, \boldsymbol{\theta}_0) &= 2D + 2\left[\log p(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\psi}}) - \log p(\bar{\boldsymbol{\psi}}|\boldsymbol{\theta}_0)\right] \\ &\quad - 2\left[\int \log p(\boldsymbol{\theta}|\boldsymbol{\psi}) p(\boldsymbol{\vartheta}|\mathbf{y}) \mathrm{d}\boldsymbol{\vartheta}\right] \\ &\quad - \left[p + q - \mathbf{tr}[-L_{0n}^{(2)}(\bar{\boldsymbol{\psi}}) V_{22}(\bar{\boldsymbol{\vartheta}})]\right], \end{aligned}$$

where $\bar{\boldsymbol{\vartheta}} = (\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\psi}})'$ is the posterior mean of $\boldsymbol{\vartheta}$ under H_1, $\bar{\boldsymbol{\vartheta}}_* = (\boldsymbol{\theta}_0, \bar{\boldsymbol{\psi}})'$, $\bar{\boldsymbol{\vartheta}}_b = (1-b)\bar{\boldsymbol{\vartheta}}_* + b\bar{\boldsymbol{\vartheta}}$, for $b \in [0, 1]$, $S(\mathbf{x}|\boldsymbol{\vartheta}) = \partial \log p(\mathbf{x}|\boldsymbol{\vartheta})/\partial\boldsymbol{\vartheta}$, $D = \int_0^1 \left\{(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \left[E_{\mathbf{z}|\mathbf{y}, \bar{\boldsymbol{\vartheta}}_b}\left(S_1(\mathbf{x}|\bar{\boldsymbol{\vartheta}}_b)\right)\right]\right\} \mathrm{d}b$ the subvector of $S(\mathbf{x}|\boldsymbol{\vartheta})$ corresponding to $\boldsymbol{\theta}$, $V_{22}(\bar{\boldsymbol{\vartheta}}) = E[(\boldsymbol{\psi} - \bar{\boldsymbol{\psi}})(\boldsymbol{\psi} - \bar{\boldsymbol{\psi}})'|\mathbf{y}, H_1]$, the submatrix of $V(\bar{\boldsymbol{\vartheta}})$ corresponding to $\boldsymbol{\psi}$, and $L_{0n}^{(2)}(\boldsymbol{\psi}) = \partial^2 \log p(\mathbf{y}, \boldsymbol{\psi}|\boldsymbol{\theta}_0)/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'$.

Remark 2.4. As shown in Li et al. (2014), $\mathbf{T}^*_{LZY}(\mathbf{y}, \boldsymbol{\theta}_0)$ appeals in four aspects. First, it is well-defined under improper priors. Second, it does not suffer from Jeffreys–Lindley's paradox and, hence, can be used under non-informative vague priors. Third, it is easy to compute. Furthermore, for latent variable models, $\mathbf{T}^*_{LZY}(\mathbf{y}, \boldsymbol{\theta}_0)$ only involves the first and the second derivatives which is easy to evaluate from the MCMC output with the help of the EM algorithm. Finally, Li et al. (2014) derived the asymptotic distribution of $\mathbf{T}^*_{LZY}(\mathbf{y}, \boldsymbol{\theta}_0)$. When $\boldsymbol{\theta}$ and $\bar{\boldsymbol{\psi}}$ are orthogonal, the asymptotic distribution is determined by the chi-squared distribution. In this case the test is asymptotically pivotal and the thresholds can be obtained form the asymptotic distribution. Unfortunately, in general the test is not asymptotically pivotal because the asymptotic distribution depends on some unknown population parameters.

3. Bayesian hypothesis testing based on a quadratic loss

3.1. The test statistic

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

To fix the idea, let

$$s(\boldsymbol{\vartheta}) = \frac{\partial \log p(\mathbf{y}|\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}, \qquad s_0(\boldsymbol{\vartheta}) = \frac{\partial \log p(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}},$$
$$\tilde{s}(\boldsymbol{\vartheta}) = s(\boldsymbol{\vartheta}) + s_0(\boldsymbol{\vartheta}), \qquad C(\boldsymbol{\vartheta}) = \tilde{s}(\boldsymbol{\vartheta})\tilde{s}(\boldsymbol{\vartheta})',$$

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

$$\triangle\mathcal{L}[H_0, \boldsymbol{\vartheta}] = (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})' C_{\theta\theta}(\bar{\boldsymbol{\vartheta}}_0)(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}), \tag{5}$$

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

$$\begin{aligned}\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0) &= \int \triangle\mathcal{L}[H_0, \boldsymbol{\vartheta}] p(\boldsymbol{\vartheta}|\mathbf{y}) \mathrm{d}\boldsymbol{\vartheta} \\ &= \int (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})' C_{\theta\theta}(\bar{\boldsymbol{\vartheta}}_0)(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) p(\boldsymbol{\vartheta}|\mathbf{y}) \mathrm{d}\boldsymbol{\vartheta},\end{aligned} \tag{6}$$

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

Remark 3.1. Clearly $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ depends on the posterior distribution directly. The prior information only influences the test statistic via the posterior distribution.

Remark 3.2. Since the posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y})$ is independent of an arbitrary constant in the prior distributions, both $s(\boldsymbol{\vartheta})$ and $C_{\theta\theta}(\bar{\boldsymbol{\vartheta}}_0)$ are independent of the arbitrary constant. As a result, $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ is well-defined under improper priors.

Remark 3.3. Under some regular condition, we will show in Theorem 3.1 that the proposed test converges to the χ^2 distribution and hence it is not subject to Jeffreys–Lindley's paradox, at least when the sample size is large. To see how it can avoid Jeffreys–Lindley's paradox, consider the example discussed in Li et al. (2014). Let $y \sim N(\theta, \sigma^2)$ with a known σ^2 and we test the null hypothesis $H_0 : \theta = 0$. Let the prior distribution of θ be $N(\mu, \tau^2)$ with $\mu = 0$. LZY showed that the posterior distribution of θ is $N(\mu(y), \omega^2)$ with

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

and BF is

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

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

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

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

Remark 3.4. To calculate $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$, the first derivatives of the observed-data likelihood function must be evaluated. For most latent variable models, the first derivatives are difficult to evaluate directly because the observed-data likelihood function is not available in closed-form. There are several approaches to calculate the first derivatives from the MCMC output.

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

$$\begin{aligned}s(\boldsymbol{\vartheta}) &= \frac{\partial \log p(\mathbf{y}|\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} = \frac{\partial \mathcal{Q}(\boldsymbol{\vartheta}|\tilde{\boldsymbol{\vartheta}})}{\partial \boldsymbol{\vartheta}}\Big|_{\tilde{\boldsymbol{\vartheta}}=\boldsymbol{\vartheta}} \\ &= \int \frac{\partial \log p(\mathbf{y}, \mathbf{z}|\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} p(\mathbf{z}|\mathbf{y}, \boldsymbol{\vartheta})\mathrm{d}\mathbf{z}.\end{aligned}$$

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

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

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

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

Remark 3.5. It is known that the BF is the ratio of two marginal likelihoods. For model M (corresponding to either the null hypothesis or the alternative hypothesis), as shown in Chib (1995) based on Bayes' theorem, the log-marginal likelihood may be calculated by

$$\log p(\mathbf{y}|\boldsymbol{\vartheta}, M) + \log p(\boldsymbol{\vartheta}|M) - \log p(\boldsymbol{\vartheta}|\mathbf{y}, M), \tag{7}$$

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

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

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

Often integration is of high-dimension and has to be evaluated numerically. Unfortunately, mimicking the strategy in Remark 3.4 by averaging $p(\mathbf{y}, \mathbf{z}^{(g)}|\boldsymbol{\vartheta}, M)$ over the effective draws $\{\mathbf{z}^{(g)}, g = 1, 2, \ldots, G\}$ from $p(\mathbf{z}|\boldsymbol{\vartheta}, M)$ is numerically unstable because the expectation is taken with respect to the prior distribution. Whereas, computing $s(\boldsymbol{\vartheta})$ in Remark 3.4 is taken with respect to the posterior distribution. All these problems make it difficult to evaluate the marginal likelihood $\log p(\mathbf{y}|M)$ and BF. To calculate $T(\mathbf{y}, \boldsymbol{\theta}_0)$, the main computational effort is to evaluate the first derivatives of $\log p(\mathbf{y}|\boldsymbol{\vartheta}, M)$, which can be achieved by the EM algorithm, the Kalman filter or the particle filter, as remarked earlier. Thus, there is a computational advantage in the proposed test over the BF.

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

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

Then, we have

$$\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0) = E_{\boldsymbol{\vartheta}|\mathbf{y}}[f(\boldsymbol{\theta})|\mathbf{y}], \qquad \widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0) = \frac{1}{G}\sum_{g=1}^{G} f\left(\boldsymbol{\theta}^{(g)}\right),$$

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

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

$$Var\left(\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)\right) = Var\left(G^{-1}\sum_{g=1}^{G} f\left(\boldsymbol{\theta}^{(g)}\right)\right) = \frac{1}{G} Var\left(f\left(\boldsymbol{\theta}^{(g)}\right)\right).$$

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

$$G^{-1}\sum_{g=1}^{G}\left(f\left(\boldsymbol{\theta}^{(g)}\right) - \widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)\right)\left(f\left(\boldsymbol{\theta}^{(g)}\right) - \widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)\right)'.$$

If $\boldsymbol{\theta}^{(g)}, g = 1, 2, \ldots, G$ are dependent random samples, following Newey and West (1987), a consistent estimator of $Var\left(\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)\right)$ is

$$\frac{1}{G}\left[\Omega_0 + \sum_{k=1}^{q}\left(1 - \frac{k}{q+1}\right)\left(\Omega_k + \Omega_k'\right)\right], \tag{8}$$

where

$$\Omega_k = G^{-1}\sum_{g=k+1}^{G}\left(f\left(\boldsymbol{\theta}^{(g)}\right) - \widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)\right)\left(f\left(\boldsymbol{\theta}^{(g)}\right) - \widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)\right)',$$

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

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

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

and that

$$\tilde{s}_{\theta}\left(\boldsymbol{\vartheta}\right) = \int \frac{\partial \log p\left(\mathbf{y}, \mathbf{z}, \boldsymbol{\vartheta}\right)}{\partial \boldsymbol{\theta}} p\left(\mathbf{z}|\mathbf{y}, \boldsymbol{\vartheta}\right)\mathrm{d}\mathbf{z}.$$

We can estimate $\tilde{s}_{\theta}\left(\bar{\boldsymbol{\vartheta}}_0\right)$ by

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

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

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

Then, we have

$$\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0) = \mathbf{tr}\left(\widehat{\mathbf{h}}_1\widehat{\mathbf{h}}_1'\widehat{\mathbf{H}}_2\right).$$

Following the notations for matrix derivatives in Magnus and Neudecker (2002), let

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

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

$$\begin{aligned}\frac{\partial \widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)}{\partial \widehat{\mathbf{h}}} &= vec(I_p)'\left[\left(\left(\widehat{\mathbf{h}}_1'\widehat{\mathbf{H}}_2\right)' \otimes I_p\right)\frac{\partial \widehat{\mathbf{h}}_1}{\partial \widehat{\mathbf{h}}}\right. \\ &\qquad \left. + \left(\widehat{\mathbf{H}}_2' \otimes \widehat{\mathbf{h}}_1\right)\frac{\partial \widehat{\mathbf{h}}_1'}{\partial \widehat{\mathbf{h}}} + \left(I_p \otimes \widehat{\mathbf{h}}_1\widehat{\mathbf{h}}_1'\right)\frac{\partial \widehat{\mathbf{H}}_2}{\partial \widehat{\mathbf{h}}}\right] \\ &= vec(I_p)'\left[\left(\widehat{\mathbf{H}}_2'\widehat{\mathbf{h}}_1 \otimes I_p + \widehat{\mathbf{H}}_2' \otimes \widehat{\mathbf{h}}_1\right)\frac{\partial \widehat{\mathbf{h}}_1}{\partial \widehat{\mathbf{h}}}\right. \\ &\qquad \left. + \left(I_p \otimes \widehat{\mathbf{h}}_1\widehat{\mathbf{h}}_1'\right)\frac{\partial \widehat{\mathbf{H}}_2}{\partial \widehat{\mathbf{h}}}\right]\end{aligned}$$

where I_p is the p-dimensional identity matrix and

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

By the Delta method,

$$Var\left(\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)\right) = \frac{\partial \widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)}{\partial \widehat{\mathbf{h}}} Var\left(\widehat{\mathbf{h}}\right)\left(\frac{\partial \widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)}{\partial \widehat{\mathbf{h}}}\right)'.$$

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

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

where

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

Remark 3.6. Based on (7), Chib (1995) provided a method to calculate the NSE for estimating $\log p(\boldsymbol{\vartheta}|\mathbf{y}, M)$. When $\log p(\mathbf{y}|\boldsymbol{\vartheta}, M)$ is available in closed-form, the NSE of the estimate of $\log p(\mathbf{y}|M)$ is the same as that of $\log p(\boldsymbol{\vartheta}|\mathbf{y}, M)$ because both $p(\mathbf{y}|\boldsymbol{\vartheta}, M)$ and $p(\boldsymbol{\vartheta}|M)$ can be computed without incurring simulation errors. However, when $p(\mathbf{y}|\boldsymbol{\vartheta}, M)$ does not have a closed-form expression, it has to be calculated by a simulation-based method (such as the EM algorithm or the particle filters) and there will be the NSE for estimating it. In this case, it will be difficult to obtain the NSE of $\log p(\mathbf{y}|\boldsymbol{\vartheta}, M)$. Relative to $\log p(\boldsymbol{\vartheta}|\mathbf{y}, M)$ whose order of magnitude is often $O_p(1)$, $\log p(\mathbf{y}|\boldsymbol{\vartheta}, M)$ is typically $O_p(n)$ so that $\log p(\mathbf{y}|\boldsymbol{\vartheta}, M)$ is dominant in $\log p(\mathbf{y}|M)$. Consequently, one cannot ignore the NSE of $\log p(\mathbf{y}|\boldsymbol{\vartheta}, M)$ when calculating the NSE of $\log p(\mathbf{y}|M)$. As a result, it will be very difficult to obtain the NSE of the estimate of $\log p(\mathbf{y}|M)$ and hence that of the BF. The ease with which one can calculate the NSE of the estimate of $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ is another important advantage of the proposed test over the BF.

3.2. The threshold value

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

Accept H_0 if $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0) \le c$; Reject H_0 if $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0) > c$.

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

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

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

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

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

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

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

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

where $\boldsymbol{\Omega}$ is the support space of $\boldsymbol{\vartheta}$.

Assumption 5. The likelihood function under both the null hypothesis and the alternative hypothesis is regular so that the standard ML theory can be applied. Furthermore, if the null hypothesis is true, let $\boldsymbol{\vartheta}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\psi}_0)$ be true value of $\boldsymbol{\vartheta}$, as $n \to \infty$, for any null sequence $k_n \to 0$, so that,

$$\sup_{\|\vartheta - \vartheta_0\| < k_n} n^{-1}\|\mathbf{I}(\boldsymbol{\vartheta}) - \mathbf{I}(\boldsymbol{\vartheta}_0)\| \xrightarrow{p} 0,$$

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

Remark 3.7. In the literature, Assumptions 1–4 have been used to develop the Bayesian large sample theory; see, for example, Chen (1985). Assumption 5 is a fundamental regularity condition for developing the standard ML theory. Based on these regularity conditions, Li et al. (2014) showed that

$$\bar{\boldsymbol{\vartheta}} = E\left[\boldsymbol{\vartheta}|\mathbf{y}, H_1\right] = \int \boldsymbol{\vartheta} p(\boldsymbol{\vartheta}|\mathbf{y})d\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}} + o_p(n^{-1/2}),$$

$$V(\widehat{\boldsymbol{\vartheta}}) = E\left[(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}})'|\mathbf{y}, H_1\right] = -L_n^{-(2)}(\widehat{\boldsymbol{\vartheta}}) + o_p(n^{-1}).$$

When the null hypothesis holds, we also have

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

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

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

Lemma 3.1. *Let*

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

When the null hypothesis is true, and $\boldsymbol{\vartheta}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\psi}_0)$ is the true value of $\boldsymbol{\vartheta}$, for any consistent estimator $\tilde{\boldsymbol{\vartheta}}$ of $\boldsymbol{\vartheta}$, we have

$$\mathbf{I}(\boldsymbol{\vartheta}_0) = O_p(n), \mathbf{I}(\tilde{\boldsymbol{\vartheta}}) = \mathbf{I}(\boldsymbol{\vartheta}_0) + o_p(n) = O_p(n),$$

$$\mathbf{J}(\boldsymbol{\vartheta}_0) = O_p(n^{-1}), \mathbf{J}(\tilde{\boldsymbol{\vartheta}}) = \mathbf{J}(\boldsymbol{\vartheta}_0) + o_p(n^{-1}) = O_p(n^{-1}).$$

Lemma 3.2. *Let $\widehat{\boldsymbol{\vartheta}}_0 = (\boldsymbol{\theta}_0, \widehat{\boldsymbol{\psi}}_0)$ be the posterior mode of $\boldsymbol{\vartheta}$ under the null hypothesis. Under Assumptions 1–5 and when the null hypothesis is true and the likelihood dominates the prior, we have*

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

$$C(\bar{\boldsymbol{\vartheta}}_0) = C(\widehat{\boldsymbol{\vartheta}}_0) + o_p(n) = O_p(n).$$

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

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

where $\widehat{\boldsymbol{\vartheta}}_{m0} = (\boldsymbol{\theta}_0, \widehat{\boldsymbol{\psi}}_{m0})$ is the ML estimator of $\boldsymbol{\vartheta}$ under the null hypothesis, $s_\theta(\boldsymbol{\vartheta})$ is the score function corresponding to $\boldsymbol{\theta}$, $\mathbf{J}_{\theta\theta}(\boldsymbol{\vartheta})$ is the submatrix of $\mathbf{J}(\boldsymbol{\vartheta})$ corresponding to $\boldsymbol{\theta}$.

Theorem 3.1. *Under Assumptions 1–5, we can show that*

$$\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0) = s_\theta(\bar{\boldsymbol{\vartheta}}_0)\left[-L_{n,\theta\theta}^{-(2)}(\bar{\boldsymbol{\vartheta}})\right]s_\theta(\bar{\boldsymbol{\vartheta}}_0) + o_p(1), \quad (9)$$

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

$$\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0) = LM + o_p(1) \xrightarrow{d} \chi^2(p). \quad (10)$$

Table 1
Comparison of $2\log BF_{10}$, $\mathbf{T}(\mathbf{y},\theta_0)$, and LM.

Prior	$N(0.25, 10^{-4})$				$N(0, 10^4)$			
n	10	100	1000	10,000	10	100	1000	10000
$2\log BF_{10}$	624.69	643.11	753.13	2601.01	−11.56	−13.45	−15.87	−18.16
$\mathbf{T}(\mathbf{y},\theta_0)$	624.13	636.81	684.81	1300.98	0.025	13.03	59.84	676.49
LM	0.025	13.03	59.84	676.49	0.025	13.03	59.84	676.49

Remark 3.8. From Eq. (10), $\mathbf{T}(\mathbf{y},\boldsymbol{\theta}_0)$ may be regarded as the Bayesian version of the LM statistic. However, the LM test is a frequentist test which is based on ML estimation of the model in the null hypothesis whereas our test is a Bayesian test which is based on the posterior quantities of the models under both the null hypothesis as well as the alternative hypothesis.

Remark 3.9. In Theorem 3.1, we can see that under the null hypothesis, the asymptotic distribution of $\mathbf{T}(\mathbf{y},\boldsymbol{\theta}_0)$ always follows the χ^2 distribution and, hence, is independent of the nuisance parameters. This suggests that the new test is asymptotically pivotal, a property that compares favorably with the use of the subjective threshold values as in Bernardo and Rueda (2002) and Li and Yu (2012).

Remark 3.10. When the likelihood dominates the prior, the posterior mode, $\widehat{\boldsymbol{\vartheta}}$, reduces to the ML estimator of $\boldsymbol{\vartheta}$ under the alternative hypothesis, and the posterior mode, $\widehat{\boldsymbol{\vartheta}}_0 = (\boldsymbol{\theta}_0, \widehat{\boldsymbol{\psi}}_0)$, reduces to the ML estimator of $\boldsymbol{\vartheta}$ under the null hypothesis. From Eq. (9), we can see that

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

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

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

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

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

Remark 3.11. $\mathbf{T}(\mathbf{y},\theta_0)$ can incorporate the prior information to improve statistical inference when the sample size is small. This property is shared by the BF but not by the LM test. To illustrate the idea, consider a simple example, where $y_1,\ldots,y_n \sim N(\theta,\sigma^2)$ with a known variance $\sigma^2 = 1$. The true value of θ is set at $\theta_0 = 0.25$. The prior distribution of θ is set as $N(\mu_0,\tau^2)$. The simple point null hypothesis is $H_0: \theta = 0$. It can be shown that

$$2\log BF_{10} = \frac{(n\bar{y}\tau^2+\mu_0\sigma^2)^2}{(\sigma^2+\tau^2)(\sigma^2\tau^2)} + \log\frac{\sigma^2}{n\tau^2+\sigma^2},$$

$$\mathbf{T}(\mathbf{y},\theta_0) = \frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}\left[\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right]^2, \qquad LM = \frac{n\bar{y}^2}{\sigma^2},$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$. When $n \longrightarrow \infty$, $\mathbf{T}(\mathbf{y},\theta_0) \longrightarrow LM$ and the asymptotic distribution for both $\mathbf{T}(\mathbf{y},\theta_0)$ and LM is $\chi^2(1)$. Let us consider the case that corresponds to an informative prior $N(0.25, 10^{-4})$ and compare it to the case that corresponds to a non-informative prior $N(0, 10^4)$. Table 1 reports $2\log BF_{10}$, $T(\mathbf{y},\theta_0)$, and LM when $n = 10, 100, 1000, 10{,}000$ under these two priors. It can be seen that both the BF and the new test depend on the prior (although the BF tends to choose the wrong model under the vague prior even when the sample size is very large) while the LM test is independent of the prior. When $n = 10$, $\mathbf{T}(\mathbf{y},\theta_0)$ correctly rejects the null hypothesis when the prior is informative but fails to reject it when the prior is vague. In this case, the LM test fails to reject the null hypothesis under both priors.[1]

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

n	10	100	1000	10,000
$2\log BF_{10}$	3.63383	1.95408	−0.28049	−2.57621
Decision	Positive	Not worth mention	Negative	Negative
$\mathbf{T}(\mathbf{y},\theta_0)$	6.03170	6.56920	6.62830	6.63420
LM	6.63490	6.63490	6.63490	6.63490

Remark 3.12. It is well known that the BF is conservative compared to the likelihood ratio test; see, for example, Edwards et al. (1963), Kass and Raftery (1995) and Li et al. (2014). Our test is also less conservative than the BF since it is asymptotically pivotal. To illustrate this property, we consider the example in Remark 3.12 of Li et al. (2014). Let $y_1,\ldots,y_n \sim N(\theta,1)$. The prior distribution of θ can be set as $N(0,\tau^2)$. We want to test the simple point null hypothesis $H_0: \theta = 0$. Suppose $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i = \sqrt{6.634897/n}$ so that the critical level of the LM test is always kept at 99%. In this case, it can be shown that $2\log BF_{10} = \frac{n\tau^2}{n\tau^2+1}(\sqrt{n}\bar{y})^2 - \log(n\tau^2+1)$, $\mathbf{T}(\mathbf{y},\theta_0) = \frac{n\tau^2}{n\tau^2+1}(\sqrt{n}\bar{y})^2$ and $LM = (\sqrt{n}\bar{y})^2$. According to Fisher's scale, we have "strong" evidence for the alternative hypothesis based on the LM test. Table 2 reports $2\log BF_{10}$, $\mathbf{T}(\mathbf{y},\theta_0)$, LM when $\tau = 1$. It can be seen that the BF finds the evidence for the alternative hypothesis to be "positive" when $n = 10$. The evidence turns to be "not worth more than a bare mention" when $n = 100$, but to "negative" when $n = 1000, 10{,}000$. This result is consistent with the conservative property of the BF relative to the LM test. In the meantime, our test statistic is slightly more conservative than the LM test although the difference is smaller and the two statistics converge to each other as the sample size grows. When the user is conservative and has a highly informative prior, we caution against the idea of basing the hypothesis testing solely on the proposed test.

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

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

Remark 3.13. We have shown that when the alternative hypothesis is correct, our test statistic is not close to the LM test. In this case, our test continues to take a nonnegative value whereas the LM test can take a negative value. This is because, in our test, the weight matrix $C_{\theta\theta}(\bar{\boldsymbol{\vartheta}}_0)$ remains at least semi-positive definite so that $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ is not negative. When $\boldsymbol{\theta}_0$ is further away from the true value of $\boldsymbol{\theta}$, $s_\theta(\bar{\boldsymbol{\vartheta}}_0)$ and $C_{\theta\theta}(\bar{\boldsymbol{\vartheta}}_0)$ will be further way from zero. Consequently, $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ will be larger so that it can discriminate H_0 against H_1. Whereas, when $\boldsymbol{\theta}_0$ is further away from the true value of $\boldsymbol{\theta}$, the weight matrix $-\mathbf{J}(\widehat{\boldsymbol{\vartheta}}_{m0})$ in the LM statistic may not be positive definite. This may cause some difficulties in using the LM test.

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

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

In this case, we have

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

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

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

$$-\mathbf{J}(\widehat{\boldsymbol{\vartheta}}_{m0}) = \frac{1}{-\mathbf{I}(\widehat{\boldsymbol{\vartheta}}_{m0})} = -\frac{1}{0.4n} < 0.$$

Consequently, the LM statistic is negative. Whereas, for our statistic, if the uninformative prior, $p(\sigma^2) \propto 1$, is used, we have

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

$$\mathbf{T}(\mathbf{y}, \sigma^2 = 1) = \int (\sigma^2 - \bar{\sigma}^2)^2 C_{\theta\theta}(\bar{\boldsymbol{\vartheta}}_0) p(\sigma^2|\mathbf{y})\mathrm{d}\sigma^2.$$

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

Remark 3.14. The implementation of the LM test requires the inversion of $-\mathbf{I}(\boldsymbol{\vartheta}_0)$. When the dimension of $\boldsymbol{\vartheta}$ is high, such an inversion may be numerically challenging. Whereas, to calculate $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$, one does not need to invert any matrix.

4. Empirical illustrations

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

4.1. Hypothesis testing in linear regression models

The first example is the simple linear regression model:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon_i \sim i.i.d.\ N\left(0, \sigma^2\right), \ i = 1, \ldots, n. \tag{11}$$

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

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

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

The marginal likelihood for the model under H_0 is given by

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

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

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

where $V^* = \left(\tilde{V}^{-1} + X'X\right)^{-1}$, $\mu^* = V^*\left(\tilde{V}^{-1}\tilde{\mu} + X'\mathbf{y}\right)$, $X = (\iota, \mathbf{x})$. The derivation is given in Appendix A.4. Hence, in this simple model, $\mathbf{BF}_{10} = p\left(\mathbf{y}|M_1\right)/p\left(\mathbf{y}|M_0\right)$ has an analytical expression. Furthermore, the analytical expression of the proposed statistic can be given by

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

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

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

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

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

The hyperparameters are set at

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

In addition, we allow the prior variance of β, V_β, to vary so that we can examine how the prior influences the BF and $\mathbf{T}\left(\mathbf{y}, \beta_0\right)$.

Table 3
$\log BF_{10}$, $\mathbf{T}(\mathbf{y}, \beta_0)$, the posterior means and standard errors of $\boldsymbol{\vartheta}$ under H_1.

	$V_\beta = 0.1$	$V_\beta = 100$	$V_\beta = 10^5$	$V_\beta = 10^{22}$	$V_\beta = 10^{25}$	$V_\beta = 10^{35}$
$\log BF_{10}$	14.7354	11.2948	7.8409	−11.7311	−15.1849	−26.6979
$\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$	14.9596	15.1693	15.1696	15.1696	15.1696	15.1696
$\bar{\beta}$	0.1220	0.1232	0.1248	0.1243	0.1236	0.1233
$SE(\bar{\beta})$	0.1331	0.1338	0.1334	0.1347	0.1343	0.1337
$\bar{\alpha}$	0.3603	0.3633	0.3587	0.3655	0.3634	0.3622
$SE(\bar{\alpha})$	0.4438	0.4445	0.4423	0.4449	0.4477	0.4435
$\bar{\sigma}^2$	12.5972	12.5792	12.5790	12.5621	12.5741	12.5817
$SE(\bar{\sigma}^2)$	2.2913	2.2768	2.2941	2.2693	2.2785	2.2936

Table 4
$\log BF_{10}$, $\mathbf{T}(\mathbf{y}, \beta_0)$, the posterior means and standard errors of $\boldsymbol{\vartheta}$ under different hyperparameters pairs (a, b).

	(0.001, 0.001)	(0.1, 0.1)	(0.1, 0.01)	(1, 0.1)	(2, 0.001)
$\log BF_{10}$	−0.21813	−0.17002	−0.16703	0.29718	0.81978
$\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$	15.2158	15.2628	15.2635	15.8318	16.3386
$\bar{\beta}$	0.12378	0.12378	0.12378	0.12291	0.12295
$SE(\bar{\beta})$	0.13399	0.1338	0.13378	0.13249	0.13056
$\bar{\alpha}$	0.35948	0.35948	0.35948	0.36327	0.36223
$SE(\bar{\alpha})$	0.13399	0.1338	0.13378	0.13249	0.13056
$\bar{\sigma}^2$	12.5948	12.5584	12.5556	12.2152	11.84
$SE(\bar{\sigma}^2)$	2.2909	2.2805	2.28	2.1805	2.0787

Since both the priors and the likelihood function are in the Normal-Gamma form, we can directly draw samples from their posterior joint distributions under H_0 and H_1. In particular, 35,000 random draws are sampled from the posterior distributions for Bayesian statistical inference.

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

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

4.2. Hypothesis testing in discrete choice models

The probit model is widely used to analyze binary choice data. In this section, we fit the probit model to a dataset originally used in Mroz (1987). Since the observed data likelihood in the probit model is available in closed-form, we can directly compute the proposed Bayesian test statistic $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ based on the MCMC output. Also, the LM test can be easily obtained.

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

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

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

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

Following Koop (2003), 35,000 draws are obtained using the Gibbs sampler under H_0 and H_1 with the first 10,000 samples discarded as burning-in samples. The convergence of Markov chains is monitored using the statistic of Heidelberger and Welch (1983). The parameter estimates and their corresponding standard errors under H_1 for both the Bayesian method and the ML method are reported in Table 5. For the Bayesian method, we report the posterior means and the posterior standard errors. For the ML method, we report the ML estimates and the asymptotic standard errors. Clearly, the difference between the two sets of results is small.

Since $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ does not have a closed-form expression, we can obtain its estimate, $\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)$, from the MCMC outputs. The estimate and the NSE (in the bracket) are reported in Table 6. Since the observed likelihood function has an analytical expression, the LM test can be easily obtained and is reported in Table 6. In addition, the estimator of $\log BF_{10}$ and its NSE are also reported in Table 6. The details about the derivation of these statistics are given in Appendix A.5.

For the individual test, the asymptotic distribution of $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ under H_0 is $\chi^2(1)$ whose 95 percentile is 3.8415. According to $\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)$ and the LM statistic, the hypothesis $\vartheta_7 = 0$ cannot be rejected, suggesting that *kidsge6* does not have a significant explanatory power on y. Furthermore, these two values are very close to each other, consistent with the result in Theorem 3.1. What is more, the BF also strongly support the null hypothesis, reinforcing the conclusion drawn from the other two statistics. The NSEs of the new test and $\widehat{\log BF_{10}}$ are of smaller order of magnitude than the corresponding statistics.

For the joint test, the asymptotic distribution of $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ under H_0 is $\chi^2(2)$ whose 99.99 percentile is 18.42. $\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)$ is much larger than 18.42, suggesting that the null hypothesis is rejected under the 99.99% probability level. Similarly, the LM statistic is much larger than the 99.99 percentile of $\chi^2(2)$ and rejects the null hypothesis. The BF also strongly supports the alternative hypothesis. The three statistics all provide the "strong" evidence that *exper* and *expersq* have the joint explanatory power on y. Furthermore, the difference between $\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)$ and the LM statistic is significant. It suggests that these two test statistics may differ significantly when the null hypothesis is not held, consistent with

Table 5
Bayesian and ML estimates and their standard errors.

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

Table 6
The proposed test statistic, the LM test statistic, $\widehat{\log BF_{10}}$, and the numerical standard errors of the proposed test statistic and $\widehat{\log BF_{10}}$ (in the bracket).

H_0	$\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)$	LM	$\widehat{\log BF_{10}}$
$\vartheta_7 = 0$	0.6805 (0.0204)	0.6861	-12.1454 (0.0226)
$\vartheta_3 = \vartheta_4 = 0$	126.7931 (3.7603)	99.088	21.9721 (0.021)

Remark 3.10. The NSEs of the new test and $\widehat{\log BF_{10}}$ are of smaller order of magnitude than the corresponding statistics.

4.3. Hypothesis testing in stochastic conditional duration models

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

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

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

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

Following Engle and Russell (1998), we adjust the raw durations using the daily season factor $\Psi(t_i)$ which is assumed to be a cubic spline with each node being the average duration on each half hour from 9:30 to 16:00, i.e.,

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

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

Because the observed-data likelihood function is not available in closed-form, it is very hard to calculate the LM statistic even for the model under the null hypothesis. However, since the complete-data likelihood function has an analytical expression, the data augmentation technique facilitates the Bayesian MCMC estimation of the models. As a result, the proposed statistic is easy to calculate and the detailed derivation of $\widehat{\mathbf{T}}(\mathbf{d}, \boldsymbol{\theta}_0)$ is reported in Appendix A.6. The prior distributions for parameters are given as follows,

$$\phi = 2\phi^* - 1, \quad \phi^* \sim Beta(1, 1), \qquad \sigma^2 \sim IG(0.01, 0.01),$$
$$(\alpha, \boldsymbol{\beta}')' | \sigma^2 \sim N\left(\mathbf{0}, 100 \times \sigma^2 I_3\right)$$

where I_3 is 3×3 identity matrix. 55,000 MCMC draws are obtained with the first 15,000 being treated as the burn-in samples. Again, we use the statistic of Heidelberger and Welch (1983) to check the convergence of all the chains. The posterior means and posterior standard errors of all the parameters under the three null hypotheses and the alternative hypothesis are reported in Table 7.

Table 8 reports the values of the new statistic and the BF and the computing time (in seconds) of the new test in the three cases. For hypotheses $\beta_1 = \beta_2 = 0$ and $\beta_2 = 0$, $\widehat{\mathbf{T}}(\mathbf{d}, \boldsymbol{\theta}_0)$ strongly reject the null hypothesis, even under the 99.9% probability level. This is consistent with the BFs, which also strongly support the model under H_1. For hypothesis $\beta_1 = 0$, the BF does not find strong evidence for the alternative hypothesis with "not worth more than a bare mention" evidence. Our proposed statistic also fails to reject the null hypothesis at the 95% probability level.

Finally, from Table 8, we can show that the new statistic takes less time to compute than the BF. Moreover, the NSEs of the new test are of a smaller order of magnitude than the corresponding statistics. However, the NSEs of the BFs are difficult to obtain because the log-likelihood is not available in closed-form for the SCD model.

5. Conclusion

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

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

Appendix

A.1. Proof of Lemma 3.1

When the likelihood information dominates the prior information, the posterior mean $\bar{\boldsymbol{\vartheta}}$ reduces to the ML estimator $\widehat{\boldsymbol{\vartheta}}$, under

Table 7
The posterior means and posterior standard errors of all the parameters under the three null hypotheses and the alternative hypothesis for the SCD model.

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

Table 8
The proposed test statistic, $\widehat{\log BF_{10}}$, their computing time (in seconds), and the numerical standard errors of the proposed test statistic (in the bracket).

	$\beta_1 = \beta_2 = 0$	$\beta_1 = 0$	$\beta_2 = 0$
$\widehat{\mathbf{T}}(\mathbf{d}, \boldsymbol{\theta}_0)$	17.8312 (0.6262)	2.3209 (0.2979)	14.8087 (0.4107)
Time for $\widehat{\mathbf{T}}(\mathbf{d}, \boldsymbol{\theta}_0)$ (s)	4116.1709	4634.9620	3840.8727
$\widehat{\log BF_{10}}$	17.9863	0.8196	18.9603
Time for $\log BF_{10}$ (s)	6889.1324	6913.0687	6510.5200

the alternative hypothesis. When H_0 is true, let $\boldsymbol{\vartheta}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\psi}_0)$ be the true value of $\boldsymbol{\vartheta}$. According to the standard ML theory and the central limit theorem, it can be shown that

$$\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \xrightarrow{d} N\left[0, F(\boldsymbol{\vartheta}_0)\right],$$

where $F(\boldsymbol{\vartheta}_0) = n\mathcal{I}^{-1}(\boldsymbol{\vartheta}_0)$, $\mathcal{I}(\boldsymbol{\vartheta}_0) = -E\left[\mathbf{I}(\boldsymbol{\vartheta}_0)\right]$ is the Fisher information matrix, and

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

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

$$-n\mathbf{J}(\boldsymbol{\vartheta}_0) \xrightarrow{p} F(\boldsymbol{\vartheta}_0),$$

where $\mathbf{J}(\boldsymbol{\vartheta}_0)$ is the inverse matrix of $\mathbf{I}(\boldsymbol{\vartheta}_0)$. Therefore, it can be shown that

$$\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0 = O_p(n^{-\frac{1}{2}}),$$
$$\mathbf{J}(\boldsymbol{\vartheta}_0) = O_p(n^{-1}), \qquad \mathbf{I}(\boldsymbol{\vartheta}_0) = O_p(n).$$

For any consistent estimator of $\boldsymbol{\vartheta}$, say $\tilde{\boldsymbol{\vartheta}}$, there exists a positive sequence $k_n^* \to 0$ such that $p(\|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\| \le k_n^*) \ge 1 - k_n^*$. Hence, when n is large enough, we can find some $N > 0$, and $n > N$ to make $\|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\| \le k_n^*$. Under Assumption 5, we have

$$\frac{1}{n}\|\mathbf{I}(\tilde{\boldsymbol{\vartheta}}) - \mathbf{I}(\boldsymbol{\vartheta}_0)\| \le \sup_{\|\vartheta - \vartheta_0\| < k_n} \frac{1}{n}\|\mathbf{I}(\boldsymbol{\vartheta}) - \mathbf{I}(\boldsymbol{\vartheta}_0)\| \xrightarrow{p} 0.$$

Hence, for any consistent estimator $\tilde{\boldsymbol{\vartheta}}$, $\frac{1}{n}\left[\mathbf{I}(\tilde{\boldsymbol{\vartheta}}) - \mathbf{I}(\boldsymbol{\vartheta}_0)\right] = o_p(1)$ so that $\mathbf{I}(\tilde{\boldsymbol{\vartheta}}) = \mathbf{I}(\boldsymbol{\vartheta}_0) + o_p(n)$ and that $\mathbf{I}(\tilde{\boldsymbol{\vartheta}}) = O_p(n)$. Similarly, $\mathbf{J}(\tilde{\boldsymbol{\vartheta}}) = \mathbf{J}(\boldsymbol{\vartheta}_0) + o_p(n^{-1})$ and $\mathbf{J}(\tilde{\boldsymbol{\vartheta}}) = O_p(n^{-1})$.

A.2. Proof of Lemma 3.2

When the likelihood information dominates the prior information, the posterior mode $\bar{\boldsymbol{\vartheta}}_0$ of $\boldsymbol{\vartheta}$ under the null hypothesis reduces to the ML estimator of $\boldsymbol{\vartheta}$ under the null hypothesis and $\tilde{s}(\boldsymbol{\vartheta}) = s(\boldsymbol{\vartheta})$. Similar to Lemma 3.1, when H_0 is true, according to the standard ML theory, we have

$$\frac{1}{\sqrt{n}} s(\boldsymbol{\vartheta}_0) \sim N[0, F(\boldsymbol{\vartheta}_0)],$$
$$\sqrt{n}(\widehat{\boldsymbol{\psi}}_0 - \boldsymbol{\psi}_0) \sim N[0, F_{\psi\psi}(\boldsymbol{\vartheta}_0)],$$

where $F_{\psi\psi}(\boldsymbol{\vartheta}_0)$ is the submatrix of $F(\boldsymbol{\vartheta}_0)$ corresponding to $\boldsymbol{\psi}$. Hence, we have

$$s(\boldsymbol{\vartheta}_0) = O_p(n^{1/2}), \qquad \widehat{\boldsymbol{\psi}}_0 - \boldsymbol{\psi}_0 = O_p(n^{-1/2}),$$
$$\widehat{\boldsymbol{\vartheta}}_0 - \boldsymbol{\vartheta}_0 = O_p(n^{-1/2}).$$

Furthermore, based on Remark 3.7, it can be shown that

$$\bar{\boldsymbol{\psi}}_0 - \widehat{\boldsymbol{\psi}}_0 = o_p(n^{-1/2}), \qquad \bar{\boldsymbol{\vartheta}}_0 - \widehat{\boldsymbol{\vartheta}}_0 = o_p(n^{-1/2}),$$
$$\begin{aligned}\bar{\boldsymbol{\psi}}_0 - \boldsymbol{\psi}_0 &= \bar{\boldsymbol{\psi}}_0 - \widehat{\boldsymbol{\psi}}_0 + \widehat{\boldsymbol{\psi}}_0 - \boldsymbol{\psi}_0 \\ &= o_p(n^{-1/2}) + O_p(n^{-1/2}) = O_p(n^{-1/2}),\end{aligned}$$
$$\bar{\boldsymbol{\vartheta}}_0 - \boldsymbol{\vartheta}_0 = O_p(n^{-1/2}).$$

Using the first-order Taylor expansion, we have

$$s(\widehat{\boldsymbol{\vartheta}}_0) = s(\boldsymbol{\vartheta}_0) + \mathbf{I}(\tilde{\boldsymbol{\vartheta}}_0)(\widehat{\boldsymbol{\vartheta}}_0 - \boldsymbol{\vartheta}_0),$$

where $\tilde{\boldsymbol{\theta}}_0$ lies on the segment between $\widehat{\boldsymbol{\vartheta}}_0$ and $\boldsymbol{\vartheta}_0$. Since $\widehat{\boldsymbol{\vartheta}}_0 - \boldsymbol{\vartheta}_0 = O_p(n^{-1/2})$, it means that $\widehat{\boldsymbol{\vartheta}}_0$ is a consistent estimator of $\boldsymbol{\vartheta}_0$ so that $\tilde{\boldsymbol{\vartheta}}_0$ is also a consistent estimator of $\boldsymbol{\vartheta}_0$. Hence, we get

$$\begin{aligned}s(\widehat{\boldsymbol{\vartheta}}_0) &= s(\boldsymbol{\vartheta}_0) + \mathbf{I}(\tilde{\boldsymbol{\vartheta}}_0)(\widehat{\boldsymbol{\vartheta}}_0 - \boldsymbol{\vartheta}_0) \\ &= s(\boldsymbol{\vartheta}_0) + [\mathbf{I}(\boldsymbol{\vartheta}_0) + o_p(n)](\widehat{\boldsymbol{\vartheta}}_0 - \boldsymbol{\vartheta}_0) \\ &= s(\boldsymbol{\vartheta}_0) + \mathbf{I}(\boldsymbol{\vartheta}_0)(\widehat{\boldsymbol{\vartheta}}_0 - \boldsymbol{\vartheta}_0) + o_p(n)(\widehat{\boldsymbol{\vartheta}}_0 - \boldsymbol{\vartheta}_0) \\ &= s(\boldsymbol{\vartheta}_0) + \mathbf{I}(\boldsymbol{\vartheta}_0)(\widehat{\boldsymbol{\vartheta}}_0 - \boldsymbol{\vartheta}_0) + o_p(n)O_p(n^{-1/2}) \\ &= s(\boldsymbol{\vartheta}_0) + \mathbf{I}(\boldsymbol{\vartheta}_0)(\widehat{\boldsymbol{\vartheta}}_0 - \boldsymbol{\vartheta}_0) + o_p(n^{1/2}) \\ &= O_p(n^{1/2}) + O_p(n)O_p(n^{-1/2}) + o_p(n^{1/2}) = O_p(n^{1/2}),\end{aligned}$$
$$C(\widehat{\boldsymbol{\vartheta}}_0) = s(\widehat{\boldsymbol{\vartheta}}_0)s(\widehat{\boldsymbol{\vartheta}}_0)' = O_p(n^{1/2})O_p(n^{1/2}) = O_p(n).$$

Similarly, since $\bar{\boldsymbol{\vartheta}}_0 - \boldsymbol{\vartheta}_0 = O_p(n^{-1/2})$, it means that $\bar{\boldsymbol{\vartheta}}_0$ is a consistent estimator of $\boldsymbol{\vartheta}_0$ so that $\tilde{\boldsymbol{\vartheta}}_0$ is also a consistent estimator of $\boldsymbol{\vartheta}_0$. Hence, we can get

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

Furthermore, we can show that

$$s(\bar{\boldsymbol{\vartheta}}_0) = s(\widehat{\boldsymbol{\vartheta}}_0) + \mathbf{I}(\tilde{\boldsymbol{\vartheta}}_0)(\bar{\boldsymbol{\vartheta}}_0 - \widehat{\boldsymbol{\vartheta}}_0),$$

where $\tilde{\boldsymbol{\vartheta}}_0$ lies on the segment between $\bar{\boldsymbol{\vartheta}}_0$ and $\widehat{\boldsymbol{\vartheta}}_0$. Because both $\widehat{\boldsymbol{\vartheta}}_0$ and $\bar{\boldsymbol{\vartheta}}_0$ are consistent estimators of $\boldsymbol{\vartheta}_0$, $\tilde{\boldsymbol{\vartheta}}_0$ is also a consistent estimator of $\boldsymbol{\vartheta}_0$. Using Lemma 3.1, we get

$$\begin{aligned}C(\bar{\boldsymbol{\vartheta}}_0) &= s(\bar{\boldsymbol{\vartheta}}_0)s(\bar{\boldsymbol{\vartheta}}_0)' \\ &= [s(\widehat{\boldsymbol{\vartheta}}_0) + \mathbf{I}(\tilde{\boldsymbol{\vartheta}}_0)(\bar{\boldsymbol{\vartheta}}_0 - \widehat{\boldsymbol{\vartheta}}_0)][s(\widehat{\boldsymbol{\vartheta}}_0) + \mathbf{I}(\tilde{\boldsymbol{\vartheta}}_0)(\bar{\boldsymbol{\vartheta}}_0 - \widehat{\boldsymbol{\vartheta}}_0)]' \\ &= s(\widehat{\boldsymbol{\vartheta}}_0)s(\widehat{\boldsymbol{\vartheta}}_0)' + 2\mathbf{I}(\tilde{\boldsymbol{\vartheta}}_0)(\bar{\boldsymbol{\vartheta}}_0 - \widehat{\boldsymbol{\vartheta}}_0)s(\widehat{\boldsymbol{\vartheta}}_0) \\ &\quad + \mathbf{I}(\tilde{\boldsymbol{\vartheta}}_0)(\bar{\boldsymbol{\vartheta}}_0 - \widehat{\boldsymbol{\vartheta}}_0)(\bar{\boldsymbol{\vartheta}}_0 - \widehat{\boldsymbol{\vartheta}}_0)'\mathbf{I}(\tilde{\boldsymbol{\vartheta}}_0) \\ &= s(\widehat{\boldsymbol{\vartheta}}_0)s(\widehat{\boldsymbol{\vartheta}}_0)' + 2O_p(n)o_p(n^{-1/2})O_p(n^{1/2}) \\ &\quad + O_p(n)o_p(n^{-1/2})o_p(n^{-1/2})O_p(n) \\ &= s(\widehat{\boldsymbol{\vartheta}}_0)s(\widehat{\boldsymbol{\vartheta}}_0)' + o_p(n) = C(\widehat{\boldsymbol{\vartheta}}_0) + o_p(n).\end{aligned}$$

A.3. Proof of Theorem 3.1

Using the Bayesian large sample theory, we have

$$\begin{aligned}
&E\left[(\boldsymbol{\vartheta}-\bar{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta}-\bar{\boldsymbol{\vartheta}})'|\mathbf{y}\right] \\
&= E\left[(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}}+\widehat{\boldsymbol{\vartheta}}-\bar{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}}+\widehat{\boldsymbol{\vartheta}}-\bar{\boldsymbol{\vartheta}})'|\mathbf{y}\right] \\
&= E\left[(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}})'|\mathbf{y}\right] \\
&\quad +2E\left[(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}})|\mathbf{y}\right](\widehat{\boldsymbol{\vartheta}}-\bar{\boldsymbol{\vartheta}})+(\widehat{\boldsymbol{\vartheta}}-\bar{\boldsymbol{\vartheta}})(\widehat{\boldsymbol{\vartheta}}-\bar{\boldsymbol{\vartheta}})' \\
&= E\left[(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}})'|\mathbf{y}\right] \\
&\quad -2(\widehat{\boldsymbol{\vartheta}}-\bar{\boldsymbol{\vartheta}})(\widehat{\boldsymbol{\vartheta}}-\bar{\boldsymbol{\vartheta}})+(\widehat{\boldsymbol{\vartheta}}-\bar{\boldsymbol{\vartheta}})(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}_{\mathbf{0}})' \\
&= E\left[(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}})'|\mathbf{y}\right]-(\widehat{\boldsymbol{\vartheta}}-\bar{\boldsymbol{\vartheta}})(\widehat{\boldsymbol{\vartheta}}-\bar{\boldsymbol{\vartheta}}) \\
&= -L_n^{-(2)}(\widehat{\boldsymbol{\vartheta}})+o_p(n^{-1})+o_p(n^{-1/2})o_p(n^{-1/2}).
\end{aligned}$$

The last equality $E\left[(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\vartheta}})'|\mathbf{y}\right] = -L_n^{-(2)}(\widehat{\boldsymbol{\vartheta}})+o_p(n^{-1})$ follows Li et al. (2012) based on the assumptions listed in Section 3.2. Hence, we have

$$\begin{aligned}
\mathbf{T}(\mathbf{y},\boldsymbol{\theta}_0) &= \int(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})'C_{\theta\theta}(\bar{\boldsymbol{\vartheta}}_0)(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})p(\boldsymbol{\vartheta}|\mathbf{y})d\boldsymbol{\vartheta} \\
&= \mathbf{tr}\left[C_{\theta\theta}(\bar{\boldsymbol{\vartheta}}_0)E[(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})'|\mathbf{y}]\right] \\
&= \mathbf{tr}\left[C_{\theta\theta}(\bar{\boldsymbol{\vartheta}}_0)[-L_{n,\theta\theta}^{-(2)}(\widehat{\boldsymbol{\vartheta}})+o_p(n^{-1})]\right] \\
&= \mathbf{tr}\left[\left(C_{\theta\theta}(\widehat{\boldsymbol{\vartheta}}_0)+o_p(n)\right)[-L_{n,\theta\theta}^{-(2)}(\widehat{\boldsymbol{\vartheta}})]\right] \\
&\quad +\mathbf{tr}\left[C_{\theta\theta}(\bar{\boldsymbol{\vartheta}}_0)o_p(n^{-1})\right] \\
&= \mathbf{tr}\left[C_{\theta\theta}(\widehat{\boldsymbol{\vartheta}}_0)[-L_{n,\theta\theta}^{-(2)}(\widehat{\boldsymbol{\vartheta}})]\right] \\
&\quad +o_p(n)[-L_{n,\theta\theta}^{-(2)}(\widehat{\boldsymbol{\vartheta}})]+O_p(n)o_p(n^{-1}) \\
&= \mathbf{tr}\left[\tilde{s}_\theta(\widehat{\boldsymbol{\vartheta}}_0)\tilde{s}_\theta(\widehat{\boldsymbol{\vartheta}}_0)'[-L_{n,\theta\theta}^{-(2)}(\widehat{\boldsymbol{\vartheta}})]\right] \\
&\quad +o_p(n)O_p(n^{-1})+o_p(1) \\
&= \mathbf{tr}\left[\tilde{s}_\theta(\widehat{\boldsymbol{\vartheta}}_0)\tilde{s}_\theta(\widehat{\boldsymbol{\vartheta}}_0)'[-L_{n,\theta\theta}^{-(2)}(\widehat{\boldsymbol{\vartheta}})]\right]+o_p(1) \\
&= \tilde{s}_\theta(\widehat{\boldsymbol{\vartheta}}_0)'[-L_{n,\theta\theta}^{-(2)}(\widehat{\boldsymbol{\vartheta}})]\tilde{s}_\theta(\widehat{\boldsymbol{\vartheta}}_0)+o_p(1).
\end{aligned}$$

This proves Eq. (9) in the theorem.

When the likelihood information dominates the prior information, the posterior mode $\widehat{\boldsymbol{\vartheta}}$ reduces to the ML estimator of $\boldsymbol{\vartheta}$ under the alternative hypothesis, the posterior mode $\widehat{\boldsymbol{\psi}}_0$ to the ML estimator of $\boldsymbol{\psi}$ under the null hypothesis, and $L_n^{(2)}(\boldsymbol{\vartheta})$ to $\mathbf{I}(\boldsymbol{\vartheta})$, $\tilde{s}(\boldsymbol{\vartheta})$ to $s(\boldsymbol{\vartheta})$. Under H_0, let $\boldsymbol{\vartheta}_0 = (\boldsymbol{\theta}_0,\boldsymbol{\psi}_0)$ be the true value of $\boldsymbol{\vartheta}$, and $\widehat{\boldsymbol{\vartheta}}_0 = (\boldsymbol{\theta}_0,\widehat{\boldsymbol{\psi}})$ be the ML estimator of $\boldsymbol{\vartheta}$. Then, when the null hypothesis is true, $\widehat{\boldsymbol{\vartheta}}$ and $\widehat{\boldsymbol{\vartheta}}_0$ are both consistent estimators of $\boldsymbol{\vartheta}$. Hence, based on Lemmas 3.1 and 3.2, we get

$$\begin{aligned}
\mathbf{J}(\widehat{\boldsymbol{\vartheta}}) &= \mathbf{I}^{-1}(\widehat{\boldsymbol{\vartheta}}) = \left[\mathbf{I}(\boldsymbol{\vartheta}_0)+o_p(n)\right]^{-1}+o_p(n^{-1}) \\
&= \mathbf{J}(\boldsymbol{\vartheta}_{\mathbf{0}})+o_p(n^{-1}), \\
\mathbf{J}(\widehat{\boldsymbol{\vartheta}}_0) &= \mathbf{I}^{-1}(\boldsymbol{\vartheta}_{\mathbf{0}}) = \left[\mathbf{I}(\boldsymbol{\vartheta}_0)+o_p(n)\right]^{-1}+o_p(n^{-1}) \\
&= \mathbf{J}(\boldsymbol{\vartheta}_{\mathbf{0}})+o_p(n^{-1}).
\end{aligned}$$

Then, we can further derive that

$$\begin{aligned}
\mathbf{T}(\mathbf{y},\boldsymbol{\theta}_0) &= \int(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})'C_{\theta\theta}(\bar{\boldsymbol{\vartheta}}_0)(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})p(\boldsymbol{\vartheta}|\mathbf{y})d\boldsymbol{\vartheta} \\
&= s_\theta(\widehat{\boldsymbol{\vartheta}}_0)'[-L_{n,\theta\theta}^{-(2)}(\widehat{\boldsymbol{\vartheta}})]s_\theta(\widehat{\boldsymbol{\vartheta}}_0)+o_p(1) \\
&= -s_\theta(\widehat{\boldsymbol{\vartheta}}_0)'\mathbf{J}_{\theta\theta}(\widehat{\boldsymbol{\vartheta}})s_\theta(\widehat{\boldsymbol{\vartheta}}_0)+o_p(1) \\
&= -s_\theta(\widehat{\boldsymbol{\vartheta}}_0)'\mathbf{J}_{\theta\theta}(\widehat{\boldsymbol{\vartheta}})s_\theta(\widehat{\boldsymbol{\vartheta}}_0)+o_p(1) \\
&= -s_\theta(\widehat{\boldsymbol{\vartheta}}_0)'\left[\mathbf{J}_{\theta\theta}(\boldsymbol{\vartheta}_0)+o_p(n^{-1})\right]s_\theta(\widehat{\boldsymbol{\vartheta}}_0)+o_p(1) \\
&= -s_\theta(\widehat{\boldsymbol{\vartheta}}_0)'\left[\mathbf{J}_{\theta\theta}(\boldsymbol{\vartheta}_0)\right]s_\theta(\widehat{\boldsymbol{\vartheta}}_0) \\
&\quad +s_\theta(\widehat{\boldsymbol{\vartheta}}_0)'o_p(n^{-1})s_\theta(\widehat{\boldsymbol{\vartheta}}_0)+o_p(1) \\
&= -s_\theta(\widehat{\boldsymbol{\vartheta}}_0)'\left[\mathbf{J}_{\theta\theta}(\boldsymbol{\vartheta}_0)\right]s_\theta(\widehat{\boldsymbol{\vartheta}}_0) \\
&\quad +O_p(n^{1/2})o_p(n^{-1})O_p(n^{1/2})+o_p(1) \\
&= -s_\theta(\widehat{\boldsymbol{\vartheta}}_0)'\left[\mathbf{J}_{\theta\theta}(\boldsymbol{\vartheta}_0)\right]s_\theta(\widehat{\boldsymbol{\vartheta}}_0)+o_p(1) \\
&= -s_\theta(\widehat{\boldsymbol{\vartheta}}_0)'\left[\mathbf{J}_{\theta\theta}(\widehat{\boldsymbol{\vartheta}}_0)+o_p(n^{1/2})\right]s_\theta(\widehat{\boldsymbol{\vartheta}}_0)+o_p(1) \\
&= -s_\theta(\widehat{\boldsymbol{\vartheta}}_0)'\mathbf{J}_{\theta\theta}(\widehat{\boldsymbol{\vartheta}}_0)s_\theta(\widehat{\boldsymbol{\vartheta}}_0) \\
&\quad +O_p(n^{1/2})o_p(n^{-1})O_p(n^{1/2})+o_p(1) \\
&= -s_\theta(\widehat{\boldsymbol{\vartheta}}_0)'\mathbf{J}_{\theta\theta}(\widehat{\boldsymbol{\vartheta}}_0)s_\theta(\widehat{\boldsymbol{\vartheta}}_0)+o_p(1) \\
&= LM+o_p(1).
\end{aligned}$$

According to the standard ML theory, under the null hypothesis, $\text{LM} \xrightarrow{d} \chi^2(p)$. Therefore, $\mathbf{T}(\mathbf{y},\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2(p)$ and the theorem is proved.

A.4. Derivation of $\mathbf{T}(\mathbf{y},\boldsymbol{\theta}_0)$ and the BF in linear regression model

It is known that the $\log BF_{10}$ can be expressed as

$$\log BF_{10} = \log p\left(\mathbf{y}|M_1\right) - \log p\left(\mathbf{y}|M_0\right).$$

In the simple linear regression model, under H_0, the marginal likelihood $p(\mathbf{y}|M_0)$ is given by

$$\begin{aligned}
p(\mathbf{y}|M_0) &= \int\int p\left(\mathbf{y}|\alpha,\beta_0\right)p\left(\alpha|\sigma^2\right)p\left(\sigma^2\right)d\alpha d\sigma^2 \\
&= \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)}\int\int\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\alpha-\beta_0x_i)^2\right) \\
&\quad\times\frac{1}{\sqrt{2\pi V_\alpha}\sigma}\exp\left(-\frac{(\alpha-\mu_\alpha)^2}{2\sigma^2V_\alpha}\right) \\
&\quad\times\left(\sigma^2\right)^{-a-\frac{n}{2}-1}\exp\left(-\frac{b}{\sigma^2}\right)d\alpha d\sigma^2 \\
&= \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)}\int\int\frac{1}{\sqrt{2\pi V_\alpha}\sigma} \\
&\quad\times\exp\left\{-\frac{1}{2\sigma^2}\left[-2\alpha\sum_{i=1}^n(y_i-\beta_0x_i)+n\alpha^2\right]\right\} \\
&\quad\times\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\beta_0x_i)^2\right) \\
&\quad\times\exp\left[-\frac{1}{2\sigma^2V_a}\left(\alpha^2-2\mu_\alpha\alpha\right)\right]\exp\left(-\frac{\mu_\alpha^2}{2\sigma^2V_\alpha}\right)d\alpha d\sigma^2 \\
&= \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)}\int\int\frac{1}{\sqrt{2\pi V_\alpha}\sigma} \\
&\quad\times\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\beta_0x_i)^2\right)\exp\left(-\frac{\mu_\alpha^2}{2\sigma^2V_\alpha}\right) \\
&\quad\times\exp\left\{-\frac{1}{2\sigma^2}\left[\left(n+\frac{1}{V_\alpha}\right)\alpha^2\right.\right. \\
&\quad\left.\left.-2\alpha\left(\sum_{i=1}^n(y_i-\beta_0x_i)+\frac{\mu_\alpha}{V_\alpha}\right)\right]\right\}d\alpha d\sigma^2 \\
&= \frac{b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)}\sqrt{\frac{1}{nV_\alpha+1}}\int_0^{+\infty}\left(\sigma^2\right)^{-a-\frac{n}{2}-1} \\
&\quad\times\exp\left\{-\frac{1}{\sigma^2}\left[b+\frac{1}{2}\left(\sum_{i=1}^n(y_i-\beta_0x_i)^2\right.\right.\right.
\end{aligned}$$

$$
\begin{aligned}
&\left.\left.\left.+\frac{\mu_\alpha^2}{V_\alpha}-\frac{\mu_\alpha^{*2}}{V_\alpha^*}\right)\right]\right\} d\sigma^2 \\
&=\frac{b^a \Gamma\left(a+\frac{n}{2}\right)}{(2\pi)^{\frac{n}{2}}\,\Gamma(a)}\sqrt{\frac{1}{nV_\alpha+1}} \\
&\quad\times\left[b+\frac{1}{2}\left(\sum_{i=1}^n\left(y_i-\beta_0x_i\right)^2+\frac{\mu_\alpha^2}{V_\alpha}-\frac{\mu_\alpha^*}{V_\alpha^*}\right)\right]^{-\left(a+\frac{n}{2}\right)} \\
&=\frac{b^a \Gamma\left(a+\frac{n}{2}\right)}{(2\pi)^{\frac{n}{2}}\,\Gamma(a)}\sqrt{\frac{1}{nV_\alpha+1}} \\
&\quad\times\left[b+\frac{1}{2}\left(\left(\mathbf{y}-\beta_0\mathbf{x}\right)'\left(\mathbf{y}-\beta_0\mathbf{x}\right)+\frac{\mu_\alpha^2}{V_\alpha}-\frac{\mu_\alpha^*}{V_\alpha^*}\right)\right]^{-\left(a+\frac{n}{2}\right)}.
\end{aligned}
$$

Under the alternative hypothesis H_1, we rewrite the equation in a matrix form:

$$\mathbf{y}=X\gamma+\boldsymbol{\varepsilon},$$

where $\gamma=(\alpha,\beta)'$, $X=(\iota,\mathbf{x})$. The prior for γ is $N\left(\tilde{\mu},\sigma^2\tilde{V}\right)$, where $\tilde{\mu}=\left(\mu_\alpha,\mu_\beta\right)'$, $\tilde{V}=diag\left(V_\alpha,V_\beta\right)$. Similarly, the marginal likelihood $p(\mathbf{y}|M_1)$ is

$$
\begin{aligned}
p(\mathbf{y}|M_1)&=\int\int p\left(\mathbf{y}|\beta,\alpha\right)p\left(\gamma|\sigma^2\right)p\left(\sigma^2\right)d\gamma d\sigma^2 \\
&=\frac{b^a}{(2\pi)^{\frac{n}{2}}\,\Gamma(a)}\int\int\left(\sigma^2\right)^{-a-\frac{n}{2}-1}\exp\left(-\frac{b}{\sigma^2}\right) \\
&\quad\times\exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{y}-X\gamma\right)'\left(\mathbf{y}-X\gamma\right)\right)\frac{1}{2\pi|\tilde{V}|^{\frac{1}{2}}\sigma^2} \\
&\quad\times\exp\left(-\frac{1}{2\sigma^2}\left(\gamma-\tilde{\mu}\right)'\tilde{V}^{-1}\left(\gamma-\tilde{\mu}\right)\right)d\gamma d\sigma^2 \\
&=\frac{b^a}{(2\pi)^{\frac{n}{2}}\,\Gamma(a)\sqrt{|\tilde{V}|}}\int\int\frac{1}{2\pi\sigma^2}\left(\sigma^2\right)^{-a-\frac{n}{2}-1} \\
&\quad\times\left\{\left(-\frac{1}{\sigma^2}\left[b+\frac{1}{2}\left(\mathbf{y}'\mathbf{y}+(\tilde{\mu})'\tilde{V}^{-1}\tilde{\mu}\right)\right]\right)\right\} \\
&\quad\times\exp\left\{-\frac{1}{2\sigma^2}\left(\gamma'\left(X'X+\tilde{V}^{-1}\right)\gamma-\gamma'\left(X'\mathbf{y}+\tilde{V}^{-1}\tilde{\mu}\right)\right.\right. \\
&\quad\left.\left.-\left(X'\mathbf{y}+\tilde{V}^{-1}\tilde{\mu}\right)'\gamma\right)\right\}d\gamma d\sigma^2 \\
&=\frac{b^a}{(2\pi)^{\frac{n}{2}}\,\Gamma(a)\sqrt{|\tilde{V}|}}\int\int\frac{1}{2\pi\sigma^2} \\
&\quad\times\exp\left\{-\frac{1}{2\sigma^2}\left(\gamma-\mu^*\right)'V^{*-1}\left(\gamma-\mu^*\right)\right\} \\
&\quad\times\exp\left(-\frac{1}{2\sigma^2}\left((\tilde{\mu})'\tilde{V}^{-1}\tilde{\mu}+\mathbf{y}'\mathbf{y}-(\mu^*)'V^{*-1}\mu^*\right)\right) \\
&\quad\times\left(\sigma^2\right)^{-a-\frac{n}{2}-1}\exp\left(-\frac{b}{\sigma^2}\right)d\gamma d\sigma^2 \\
&=\frac{b^a\sqrt{|V^*|}}{(2\pi)^{\frac{n}{2}}\,\Gamma(a)\sqrt{|\tilde{V}|}}\int\left(\sigma^2\right)^{-a-\frac{n}{2}-1}\exp\left\{-\frac{1}{\sigma^2}\left[b\right.\right. \\
&\quad\left.\left.+\frac{1}{2}\left((\tilde{\mu})'\tilde{V}^{-1}\tilde{\mu}+\mathbf{y}'\mathbf{y}-(\mu^*)'V^{*-1}\mu^*\right)\right]\right\}d\sigma^2 \\
&=\frac{b^a\sqrt{|V^*|}\Gamma\left(a+\frac{n}{2}\right)}{(2\pi)^{\frac{n}{2}}\,\Gamma(a)\sqrt{|\tilde{V}|}} \\
&\quad\times\left[b+\frac{1}{2}\left((\tilde{\mu})'\tilde{V}^{-1}\tilde{\mu}+\mathbf{y}'\mathbf{y}-(\mu^*)'V^{*-1}\mu^*\right)\right]^{-\left(a+\frac{n}{2}\right)}.
\end{aligned}
$$

In the following, we show how to calculate $\mathbf{T}\left(\mathbf{y},\boldsymbol{\theta}_0\right)$. It is noted that the log-likelihood function is:

$$
\begin{aligned}
\log p\left(\mathbf{y}|\boldsymbol{\vartheta}\right)&=-\frac{n}{2}\log\left(2\pi\right)-\frac{n}{2}\log\sigma^2 \\
&\quad-\frac{1}{2\sigma^2}\sum_{i=1}^n\left(y_i-\alpha-\beta x_i\right)^2.
\end{aligned}
$$

Hence, given $\boldsymbol{\vartheta}=\left(\alpha,\beta,\sigma^2\right)'$, for H_0 of $\boldsymbol{\theta}=\beta$, we have

$$
\begin{aligned}
s(\boldsymbol{\vartheta})&=\left(\frac{1}{\sigma^2}\sum_{i=1}^n\left(y_i-\alpha-\beta x_i\right),\frac{1}{\sigma^2}\sum_{i=1}^n x_i\left(y_i-\alpha-\beta x_i\right),\right. \\
&\quad\left.-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^n\left(y_i-\alpha-\beta x_i\right)^2\right)',
\end{aligned}
$$

and

$$
\begin{aligned}
C_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}_0\right)&=\left[\frac{1}{\bar{\sigma}_0^2}\sum_{i=1}^n x_i\left(y_i-\bar{\alpha}_0-\beta_0x_i\right)-\frac{1}{V_\beta}\left(\beta_0-\mu_\beta\right)\right]^2 \\
&=\left[\frac{1}{\bar{\sigma}_0^2}\mathbf{x}'\left(\mathbf{y}-\bar{\alpha}_0\iota-\beta_0\mathbf{x}\right)-\frac{1}{V_\beta}\left(\beta_0-\mu_\beta\right)\right]^2,
\end{aligned}
$$

where $\bar{\alpha}_0$ and $\bar{\sigma}_0^2$ are the posterior means of α and σ^2 under H_0.

Since the likelihood and the prior are both in the Normal-Gamma form, based on the previous derivation of $p\left(\mathbf{y}|M_1\right)$, if we integrate the σ^2, we can have the posterior density of $\gamma=(\alpha,\beta)'$

$$
\begin{aligned}
\pi\left(\gamma|\mathbf{y}\right)&\propto\left[b+\frac{1}{2}\left((\tilde{\mu})'\tilde{V}^{-1}\tilde{\mu}+\mathbf{y}'\mathbf{y}-(\mu^*)'V^{*-1}\mu^*\right)\right. \\
&\quad\left.+\frac{1}{2}\left(\gamma-\mu^*\right)'V^{*-1}\left(\gamma-\mu^*\right)\right]^{\frac{2a+n}{2}+1} \\
&\propto\left[1+\frac{1}{2\nu s}\left(\gamma-\mu^*\right)'V^{*-1}\left(\gamma-\mu^*\right)\right]^{\frac{\nu}{2}+1},
\end{aligned}
$$

which is a density function of multivariate t distribution with degrees of freedom $\nu=2a+n$, mean μ^*, and a positive definite symmetric matrix, V^*. That is,

$$\gamma|\mathbf{y}\sim t\left(\mu^*,2sV^*,\nu\right).$$

Let

$$\mu^*=\begin{pmatrix}\mu_1^*\\ \mu_2^*\end{pmatrix},\qquad V^*=\begin{pmatrix}V_{11}^* & V_{12}^*\\ V_{21}^* & V_{22}^*\end{pmatrix}.$$

It is easy to show that $\beta|\mathbf{y}\sim t\left(\mu_2^*,2sV_{22}^*,\nu\right)$. Then, the posterior variance of β is $Var\left(\beta|\mathbf{y}\right)=\frac{2sV_{22}^*}{\nu-2}$. Hence, the proposed test statistic can be calculated analytically as

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

A.5. Derivation of the BF and $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ in the probit model

In the binary probit model, for each y_i, $i = 1, 2, \ldots, n$, there is a corresponding latent variable z_i that satisfies:

$$\begin{cases} y_i = 1 & \text{if } z_i \geq 0 \\ y_i = 0 & \text{if } z_i < 0, \end{cases}$$

and

$$z_i = \mathbf{x}_i'\boldsymbol{\vartheta} + e_i,$$

where $\boldsymbol{\vartheta}$ is the $(p+q)\times 1$ parameter vector measuring the marginal effects and $e_i \sim N(0, 1)$ for $i = 1, \ldots, n$.

Rewrite the above equation as:

$$z_i = \mathbf{x}_{i1}'\boldsymbol{\psi} + \mathbf{x}_{i2}'\boldsymbol{\theta} + e_i.$$

For each i, we have

$$\begin{cases} p(y_i = 1|\boldsymbol{\vartheta}) = p(z_i \geq 0|\boldsymbol{\vartheta}) = p\left(e_i \geq -\left(\mathbf{x}_{i1}'\boldsymbol{\psi} + \mathbf{x}_{i2}'\boldsymbol{\theta}\right)|\boldsymbol{\vartheta}\right) \\ \quad = \Phi\left[(2y_i - 1)\left(\mathbf{x}_{i1}'\boldsymbol{\psi} + \mathbf{x}_{i2}'\boldsymbol{\theta}\right)\right] \\ p(y_i = 0|\boldsymbol{\vartheta}) = p(z_i < 0|\boldsymbol{\vartheta}) = p\left(e_i < -\left(\mathbf{x}_{i1}'\boldsymbol{\psi} + \mathbf{x}_{i2}'\boldsymbol{\theta}\right)|\boldsymbol{\vartheta}\right) \\ \quad = \Phi\left[(2y_i - 1)\left(\mathbf{x}_{i1}'\boldsymbol{\psi} + \mathbf{x}_{i2}'\boldsymbol{\theta}\right)\right], \end{cases}$$

where the $\Phi(\cdot)$ is the standard normal cumulative distribution function. Note that the log-likelihood function is:

$$\log p(\mathbf{y}|\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \log \Phi\left[q_i\left(\mathbf{x}_{i1}'\boldsymbol{\psi} + \mathbf{x}_{i2}'\boldsymbol{\theta}\right)\right],$$

where $q_i = 2y_i - 1$.

- The estimator of $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ and its NSE.
  For H_0 of $\boldsymbol{\theta} = \mathbf{0}$, note that,

$$\frac{\partial \log p(\mathbf{y}|\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} q_i \frac{\phi\left[q_i\left(\mathbf{x}_{i1}'\boldsymbol{\psi} + \mathbf{x}_{i2}'\boldsymbol{\theta}\right)\right]\mathbf{x}_{i2}}{\Phi\left[q_i\left(\mathbf{x}_{i1}'\boldsymbol{\psi} + \mathbf{x}_{i2}'\boldsymbol{\theta}\right)\right]},$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution. The proposed test statistic is

$$\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_\mathbf{0}) = \int \left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right)' C_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}_0\right)\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right) p(\boldsymbol{\vartheta}|y)\, d\boldsymbol{\vartheta},$$

where

$$\begin{aligned} C_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}_0\right) &= \left(\frac{\partial \log p(\mathbf{y}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log p(\mathbf{y}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}}\right)'\Bigg|_{\boldsymbol{\vartheta} = \bar{\boldsymbol{\vartheta}}_0} \\ &= \left(\sum_{i=1}^{n} \frac{\phi\left[q_i\left(\mathbf{x}_{i1}'\bar{\boldsymbol{\psi}}_0\right)\right] q_i\mathbf{x}_{i2}}{\Phi\left[q_i\left(\mathbf{x}_{i1}'\bar{\boldsymbol{\psi}}_0\right)\right]}\right) \\ &\quad \times \left(\sum_{i=1}^{n} \frac{\phi\left[q_i\left(\mathbf{x}_{i1}'\bar{\boldsymbol{\psi}}_0\right)\right] q_i\mathbf{x}_{i2}}{\Phi\left[q_i\left(\mathbf{x}_{i1}'\bar{\boldsymbol{\psi}}_0\right)\right]}\right)', \end{aligned}$$

where $\bar{\boldsymbol{\vartheta}}_0 = (\boldsymbol{\theta}_0, \bar{\boldsymbol{\psi}}_0)$ and $\bar{\boldsymbol{\psi}}_0$ is the posterior mean of $\boldsymbol{\psi}$ under H_0.
To sum up, to compute the $\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)$, we firstly draw MCMC samples for the model under H_0 and calculate $C_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}_0\right)$. We then draw MCMC samples for the model under H_1 to obtain $\left\{\boldsymbol{\vartheta}^{(g)}\right\}_{g=1}^{G} = \left\{\boldsymbol{\theta}^{(g)}, \boldsymbol{\psi}^{(g)}\right\}_{g=1}^{G}$. Naturally, the estimator of the statistic is

$$\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0) = \frac{1}{G}\sum_{g=1}^{G} f\left(\boldsymbol{\theta}^{(g)}\right),$$

where,

$$f\left(\boldsymbol{\theta}^{(g)}\right) = \left(\boldsymbol{\theta}^{(g)} - \bar{\boldsymbol{\theta}}\right)' C_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}_0\right)\left(\boldsymbol{\theta}^{(g)} - \bar{\boldsymbol{\theta}}\right),$$

where $\bar{\boldsymbol{\theta}}$ is the posterior mean of $\boldsymbol{\theta}$ for the model under H_1.
Following the discussion about the NSE in Section 3, the numerical variance of $\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)$ is

$$Var\left(\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)\right) = \frac{1}{G}\left[\Omega_0 + 2\sum_{k=1}^{q}\left(1 - \frac{k}{q+1}\right)\Omega_k\right],$$

where

$$\Omega_k = \frac{1}{G}\sum_{g=k+1}^{G}\left(f\left(\boldsymbol{\theta}^{(g)}\right) - \widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)\right)^2.$$

- The estimator of the BF and its NSE.
  We know that the logarithmic observed likelihood function, $\log p(\mathbf{y}|\boldsymbol{\vartheta})$, is given by

$$\log p(\mathbf{y}|\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \log \Phi\left[q_i\left(\mathbf{x}_{i1}'\boldsymbol{\psi} + \mathbf{x}_{i2}'\boldsymbol{\theta}\right)\right],$$

which is easy to compute.
Based on Chib (1995), the logarithmic marginal likelihood under H_1, $\log p(\mathbf{y}|M_1)$, is given by

$$\log p(\mathbf{y}|M_1) = \log p\left(\mathbf{y}|\bar{\boldsymbol{\vartheta}}\right) + \log p\left(\bar{\boldsymbol{\vartheta}}\right) - \log p\left(\bar{\boldsymbol{\vartheta}}|\mathbf{y}\right),$$

where $p\left(\bar{\boldsymbol{\vartheta}}\right)$ is the pdf of the prior evaluated at $\bar{\boldsymbol{\vartheta}}$, $p\left(\bar{\boldsymbol{\vartheta}}|\mathbf{y}\right)$ is the pdf of the posterior distribution evaluated at $\bar{\boldsymbol{\vartheta}}$. The posterior quantity can be approximated by

$$\widehat{p}\left(\bar{\boldsymbol{\vartheta}}|\mathbf{y}\right) = \frac{1}{G}\sum_{g=1}^{G} p\left(\bar{\boldsymbol{\vartheta}}|\mathbf{z}_1^{(g)}\right),$$

where $\{\mathbf{z}_1^{(g)}, g = 1, 2, \ldots, G\}$ are efficient random draws from $p\left(\mathbf{z}_1|\mathbf{y}, \bar{\boldsymbol{\vartheta}}\right)$ and the posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{z})$ has a closed-form expression in this model. The logarithmic marginal likelihood under H_0, $\log p(\mathbf{y}|M_0)$, is given by

$$\log p(\mathbf{y}|M_0) = \log p\left(\mathbf{y}|\bar{\boldsymbol{\vartheta}}_0\right) + \log p\left(\bar{\boldsymbol{\psi}}_0\right) - \log p\left(\bar{\boldsymbol{\psi}}_0|\mathbf{y}, \boldsymbol{\theta}_0\right).$$

Similarly, $\widehat{p}\left(\bar{\boldsymbol{\psi}}_0|\mathbf{y}, \boldsymbol{\theta}_0\right) = \frac{1}{G}\sum_{g=1}^{G} p\left(\bar{\boldsymbol{\psi}}_0|\mathbf{z}_0^{(g)}, \boldsymbol{\theta}_0\right)$, and $\{\mathbf{z}_0^{(g)}, g = 1, 2, \ldots, G\}$ are efficient random draws from $p\left(\mathbf{z}_0|\mathbf{y}, \bar{\boldsymbol{\vartheta}}_0\right)$.
Hence, the logarithmic BF can be estimated by

$$\begin{aligned} \widehat{\log BF_{10}} &= \left[\log p\left(\mathbf{y}|\bar{\boldsymbol{\vartheta}}\right) + \log p\left(\bar{\boldsymbol{\vartheta}}\right) - \log \widehat{p}\left(\bar{\boldsymbol{\vartheta}}|\mathbf{y}\right)\right] \\ &\quad - \left[\log p\left(\mathbf{y}|\bar{\boldsymbol{\vartheta}}_0\right) + \log p\left(\bar{\boldsymbol{\psi}}_0\right) - \log \widehat{p}\left(\bar{\boldsymbol{\psi}}_0|\mathbf{y}, \boldsymbol{\theta}_0\right)\right]. \end{aligned}$$

To calculate the NSE, following Chib (1995), let $h_1^{(g)} = p\left(\bar{\boldsymbol{\vartheta}}|\mathbf{z}^{(g)}\right)$, $h_0^{(g)} = p\left(\bar{\boldsymbol{\psi}}_0|\mathbf{z}_0^{(g)}, \boldsymbol{\theta}_0\right)$, $h^{(g)} = \left(h_1^{(g)}, h_0^{(g)}\right)'$, $\widehat{h} = \left(\widehat{h}_1, \widehat{h}_0\right)$, $\widehat{h}_0 = \frac{1}{G}\sum_{g=1}^{G} h_0^{(g)}$, $\widehat{h}_1 = \frac{1}{G}\sum_{g=1}^{G} h_1^{(g)}$. Then the numerical variance is

$$Var\left(\widehat{\log BF_{10}}\right) = \left(\frac{\partial \widehat{\log BF_{10}}}{\partial \widehat{h}}\right)' Var(h)\left(\frac{\partial \widehat{\log BF_{10}}}{\partial \widehat{h}}\right),$$

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

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

$$\frac{\partial \widehat{\log BF_{10}}}{\partial \widehat{h}} = \begin{pmatrix} -\widehat{p}\left(\bar{\boldsymbol{\vartheta}}|\mathbf{y}\right)^{-1} \\ \widehat{p}\left(\bar{\boldsymbol{\psi}}_0|\mathbf{y}, \boldsymbol{\theta}_0\right)^{-1} \end{pmatrix}.$$

A.6. Derivation of the BF and $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ in the stochastic conditional duration model

To save the space, here we only discuss the most specification corresponding to H_1. For the SDC model under H_1, denoted as M_1, given by

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

we want to test whether $\boldsymbol{\beta} = \mathbf{0}$ (hence $\boldsymbol{\theta} = \boldsymbol{\beta}$ in this case). As a result, the nuisance parameter $\boldsymbol{\psi} = (\alpha, \phi, \sigma^2)'$ and $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \boldsymbol{\psi}')'$.

- The estimator of $\mathbf{T}(\mathbf{y}, \boldsymbol{\theta}_0)$ and its NSE.
  The proposed statistic is given by:

$$\mathbf{T}(\mathbf{d}, \boldsymbol{\theta}_0) = \int \left(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\right)' C_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}_0\right)\left(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\right) p(\boldsymbol{\vartheta}|\mathbf{d})\, d\boldsymbol{\vartheta}$$
$$= \mathbf{tr}\left[C_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}_0\right) E\left(\left(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\right)'|\mathbf{y}\right)\right],$$

where $\mathbf{d} = \{d_t\}_{t=1}^T$, $\bar{\boldsymbol{\vartheta}}_0 = (\mathbf{0}, \bar{\boldsymbol{\psi}}_0)$, $\bar{\boldsymbol{\psi}}_0$ is the posterior mean of $\boldsymbol{\psi}$ under H_0, $\bar{\boldsymbol{\beta}}$ is the posterior mean of $\boldsymbol{\beta}$ under H_1, and

$$C_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}_0\right) = \left[\left(\frac{\partial \log p(\mathbf{d}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log p(\mathbf{d}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}}\right)'\right]\Bigg|_{\boldsymbol{\vartheta} = \bar{\boldsymbol{\vartheta}}_0}$$
$$= s_\theta\left(\bar{\boldsymbol{\vartheta}}_0\right) s_\theta\left(\bar{\boldsymbol{\vartheta}}_0\right)'.$$

According to Remark 3.4, the partial derivative of log-likelihood function with respect to $\boldsymbol{\theta}$ can be approximated based on the $\mathcal{Q}$-function. That is,

$$s_\theta\left(\bar{\boldsymbol{\vartheta}}_0\right) \approx \frac{1}{G}\sum_{g=1}^{G}\left[-\frac{1}{2\bar{\sigma}_0^2}\tilde{X}'\left(\tilde{\mathbf{y}}^{(g)} - \bar{\alpha}_0\iota\right)\right]$$
$$= \frac{1}{G}\sum_{g=1}^{G}\mathbf{h}_1^{(g)} = \widehat{\mathbf{h}}_1,$$

where $\iota = (1, \ldots, 1)'$, $\tilde{\mathbf{y}}^{(g)} = (\sqrt{1-\bar{\phi}_0^2}\varphi_1^{(g)}, \varphi_2^{(g)} - \bar{\phi}_0\varphi_1^{(g)}, \ldots, \varphi_T^{(g)} - \bar{\phi}_0\varphi_{T-1}^{(g)})'$, $\tilde{X} = \left(\sqrt{\frac{1+\bar{\phi}_0}{1-\bar{\phi}_0}}x_1', x_2', \ldots, x_T'\right)$, $(\bar{\alpha}_0, \bar{\phi}_0, \bar{\sigma}_0^2)$ is the Bayesian estimator under H_0, $\{\varphi_t^{(g)}, g = 1, 2, \ldots, G, t = 1, 2, \ldots, T\}$ are effective draws of the latent variables from the posterior distribution $p(\boldsymbol{\varphi}|\mathbf{d}, \bar{\boldsymbol{\vartheta}}_0)$. Hence, $\mathbf{T}(\mathbf{d}, \boldsymbol{\theta}_0)$ can be approximated by

$$\widehat{\mathbf{T}}(\mathbf{d}, \boldsymbol{\theta}_0) = \mathbf{tr}\left\{\left[\widehat{C}_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}_0\right)\right]\left[\frac{1}{G}\sum_{g=1}^{G}\left(\boldsymbol{\beta}^{(g)} - \bar{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta}^{(g)} - \bar{\boldsymbol{\beta}}\right)'\right]\right\}$$
$$= \mathbf{tr}\left(\widehat{\mathbf{h}}_1\widehat{\mathbf{h}}_1'\widehat{\mathbf{H}}_2\right),$$

where $\bar{\boldsymbol{\beta}}$ is the posterior mean of $\boldsymbol{\beta}$ under H_1, $\left\{\boldsymbol{\beta}^{(g)}\right\}_{g=1}^G$ are the MCMC draws from the posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y})$, and

$$\widehat{C}_{\theta\theta}\left(\bar{\boldsymbol{\vartheta}}_0\right) = \widehat{\mathbf{h}}_1\widehat{\mathbf{h}}_1', \widehat{\mathbf{H}}_2 = \frac{1}{G}\sum_{g=1}^{G}\left(\boldsymbol{\beta}^{(g)} - \bar{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta}^{(g)} - \bar{\boldsymbol{\beta}}\right)'$$
$$= \frac{1}{G}\sum_{g=1}^{G}\mathbf{H}_2^{(g)}.$$

To calculate the NSE, let $\mathbf{h}_2^{(g)} = vech\left(\mathbf{H}_2^{(g)}\right)$, $\mathbf{h}^{(g)} = (\mathbf{h}_1^{(g)\prime}, \mathbf{h}_2^{(g)\prime})'$, $\widehat{\mathbf{h}} = \left(\widehat{\mathbf{h}}_1', \widehat{\mathbf{h}}_2'\right)'$. We have

$$\frac{\partial\widehat{\mathbf{h}}_1}{\partial\widehat{\mathbf{h}}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \quad \frac{\partial\widehat{\mathbf{H}}_2}{\partial\widehat{\mathbf{h}}} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Hence,

$$\frac{\partial\widehat{\mathbf{T}}(\mathbf{d}, \boldsymbol{\theta}_0)}{\partial\widehat{\mathbf{h}}} = vec(I_p)'\left[\left(\widehat{\mathbf{H}}_2'\widehat{\mathbf{h}}_1 \otimes I_p + \widehat{\mathbf{H}}_2' \otimes \widehat{\mathbf{h}}_1\right)\frac{\partial\widehat{\mathbf{h}}_1}{\partial\widehat{\mathbf{h}}} + \left(I_p \otimes \widehat{\mathbf{h}}_1\widehat{\mathbf{h}}_1'\right)\frac{\partial\widehat{\mathbf{H}}_2}{\partial\widehat{\mathbf{h}}}\right],$$

$$Var\left(\widehat{\mathbf{T}}(\mathbf{d}, \boldsymbol{\theta}_0)\right) = \frac{\partial\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)}{\partial\widehat{\mathbf{h}}} Var\left(\widehat{\mathbf{h}}\right)\left(\frac{\partial\widehat{\mathbf{T}}(\mathbf{y}, \boldsymbol{\theta}_0)}{\partial\widehat{\mathbf{h}}}\right)',$$

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

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

where I_p is the p-dimension identity matrix.

- The estimator of the BF.
  Let $\log p(\mathbf{d}|M_0)$ and $\log p(\mathbf{d}|M_1)$ be the marginal likelihood under H_0 and H_1 respectively. Hence,

$$\log BF_{10} = \log p(\mathbf{d}|M_1) - \log p(\mathbf{d}|M_0).$$

The marginal likelihood under H_1 is

$$\log p(\mathbf{d}|M_1) = \log p\left(\mathbf{d}|\bar{\boldsymbol{\vartheta}}\right) + \log p\left(\bar{\boldsymbol{\vartheta}}\right) - \log p\left(\bar{\boldsymbol{\vartheta}}|\mathbf{y}\right),$$

where $p\left(\bar{\boldsymbol{\vartheta}}\right)$ is the prior density function evaluated at $\bar{\boldsymbol{\vartheta}}$, $p\left(\bar{\boldsymbol{\vartheta}}|\mathbf{y}\right)$ is the posterior density function evaluated at $\bar{\boldsymbol{\vartheta}}$. The marginal likelihood under H_0 is

$$\log p(\mathbf{d}|M_0) = \log p\left(\mathbf{y}|\bar{\boldsymbol{\vartheta}}_0\right) + \log p\left(\bar{\boldsymbol{\psi}}_0\right) - \log p\left(\bar{\boldsymbol{\psi}}_0|\mathbf{d}, \boldsymbol{\theta}_0\right).$$

Following Chib (1995), we can approximate the quantities at the right hand side of the marginal likelihood equations as follows,

  – We use the auxiliary particle filter method proposed by Pitt and Shephard (1999) to estimate $\log p\left(\mathbf{y}|\bar{\boldsymbol{\vartheta}}\right)$ and $\log p\left(\mathbf{y}|\bar{\boldsymbol{\vartheta}}_0\right)$. The code is provided by Creal (2012).
  – $\log p\left(\bar{\boldsymbol{\vartheta}}\right)$ and $\log p\left(\bar{\boldsymbol{\psi}}_0\right)$ are easy to evaluate since the prior distributions are standard statistical distributions.
  – $\log p\left(\bar{\boldsymbol{\vartheta}}|\mathbf{y}\right)$ and $\log p\left(\bar{\boldsymbol{\psi}}_0|\mathbf{d}, \boldsymbol{\theta}_0\right)$ can be estimated via the approach of Chib (1995).

  However, since the NSE of the logarithmic observed likelihood function dominates that of the logarithmic marginal likelihood which is estimated by particle filters, the NSE of the BF cannot be obtained.

References

An, S., Schorfheide, F., 2007. Bayesian analysis of DSGE models. Econom. Rev. 26, 211–219.
Bauwens, L., Veredas, D., 2004. The stochastic conditional duration model: a latent variable model for the analysis of financial durations. J. Econometrics 119, 381–412.
Berger, J.O., Perrichi, L.R., 1996. The intrinsic Bayes factor for model selection and prediction. J. Amer. Statist. Assoc. 91, 109–122.
Bernardo, J.M., Rueda, R., 2002. Bayesian hypothesis testing: A reference approach. Internat. Statist. Rev. 70, 351–372.
Breusch, T.S., Pagan, A.R., 1980. The Lagrange multiplier test and its applications to model specification in econometrics. Rev. Econom. Stud. 47, 239–253.
Brooks, C., 2008. Introductory Econometrics for Finance. Cambridge university press.
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. Amer. Statist. Assoc. 90, 1313–1321.
Chib, S., Jeliazkov, I., 2001. Marginal likelihood from the Metropolis–Hastings output. J. Amer. Statist. Assoc. 96, 270–281.
Creal, D., 2012. A survey of sequential Monte Carlo methods for economics and finance. Econom. Rev. 31, 245–296.
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 1, 1–38.
Doucet, A., Johansen, A.M., 2011. A tutorial on particle filtering and smoothing: Fifteen years later. In: The Oxford Handbook of Nonlinear Filtering.

Doucet, A., Shephard, N., 2012. Robust inference on parameters via particle filters and sandwich covariance matrices, Working Paper, Harvard University.
Edwards, W., Lindman, H., Savage, L.J., 1963. Bayesian statistical inference for psychological research. Psychol. Rev. 70, 193–242.
Engle, R.F., Russell, J.R., 1998. Autoregressive conditional duration: a new model for irregularly spaced transaction data. Econometrica 66, 1127–1162.
Geweke, J., 2007. Bayesian model comparison and validation. Amer. Econom. Rev. 97, 60–64.
Han, C., Carlin, B.P., 2001. Markov chain Monte Carlo methods for computing Bayes factor: a comparative review. J. Amer. Statist. Assoc. 96, 1122–1132.
Heidelberger, P., Welch, P.W., 1983. Simulation run length control in the presence of an initial transient. Oper. Res. 31, 1109–1144.
Imai, S., Jain, N., Ching, A., 2009. Bayesian estimation of dynamic discrete choice models. Econometrica 77, 1865–1899.
Jeffreys, H., 1961. Theory of Probability, 3rd ed. Oxford University Press, Oxford.
Kass, R.E., Raftery, A.E., 1995. Bayes factor. J. Amer. Stat. Assoc. 90, 773–795.
Koop, G., 2003. Bayesian Econometrics. Wiley.
Li, Y., Yu, J., 2012. Bayesian hypothesis testing in latent variable models. J. Econometrics 166, 237–246.
Li, Y., Zeng, T., Yu, J., 2014. A new approach to Bayesian hypothesis testing. J. Econometrics 178, 602–612.
Li, Y., Zeng, T., Yu, J., 2012. Robust deviation information criterion for latent variable models. *Working paper*, Singapore Management University.
Magnus, J.R., Neudecker, H., 2002. Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, Chichester.
McCulloch, R.E., 1989. Local model influence. J. Amer. Statist. Assoc. 84, 473–478.
Mroz, T.A., 1987. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. Econometrica 55, 765–799.
Newey, W.K., West, K.D., 1987. A simple positive semi-positive heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.
O'Hagan, A., 1995. Fractional bayes factors for model comparison (with discussion). J. R. Stat. Soc. Ser. B 57, 99–138.
Pitt, M.K., Shephard, N., 1999. Filtering via simulation: Auxiliary particle filters. Journal of the American statistical association 94, 590–599.
Poirier, D.J., 1995. Intermediate Statistics and Econometrics: A Comparative Approach. MIT Press, Cambridge, MA.
Poyiadjis, G., Doucet, A., Singh, S., 2011. Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. Biometrika 98, 65–80.
Robert, C., 1993. A note on Jeffreys–Lindley Paradox. Statistica Sinica 3, 601–608.
Robert, C., 2001. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, 2nd version. Springer Texts in Statistics.
Shephard, N., 2005. Stochastic Volatility: Selective Readings. Oxford University Press.
Wooldridge, J.M., 2002. Introductory Econometrics: A Modern Approach, Third ed. Cengage Learning.