Journal of Econometrics 234 (2023) 28–52

Contents lists available at ScienceDirect

Journal of Econometrics

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

Improved marginal likelihood estimation via power posteriors and importance sampling ☆

Yong Li [a], Nianling Wang [b,*], Jun Yu [c,*]

[a] School of Economics, Renmin University of China, China
[b] School of Finance, Capital University of Economics and Business, China
[c] School of Economics and Lee Kong Chian School of Business, Singapore Management University, Singapore

ARTICLE INFO

Article history:
Received 22 July 2019
Received in revised form 16 November 2021
Accepted 17 November 2021
Available online 18 December 2021

JEL classification:
C11
C12

Keywords:
Bayes factor
Marginal likelihood
Markov Chain Monte Carlo
Model choice
Power posteriors
Importance sampling

ABSTRACT

Power posteriors have become popular in estimating the marginal likelihood of a Bayesian model. A power posterior is referred to as the posterior distribution that is proportional to the likelihood raised to a power $b \in [0, 1]$. Important power-posterior-based algorithms include thermodynamic integration (TI) of Friel and Pettitt (2008) and steppingstone sampling (SS) of Xie et al. (2011). In this paper, it is shown that the Bernstein–von Mises (BvM) theorem holds for power posteriors under regularity conditions. Due to the BvM theorem, power posteriors, when adjusted by the square root of the auxiliary constant, have the same limit distribution as the original posterior distribution, facilitating the implementation of the modified TI and SS methods via importance sampling. Unlike the TI and SS methods that require repeated sampling from the power posteriors, the modified methods only need the original posterior output and hence, are computationally more efficient. Moreover, they completely avoid the coding efforts associated with sampling from the power posteriors. Primitive conditions, under which the TI and modified TI algorithms can produce consistent estimators of the marginal likelihood, are provided. The numerical efficiency of the proposed methods is illustrated using two models.

1. Introduction

A highly important statistical decision faced by practitioners is the model comparison. In the Bayesian paradigm, the Bayes factors (BFs) are arguably the most widely used Bayesian statistic for comparing models (Kass and Raftery, 1995; Young and Pettit, 1996). BFs have received many applications in practice. For example, since they are derived from the probabilities of the null hypothesis and the alternative hypothesis conditional on the data (i.e., the two marginal likelihoods), it is a natural tool for Bayesian hypothesis testing. Furthermore, it is tipped as a statistic to avoid the so-called *p*-hacking problem (Harvey, 2017; Brodeur et al., 2020). This is not surprising because BFs are not based on the argument

☆ We would like to thank Xiaohong Chen (the co-editor), an AE, two referees, Nial Friel, Peter Phillips, Qiman Shao, Tao Zeng for helpful discussions. Li gratefully acknowledges the financial support of the Chinese Natural Science fund (No. 71773130, No. 71971109), and fund for building world-class universities (disciplines) of Renmin University of China, and the Digital Economy Platform, Major Innovation & Planning Interdisciplinary Platform for the Double-First Class Initiative, Renmin University of China. Wang gratefully acknowledges the hospitality during her research visits to Singapore Management University. Yu would like to acknowledge that this research/project is supported by the Ministry of Education, Singapore, under its Academic Research Fund (AcRF) Tier 2 (Award Number MOE-T2EP402A20-0002). He also acknowledges the financial support from the Lee Foundation.

* Corresponding authors.
E-mail addresses: nianling.wang@cueb.edu.cn (N. Wang), yujun@smu.edu.sg (J. Yu).

https://doi.org/10.1016/j.jeconom.2021.11.009
0304-4076/

of repeated sampling. Moreover, BFs can be used to choose weights in model averaging and forecast combinations (Sala-i Martin et al., 2004; Moral-Benito, 2013).

The calculation of BFs generally requires the calculation of the marginal likelihood values of competing models. Marginal likelihood conducts integrations over the entire parameter space. In practice, the integrations can impose serious computational challenges. In the Bayesian literature, various approaches have been proposed to compute the marginal likelihood. Some excellent reviews are found in DiCiccio et al. (1997) and Han and Carlin (2001).

In this paper, we plan to improve some existing power-posterior-based methods to estimate the marginal likelihood based on posterior output. A power posterior is referred to as the posterior distribution that is proportional to the likelihood raised to a power $b \in [0, 1]$. Existing algorithms include thermodynamic integration (TI) of Friel and Pettitt (2008) and steppingstone sampling (SS) of Xie et al. (2011). Compared with other posterior-output-based approaches, the power-posterior-based approaches require very little tuning, are easy to implement, and lead to small Monte Carlo errors.

However, there are several drawbacks to the existing power-posterior-based approaches. First and foremost, the sampling cost is very high. To implement the existing power-posterior-based methods, one needs MCMC output at grid points over the interval [0, 1], denoted by $\{b_s\}_{s=0}^{S}$ where $S+1$ is the number of grid points.[1] Hence, MCMC sampling has to be repeated for $S+1$ times, greatly increasing the computational cost when S is moderate or large. Second, to calculate the Monte Carlo standard error (MCSE) of the marginal likelihood estimate, independent MCMC chains, at all grid points, have to be obtained. As a result, the computational cost would inevitably increase sharply. Third, for many standard models with regular distributions, the power posteriors may lead to non-standard distributions, so that standard Bayesian software such as WinBUGS (Spiegelhalter et al., 2003) is difficult to use. As a result, extra coding efforts are needed to draw random samples from power posteriors. Finally, it is not known when the algorithms provide consistent estimators of the marginal likelihood.

To overcome these disadvantages in the existing power-posterior-based approaches, the present paper develops new approaches to estimate the marginal likelihood. The theoretical underpinning of the proposed approaches is the Bernstein–von Mises (BvM) theorem that we manage to develop for the power posteriors. Due to the BvM theorem, we show that the power posteriors, when adjusted by the square root of the grid points, have the same asymptotic normal distribution as the original posterior distribution. This property suggests that we can use the original posterior distribution, adjusted by a simple linear transformation, to design importance sampling techniques to estimate the marginal likelihood.

Using the self-normalized importance sampling technique, we propose the modified TI and SS algorithms. The new algorithms avoid the need to make random draws from the power posterior at any grid point. Therefore, new methods significantly reduce computational costs. Furthermore, the coding efforts to draw random samples from the power posteriors are completely avoided, and hence, our methods are easy to implement using popular Bayesian software. It is important to note that since our approach is based on importance sampling, we do not need the sample size to go to infinity. Under a set of primitive conditions, we show that the modified TI algorithm (as well as the TI algorithm) can provide consistent estimators of the marginal likelihood.[2]

The rest of the paper is organized as follows. Section 2 reviews the TI and SS algorithms. Section 3 introduces our new methods. Section 4 establishes the BvM theorem for the power posterior and obtains consistency of the TI and modified TI methods. In Section 5, we compare the performance of the proposed methods with that of the original TI and SS algorithms using two examples. Section 6 concludes the paper. The appendix collects the proof of theoretical results in the paper. An online appendix contains the detailed proof of Theorem 4.1 and the proof of four lemmas that are useful to prove Theorems 4.2–4.3.

We use the following notations throughout the paper: O, o, O_p, o_p, $\xrightarrow{p}$, $\xrightarrow{d}$, $\xrightarrow{L^2}$, $\sim$, and $a := b$, to denote the same order, the smaller order, the same order in probability, the smaller order in probability, convergence in probability, convergence in distribution, convergence in mean square, asymptotic equivalence, and defining a as b, respectively. We denote $P(A)$ the probability of event A and $p(\cdot)$ a probability density function. We denote $\|A\|$ the Euclidean norm of a vector A, $vec(B)$ the vec operator that stacks the columns of matrix B, $\otimes$ the Kronecker product, $\nabla_x^k f(x)$ the kth derivative of f with respect to x. Let $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{y})$ denote the prior density and the posterior density, respectively, where $\boldsymbol{\theta}$ is the parameter in the original model and $\mathbf{y}$ is the observed data. To distinguish the parameter in the original posterior and that in the power posterior, we let $\boldsymbol{\theta}_b$ denote the parameter in the model whose likelihood function is $p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b$ and hence, $p\left(\boldsymbol{\theta}_b|\mathbf{y}, b\right)$ is the corresponding (power) posterior density. Let $p_A\left(\boldsymbol{\theta}_b|\mathbf{y}, b\right)$ denote approximate power posterior density. Furthermore, we let $\boldsymbol{\theta}_{0,(j)}$, $\boldsymbol{\theta}_{1,(j)}$, $\boldsymbol{\theta}_{b,(j)}$, $\boldsymbol{\theta}_{b,(j)}^{tr}$ denote random samples from $p(\boldsymbol{\theta})$, $p(\boldsymbol{\theta}|\mathbf{y})$, $p\left(\boldsymbol{\theta}_b|\mathbf{y}, b\right)$, $p_A\left(\boldsymbol{\theta}_b|\mathbf{y}, b\right)$, respectively.[3] Let $E_0(X)$, $E_{\boldsymbol{\theta}|\mathbf{y}}(X)$, $E_{\boldsymbol{\theta}_b|\mathbf{y},b}(X)$ and $E_A(X)$ (or $Var_0(X)$, $Var_{\boldsymbol{\theta}|\mathbf{y}}(X)$, $Var_{\boldsymbol{\theta}_b|\mathbf{y},b}(X)$ and $Var_A(X)$) denote the expectation (or variance) of X with respect to $p(\boldsymbol{\theta})$, $p(\boldsymbol{\theta}|\mathbf{y})$, $p\left(\boldsymbol{\theta}_b|\mathbf{y}, b\right)$, $p_A\left(\boldsymbol{\theta}_b|\mathbf{y}, b\right)$, respectively. Finally, we use I_q to denote the q-dimensional identity matrix.

1 The TI and SS methods are mainly implemented together with MCMC output due to the popularity and versatility of MCMC. However, the TI and SS methods (and hence our modified methods) are applicable to other samplers.

2 A similar set of conditions can be specified, under which the modified SS algorithm (as well as the SS algorithm) can provide consistent estimators of the marginal likelihood. To keep the paper at a reasonable length, we omit this development.

3 Without loss of generality and being consistent with the common practice, we assume random draws from the prior are independent and identically distributed (iid) while random draws from the posterior and power posterior are stationary and ergodic. The theory developed in this paper remains valid when random draws from the prior are stationary and ergodic while the random draws from the posterior and power posterior are iid.

2. Review of marginal likelihood estimation based on power posteriors

Let $\mathbf{y}$ be data, $p(\mathbf{y}|\boldsymbol{\theta})$ be the likelihood function of a model that depends on the set of parameters $\boldsymbol{\theta} \in \Theta$, where Θ is the parameter space. Let $p(\boldsymbol{\theta})$ be the density function for a proper prior distribution. The posterior distribution of parameters is given by

$$p\left(\boldsymbol{\theta}|\mathbf{y}\right) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{m(\mathbf{y})}, \text{ where } m(\mathbf{y}) = \int_{\Theta} p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{1}$$

The normalization constant $m(\mathbf{y})$ is called the marginal likelihood, and our purpose is to estimate it given the data $\mathbf{y}$ and the model. Moreover, let b be an auxiliary constant from the interval $[0, 1]$, $p\left(\boldsymbol{\theta}_b|\mathbf{y}, b\right)$ be the power posterior density function, and $m(\mathbf{y}|b)$ be the corresponding power marginal likelihood. Friel and Pettitt (2008) define the power posterior as:

$$p\left(\boldsymbol{\theta}_b|\mathbf{y}, b\right) = \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b p\left(\boldsymbol{\theta}_b\right)}{m(\mathbf{y}|b)}, \text{ where } m(\mathbf{y}|b) = \int_{\Theta} p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b p\left(\boldsymbol{\theta}_b\right) \mathrm{d}\boldsymbol{\theta}_b. \tag{2}$$

In other words, the statistical model corresponding to the power posterior essentially raises the original likelihood function to power b, that is, $p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b$.

Below we first provide a brief review of the TI algorithm of Friel and Pettitt (2008) and the SS algorithm of Xie et al. (2011) and then discuss their properties.

2.1. TI algorithm

When $b = 0$ or 1, Friel and Pettitt (2008) note that

$$m(\mathbf{y}|1) = \int_{\Theta} p\left(\mathbf{y}|\boldsymbol{\theta}_b\right) p\left(\boldsymbol{\theta}_b\right) \mathrm{d}\boldsymbol{\theta}_b = m(\mathbf{y}) \text{ and } m(\mathbf{y}|0) = \int_{\Theta} p\left(\boldsymbol{\theta}_b\right) \mathrm{d}\boldsymbol{\theta}_b = 1, \tag{3}$$

The first derivative of $\ln m(\mathbf{y}|b)$ is

$$\begin{aligned}\mathcal{U}(b) &:= \frac{\partial \ln m(\mathbf{y}|b)}{\partial b} = \frac{1}{m(\mathbf{y}|b)}\frac{\partial m(\mathbf{y}|b)}{\partial b} = \int_{\Theta} \frac{\partial \ln p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b}{\partial b} \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b p\left(\boldsymbol{\theta}_b\right)}{m(\mathbf{y}|b)} \mathrm{d}\boldsymbol{\theta}_b \\ &= \int_{\Theta} \ln p\left(\mathbf{y}|\boldsymbol{\theta}_b\right) \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b p\left(\boldsymbol{\theta}_b\right)}{m(\mathbf{y}|b)} \mathrm{d}\boldsymbol{\theta}_b = \int_{\Theta} \ln p\left(\mathbf{y}|\boldsymbol{\theta}_b\right) p\left(\boldsymbol{\theta}_b|\mathbf{y}, b\right) \mathrm{d}\boldsymbol{\theta}_b = E_{\boldsymbol{\theta}_b|\mathbf{y},b} \ln p\left(\mathbf{y}|\boldsymbol{\theta}_b\right).\end{aligned} \tag{4}$$

Based on (3) and (4), one can recover the integral from the first-order derivative as

$$\ln m(\mathbf{y}) = \ln m(\mathbf{y}|1) - \ln m(\mathbf{y}|0) = \int_0^1 \mathcal{U}(b)\mathrm{d}b = \int_0^1 E_{\boldsymbol{\theta}_b|\mathbf{y},b} \ln p\left(\mathbf{y}|\boldsymbol{\theta}_b\right) \mathrm{d}b. \tag{5}$$

Eq. (5) suggests an approach to estimation of the log marginal likelihood via the power posteriors as shown in Friel and Pettitt (2008).

In general, the integral $\int_0^1 \mathcal{U}(b)\mathrm{d}b$ does not have an analytical solution. Friel and Pettitt (2008) propose to numerically approximate it using the trapezoidal rule. In particular, based on the grid $\{b_s = (s/S)^c\}_{s=0}^S$ with $c \geq 1$, $\ln m(\mathbf{y})$ is approximated by

$$\sum_{s=0}^{S-1}(b_{s+1} - b_s)\frac{\mathcal{U}(b_{s+1}) + \mathcal{U}(b_s)}{2}. \tag{6}$$

Furthermore, since $\mathcal{U}(b_s) = E_{\boldsymbol{\theta}_{b_s}|\mathbf{y},b_s} \ln p\left(\mathbf{y}|\boldsymbol{\theta}_{b_s}\right)$ does not have an analytical expression, it can be estimated by

$$\widehat{\mathcal{U}}(b_s) := \frac{1}{J}\sum_{j=1}^{J} \ln p\left(\mathbf{y}|\boldsymbol{\theta}_{b_s,(j)}\right), \tag{7}$$

where $\left\{\boldsymbol{\theta}_{b_s,(j)}\right\}_{j=1}^J$ are effective random draws from the power posterior $p\left(\boldsymbol{\theta}_{b_s}|\mathbf{y}, b_s\right)$. Combining (6) and (7), one may estimate the log marginal likelihood by

$$\widehat{\ln m}_{TI}(\mathbf{y}) := \sum_{s=0}^{S-1}\left(b_{s+1} - b_s\right)\frac{\widehat{\mathcal{U}}(b_{s+1}) + \widehat{\mathcal{U}}(b_s)}{2}. \tag{8}$$

In the literature, this approach is referred to as the TI algorithm.

As indicated in Eq. (8), the TI algorithm estimates the log marginal likelihood rather than the marginal likelihood to ensure computational stability. The TI algorithm can be summarized as follows.

TI Algorithm

1. Choose grid points $\left\{b_s = (s/S)^c\right\}_{s=0}^{S}$ with $c \geq 1$.
2. For each b_s, draw J random samples $\left\{\boldsymbol{\theta}_{b_s,(j)}\right\}_{j=1}^{J}$ (such as MCMC samples) from the power posterior $p\left(\boldsymbol{\theta}_{b_s}|\mathbf{y}, b_s\right)$.
3. For each b_s, calculate $\widehat{\mathcal{U}}(b_s)$ based on Eq. (7).
4. The log marginal likelihood is estimated by Eq. (8).

2.2. SS algorithm

Xie et al. (2011) propose another algorithm to estimate the marginal likelihood based on the power posteriors. The basic idea is to explore the following identity:

$$m(\mathbf{y}) = \frac{m(\mathbf{y}|b=1)}{m(\mathbf{y}|b=0)} = \prod_{s=0}^{S-1} \frac{m\left(\mathbf{y}|b_{s+1}\right)}{m\left(\mathbf{y}|b_s\right)} = \prod_{s=0}^{S-1} r\left(b_s\right), \tag{9}$$

where

$$\begin{aligned} r\left(b_s\right) &= \frac{m\left(\mathbf{y}|b_{s+1}\right)}{m\left(\mathbf{y}|b_s\right)} = \frac{\int_\Theta p\left(\mathbf{y}|\boldsymbol{\theta}_{b_{s+1}}\right)^{b_{s+1}} p\left(\boldsymbol{\theta}_{b_{s+1}}\right) d\boldsymbol{\theta}_{b_{s+1}}}{m\left(\mathbf{y}|b_s\right)} \qquad (10)\\ &= \int_\Theta \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_{b_{s+1}}\right)^{b_{s+1}}}{p\left(\mathbf{y}|\boldsymbol{\theta}_{b_{s+1}}\right)^{b_s}} \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_{b_{s+1}}\right)^{b_s} p\left(\boldsymbol{\theta}_{b_{s+1}}\right)}{\int_\Theta p\left(\mathbf{y}|\boldsymbol{\theta}_{b_{s+1}}\right)^{b_s} p\left(\boldsymbol{\theta}_{b_{s+1}}\right) d\boldsymbol{\theta}_{b_{s+1}}} d\boldsymbol{\theta}_{b_{s+1}} \\ &= \int_\Theta p\left(\mathbf{y}|\boldsymbol{\theta}_{b_s}\right)^{b_{s+1}-b_s} p\left(\boldsymbol{\theta}_{b_s}|\mathbf{y}, b_s\right) d\boldsymbol{\theta}_{b_s} = \int_\Theta \exp\left[\left(b_{s+1}-b_s\right) \ln p\left(\mathbf{y}|\boldsymbol{\theta}_{b_s}\right)\right] p\left(\boldsymbol{\theta}_{b_s}|\mathbf{y}, b_s\right) d\boldsymbol{\theta}_{b_s} \\ &= E_{\boldsymbol{\theta}_{b_s}|\mathbf{y},b_s} \exp\left[\left(b_{s+1}-b_s\right) \ln p\left(\mathbf{y}|\boldsymbol{\theta}_{b_s}\right)\right]. \qquad (11) \end{aligned}$$

In general $E_{\boldsymbol{\theta}_{b_s}|\mathbf{y},b_s} \exp\left[\left(b_{s+1}-b_s\right) \ln p\left(\mathbf{y}|\boldsymbol{\theta}_{b_s}\right)\right]$ does not have an analytical expression. It can be estimated by

$$\widehat{r}\left(b_s\right) := \frac{1}{J} \sum_{j=1}^{J} \exp\left[\left(b_{s+1}-b_s\right) \ln p\left(\mathbf{y}|\boldsymbol{\theta}_{b_s,(j)}\right)\right], \tag{12}$$

where $\left\{\boldsymbol{\theta}_{b_s,(j)}\right\}_{j=1}^{J}$ are effective random samples from the power posterior $p\left(\boldsymbol{\theta}_{b_s}|\mathbf{y}, b_s\right)$. Unfortunately, the right-hand side of Eq. (12) can easily explode with the exponential term. A numerically more stable estimate of $E_{\boldsymbol{\theta}_{b_s}|\mathbf{y},b_s} \exp\left[\left(b_{s+1}-b_s\right) \ln p(\mathbf{y}|\boldsymbol{\theta}_{b_s})\right]$ is

$$\begin{aligned} \widehat{r}\left(b_s\right) &= \frac{1}{J} \sum_{j=1}^{J} \exp\left\{\left(b_{s+1}-b_s\right)\left(\ln p\left(\mathbf{y}|\boldsymbol{\theta}_{b_s,(j)}\right) - \bar{L}_{b_s}\right) + \left(b_{s+1}-b_s\right)\bar{L}_{b_s}\right\} \\ &= \exp\left[\left(b_{s+1}-b_s\right)\bar{L}_{b_s}\right] \left\{\frac{1}{J} \sum_{j=1}^{J} \exp\left[\left(b_{s+1}-b_s\right)\left(\ln p\left(\mathbf{y}|\boldsymbol{\theta}_{b_s,(j)}\right) - \bar{L}_{b_s}\right)\right]\right\}, \end{aligned} \tag{13}$$

where $\bar{L}_{b_s} = \max_{j\in\{1,\ldots,J\}} \left\{\ln p\left(\mathbf{y}|\boldsymbol{\theta}_{b_s,(j)}\right)\right\}$. To estimate the log marginal likelihood, Xie et al. (2011) propose to use

$$\begin{aligned} \sum_{s=0}^{S-1} \ln \widehat{r}\left(b_s\right) = &\sum_{s=0}^{S-1} \ln\left(\frac{1}{J} \sum_{j=1}^{J} \exp\left[\left(b_{s+1}-b_s\right)\left(\ln p\left(\mathbf{y}|\boldsymbol{\theta}_{b_s,(j)}\right) - \bar{L}_{b_s}\right)\right]\right) \\ &+ \sum_{s=0}^{S-1} \left[\left(b_{s+1}-b_s\right)\bar{L}_{b_s}\right]. \end{aligned} \tag{14}$$

The SS algorithm can be summarized as follows.

SS Algorithm

1. Choose grid points $\left\{b_s = (s/S)^c\right\}_{s=0}^{S}$ with $c \geq 1$.
2. For each b_s, draw J random samples $\left\{\boldsymbol{\theta}_{b_s,(j)}\right\}_{j=1}^{J}$ (such as MCMC samples) from the power posterior $p\left(\boldsymbol{\theta}_{b_s}|\mathbf{y}, b_s\right)$.
3. For each b_s, calculate $\widehat{r}\left(b_s\right)$ based on Eq. (13).
4. The log marginal likelihood is estimated by Eq. (14).

2.3. Discussion of two algorithms

It is clear that both algorithms require repeated sampling from the power posteriors corresponding to all grid points $\left\{b_s = (s/S)^c\right\}_{s=0}^{S}$. When MCMC sampling is time-consuming, obtaining MCMC samples for $S+1$ times will make the two algorithms even more time-consuming.

There are two sources of errors in the TI algorithm: (i) the estimation error when using the sample mean to estimate the population mean; (ii) the discretization error associated with the numerical integration based on the grid points $\left\{b_s = (s/S)^c\right\}_{s=0}^{S}$. One can reduce the estimation error by increasing the number of MCMC samples (i.e., J) for each power posterior. To reduce the discretization error, a simple solution is to increase the number of grid points (i.e., S). Annis et al. (2019) suggest using 30–35 or more grid points. Unfortunately, a larger S leads to a higher computational cost. Moreover, it is unclear how to choose c. As pointed in Annis et al. (2019), currently, there is no known solution to the choice of the grid points.

An interesting feature in the SS algorithm is that although $\prod_{s=0}^{S-1}\widehat{r}(b_s)$ is an unbiased estimate of $m(\mathbf{y})$, $\sum_{s=0}^{S-1}\ln\widehat{r}(b_s)$ is a biased estimate of $\ln m(\mathbf{y})$ due to the Jensen inequality. While the bias reduces as the number of grid points increases, a larger S leads to higher computational cost.

The TI algorithm estimates the marginal likelihood on the log scale, while the SS algorithm estimates the marginal likelihood on the exponential scale. Hence, the TI algorithm may enjoy better computational stability. However, the SS algorithm is not subject to the discretization error associated with numerical integration.

Both algorithms require obtaining MCMC samples for $S+1$ times. As these MCMC samples are independent, the task can be parallelized but may nonetheless be computationally expensive.

Sampling from the power posteriors, which almost always correspond to non-standard distributions, cannot be done by simply revising the code that samples from the original posterior. Extra coding efforts are required. For example, the Student t distribution is often represented as a normal-gamma mixture distribution, facilitating Gibbs sampling. Unfortunately, in the power likelihood, this normal-gamma mixture representation is not available anymore, making Gibbs sampling more difficult. Furthermore, in popular Bayesian software, such as WinBUGS, it may not be possible to use built-in functions to obtain MCMC samples from non-standard distributions. Hence, researchers must first derive expressions for model-specific power posteriors and then write a new code to draw MCMC samples from the power posteriors. This extra coding effort often poses a challenge to applied researchers.

Last but not least, although both algorithms can provide reliable estimates to the marginal likelihood as shown in Annis et al. (2019), the literature has not established conditions under which the estimators are consistent.

3. New approaches

To overcome the aforementioned disadvantages in the existing TI and SS algorithms, we now develop new approaches to the estimation of the marginal likelihood based on the power posteriors.

3.1. BvM theorem for the original posterior and power posteriors

In this subsection, we first review the standard BvM theorem for the original posterior and then present the BvM theorem for the power posterior without proving it. Let $p\left(\mathbf{y}|\boldsymbol{\theta}\right)$ be a parametric model for the observed data $\mathbf{y} = \{y_t\}_{t=0}^{n}$. The set of q parameters in the model is $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^q$. For $0 \leq t \leq n$, we denote $\mathbf{y}^t = (y_0, y_1, \ldots, y_t)$, $l_t\left(\boldsymbol{\theta}\right) = \ln p\left(\mathbf{y}^t|\boldsymbol{\theta}\right) - \ln p\left(\mathbf{y}^{t-1}|\boldsymbol{\theta}\right)$ the conditional log-likelihood for the tth observation for any $1 \leq t \leq n$. and $l_t^{(j)}\left(\boldsymbol{\theta}\right)$ the jth derivative of $l_t\left(\boldsymbol{\theta}\right)$. Hence, the log-likelihood function $\mathcal{L}_n(\boldsymbol{\theta})$ ($= \ln p(\mathbf{y}|\boldsymbol{\theta})$ conditional on the initial observation) and its kth order derivative can be expressed as

$$\mathcal{L}_n(\boldsymbol{\theta}) = \ln p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{t=1}^{n} l_t(\boldsymbol{\theta}) + \ln p\left(\mathbf{y}^0|\boldsymbol{\theta}\right) = \sum_{t=1}^{n} l_t(\boldsymbol{\theta}),$$

$$\mathcal{L}_n^{(k)}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}^k \ln p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{t=1}^{n} l_t^{(k)}(\boldsymbol{\theta}).$$

where $\ln p\left(\mathbf{y}^0|\boldsymbol{\theta}\right)$ is assumed to be zero without loss of generality.

Let $\boldsymbol{\theta}^0$ be the pseudo true value that minimizes the Kullback–Leibler loss between the data generating process (DGP) and the candidate model, that is,

$$\boldsymbol{\theta}^0 = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int \ln \frac{p_0(\mathbf{y})}{p(\mathbf{y}|\boldsymbol{\theta})} p_0(\mathbf{y}) d\mathbf{y}, \tag{15}$$

where $p_0(\mathbf{y})$ is the density of the DGP. Let P_0 denote the probability distribution corresponding to p_0. Note that we do not require $p_0(\mathbf{y})$ and $p(\mathbf{y}|\boldsymbol{\theta})$ to be the same or even $p(\mathbf{y}|\boldsymbol{\theta})$ to be a good model in the sense of Li et al. (2020). In other words, model misspecification is allowed.

Let $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_b$ be the maximum likelihood (ML) estimate of parameters corresponding to the original model whose likelihood function is $p(\mathbf{y}|\boldsymbol{\theta})$ and the model whose likelihood function is $p(\mathbf{y}|\boldsymbol{\theta}_b)^b$, respectively. It is easy to show that,

$$\widehat{\boldsymbol{\theta}}_b = \arg\min_{\boldsymbol{\theta}_b \in \boldsymbol{\Theta}} \ln\left(p(\mathbf{y}|\boldsymbol{\theta}_b)^b\right) = \arg\min_{\boldsymbol{\theta}_b \in \boldsymbol{\Theta}} b \ln p(\mathbf{y}|\boldsymbol{\theta}_b) = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ln p(\mathbf{y}|\boldsymbol{\theta}) = \widehat{\boldsymbol{\theta}}. \tag{16}$$

Furthermore, let

$$\boldsymbol{\Sigma}_n = \left(-\mathcal{L}_n^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)\right)^{-1} = \left(-\mathcal{L}_n^{(2)}\left(\widehat{\boldsymbol{\theta}}_b\right)\right)^{-1}. \tag{17}$$

According to the standard BvM, as $n \to \infty$, $\boldsymbol{\Sigma}_n^{-\frac{1}{2}}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)|\mathbf{y}$ converges in total variation to $N\left(\mathbf{0}, I_q\right)$ in P_0-probability, where $\boldsymbol{\theta}$ has density $p(\boldsymbol{\theta}|\mathbf{y})$. With a slight but important adjustment, as shown in Theorem 4.1, the BvM theorem applies to the power posterior. In particular, for any $b \in (0, 1]$, as $n \to \infty$, $\sqrt{b}\boldsymbol{\Sigma}_n^{-\frac{1}{2}}\left(\boldsymbol{\theta}_b - \widehat{\boldsymbol{\theta}}\right)|\mathbf{y}, b$ converges in total variation to $N\left(\mathbf{0}, I_q\right)$ in P_0-probability, where $\boldsymbol{\theta}_b$ has density $p(\boldsymbol{\theta}_b|\mathbf{y}, b)$.

This new BvM theorem motivates us to propose the following linear transformation and obtain an approximation to the power posterior via

$$\boldsymbol{\theta}_b = \frac{1}{\sqrt{b}}\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right) + \bar{\boldsymbol{\theta}}, \tag{18}$$

where $\boldsymbol{\theta}$ has density $p(\boldsymbol{\theta}|\mathbf{y})$, $\bar{\boldsymbol{\theta}} = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$ is the posterior mean of $\boldsymbol{\theta}$. Eq. (18) implies that $\boldsymbol{\theta} = \sqrt{b}\left(\boldsymbol{\theta}_b - \bar{\boldsymbol{\theta}}\right) + \bar{\boldsymbol{\theta}}$. We denote the probability density function of $\boldsymbol{\theta}_b$ conditional on $\mathbf{y}$ and b defined in Eq. (18) as $p_A(\boldsymbol{\theta}_b|\mathbf{y}, b)$, and call it approximate power posterior density. By the change-of-variable technique, $p_A(\boldsymbol{\theta}_b|\mathbf{y}, b)$ can be expressed as:

$$p_A(\boldsymbol{\theta}_b|\mathbf{y}, b) = p(\boldsymbol{\theta}|\mathbf{y})\left|\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\theta}_b}\right| = p(\boldsymbol{\theta}|\mathbf{y}) b^{\frac{q}{2}}. \tag{19}$$

where $\boldsymbol{\theta} = \sqrt{b}\left(\boldsymbol{\theta}_b - \bar{\boldsymbol{\theta}}\right) + \bar{\boldsymbol{\theta}}$. According to the new BvM theorem, $p_A(\boldsymbol{\theta}_b|\mathbf{y}, b)$ converges to the same normal distribution as $p(\boldsymbol{\theta}_b|\mathbf{y}, b)$. Hence, when n is moderate or large, $p_A(\boldsymbol{\theta}_b|\mathbf{y}, b)$ provides a good approximation to $p(\boldsymbol{\theta}_b|\mathbf{y}, b)$ and random samples from $p_A(\boldsymbol{\theta}_b|\mathbf{y}, b)$ serve as good approximations to those from $p(\boldsymbol{\theta}_b|\mathbf{y}, b)$. We are now in the position to apply the BvM theorem to modify the TI and SS algorithms.

3.2. TI-LWY algorithm

Based on the linear transformation given in Eq. (18), we are now in the position to design our TI-LWY algorithm to avoid many repeated MCMC sampling from a sequence of power posteriors as in TI algorithm.

Before we do that, a trivial but important step is needed, that is, a reparameterization step. In practice, it is often the case that the parameters have certain boundaries. For instance, the precision parameter, defined as the inverse of variance parameter, $h = 1/\sigma^2$, is naturally restricted to be positive. Therefore, applying the linear transformation in Eq. (18) to h may obtain a h_b being negative and hence not well-defined. Therefore, to avoid the possible boundary issues in the linear transformation in Eq. (18), we need to do a reparameterization to $\boldsymbol{\theta}$ first. Let $\boldsymbol{\phi} = g^{-1}(\boldsymbol{\theta})$ be the reparameterization of $\boldsymbol{\theta}$, where $g^{-1}(\cdot): \boldsymbol{\Theta} \to \boldsymbol{\Phi}$ is some one-to-one mapping function from $\boldsymbol{\theta}$ to $\boldsymbol{\phi}$, where $\boldsymbol{\Phi} = g^{-1}(\boldsymbol{\Theta})$ is the parameter space of $\boldsymbol{\phi}$. Let $p_{\boldsymbol{\phi}}(\boldsymbol{\phi})$, $p_{\boldsymbol{\phi}}(\mathbf{y}|\boldsymbol{\phi})$, $p_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\mathbf{y})$, $p_{\boldsymbol{\phi}}\left(\boldsymbol{\phi}_b|\mathbf{y}, b\right)$, $p_{A\boldsymbol{\phi}}\left(\boldsymbol{\phi}_b|\mathbf{y}, b\right)$, $m_{\boldsymbol{\phi}}(\mathbf{y})$, $m_{\boldsymbol{\phi}}(\mathbf{y}|b)$ denote the prior density function, the likelihood function, the posterior distribution, the power posterior distribution, the approximate power posterior distribution, the marginal likelihood function, and the power marginal likelihood function after reparameterization, respectively. The reparameterization aims to ensure that, the linear transformation based on parameter $\boldsymbol{\phi}$ does not suffer from the boundary problem, that is, $\boldsymbol{\phi}_b = \frac{1}{\sqrt{b}}\left(\boldsymbol{\phi} - \bar{\boldsymbol{\phi}}\right) + \bar{\boldsymbol{\phi}}$, satisfy $\boldsymbol{\phi}_b \in \boldsymbol{\Phi}$ for any $b \in (0, 1]$.

An easy way to conduct such a reparameterization is to map the original parameter space $\boldsymbol{\Theta}$ to a complete space, for instance, the whole real number space. Again, taking the precision parameter h as an example, the reparameterization could be $g^{-1}(h) = \ln h \in \mathbb{R}$. For more example illustrations, see Remark 4.6 in Section 4.2.

Now, suppose $\left\{\boldsymbol{\theta}_{1,(j)}\right\}_{j=1}^{J}$ are random samples from the original posterior $p(\boldsymbol{\theta}|\mathbf{y})$. Based on the reparameterization $\boldsymbol{\phi}_{1,(j)} = g^{-1}\left(\boldsymbol{\theta}_{1,(j)}\right)$, we obtain $\left\{\boldsymbol{\phi}_{1,(j)}\right\}_{j=1}^{J}$ random samples from the posterior $p_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\mathbf{y})$.[4] For any $b \in (0, 1]$, by the linear transformation

$$\boldsymbol{\phi}_{b,(j)}^{tr} = \frac{1}{\sqrt{b}}\left(\boldsymbol{\phi}_{1,(j)} - \bar{\boldsymbol{\phi}}_J\right) + \bar{\boldsymbol{\phi}}_J, \text{ where } \bar{\boldsymbol{\phi}}_J = \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{\phi}_{1,(j)}, \tag{20}$$

we obtain $\left\{\boldsymbol{\phi}_{b,(j)}^{tr}\right\}_{j=1}^{J}$, which are random samples from $p_{A\boldsymbol{\phi}}\left(\boldsymbol{\phi}_b|\mathbf{y}, b\right)$.

Following the TI algorithm, we now show how to apply the linear transformation and the importance sampling technique to obviate the need to sample from the power posteriors.

[4] The reparameterization is to ensure that the linear transformation in Eq. (18) is well-defined. See Remark 4.6 in Section 4.2 for details about this reparameterization.

Then, from Eq. (4), with the reparameterization, we have

$$\begin{aligned}\mathcal{U}(b) &= \int_{\Phi} \ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_b\right) p_{\phi}\left(\boldsymbol{\phi}_b|\mathbf{y}, b\right) \mathrm{d}\boldsymbol{\phi}_b \\ &= \int_{\Phi} \ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_b\right) \frac{p_{\phi}\left(\boldsymbol{\phi}_b|\mathbf{y}, b\right)}{p_{A\phi}\left(\boldsymbol{\phi}_b|\mathbf{y}, b\right)} p_{A\phi}\left(\boldsymbol{\phi}_b|\mathbf{y}, b\right) \mathrm{d}\boldsymbol{\phi}_b \\ &= \int_{\Phi} \ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_b\right) w_{\phi}\left(\boldsymbol{\phi}_b\right) p_{A\phi}\left(\boldsymbol{\phi}_b|\mathbf{y}, b\right) \mathrm{d}\boldsymbol{\phi}_b,\end{aligned}$$

where

$$\begin{aligned}w_{\phi}\left(\boldsymbol{\phi}_b\right) &= \frac{p_{\phi}\left(\boldsymbol{\phi}_b|\mathbf{y}, b\right)}{p_{A\phi}\left(\boldsymbol{\phi}_b|\mathbf{y}, b\right)} = \frac{p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_b\right)^b p_{\phi}\left(\boldsymbol{\phi}_b\right)/m_{\phi}(\mathbf{y}|b)}{b^{\frac{q}{2}} p_{\phi}(\mathbf{y}|\boldsymbol{\phi}) p_{\phi}(\boldsymbol{\phi})/m_{\phi}(\mathbf{y})} \\ &= \frac{p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_b\right)^b p_{\phi}\left(\boldsymbol{\phi}_b\right)}{p_{\phi}(\mathbf{y}|\boldsymbol{\phi}) p_{\phi}(\boldsymbol{\phi})} \frac{m_{\phi}(\mathbf{y})}{b^{\frac{q}{2}} m_{\phi}(\mathbf{y}|b)}.\end{aligned} \tag{21}$$

where $\boldsymbol{\phi}_b$ has density $p_{A\phi}\left(\boldsymbol{\phi}_b|\mathbf{y}, b\right)$, and $\boldsymbol{\phi} = \sqrt{b}\left(\boldsymbol{\phi}_b - \bar{\boldsymbol{\phi}}\right) + \bar{\boldsymbol{\phi}}$. We can estimate $\mathcal{U}(b)$ using either the ordinary importance sampler or the self-normalized importance sampler, defined, respectively, by[5]

$$\begin{aligned}\widehat{\mathcal{U}}_w(b) &= \frac{1}{J} \sum_{j=1}^{J} \ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{b,(j)}^{tr}\right) w_{\phi}\left(\boldsymbol{\phi}_{b,(j)}^{tr}\right), \\ \widehat{\mathcal{U}}_{LWY}(b) &= \sum_{j=1}^{J} \ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{b,(j)}^{tr}\right) \widehat{W}_{\phi}\left(\boldsymbol{\phi}_{b,(j)}^{tr}\right),\end{aligned} \tag{22}$$

where

$$\begin{aligned}\widehat{W}_{\phi}\left(\boldsymbol{\phi}_{b,(j)}^{tr}\right) &= \frac{w_{\phi}\left(\boldsymbol{\phi}_{b,(j)}^{tr}\right)}{\sum_{j=1}^{J} w_{\phi}\left(\boldsymbol{\phi}_{b,(j)}^{tr}\right)} \\ &= \frac{p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{b,(j)}^{tr}\right)^b p_{\phi}\left(\boldsymbol{\phi}_{b,(j)}^{tr}\right) / \left[p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{1,(j)}\right) p_{\phi}\left(\boldsymbol{\phi}_{1,(j)}\right)\right]}{\sum_{j=1}^{J} p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{b,(j)}^{tr}\right)^b p_{\phi}\left(\boldsymbol{\phi}_{b,(j)}^{tr}\right) / \left[p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{1,(j)}\right) p_{\phi}\left(\boldsymbol{\phi}_{1,(j)}\right)\right]} \\ &= \frac{\exp\left\{b \ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{b,(j)}^{tr}\right) - \ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{1,(j)}\right) + \ln p_{\phi}\left(\boldsymbol{\phi}_{b,(j)}^{tr}\right) - \ln p_{\phi}\left(\boldsymbol{\phi}_{1,(j)}\right)\right\}}{\sum_{j=1}^{J} \exp\left\{b \ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{b,(j)}^{tr}\right) - \ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{1,(j)}\right) + \ln p_{\phi}\left(\boldsymbol{\phi}_{b,(j)}^{tr}\right) - \ln p_{\phi}\left(\boldsymbol{\phi}_{1,(j)}\right)\right\}}.\end{aligned} \tag{23}$$

Remark 3.1. It is important to note that the BvM theorem is only used to find a reasonable proposal distribution for importance sampling. There is no need for the proposal distribution to be very close to the normal distribution, and hence, n does not need to go to infinity for our methods to consistently estimate the marginal likelihood. As pointed out by a referee, our idea bears some resemblance to bootstrap where the asymptotic normality is used to justify bootstrap, but bootstrap samples do not come from the normal distribution.

Remark 3.2. While we do not use the normal distribution for importance sampling, as suggested by a referee, the BvM theorem suggests that the normal distribution can also be used as a proposal distribution. This alternative importance sampler requires additional sampling from the normal distribution, whereas in the proposed methods, we only need the random samples from the original posterior.

Remark 3.3. In practice, with a finite n, when the auxiliary constant b is very small, for example, when $b \leq 1/n$, the power posterior may be far away from the normal distribution.[6] In this case, the prior distribution provides a good approximation to the power posterior $p\left(\boldsymbol{\theta}_b|\mathbf{y}, b\right)$ and hence, can be used as the proposal distribution. Let $\left\{\boldsymbol{\theta}_{0,(j)}\right\}_{j=1}^{J_0}$ be random samples (usually iid) from prior $p(\boldsymbol{\theta})$. Note that

$$\mathcal{U}(b) = \int_{\Theta} \ln p\left(\mathbf{y}|\boldsymbol{\theta}\right) w_{0b}\left(\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta},$$

5 While $\widehat{\mathcal{U}}_w(b)$ is an unbiased estimator of $U(b)$, it is infeasible because of an unknown constant involved. Although $\widehat{\mathcal{U}}_{LWY}(b)$ is a slightly biased estimator of $U(b)$, it is consistent as we will show in the appendix.

6 We can prove that the BvM theorem shown in next section remains valid when b is made to be dependent on n such that $n \times b \sim O(n^{\alpha})$ with $\alpha > 0$.

where

$$w_{0b}\left(\boldsymbol{\theta}\right)=\frac{p\left(\boldsymbol{\theta}|\mathbf{y},b\right)}{p\left(\boldsymbol{\theta}\right)}=\frac{\frac{p(\mathbf{y}|\boldsymbol{\theta})^{b}p(\boldsymbol{\theta})}{m(\mathbf{y}|b)}}{p\left(\boldsymbol{\theta}\right)}=\frac{p\left(\mathbf{y}|\boldsymbol{\theta}\right)^{b}}{m(\mathbf{y}|b)}. \tag{24}$$

where $\boldsymbol{\theta}$ has density $p(\boldsymbol{\theta})$. We can estimate $\mathcal{U}(b)$ using either the ordinary importance sampler or the self-normalized importance sampler, defined, respectively, by

$$\widehat{\mathcal{U}}_{w}(b)=\frac{1}{J_0}\sum_{j=1}^{J_0}\ln p\left(\mathbf{y}|\boldsymbol{\theta}_{0,(j)}\right)w_{0b}\left(\boldsymbol{\theta}_{0,(j)}\right),$$

$$\widehat{\mathcal{U}}_{LWY}(b)=\sum_{j=1}^{J_0}\ln p\left(\mathbf{y}|\boldsymbol{\theta}_{0,(j)}\right)\widehat{W}_{0b}\left(\boldsymbol{\theta}_{0,(j)}\right), \tag{25}$$

where

$$\widehat{W}_{0b}\left(\boldsymbol{\theta}_{0,(j)}\right)=\frac{w_{0b}\left(\boldsymbol{\theta}_{0,(j)}\right)}{\sum_{j=1}^{J_0}w_{0b}\left(\boldsymbol{\theta}_{0,(j)}\right)}=\frac{p\left(\mathbf{y}|\boldsymbol{\theta}_{0,(j)}\right)^{b}}{\sum_{j=1}^{J_0}p\left(\mathbf{y}|\boldsymbol{\theta}_{0,(j)}\right)^{b}}=\frac{\exp\left\{b\ln p\left(\mathbf{y}|\boldsymbol{\theta}_{0,(j)}\right)\right\}}{\sum_{j=1}^{J_0}\exp\left\{b\ln p\left(\mathbf{y}|\boldsymbol{\theta}_{0,(j)}\right)\right\}}. \tag{26}$$

In practice, when the sample size n is moderate or large, $1/n$ is small and there are very few grid points in $[0, \frac{1}{n}]$.

In this paper, we denote the new algorithm the TI-LWY algorithm, which is summarized as follows:

TI-LWY Algorithm

1. Choose grid points $\left\{b_s=(s/S)^c\right\}_{s=0}^{S}$ with $c\geq 1$.
2. Draw J_0 samples $\left\{\boldsymbol{\theta}_{0,(j)}\right\}_{j=1}^{J_0}$ from the prior $p(\boldsymbol{\theta})$ and compute $\widehat{\mathcal{U}}_{LWY}(0)=\widehat{\mathcal{U}}(0)=\frac{1}{J_0}\sum_{j=1}^{J_0}\ln p\left(\mathbf{y}|\boldsymbol{\theta}_{0,(j)}\right)$.
3. When $0<b_s\leq 1/n$, compute $\widehat{\mathcal{U}}_{LWY}(b_s)$ from Eq. (25).
4. Draw J samples $\left\{\boldsymbol{\theta}_{1,(j)}\right\}_{j=1}^{J}$ from the posterior $p(\boldsymbol{\theta}|\mathbf{y})$.
5. Based on the reparameterization $\boldsymbol{\phi}_{1,(j)}=g^{-1}\left(\boldsymbol{\theta}_{1,(j)}\right)$, obtain J samples $\left\{\boldsymbol{\phi}_{1,(j)}\right\}_{j=1}^{J}$ from the original posterior distribution $p_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\mathbf{y})$.
6. When $b_s>1/n$, we get

$$\boldsymbol{\phi}_{b_s,(j)}^{tr}=\frac{1}{\sqrt{b_s}}\left(\boldsymbol{\phi}_{1,(j)}-\bar{\boldsymbol{\phi}}_J\right)+\bar{\boldsymbol{\phi}}_J,\text{ where }\bar{\boldsymbol{\phi}}_J=\frac{1}{J}\sum_{j=1}^{J}\boldsymbol{\phi}_{1,(j)}.$$

   For each $b_s>1/n$, compute $\widehat{\mathcal{U}}_{LWY}(b_s)$ by Eqs. (22).
7. The log marginal likelihood is estimated by

$$\widehat{\ln m}_{TI-LWY}(\mathbf{y}):=\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})+\widehat{\mathcal{U}}_{LWY}(b_s)}{2}. \tag{27}$$

Remark 3.4. In practice, when computing the weights in Eqs. (23) and (26), to ensure computational stability, we suggest demeaning the log-likelihood shown in the exponential term first.

Remark 3.5. Based on the simple linear transformation in step 6, the TI-LWY algorithm only requires the availability of $\left\{\boldsymbol{\theta}_{1,(j)}\right\}_{j=1}^{J}$ which are effective random samples generated from the original posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. There is no need to draw MCMC samples from the power posteriors. Hence, our method significantly reduces the computational cost. Moreover, coding efforts associated with drawing MCMC samples from the power posteriors are completely avoided.

Remark 3.6. An important difference between the TI algorithm and the TI-LWY algorithm is that the latter is based on the importance sampling approach where the proposal distribution is developed based on the BvM theorem. When the sample size is very small such that the posterior distribution is very far away from the normal distribution, the TI algorithm is still a good choice. However, TI-LWY can reduce a lot of computational efforts when the sample size is moderate or large.

Remark 3.7. Hoehna et al. (2017) explain how to use parallel computing to implement the TI algorithm. One can also use parallel computing to implement the TI-LWY algorithm. This is because, while our method does not require MCMC sampling from the power posteriors, we need to evaluate the likelihood function and obtain the importance weights at each grid point. These calculations can be parallelized too. In the examples discussed below, we only report the CPU time without resorting to parallel computing. If parallel computing is used, the computing time of both algorithms can be reduced, but the relative computational cost would remain the same.

3.3. SS-LWY algorithm

Based on Eq. (11), with the reparameterization, we have

$$\begin{aligned}
r(b_s) &= \int_{\Phi} \exp\left[(b_{s+1}-b_s)\ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{b_s}\right)\right] p_{\phi}\left(\boldsymbol{\phi}_{b_s}|\mathbf{y}, b_s\right) d\boldsymbol{\phi}_{b_s} \\
&= \int_{\Phi} \exp\left[(b_{s+1}-b_s)\ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{b_s}\right)\right] \frac{p_{\phi}\left(\boldsymbol{\phi}_{b_s}|\mathbf{y}, b_s\right)}{p_{A\phi}\left(\boldsymbol{\phi}_{b_s}|\mathbf{y}, b_s\right)} p_{A\phi}\left(\boldsymbol{\phi}_{b_s}|\mathbf{y}, b_s\right) \mathrm{d}\boldsymbol{\phi}_{b_s} \\
&= \int_{\Phi} \exp\left[(b_{s+1}-b_s)\ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}_{b_s}\right)\right] w_{\phi}\left(\boldsymbol{\phi}_{b_s}\right) p_{A\phi}\left(\boldsymbol{\phi}_{b_s}|\mathbf{y}, b_s\right) \mathrm{d}\boldsymbol{\phi}_{b_s},
\end{aligned}$$

where $w_{\phi}\left(\boldsymbol{\phi}_{b_s}\right)$ is the same as in Eq. (21).

Let $\left\{\boldsymbol{\theta}_{1,(j)}, \boldsymbol{\phi}_{1,(j)}, \boldsymbol{\phi}^{tr}_{b_s,(j)}\right\}_{j=1}^{J}$ be draws as defined in Section 3.2. We can estimate $r(b_s)$ using the self-normalized importance sampler defined by

$$\begin{aligned}
\widehat{r}_{LWY}(b_s) &= \sum_{j=1}^{J} \exp\left[(b_{s+1}-b_s)\left(\ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}^{tr}_{b_s,(j)}\right) - \overline{L}^{tr}_{b_s}\right) + (b_{s+1}-b_s)\overline{L}^{tr}_{b_s}\right] \widehat{W}_{\phi}\left(\boldsymbol{\phi}^{tr}_{b_s,(j)}\right) \\
&= \exp\left[(b_{s+1}-b_s)\,\overline{L}^{tr}_{b_s}\right] \sum_{j=1}^{J} \exp\left[(b_{s+1}-b_s)\left(\ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}^{tr}_{b_s,(j)}\right) - \overline{L}^{tr}_{b_s}\right)\right] \widehat{W}_{\phi}\left(\boldsymbol{\phi}^{tr}_{b_s,(j)}\right),
\end{aligned} \tag{28}$$

where $\overline{L}^{tr}_{b_s} = \max_{j\in\{1,\ldots,J\}}\left\{\ln p_{\phi}\left(\mathbf{y}|\boldsymbol{\phi}^{tr}_{b_s,(j)}\right)\right\}$ and $\left\{\widehat{W}_{\phi}\left(\boldsymbol{\phi}^{tr}_{b_s,(j)}\right)\right\}_{j=1}^{J}$ are the same as in Eq. (23).

Remark 3.8. Similar to the TI algorithm, when b_s is very small, we suggest using the prior distribution as the proposal distribution. For $b_s \in [0, 1/n]$, using notations in Remark 3.3, we have

$$r(b_s) = \int_{\Theta} \exp\left[(b_{s+1}-b_s)\ln p\left(\mathbf{y}|\boldsymbol{\theta}\right)\right] w_{0b_s}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

where $w_{0b_s}(\boldsymbol{\theta})$ is the same as defined in Eq. (24). We can estimate $r(b_s)$ using the self-normalized importance sampler defined by

$$\begin{aligned}
\widehat{r}_{LWY}(b_s) &= \sum_{j=1}^{J_0} \exp\left[(b_{s+1}-b_s)\ln p\left(\mathbf{y}|\boldsymbol{\theta}_{0,(j)}\right)\right] \widehat{W}_{0b_s}\left(\boldsymbol{\theta}_{0,(j)}\right) \\
&= \exp\left[(b_{s+1}-b_s)\overline{L}_0\right] \sum_{j=1}^{J_0} \exp\left[(b_{s+1}-b_s)\left(\ln p\left(\mathbf{y}|\boldsymbol{\theta}_{0,(j)}\right) - \overline{L}_0\right)\right] \widehat{W}_{0b_s}\left(\boldsymbol{\theta}_{0,(j)}\right),
\end{aligned} \tag{29}$$

where $\overline{L}_0 = \max_{j\in\{1,\ldots,J_0\}}\left\{\ln p\left(\mathbf{y}|\boldsymbol{\theta}_{0,(j)}\right)\right\}$.

In this paper, we denote the new algorithm the SS-LWY algorithm, which is summarized as follows:

SS-LWY Algorithm

1. Choose grid points $\left\{b_s = (s/S)^c\right\}_{s=0}^{S}$ with $c \geq 1$.
2. When $0 \leq b_s \leq 1/n$, draw J_0 samples $\left\{\boldsymbol{\theta}_{0,(j)}\right\}_{j=1}^{J_0}$ from the prior distribution $p(\boldsymbol{\theta})$ and compute $\widehat{r}_{LWY}(b_s)$ by Eq. (29).
3. Draw J samples $\left\{\boldsymbol{\theta}_{1,(j)}\right\}_{j=1}^{J}$ from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$.
4. Based on the reparameterization $\boldsymbol{\phi}_{1,(j)} = g^{-1}\left(\boldsymbol{\theta}_{1,(j)}\right)$, obtain samples $\left\{\boldsymbol{\phi}_{1,(j)}\right\}_{j=1}^{J}$ from the posterior distribution $p_{\phi}(\boldsymbol{\phi}|\mathbf{y})$.
5. When $b_s > 1/n$, obtain
$$\boldsymbol{\phi}^{tr}_{b_s,(j)} = \frac{1}{\sqrt{b_s}}\left(\boldsymbol{\phi}_{1,(j)} - \bar{\boldsymbol{\phi}}_J\right) + \bar{\boldsymbol{\phi}}_J, \text{ where } \bar{\boldsymbol{\phi}}_J = \frac{1}{J}\sum_{j=1}^{J}\boldsymbol{\phi}_{1,(j)}.$$
For each $b_s > 1/n$, compute $\widehat{r}_{LWY}(b_s)$ by Eq. (28).
6. The log-marginal likelihood is estimated by
$$\sum_{s=0}^{S-1} \ln \widehat{r}_{LWY}(b_s). \tag{30}$$

4. Asymptotic properties

In this section, we give a set of primitive assumptions. Under these assumptions, we obtain the BvM theorem for the power posteriors. As in the BvM literature (Schervish, 2012), our framework is probabilistic in nature, that is, under repeated sampling with a fixed pseudo true parameter value. We then give a set of primitive conditions. Under these conditions, we show the consistency of the TI and the modified TI algorithms. Following the conventional practice in the Bayesian large sample literature (Chen, 1985; Kass et al., 1990), we assume the data $\mathbf{y}$ with a finite and fixed n is given. Hence, the approach is non-probabilistic in nature.

4.1. BvM for power posteriors

Assumption 1: The pseudo-true value $\boldsymbol{\theta}^0$ is the interior point of some compact set $\boldsymbol{\Theta}$.

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

Assumption 3: For all t, $l_t(\boldsymbol{\theta})$ is three-times differentiable on $\boldsymbol{\Theta}$ almost surely.

Assumption 4: For $j = 0, 1, 2$, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, $\left\| l_t^{(j)}(\boldsymbol{\theta}) - l_t^{(j)}(\boldsymbol{\theta}') \right\| \leq c_t^j(\mathbf{y}^t) \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|$ holds with probability 1, where $c_t^j(\mathbf{y}^t)$ is a positive random variable with $\sup_t E\left\| c_t^j(\mathbf{y}^t) \right\| < \infty$ and $\frac{1}{n}\sum_{t=1}^{n}\left(c_t^j(\mathbf{y}^t) - E\left(c_t^j(\mathbf{y}^t)\right)\right) \xrightarrow{P_0} 0$ where $l_t^{(0)}(\boldsymbol{\theta}) = l_t(\boldsymbol{\theta})$, where P_0 is the probability with respect to the DGP P_0.

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

Assumption 6: $\left\{ l_t^{(j)}(\boldsymbol{\theta}) \right\}$ is L_2-near epoch dependent with respect to $\{\mathbf{y}_t\}$ of size -1 when $j = 0, 1$, and size $-\frac{1}{2}$ when $j = 2$ uniformly on $\boldsymbol{\Theta}$.

Assumption 7: For $\delta > 0$, let $N_0(\delta)$ be the open ball of radius δ around $\boldsymbol{\theta}^0$. If $N_0(\delta) \subseteq \boldsymbol{\Theta}$, there exists some $K(\delta) > 0$ such that

$$\lim_{n\to\infty} P_0\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus N_0(\delta)} \lambda_n \left[\mathcal{L}_n(\boldsymbol{\theta}) - \mathcal{L}_n\left(\boldsymbol{\theta}^0\right) \right] < -K(\delta) \right) = 1,$$

where λ_n is the smallest eigenvalue of $n\boldsymbol{\Sigma}_n = \left[-\frac{1}{n}\mathcal{L}_n^{(2)}\left(\widehat{\boldsymbol{\theta}}\right) \right]^{-1}$ and P_0 is the probability with respect to the DGP P_0.

Assumption 8: For any $\epsilon > 0$, there exists $\delta(\epsilon) > 0$ such that

$$\lim_{n\to\infty} P_0\left(\sup_{\boldsymbol{\theta}\in N_0(\delta(\epsilon)), \|\mathbf{r}_0\|=1} \left| 1 + \mathbf{r}_0' \boldsymbol{\Sigma}_n^{1/2} \mathcal{L}_n^{(2)}(\boldsymbol{\theta}) \boldsymbol{\Sigma}_n^{1/2} \mathbf{r}_0 \right| < \epsilon \right) = 1.$$

where $\mathbf{r}_0$ is a q-dimensional vector.

Assumption 9: The sequence of matrices $E\left[-\frac{1}{n}\mathcal{L}_n^{(2)}\left(\boldsymbol{\theta}^0\right) \right]$ is uniformly positive definite in n.

Assumption 10: The prior density $p(\boldsymbol{\theta})$ is continuous on the set $\boldsymbol{\Theta}$ and $0 < p\left(\boldsymbol{\theta}^0\right) < +\infty$.

Remark 4.1. These assumptions are general regularity conditions and are easy to verify. Assumption 1 is the compactness condition and allows for model misspecification. Assumption 2 implies weak dependence in $\mathbf{y}_t$. Assumption 3 is the continuity condition. Assumption 4 is the Lipschitz condition for l_t to ensure the uniform law of large numbers for dependent and heterogeneous stochastic processes. Assumption 5 is the dominance condition for l_t. Assumption 6 implies weak dependence in l_t. Assumption 7 is the identification condition. Assumption 8 is the smoothness condition. Assumption 9 is the standard condition for the information matrix. Assumption 10 is the standard condition for the prior distribution. These assumptions are imposed on developing the BvM theorem under repeated sampling with a fixed pseudo true $\boldsymbol{\theta}^0$. We need to point out other similar sets of regularity conditions are also available in the literature for developing the BvM theorem, especially in the iid case. One can refer to the textbook treatments of this subject matter in Gelman et al. (2004, Appendix B) and Schervish (2012, section 7.4.2) and many references therein.

Remark 4.2. Schervish (2012) presents the BvM theorem based on convergence in total variation for correctly specified models with iid data. Liu et al. (2021) relaxed the iid assumption by allowing for weak dependence. Kleijn and van der Vaart (2012) and Müller (2013) extend the BvM theorem to misspecified models with iid data and weakly dependent data, respectively. For iid data, Gelman et al. (2004) present a version of the BvM theorem based on convergence in distribution. In our context, let $\mathbf{z}_n = \boldsymbol{\Sigma}_n^{-1/2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)$ where $\boldsymbol{\theta}$ has density $p(\boldsymbol{\theta}|\mathbf{y})$ and $A_n = \left\{ \mathbf{z}_n : \widehat{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_n^{1/2}\mathbf{z}_n \in \boldsymbol{\Theta} \right\}$ be the support space of $\mathbf{z}_n$. Let $B \sqsubseteq A_n$ be a Borel set. Then, according to the standard BvM theorem, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P_0\left(\sup_{B\sqsubseteq A_n} \left(\left| \Pr\left(\mathbf{z}_n \in B|\mathbf{y}\right) - \Psi(B) \right| > \varepsilon \right) \right) = 0. \tag{31}$$

where $\Psi(B)$ stands for the probability that an $N_q(\mathbf{0}, I_q)$ vector lies in B. The following theorem shows that the BvM theorem is also applicable to the power posteriors.

Theorem 4.1. *Suppose Assumptions 1–10 hold. For any auxiliary constant $b \in (0, 1]$ that does not depend on n, let $\mathbf{z}_{nb} = \left(b^{-1}\boldsymbol{\Sigma}_n\right)^{-1/2}\left(\boldsymbol{\theta}_b - \widehat{\boldsymbol{\theta}}\right)$ where $\boldsymbol{\theta}_b$ has density $p\left(\boldsymbol{\theta}_b|\mathbf{y}, b\right)$ and $A_{nb} = \left\{\mathbf{z}_{nb} : \widehat{\boldsymbol{\theta}} + \left(b^{-1}\boldsymbol{\Sigma}_n\right)^{1/2}\mathbf{z}_{nb} \in \Theta\right\}$ be the support space of $\mathbf{z}_{nb}$. Let $B \sqsubseteq A_{nb}$ be a Borel set. Then, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} P_0\left(\sup_{B\sqsubseteq A_{nb}} \left(\left|\Pr\left(\mathbf{z}_{nb} \in B|\mathbf{y}, b\right) - \Psi(B)\right| > \varepsilon\right)\right) = 0. \tag{32}$$

Remark 4.3. Since convergence in total variation implies convergence in distribution (Chapter 2 of van der Vaart (2000)), according to Theorem 4.1, we expect the distribution of $\mathbf{z}_{nb}$ is close to that of $N_q(\mathbf{0}, I_q)$ for modest and large n. This limit theory is the motivation for us to find a good proposal distribution for importance sampling when estimating the marginal likelihood.

4.2. Consistency of the TI algorithm and TI-LWY algorithms

The marginal likelihood $m(\mathbf{y})$, and hence, $\ln m(\mathbf{y})$, when they are conditional on the data $\mathbf{y}$, are not random. Note that from Lemma 7.1 in the online supplement,

$$0 \le \ln p(\mathbf{y}|\widehat{\boldsymbol{\theta}}) - \ln m(\mathbf{y}) \le \ln\left[\ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right) - \mathcal{U}(0)\right].$$

This implies that $\ln m(\mathbf{y}) = O(n)$. Therefore, when $n \to +\infty$, $\ln m(\mathbf{y})$ goes to ∞. Clearly, it is only meaningful to construct a consistent estimator of $\ln m(\mathbf{y})$ when it is fixed and finite. That is why we further assume n is fixed and finite so that $\ln m(\mathbf{y})$, when it is conditional on $\mathbf{y}$, is also fixed and finite. In this sense, our framework is the same as in Chen (1985), which is not a probabilistic sampling-theory asymptotic approach. Following Chen (1985) and Kass et al. (1990), we introduce the following regularity conditions.

Condition 1. Θ is a compact subset of $\mathbb{R}^q$.

Condition 2. For all t, $l_t(\boldsymbol{\theta})$ is six-times differentiable on Θ.

Condition 3. For $j = 0, 1, 2$, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, $\left\|l_t^{(j)}(\boldsymbol{\theta}) - l_t^{(j)}(\boldsymbol{\theta}')\right\| \le c_t^j\left(\mathbf{y}^t\right)\left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|$, where $c_t^j\left(\mathbf{y}^t\right)$ is a positive constant with $\sup_t c_t^j\left(\mathbf{y}^t\right) < \infty$, and $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^n c_t^j\left(\mathbf{y}^t\right) < \infty$.

Condition 4. For $j = 0, 1, 2$, there exists a function $M_t(\mathbf{y}^t) > 0$ such that for all $\boldsymbol{\theta} \in \Theta$, $l_t^{(j)}(\boldsymbol{\theta})$ exists, $\sup_{\boldsymbol{\theta}\in\Theta}\left\|l_t^{(j)}(\boldsymbol{\theta})\right\| \leqslant M_t(\mathbf{y}^t)$, and $\sup_t M_t(\mathbf{y}^t) \le M < \infty$.

Condition 5. For $\delta > 0$, let $N(a, \delta)$ be the open ball of radius δ around a. If $N\left(\widehat{\boldsymbol{\theta}}, \varepsilon\right) \subseteq \Theta$, then

$$\limsup_{n\to\infty}\sup_{\boldsymbol{\theta}\in\Theta\setminus N\left(\widehat{\boldsymbol{\theta}},\varepsilon\right)}\frac{1}{n}\sum_{t=1}^n\left[l_t(\boldsymbol{\theta}) - l_t\left(\widehat{\boldsymbol{\theta}}\right)\right] < 0.$$

Condition 6. $-\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)$ is positive definite for all n, and the smallest eigenvalue of $-\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)$, denoted by $\lambda_{-\bar{\mathbf{H}}_n(\widehat{\boldsymbol{\theta}})}$, is greater than $\varepsilon > 0$, where $\bar{\mathbf{H}}_n(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{t=1}^n l_t^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)$.

Condition 7. The prior density $p(\boldsymbol{\theta})$ is four-times continuously differentiable and $0 < p\left(\widehat{\boldsymbol{\theta}}\right) < +\infty$.

Condition 8. $\boldsymbol{\theta}$ has finite 16th moment under the prior $p(\boldsymbol{\theta})$. For any $i, j \in \{1, \ldots, q\}$, $\int_\Theta\left[\frac{1}{n}\mathcal{L}_{n,ij}^{(2)}(\boldsymbol{\theta}_a)\right]^8 p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} < +\infty$ where $\boldsymbol{\theta}_a = a\boldsymbol{\theta} + (1-a)\widehat{\boldsymbol{\theta}}$ for any $a \in (0, 1]$, $\mathcal{L}_{n,ij}^{(2)}(\boldsymbol{\theta}_a)$ is the $(i,j)^{th}$ element of $\mathcal{L}_n^{(2)}(\boldsymbol{\theta}_a)$ and $\int_\Theta|\ln p(\boldsymbol{\theta})|\,p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} < +\infty$.

Remark 4.4. The conditions are different from the assumptions used to develop the BvM theorem in Section 4.1. A key difference is that since the posterior and the marginal likelihood are now conditional on $\mathbf{y}$, they are no longer random. That is why we remove Assumptions 2 and 6 and modify Assumptions 4, 5, 7, 9, 10 by Condition 3, 4, 5, 6, 7, respectively.

Remark 4.5. Combining with Condition 7, Conditions 1–6 are sufficient for the conditions specified in Chen (1985) and Kass et al. (1990). In particular, Conditions 1–2 imply condition (P1) of Chen (1985), which ensures the existence of MLE.

Condition 6 implies condition (P2) and the steepness condition (C1) of Chen (1985). Condition 3 implies the smoothness condition (C2) of Chen (1985) and condition (ii) of Kass et al. (1990). Condition 4 is similar to condition (i) of Kass et al. (1990). Condition 5 implies the concentration condition (C3) of Chen (1985) and condition (iii') of Kass et al. (1990). Conditions 7–8 put restrictions on the prior distribution.

Theorem 4.2. *For $\left\{b_s = (s/S)^c\right\}_{s=0}^S$ with any $c \geq 1$, under Conditions 1–8, when the data $\mathbf{y}$ is given with a fixed and finite n, as $S, J \to +\infty$,*

$$\widehat{\ln m}_{TI}(\mathbf{y}) = \ln m(\mathbf{y}) + o_p(1), \tag{33}$$

where p in $o_p(1)$ is understood as the probability measure corresponding to the joint power posterior $p\left(\boldsymbol{\theta}_{b_s}|\mathbf{y}, b_s\right)$ with $\left\{b_s = (s/S)^c\right\}_{s=0}^S$.

To avoid the boundary problem for the linear transformation defined in Eq. (18), we impose the following condition.

Condition 9. The linear transformation in Eq. (18) is well-defined in the sense that $\boldsymbol{\theta}_b (= \frac{1}{\sqrt{b}}\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right) + \bar{\boldsymbol{\theta}}) \in \boldsymbol{\Theta}$ for any $b \in [0, 1]$. Otherwise, there exists a one-to-one mapping between $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and the new parameter $\boldsymbol{\phi} \in \boldsymbol{\Phi} = g^{-1}(\boldsymbol{\Theta})$ such that $\boldsymbol{\phi} = g^{-1}(\boldsymbol{\theta})$, $\boldsymbol{\phi}_b = \frac{1}{\sqrt{b}}\left(\boldsymbol{\phi} - \bar{\boldsymbol{\phi}}\right) + \bar{\boldsymbol{\phi}}$ with $\bar{\boldsymbol{\phi}} = \int_{\boldsymbol{\Phi}} \boldsymbol{\phi} p_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\mathbf{y}) d\boldsymbol{\phi}$ and $\boldsymbol{\phi}_b \in \boldsymbol{\Phi}$ for any $b \in [0, 1]$.

Remark 4.6. In practice, it is fairly easy to find such a reparameterization to satisfy Condition 9. For example, the degrees of freedom parameter υ in the Student t distribution is constrained to be larger than 2. Then, we can use the reparameterization $\phi = \ln(\upsilon - 2) \in \mathbb{R}$ such that any linear transformation of ϕ lies in $\mathbb{R}$. In this case, $g(\phi) = \exp(\phi) + 2$ and $g^{-1}(\upsilon) = \ln(\upsilon - 2)$. For another example, the correlation coefficient δ is constrained to be in $[-1, 1]$. Then we can use the reparameterization $\phi = \tan\left(\frac{\pi}{2}\delta\right) \in \mathbb{R}$. In this case, $g(\phi) = \frac{2}{\pi}\arctan(\phi)$ and $g^{-1}(\delta) = \tan\left(\frac{\pi}{2}\delta\right)$.

According to Remark 4.6, if $\boldsymbol{\theta}_b = \frac{1}{\sqrt{b}}\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right) + \bar{\boldsymbol{\theta}} \notin \boldsymbol{\Theta}$, a reparameterization can be made to $\boldsymbol{\phi}$ to satisfy Condition 9. For the convenience of proof, we simply assume that the linear transformation in Eq. (18) is well-defined so that $\boldsymbol{\theta}_b \in \boldsymbol{\Theta}$ for any $b \in [0, 1]$.

Theorem 4.3. *For $\left\{b_s = (s/S)^c\right\}_{s=0}^S$ with any $c \geq 1$, under Conditions 1–9, when the data $\mathbf{y}$ is given with a fixed and finite n, as $S, J_0, J \to +\infty$,*

$$\widehat{\ln m}_{TI-LWY}(\mathbf{y}) = \ln m(\mathbf{y}) + o_p(1),$$

where p in $o_p(1)$ is understood as the probability measure corresponding to the joint posterior $p(\boldsymbol{\theta}|\mathbf{y})$ and prior $p(\boldsymbol{\theta})$.

Remark 4.7. Regarding the SS algorithm, no proof of consistency is available in the literature. To save space, we do not provide the proof of its consistency, nor the proof of the consistency of the modified SS-LWY algorithm in the present paper. While these proofs require additional efforts, the key idea is similar to the proofs for the TI and modified TI algorithm developed here. In particular, we need to show that the approximation error introduced by the importance sampler converges to zero in probability. We leave the detailed and complete treatment in a future study.

5. Examples

In this section, we use two examples to evaluate and compare the performance of the proposed algorithms and the original TI and SS algorithms, including estimation performance and computational efficiency.[7]

In the first example, we consider a multivariate linear regression model with Gaussian errors for which the marginal likelihood is available in closed-form. Based on the closed-form expression, we can accurately evaluate and compare the performance of candidate algorithms. We repeat all estimation procedures 100 times to calculate the mean (or bias) and the MCSE of the marginal likelihood. To further illustrate the computational advantage of the newly proposed algorithm, we also investigate the multivariate linear regression model with Student t errors. The t distribution complicates the likelihood function as well as the power posterior sampling. It allows us to highlight the computational efficiency of the proposed algorithm relative to the TI and SS algorithms. To illustrate potential extra coding efforts required by the TI and SS algorithms, we use WinBUGS to draw MCMC samples in this example. The TI-LWY and SS-LWY algorithms are easily implementable in WinBUGS without extra coding or sampling efforts from the power posteriors. Whereas, the TI and SS algorithms require users to code the likelihood density corresponding to the power posterior using the "zeros trick" technique (Chapter 9 in Spiegelhalter et al. (2003)) in WinBUGS. Compared with using existing distributions in WinBUGS, the "zeros trick" technique greatly slows down the sampling speed. In the second example, we consider several copula

[7] Computational efficiency for the TI, TI-LWY algorithms is measured based on the CPU time on a common desktop with Intel(R) Core(TM) i7-6700 CPU @3.40 GHz. Computational efficiency for the SS, SS-LWY algorithms is measured based on the CPU time on another common desktop with Intel(R) Core(TM) i7-7500 CPU @ 2.70 GHz.

Table 1
Bias and MCSE (in parenthesis) of estimates of log marginal likelihood in $\boldsymbol{M_1}$.

	$c = 1$		$c = 3$	
	TI	TI-LWY	TI	TI-LWY
$S = 20$	−495.25(4.12)	−495.25(4.14)	−2.15(0.03)	−2.14(0.17)
$S = 40$	−244.04(2.06)	−244.04(2.09)	−0.59(0.01)	−0.58(0.22)
$S = 100$	−94.37(0.82)	−94.37(0.86)	−0.08(0.01)	−0.07(0.17)
	SS	SS-LWY	SS	SS-LWY
$S = 20$	−0.54(1.19)	−0.54(1.19)	0.00(0.02)	0.01(0.13)
$S = 40$	−0.10(0.50)	−0.10(0.51)	0.00(0.02)	0.02(0.19)
$S = 100$	−0.03(0.16)	−0.02(0.17)	0.00(0.01)	0.02(0.16)

models. Most copula models do not lead to standard distributions, making it difficult to use WinBUGS to obtain MCMC samples. In this paper, we use the "mcmc" package in R to obtain MCMC samples. To use this package, one only needs to specify the posterior density directly. As a result, no extra coding effort is required to implement the TI and SS algorithms as one can conveniently raise the original likelihood to any power $b \in (0, 1]$. However, as reported below, the TI and SS algorithms are much slower to implement than the proposed algorithms.

5.1. Linear regression models

In the first example, we use a linear regression model with multiple explanatory variables to illustrate the effectiveness of the proposed approaches. The data contains the sale price of 546 houses sold in Windsor, Canada in 1987. For more details about the data, one can refer to Koop (2003). We are interested in factors that can influence house prices. There are four explanatory variables, including the size, the number of bedrooms, the number of bathrooms, and the number of stories. The following two linear models are considered:

$$\boldsymbol{M_1}: \ y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \ldots, n.$$
$$\boldsymbol{M_2}: \ y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i, \quad \varepsilon_i \sim t(0, \sigma^2, \upsilon), \quad i = 1, 2, \ldots, n.$$

For $\boldsymbol{M_1}$, we use the same priors as in Koop (2003), that is, $\beta \sim N(\beta_0, h^{-1}V_0)$, $h := \frac{1}{\sigma^2} \sim \Gamma(s, r)$, where β_0, V_0 are the prior mean vector and the prior variance of β, h is inverse of the error variance with s, r being the scale parameter and the rate parameter of a Gamma distribution. Furthermore, following Koop (2003), we set $\beta_0 = \begin{bmatrix} 0, & 10, & 5000, & 10^4, & 10^4 \end{bmatrix}'$, $V_0 = diag\begin{pmatrix} 2.4, & 6 \times 10^{-7}, & 0.15, & 0.6, & 0.6 \end{pmatrix}$, $s = 2.5$, $r = 6.25 \times 10^7$. In both models, h is the precision parameter which has to be greater than zero. To ensure Condition 9, we use the reparameterization $\phi(h) = \ln(h) \in \mathbb{R}$.

Note that for $\boldsymbol{M_1}$ the marginal likelihood is available analytically and hence, can be calculated without any error. The true marginal likelihood value is around -6151. Moreover, the power posterior has a closed-form expression, which is always the normal-gamma distribution at any grid point. Therefore, it is easy to draw from the power posteriors directly.

In $\boldsymbol{M_2}$, υ is the degrees of freedom parameter with $\upsilon > 2$ in the t distribution. To ensure Condition 9, we use the reparameterization $\phi(\upsilon) = \ln(\upsilon - 2) \in \mathbb{R}$. We assign the same prior distribution for h as in $\boldsymbol{M_1}$. We choose the prior distribution for $\upsilon - 2$ to be an exponential distribution, that is, $\upsilon - 2 \sim Exp(0.05)$. The power posteriors do not have closed-form expressions for $\boldsymbol{M_2}$.

We generate the MCMC output from the original posterior for $\boldsymbol{M_2}$ using WinBUGS. We also generate the MCMC output from the power posterior for $\boldsymbol{M_2}$ by using the "zeros trick" technique and defining the power posterior distribution corresponding to each grid point as a new distribution in WinBUGS. For each chain, we draw 100,000 samples in total, with the first 40,000 samples discarded and the next 60,000 kept as effective samples. We take one sample from every three samples to reduce the dependence of the chain so that $J = J_0 = 20{,}000$. We then estimate the log marginal likelihood using the four algorithms, namely, the TI and SS algorithm and the two proposed algorithms.

For the choice of other tuning parameters, namely c and S, we follow Friel and Pettitt (2008). As $\mathcal{U}(b_s)$ involves a higher level of non-linearity as b_s is closer to zero, to calculate $\int_0^1 \mathcal{U}(b)\mathrm{d}b$, a fine grid is needed near zero. With $c > 1$, more grid points are assigned in the region near zero.

Table 1 reports the bias and the MCSEs of the estimates of the log marginal likelihood from the four algorithms when $c = 1$, 3, and $S = 20$, 40, 100 in $\boldsymbol{M_1}$. Table 2 reports the log marginal likelihood estimates from the four algorithms when $c = 3$, and $S = 20$, 40, 100 in $\boldsymbol{M_2}$. We cannot obtain the bias in $\boldsymbol{M_2}$ as the true value of the marginal likelihood is unknown in $\boldsymbol{M_2}$.

We can see from Table 1 that both TI and TI-LWY provide good approximations to the true value when S is moderate and $c = 3$. The MCSEs always take small values, reinforcing the finding in Friel and Pettitt (2008). When $c = 1$, the quality of the approximations is much worse, confirming the suggestion that a fine grid should be used in the regions near zero. This is the reason why we only choose $c = 3$ for the rest of the paper.

Both SS and SS-LWY provide good estimates to the true value for the two values of c. This is because the SS and SS-LWY algorithms do not involve discretization errors, so that they are less sensitive to c than TI and TI-LWY. However,

Table 2
Estimates of log marginal likelihood for $\boldsymbol{M_2}$ with $c = 3$.

	TI	TI-LWY	SS	SS-LWY
$S = 20$	−6514	−6518	−6513	−6517
$S = 40$	−6513	−6518	−6513	−6518
$S = 100$	−6513	−6517	−6513	−6517

Table 3
CPU time (in minutes or in hours) in linear regression models.

	$\boldsymbol{M_1}$		$\boldsymbol{M_2}$	
	TI	TI-LWY	TI	TI-LWY
$S = 20$	19.71 min	20.31 min	3.81 h	0.35 h
$S = 40$	40.25 min	39.21 min	9.12 h	0.84 h
$S = 100$	108.96 min	93.49 min	22.80 h	1.91 h
	SS	SS-LWY	SS	SS-LWY
$S = 20$	28.45 min	28.30 min	4.69 h	0.65 h
$S = 40$	56.39 min	55.76 min	9.96 h	1.03 h
$S = 100$	142.42 min	133.96 min	25.96 h	2.29 h

SS essentially does importance sampling for $p\left(\boldsymbol{\theta}_{b_s+1}|\mathbf{y}, b_{s+1}\right)$ using a slightly more diffuse distribution $p\left(\boldsymbol{\theta}_{b_s}|\mathbf{y}, b_s\right)$ as the proposal distribution. Note that the power posterior distributions vary more significantly across small values of b_s. The closer the proposal distribution to the target distribution, the better the performance of importance sampling. Therefore, for the SS algorithm, $c > 1$ is still preferred. As we can see from Table 1, the estimates by the SS and SS-LWY algorithms under $c = 3$ have a smaller bias and a smaller MCSE than those under $c = 1$.

For all values of S and c and both models, the two proposed algorithms always provide estimates as good as the corresponding TI and SS algorithms. And based on the log marginal likelihood values of $\boldsymbol{M_1}$ and $\boldsymbol{M_2}$, one can obtain the BFs. It is evident that $\boldsymbol{M_1}$ fits the data better than $\boldsymbol{M_2}$.

In Table 3 we report the CPU time for estimating the log marginal likelihood of $\boldsymbol{M_1}$ (100 times) and $\boldsymbol{M_2}$ (once). Since $\boldsymbol{M_1}$ has a closed-form expression for the power posteriors, not surprisingly, there is not much computational gain in using the two proposed algorithms relative to the TI and SS algorithms as drawing from the power posteriors is easy. However, there is a substantial gain in the two proposed algorithms in terms of the computational cost relative to the TI and SS algorithms in $\boldsymbol{M_2}$. While not reported, the two proposed algorithms save even more of the CPU time if both methods are used to compute MCSEs in $\boldsymbol{M_2}$ than the TI and SS algorithms.

5.2. Copula models

In this subsection, following Hurn et al. (2020), we consider several copula models for stock returns. Unlike Hurn et al. (2020), where the copula models are estimated using ML, we estimate competing models using MCMC. For each competing model, we use the four algorithms to estimate the log marginal likelihood and then obtain the BFs to make a pair-wise comparison of nested and nonnested models. A comparison of nonnested models in the Bayesian paradigm is easier than in the frequentist paradigm.

Let r_{1t} and r_{2t} be daily log returns at time t. Assume

$$r_{1t} = \mu_1 + \sigma_1 z_{1t},$$
$$r_{2t} = \mu_2 + \sigma_2 z_{2t},$$

where μ_i, σ_i are the mean and the standard deviation of r_{it} for $i = 1, 2$. The joint distribution of returns is modeled by a copula function, that is,

$$F(r_{1t}, r_{2t}) = C(F_1(r_{1t}), F_2(r_{2t}); \delta),$$

where $F_i(\cdot)$ is the marginal distribution for r_{it} and $C(\cdot; \delta)$ is the copula function with parameter δ. Different assumptions about the marginal distribution of z_{it} and the copula function are made below, leading to different models. All competing models are fit to daily log returns on the S&P 100 and S&P 600 Indices for the period 17 August 1995 to 28 December 2018.[8]

We use the "mcmc" package in R to obtain the MCMC output. It requires users to provide the kernel of the likelihood function and the prior. Recall that

$$p(\boldsymbol{\theta}_b|\mathbf{y}, b) \propto p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b p(\boldsymbol{\theta}_b).$$

[8] We have extended the sample period of the same returns used in Hurn et al. (2020) to that from 17 August 1995 to 28 December 2018.

Table 4
Posterior means and posterior standard errors of parameters for the Gaussian copula normal marginals model.

Parameters	μ_1	h_1	μ_2	h_2	δ
posterior mean	0.0265	0.7058	0.0367	0.5478	0.8422
posterior sd	0.0163	0.0132	0.0186	0.0102	0.0038

Table 5
Estimates of log-marginal likelihood for the Gaussian copula normal marginals model with $c = 3$.

	TI	TI-LWY	SS	SS-LWY
$S = 20$	−15726	−15729	−15722	−15720
$S = 40$	−15720	−15721	−15720	−15719
$S = 100$	−15717	−15718	−15718	−15719

Table 6
CPU time for the four algorithms for the Gaussian copula normal marginals model.

	TI	TI-LWY	SS	SS-LWY
$S = 20$	3.80 min	0.54 min	4.85 min	0.71 min
$S = 40$	8.95 min	0.89 min	8.93 min	1.19 min
$S = 100$	19.11 min	2.15 min	21.81 min	2.88 min

The kernel of the target functions for the model corresponding to the power posterior, in the log form, is

$$b \ln p\left(\mathbf{y}|\boldsymbol{\theta}_b\right) + \ln p\left(\boldsymbol{\theta}_b\right).$$

For each b, we iterate 100,000 times in total, and the first half of the chain is discarded as burn-in. For the remaining 50,000 samples, we keep one out of every five samples to reduce the dependence of the chain so that $J = J_0 = 10{,}000$.

5.2.1. Gaussian copula normal marginals

In this model we assume $z_{1t}, z_{2t} \sim N(0, 1)$ and $C(\cdot; \delta)$ to be the Gaussian copula function. This is equivalent to assuming $(r_{1t}, r_{2t})'$ follows a bivariate normal distribution with the correlation coefficient $\delta \in [-1, 1]$. The log-likelihood function at time t is:

$$\ln L_t = -\ln 2\pi - \frac{1}{2}\ln\left(\frac{1-\delta^2}{h_1 h_2}\right) - \frac{z_{1t}^2 + z_{2t}^2 - 2\delta z_{1t} z_{2t}}{2(1-\delta^2)},$$

where $h_i = 1/\sigma_i^2$ is the precision parameter, and $z_{it} = (r_{it} - \mu_i)h_i^{1/2}$ for $i = 1, 2$. The parameters of interest are $\boldsymbol{\theta} = \begin{pmatrix} \mu_1 & h_1 & \mu_2 & h_2 & \delta \end{pmatrix}'$.

To do Bayesian analysis, we assign the following prior distributions on parameters,

$$\mu_i \sim N(0, 25),\ h_i \sim \Gamma(0.1, 1),\ i = 1, 2, \text{ and } \delta \sim U[-1, 1].$$

To validate Condition 9, we use the reparameterization $\phi(h_i) = \ln(h_i) \in \mathbb{R}$ and $\phi(\delta) = \tan\left(\frac{\pi}{2}\delta\right) \in \mathbb{R}$ when implementing the two proposed algorithms.

The posterior means and posterior standard errors of these parameters are reported in Table 4. These estimates are reasonable. For example, the posterior mean of δ is 0.8422, suggesting that there is a strong linear relationship between the two daily returns. However, the Gaussian copula implies that there is no tail dependence between the two daily returns. The estimates of the marginal likelihood by the four algorithms are reported in Table 5 while the CPU time is reported in Table 6. It is clear that all four methods provide reliable estimates. However, the proposed algorithms are much cheaper to implement computationally, using about only 10% of the CPU time.

5.2.2. Gaussian copula t marginals

In this model we assume $z_{1t}, z_{2t} \sim t(0, 1, v)$ and $C(\cdot; \delta)$ to be the Gaussian copula function. The log-likelihood function at time t is:

$$\ln L_t = -\frac{1}{2}\ln(1-\delta^2) - \frac{q_{1t}^2 + q_{2t}^2 - 2\delta q_{1t} q_{2t}}{2(1-\delta^2)} + \frac{1}{2}(q_{1t}^2 + q_{2t}^2) + \ln\left(h_1^{1/2} f(z_{1t}; v) h_2^{1/2} f(z_{2t}; v)\right),$$

where $\delta \in [-1, 1]$, $q_{it} = \Phi^{-1}(F(z_{it}; v))$, $z_{it} = (r_{it} - \mu_i)h_i^{1/2}$, $F(z_{it}; v)$, $f(z_{it})$ are the CDF and PDF of the t distribution with v degrees of freedom ($v > 2$), and $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution. The Gaussian copula t marginals model nests the Gaussian copula normal marginals model. If $v \to \infty$, the two models are the same. To validate Condition 9, we use the reparameterization $\phi(h_i) = \ln(h_i) \in \mathbb{R}$, $\phi(v) = \ln(v-2) \in \mathbb{R}$, $\phi(\delta) = \tan\left(\frac{\pi}{2}\delta\right) \in \mathbb{R}$. Since v should be larger than 2 and $\delta \in [-1, 1]$, we use an exponential prior distribution for $v - 2$ and a uniform prior for δ, that is, $v - 2 \sim Exp(1)$, $\delta \sim U[-1, 1]$.

Table 7
Posterior means and posterior standard errors of parameters for the Gaussian copula t marginals model.

Parameters	μ_1	h_1	μ_2	h_2	δ	v
Posterior mean	0.0490	1.5492	0.0618	1.0860	0.8236	3.8831
Posterior sd	0.0116	0.0401	0.0139	0.0271	0.0042	0.0957

Table 8
Estimates of log marginal likelihood for the Gaussian copula t marginals model with $c = 3$.

	TI	TI-LWY	SS	SS-LWY
$S = 20$	−14879	−14887	−14876	−14879
$S = 40$	−14879	−14881	−14880	−14879
$S = 100$	−14880	−14879	−14880	−14877

Table 9
CPU time for the four algorithms for the Gaussian copula t marginals model.

	TI	TI-LWY	SS	SS-LWY
$S = 20$	6.73 h	0.91 h	6.72 h	1.09 h
$S = 40$	12.88 h	1.54 h	13.52 h	1.77 h
$S = 100$	32.75 h	3.33 h	34.38 h	3.96 h

Table 10
Posterior means and posterior standard errors of parameters for the t copula t marginals model.

Parameters	μ_1	h_1	μ_2	h_2	δ	v	η
Posterior mean	0.0558	1.7318	0.0704	1.2037	0.8168	3.3382	3.6102
Posterior sd	0.0115	0.0557	0.0139	0.0364	0.0051	0.1334	0.1413

The likelihood function of this model is more complicated than the Gaussian copula normal marginals model. It requires a longer CPU time to do posterior sampling. For example, to sample from the posterior distribution, for the Gaussian copula normal marginals model it only takes 10 s, whereas for the Gaussian copula t marginals model it takes about 17 min. Consequently, the TI and SS algorithms require more CPU time to estimate the marginal likelihood of the Gaussian copula t marginals model.

The posterior means and posterior standard errors of the parameters are reported in Table 7. Again, these estimates are reasonable. For example, the posterior mean of v is 3.8831, suggesting evidence of very heavy tails in the daily returns. The estimates of the log marginal likelihood by the four algorithms are reported in Table 8 while the CPU time is reported in Table 9. With moderate S, all methods provide reliable estimates. Comparing the marginal likelihood values in Tables 5 and 8, it is clear that the Gaussian copula t marginals model fits the data much better than the Gaussian copula normal marginals model. This conclusion is very reasonable given the heavy tails in the daily returns. Moreover, the proposed algorithms are much cheaper to implement computationally than the TI and SS algorithms, using about only 10% of the CPU time.

5.2.3. t copula t marginals

In this model we assume $z_{1t}, z_{2t} \sim t(0, 1, v)$ and $C(\cdot; \delta, \eta)$ to be the t copula function where δ is the correlation coefficient and η captures the tail dependence. Unlike the Gaussian copula, the t copula allows for tail dependence in both tails. The log-likelihood function at time t is:

$$\ln L_t = -\ln(2\pi) - \frac{1}{2}\ln(1-\delta^2) - \frac{\eta+2}{2}\ln\left(1 + \frac{q_{1t}^2 + q_{2t}^2 - 2\delta q_{1t}q_{2t}}{\eta(1-\delta^2)}\right)$$
$$- \ln f(q_{1t}; \eta) - \ln f(q_{2t}; \eta) + \ln\left(f(z_{1t}; v)h_1^{1/2}\right) + \ln\left(f(z_{2t}; v)h_2^{1/2}\right),$$

where $\delta \in [-1, 1]$, $q_{it} = F^{-1}(F(z_{it}; v); \eta)$, $z_{it} = (r_{it} - \mu_i)h_i^{1/2}, i = 1, 2$. The t copula t marginals model nests the Gaussian copula t marginals model. If $\eta \to +\infty$, the two models are the same. To validate Condition 9, we use the reparameterization $\phi(h_i) = \ln(h_i) \in \mathbb{R}$, $\phi(v) = \ln(v-2) \in \mathbb{R}$, $\phi(\eta) = \ln(\eta - 2) \in \mathbb{R}$, $\phi(\delta) = \tan\left(\frac{\pi}{2}\delta\right) \in \mathbb{R}$. For the prior distributions, we assume $v - 2, \eta - 2 \sim Exp(1)$, $\delta \sim U[-1, 1]$.

The MCMC sampling from the posterior distribution is even more complicated for the t-copula t marginals model. It requires more CPU time (about 1 h) to draw from the original posterior and the power posteriors for once. The posterior means and posterior standard errors of the parameters are reported in Table 10. Again, these estimates are reasonable. For example, the posterior mean of η is 3.6102, suggesting evidence of strong tail dependence between the daily returns. The estimates of the log marginal likelihood by the four algorithms are reported in Table 11 while the CPU time is reported in Table 12. All methods provide reliable estimates of the log marginal likelihood. Comparing the log marginal likelihood values in Tables 8 and 11, it is clear that the t copula t marginals model fits the data much better than the Gaussian

Table 11
Estimates of log marginal likelihood for the t copula t marginals model with $c = 3$.

	TI	TI-LWY	SS	SS-LWY
$S = 20$	−14694	−14689	−14689	−14687
$S = 40$	−14691	−14683	−14690	−14689
$S = 100$	−14691	−14681	−14692	−14687

Table 12
CPU time for the four algorithms for the t copula t marginals model.

	TI	TI-LWY	SS	SS-LWY
$S = 20$	24.45 h	3.50 h	25.41 h	4.09 h
$S = 40$	49.72 h	5.78 h	52.49 h	6.28 h
$S = 100$	120.85 h	12.20 h	127.58 h	13.58 h

Table 13
Posterior means and posterior standard errors of parameters for the Clayton copula t marginals model.

Parameters	μ_1	h_1	μ_2	h_2	δ	v
Posterior mean	0.1638	1.9200	0.1920	1.3491	2.2459	2.5487
Posterior sd	0.0110	0.0590	0.0134	0.0386	0.0447	0.0682

Table 14
Estimates of log marginal likelihood for the Clayton copula t marginals model with $c = 3$.

	TI	TI-LWY	SS	SS-LWY
$S = 20$	−15279	−15284	−15276	−15277
$S = 40$	−15280	−15278	−15281	−15276
$S = 100$	−15280	−15276	−15281	−15275

copula t marginals model. This conclusion is very reasonable because there is not only a strong linear relationship but also a strong tail dependence between the two daily returns. Moreover, the two proposed algorithms are much cheaper to implement computationally than the TI and SS algorithms, using only about 10% of the CPU time. Even with $S = 20$, the computational burden is a major challenge for the TI and SS algorithms, requiring 24–25 h of CPU time once. Giving the computational cost, it is impossible to obtain the MCSE using the TI and SS algorithms.

From Table 11, it can be seen that there is a noticeable difference between the log-marginal likelihood values obtained by TI and TI-LWY. With the concern that the difference may be due to a reasonably small value of J being used, we increase MCMC iterations to 200,000, and keep the last half of draws (i.e., $J = J_0 = 100,000$). The log-marginal likelihood estimate obtained by the proposed TI-LWY algorithm is -14695, -14688, and -14686 for $S = 20, 40, 100$ respectively. However, with the increased J, we cannot obtain the estimates of the log-marginal likelihood by the TI algorithm as it is too time-consuming.

5.2.4. *Clayton copula t marginals*

In this model we assume $z_{1t}, z_{2t} \sim t(0, 1, v)$ and $C(\cdot; \delta)$ to be the Clayton copula function. The Clayton copula function is given by:

$$C(u_1, u_2; \delta) = \left(u_1^{-\delta} + u_2^{-\delta} - 1\right)^{-1/\delta}, \quad 0 < \delta < \infty,$$

where $\delta > 0$ captures the degree of left tail dependence of the two marginals. This model does not nest or is not nested by any model introduced earlier as the Clayton copula only allows for dependence at left tails. The log-likelihood function at time t is:

$$\begin{aligned}\ln L_t = & \ln(1+\delta) - (1+\delta)(\ln u_{1t} + \ln u_{2t}) - (2 + 1/\delta)\ln\left(u_{1t}^{-\delta} + u_{2t}^{-\delta} - 1\right) \\ & + \ln(f(z_{1t}; v)h_1^{1/2}) + \ln(f(z_{2t}; v)h_2^{1/2}),\end{aligned}$$

where $z_{it} = (r_{it} - \mu_i)h_i^{1/2}$ and $u_{it} = F(z_{it}; v)$ for $i = 1, 2$. To validate Condition 9, we use the reparameterization $\phi(h_i) = \ln(h_i) \in \mathbb{R}$, $\phi(v) = \ln(v - 2) \in \mathbb{R}$, $\phi(\delta) = \ln\delta \in \mathbb{R}$. As for the prior of δ, we assume $\delta \sim \Gamma(1, 1)$.

The posterior means and posterior standard errors of these parameters are reported in Table 13. Again, these estimates are reasonable. For example, the posterior mean of δ is 2.246, suggesting evidence of strong dependence in the left tails. The estimates of the marginal likelihood by the four algorithms are reported in Table 14 while the CPU time is reported in Table 15. All methods provide reliable estimates. According to the log marginal likelihood value, the Clayton copula does not fit the data so well. This is because while the Clayton copula allows for the lower tail dependence, it does not allow for any upper tail dependence. The upper tail dependence is important in our data.

Table 15
CPU time for the four algorithms for the Clayton copula t marginals model.

	TI	TI-LWY	SS	SS-LWY
$S = 20$	7.16 h	1.05 h	7.41 h	1.20 h
$S = 40$	13.86 h	1.70 h	14.89 h	1.89 h
$S = 100$	36.42 h	3.64 h	39.13 h	4.20 h

Comparing the marginal likelihood values of all four models reported in Tables 5, 8, 11 and 14 (some are nested and some are not), it is clear that the t copula t marginals model fits the data the best and by wide margins, followed by the Gaussian copula t marginals model, then by the Clayton copula t marginals model, and finally by the Gaussian copula normal marginals model. Again, the proposed algorithms are much cheaper to implement computationally than the TI and SS algorithms, using only about 10% of the CPU time.

6. Concluding remarks

In this paper, under some regularity conditions, we establish the BvM theorem for the power posteriors. Due to the BvM theorem, the power posteriors, when adjusted by the square root of the grid points, converge to the normal distribution, which is also the limit of the original posterior distribution. This large sample theory, therefore, allows us to improve the power-posterior-based methods to estimate marginal likelihood by providing a proposal distribution for importance sampling. Particularly, we apply this idea to the TI approach of Friel and Pettitt (2008) and the SS approach of Xie et al. (2011). Unlike the standard power posteriors methods that require repeated posterior sampling from the power posteriors, the new methods only require a posterior output of the original posterior. Hence, they are computationally more efficient. Moreover, for models where extra coding effort is needed to draw random samples from power posteriors, such coding effort is completely avoided.

The accuracy of the proposed methods are examined and compared with the standard power-posterior-based methods in the Gaussian linear regression model where the true value of the marginal likelihood can be obtained. It suggests that the proposed methods provide reliable estimates of the marginal likelihood. It performs as well as the standard power-posterior-based methods in terms of both bias and MCSE. A comparison of the computational efficiency of the proposed methods relative to that of the standard power-posterior-based methods is made using a linear regression model and several copula models. The comparison suggests that when a model is reasonably complicated, standard power-posterior-based methods are very time-consuming, and our methods can reduce about 90% of the CPU time of those methods.

The marginal likelihood is only well-defined under proper priors. Therefore, it is important to note that, as our methods aim to estimate the marginal likelihood, they cannot be used in connection to improper priors. Moreover, we want to point out that although this paper tries to improve the existing algorithms for marginal likelihood estimation, our idea can be used in other cases where power posteriors are used. For example, they can be extended to the sequential Monte Carlo method of Herbst and Schorfheide (2015) and the striated Metropolis–Hastings algorithm of Waggoner et al. (2016). These extensions are beyond the scope of this paper.

Although we have shown that the TI and modified TI algorithms provide consistent estimators of the log marginal likelihood, due to space constraint, we have not shown that the SS and improved SS algorithms yield consistent estimators of the marginal likelihood. Such a result will be reported in a separate study.

Appendix A

A.1. Proof of Theorem 4.1

The power posterior density of $\mathbf{z}_{nb}$, $p\left(\mathbf{z}_{nb}|\mathbf{y}, b\right)$, can be decomposed into two parts:

$$p\left(\mathbf{z}_{nb}|\mathbf{y}, b\right) = \frac{\left|b^{-1}\boldsymbol{\Sigma}_n\right|^{1/2} p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b p\left(\boldsymbol{\theta}_b\right)}{m\left(\mathbf{y}|b\right)}$$
$$= \left[\frac{\left|b^{-1}\boldsymbol{\Sigma}_n\right|^{1/2} p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b p\left(\boldsymbol{\theta}^0\right)}{m\left(\mathbf{y}|b\right)}\right]\left[\frac{p\left(\boldsymbol{\theta}_b\right)}{p\left(\boldsymbol{\theta}^0\right)}\frac{p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b}{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b}\right]. \quad (34)$$

To establish the BvM theorem for the power posterior, we need to show the convergence in total variation, that is, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P_0\left(\sup_{B\sqsubseteq A_{nb}}\left|\int_B\left[p\left(\mathbf{z}_{nb}|\mathbf{y}, b\right) - (2\pi)^{-q/2}\exp\left(-\frac{\mathbf{z}_{nb}'\mathbf{z}_{nb}}{2}\right)\right]d\mathbf{z}_{nb}\right| > \varepsilon\right) = 0, \quad (35)$$

where $B \sqsubseteq A_{nb}$ is any Borel set.

Note that, for any $B \sqsubseteq A_{nb}$,

$$\left|\int_B \left[p\left(\mathbf{z}_{nb}|\mathbf{y}, b\right) - (2\pi)^{-q/2}\exp\left(-\frac{\mathbf{z}_{nb}'\mathbf{z}_{nb}}{2}\right)\right] d\mathbf{z}_{nb}\right|$$
$$\leq \int_B \left|p\left(\mathbf{z}_{nb}|\mathbf{y}, b\right) - (2\pi)^{-q/2}\exp\left(-\frac{\mathbf{z}_{nb}'\mathbf{z}_{nb}}{2}\right)\right| d\mathbf{z}_{nb}$$
$$\leq \int_{A_{nb}} \left|p\left(\mathbf{z}_{nb}|\mathbf{y}, b\right) - (2\pi)^{-q/2}\exp\left(-\frac{\mathbf{z}_{nb}'\mathbf{z}_{nb}}{2}\right)\right| d\mathbf{z}_{nb}.$$

Hence, Eq. (35) holds if

$$\lim_{n\to\infty} P_0\left(\int_{A_{nb}} \left|p\left(\mathbf{z}_{nb}|\mathbf{y}, b\right) - (2\pi)^{-q/2}\exp\left(-\frac{\mathbf{z}_{nb}'\mathbf{z}_{nb}}{2}\right)\right| d\mathbf{z}_{nb} > \varepsilon\right) = 0. \tag{36}$$

Note that for the integrand in (36), we have

$$\begin{aligned}
&\left|p\left(\mathbf{z}_{nb}|\mathbf{y}, b\right) - (2\pi)^{-q/2}\exp\left(-\frac{\mathbf{z}_{nb}'\mathbf{z}_{nb}}{2}\right)\right| \\
&= \left|\frac{\left|b^{-1}\boldsymbol{\Sigma}_n\right|^{1/2} p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b p\left(\boldsymbol{\theta}^0\right)}{m\left(\mathbf{y}|b\right)} \times \frac{p\left(\boldsymbol{\theta}_b\right)}{p\left(\boldsymbol{\theta}^0\right)} \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b}{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b} - (2\pi)^{-q/2}\exp\left(-\frac{\mathbf{z}_{nb}'\mathbf{z}_{nb}}{2}\right)\right| \\
&\leq \left|\frac{\left|b^{-1}\boldsymbol{\Sigma}_n\right|^{1/2} p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b p\left(\boldsymbol{\theta}^0\right)}{m\left(\mathbf{y}|b\right)} \times \frac{p\left(\boldsymbol{\theta}_b\right)}{p\left(\boldsymbol{\theta}^0\right)} \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b}{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b} - (2\pi)^{-q/2} \frac{p\left(\boldsymbol{\theta}_b\right)}{p\left(\boldsymbol{\theta}^0\right)} \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b}{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b}\right| \\
&\quad + \left|(2\pi)^{-q/2} \frac{p\left(\boldsymbol{\theta}_b\right)}{p\left(\boldsymbol{\theta}^0\right)} \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b}{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b} - (2\pi)^{-q/2}\exp\left(-\frac{\mathbf{z}_{nb}'\mathbf{z}_{nb}}{2}\right)\right| \\
&= \left|\frac{\left|b^{-1}\boldsymbol{\Sigma}_n\right|^{1/2} p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b p\left(\boldsymbol{\theta}^0\right)}{m\left(\mathbf{y}|b\right)} - (2\pi)^{-q/2}\right| \times \frac{p\left(\boldsymbol{\theta}_b\right)}{p\left(\boldsymbol{\theta}^0\right)} \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b}{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b} \\
&\quad + (2\pi)^{-q/2}\left|\frac{p\left(\boldsymbol{\theta}_b\right)}{p\left(\boldsymbol{\theta}^0\right)} \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b}{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b} - \exp\left(-\frac{\mathbf{z}_{nb}'\mathbf{z}_{nb}}{2}\right)\right|.
\end{aligned} \tag{37}$$

From (37), we can see that to prove (36), it is sufficient to prove that, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P_0\left(\left|\frac{\left|b^{-1}\boldsymbol{\Sigma}_n\right|^{1/2} p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b p\left(\boldsymbol{\theta}^0\right)}{m\left(\mathbf{y}|b\right)} - (2\pi)^{-q/2}\right| > \varepsilon\right) = 0, \tag{38}$$

and

$$\lim_{n\to\infty} P_0\left(\int_{A_{nb}} \left|\frac{p\left(\boldsymbol{\theta}_b\right)}{p\left(\boldsymbol{\theta}^0\right)} \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_b\right)^b}{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b} - \exp\left(-\frac{\mathbf{z}_{nb}'\mathbf{z}_{nb}}{2}\right)\right| d\mathbf{z}_{nb} > \varepsilon\right) = 0. \tag{39}$$

Since the detailed proof of (38) and (39) is too long, we leave it to the online appendix.

A.2. Two lemmas

Before we prove Theorem 4.2, we need to establish two lemmas. All the results in the two lemmas are conditional on the data $\mathbf{y}$ with a fixed and finite n, and hence, non-probabilistic in nature. The proof of the two lemmas is given in the online appendix.

Lemma A.1. *Under Conditions 1–8, when the data $\mathbf{y}$ is given with a fixed and finite n, for any $b \in [0, 1]$,*

$$|\mathcal{U}(b)| < +\infty,$$

$$\int_\Theta \left[\ln p(\mathbf{y}|\boldsymbol{\theta})\right]^4 p(\boldsymbol{\theta}) d\boldsymbol{\theta} < +\infty,$$

$$1 \leq \frac{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right)^b}{m(\mathbf{y}|b)} \leq \exp\left[b\,\mathcal{U}^*(0)\right] < +\infty,$$

where $\mathcal{U}^*(b) = \ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right) - \mathcal{U}(b)$.

Remark A.1. The first part of Lemma A.1 suggests that $\mathcal{U}(b)$ is bounded for any $b \in [0, 1]$, and hence, $\ln m(\mathbf{y})$ is also bounded. The second part of Lemma A.1 ensures that we can apply a central limit theorem (CLT) to random samples so that the order of the error in using $\widehat{\mathcal{U}}(b)$ to approximate $\mathcal{U}(b)$ can be obtained.

Lemma A.2. *Under Conditions 1–8, when the data* $\mathbf{y}$ *is given with a fixed and finite* n*, as* $J \to +\infty$*, for any* $b \in [0, 1]$*, we have*

$$\sup_{b\in[0,1]} Var_{\boldsymbol{\theta}_b|\mathbf{y},b}\left[\widehat{\mathcal{U}}(b)\right] = O(J^{-1}),$$

where $\widehat{\mathcal{U}}(b)$ *is defined in (7). The* J *random samples are either iid draws from* $p(\boldsymbol{\theta}_b|\mathbf{y}, b)$ *or stationary and ergodic draws from a Markov chain that has* $p(\boldsymbol{\theta}_b|\mathbf{y}, b)$ *as its stationary probability density.*

A.3. Proof of Theorem 4.2

In this proof, we will understand p in o_p or O_p as the probability measure corresponding to the joint power posterior $p\left(\boldsymbol{\theta}_{b_s}|\mathbf{y}, b_s\right)$ with $\left\{b_s = (s/S)^c\right\}_{s=0}^{S}$. Note that

$$\begin{aligned}
&\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}(b_{s+1})+\widehat{\mathcal{U}}(b_s)}{2} - \ln m(\mathbf{y})\\
&= \sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}(b_{s+1})-\mathcal{U}(b_{s+1})+\widehat{\mathcal{U}}(b_s)-\mathcal{U}(b_s)}{2}\\
&\quad + \sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\mathcal{U}(b_s)+\mathcal{U}(b_{s+1})}{2} - \ln m(\mathbf{y}).
\end{aligned} \tag{40}$$

Hence, it will be sufficient if we can show that, as $J \to +\infty$,

$$\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}(b_{s+1})-\mathcal{U}(b_{s+1})+\widehat{\mathcal{U}}(b_s)-\mathcal{U}(b_s)}{2} = o_p(1), \tag{41}$$

and that, as $S \to +\infty$,

$$\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\mathcal{U}(b_s)+\mathcal{U}(b_{s+1})}{2} - \ln m(\mathbf{y}) = o(1). \tag{42}$$

First, to prove (41), let $\widehat{M}_1 = \sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}(b_{s+1})-\mathcal{U}(b_{s+1})}{2}$, $\widehat{M}_2 = \sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}(b_s)-\mathcal{U}(b_s)}{2}$, and $\widehat{M} = \sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}(b_{s+1})-\mathcal{U}(b_{s+1})+\widehat{\mathcal{U}}(b_s)-\mathcal{U}(b_s)}{2} = \widehat{M}_1+\widehat{M}_2$. Note that, since $\widehat{\mathcal{U}}(b)$ is an unbiased estimator of $\mathcal{U}(b)$ for any b, we have $E\left(\widehat{M}\right) = 0$. Consequently,

$$Var\left(\widehat{M}\right) = E\left(\widehat{M}_1+\widehat{M}_2\right)^2 \le 2E\left(\widehat{M}_1^2\right) + 2E\left(\widehat{M}_2^2\right).$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\widehat{M}_1^2 &= \left[\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right]^2\\
&= \left[\sum_{s=0}^{S-1}\left(\sqrt{b_{s+1}-b_s}\,\frac{\widehat{\mathcal{U}}(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right)\left(\sqrt{b_{s+1}-b_s}\right)\right]^2\\
&\le \left[\sum_{s=0}^{S-1}\left(\sqrt{b_{s+1}-b_s}\,\frac{\widehat{\mathcal{U}}(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right)^2\right]\sum_{s=0}^{S-1}(b_{s+1}-b_s)\\
&= \sum_{s=0}^{S-1}(b_{s+1}-b_s)\left(\frac{\widehat{\mathcal{U}}(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right)^2.
\end{aligned}$$

Hence,

$$E\left(\widehat{M}_1^2\right) \le E\left[\sum_{s=0}^{S-1}(b_{s+1}-b_s)\left(\frac{\widehat{\mathcal{U}}(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right)^2\right]$$

$$
\begin{aligned}
&= \sum_{s=0}^{S-1} (b_{s+1} - b_s)\, E\left(\frac{\widehat{\mathcal{U}}(b_{s+1}) - \mathcal{U}(b_{s+1})}{2}\right)^2 \\
&\leq \sup_{b_{s+1}\in[0,1]} E\left(\frac{\widehat{\mathcal{U}}(b_{s+1}) - \mathcal{U}(b_{s+1})}{2}\right)^2 \sum_{s=0}^{S-1} (b_{s+1} - b_s) \\
&= \sup_{b_{s+1}\in[0,1]} E\left(\frac{\widehat{\mathcal{U}}(b_{s+1}) - \mathcal{U}(b_{s+1})}{2}\right)^2 = O(J^{-1}),
\end{aligned}
$$

where the last step is based on Lemma A.2. Similarly, we can show that $E\left(\widehat{M}_2^2\right) = O(J^{-1})$. Combining them, we have, $Var\left(\widehat{M}\right) = E\left(\widehat{M}^2\right) = O(J^{-1})$. Since $E\left(\widehat{M}\right) = 0$, we have $\widehat{M} \xrightarrow{L^2} 0$ as $J \to +\infty$, which implies $\widehat{M} = o_p(1)$. This completes the proof of (41).

Second, to prove Eq. (42), apply the trapezoidal rule to the integral of $\mathcal{U}(b)$ over the interval $[b_s, b_{s+1}]$, we have

$$
\int_{b_s}^{b_{s+1}} \mathcal{U}(b)db \approx \frac{b_{s+1} - b_s}{2}(\mathcal{U}(b_s) + \mathcal{U}(b_{s+1})).
$$

The local approximation error is,

$$
\frac{b_{s+1} - b_s}{2}(\mathcal{U}(b_s) + \mathcal{U}(b_{s+1})) - \int_{b_s}^{b_{s+1}} \mathcal{U}(b)db = \int_{b_s}^{b_{s+1}} \left(b - \frac{b_{s+1} + b_s}{2}\right)\mathcal{U}'(b)db,
$$

by integration by parts. Hence, the global approximation error is

$$
\begin{aligned}
&\left|\sum_{s=0}^{S-1} \int_{b_s}^{b_{s+1}} \left(b - \frac{b_{s+1} + b_s}{2}\right)\mathcal{U}'(b)db\right| \\
&\leq \sum_{s=0}^{S-1} \int_{b_s}^{b_{s+1}} \left|\left(b - \frac{b_{s+1} + b_s}{2}\right)\mathcal{U}'(b)\right| db \\
&\leq \sum_{s=0}^{S-1} \int_{b_s}^{b_{s+1}} \frac{b_{s+1} - b_s}{2}\mathcal{U}'(b)db \\
&= \sum_{s=0}^{S-1} \frac{(b_{s+1} - b_s)}{2}\, (\mathcal{U}(b_{s+1}) - \mathcal{U}(b_s)) \\
&\leq \frac{\max_s\{b_{s+1} - b_s\}}{2} \sum_{s=0}^{S-1} [\mathcal{U}(b_{s+1}) - \mathcal{U}(b_s)] \\
&= O(S^{-1})\,[\mathcal{U}(1) - \mathcal{U}(0)] = O(S^{-1}) \to 0, \text{ as } S \to +\infty,
\end{aligned}
$$

where the second last step is due to $\max_s\{b_{s+1} - b_s\} = O(S^{-1})$ and $\mathcal{U}(1) - \mathcal{U}(0)$ being finite by Lemma A.1. Hence, (42) also holds. Based on Eqs. (41) and (42), Theorem 4.2 is proved.

A.4. Two more lemmas

Before we prove Theorem 4.3, we need to establish two more lemmas. The proof of these two lemmas is given in the online appendix.

Lemma A.3. *Under Conditions 1–9, for any $b \in (\frac{1}{n}, 1]$, there exists a positive integer number n^*, such that when $n \geq n^*$,*

$$
\int_\Theta \left[\frac{p(\boldsymbol{\theta}_b|\mathbf{y}, b)}{p_A(\boldsymbol{\theta}_b|\mathbf{y}, b)}\right]^2 p_A(\boldsymbol{\theta}_b|\mathbf{y}, b)\, d\boldsymbol{\theta}_b < +\infty, \quad \int_\Theta \left[\frac{p(\boldsymbol{\theta}_b|\mathbf{y}, b)}{p_A(\boldsymbol{\theta}_b|\mathbf{y}, b)}\right]^3 p_A(\boldsymbol{\theta}_b|\mathbf{y}, b)\, d\boldsymbol{\theta}_b < +\infty, \tag{43}
$$

where $p(\boldsymbol{\theta}_b|\mathbf{y}, b)$, $p_A(\boldsymbol{\theta}_b|\mathbf{y}, b)$ are defined in (2), (19) respectively.

Lemma A.4. *Suppose the data $\mathbf{y}$ is given with a fixed and finite n and Conditions 1–9 hold. As $J_0 \to +\infty$, for any $b \in [0, 1/n]$, we have*

$$
\sup_{b\in[0,\frac{1}{n}]} Var_0\left[\widehat{\mathcal{U}}_w(b)\right] = O(J_0^{-1}),
$$

where $\widehat{\mathcal{U}}_w(b)$ is defined in (25). The J_0 random samples are iid draws from $p(\boldsymbol{\theta})$. Moreover, as $J \to +\infty$, for any $b \in (1/n, 1]$, we have

$$\sup_{b\in(\frac{1}{n},1]} Var_A\left[\widehat{\mathcal{U}}_w(b)\right] = O(J^{-1}),$$

where $\widehat{\mathcal{U}}_w(b)$ is defined in (22) and Var_A is the variance with respect to $p_A(\boldsymbol{\theta}_b|\mathbf{y}, b)$ defined in (19). The J random samples are obtained by first making iid draws from posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ or stationary and ergodic draws from a Markov chain that has $p(\boldsymbol{\theta}|\mathbf{y})$ as its stationary probability density, and then transforming the draws via (18).

A.5. Proof of Theorem 4.3

In this proof, we will understand p in o_p or O_p as the probability measure corresponding to posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ when $b \in (\frac{1}{n}, 1]$ and prior distribution $p(\boldsymbol{\theta})$ when $b \in [0, \frac{1}{n}]$. Note that

$$\begin{aligned}
&\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})+\widehat{\mathcal{U}}_{LWY}(b_s)}{2} - \ln m(\mathbf{y})\\
&= \sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})-\mathcal{U}(b_{s+1})+\widehat{\mathcal{U}}_{LWY}(b_s)-\mathcal{U}(b_s)}{2}\\
&\quad + \sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\mathcal{U}(b_s)+\mathcal{U}(b_{s+1})}{2} - \ln m(\mathbf{y}).
\end{aligned} \tag{44}$$

To prove Theorem 4.3, we only need to prove that, as $J, J_0 \to +\infty$,

$$\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})-\mathcal{U}(b_{s+1})+\widehat{\mathcal{U}}_{LWY}(b_s)-\mathcal{U}(b_s)}{2} = o_p(1), \tag{45}$$

and that, as $S \to +\infty$,

$$\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\mathcal{U}(b_s)+\mathcal{U}(b_{s+1})}{2} - \ln m(\mathbf{y}) = o(1). \tag{46}$$

Note that (46) has been proved in Theorem 4.2. Hence, we only need to prove (45).

To prove (45), note that for $b \in (0, 1]$, while $\widehat{\mathcal{U}}_w(b)$ is an unbiased estimator of $\mathcal{U}(b)$, $\widehat{\mathcal{U}}_{LWY}(b)$, as a self-normalized importance sampler, is generally a biased estimator of $\mathcal{U}(b)$. It can be seen that

$$\begin{aligned}
&\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})-\mathcal{U}(b_{s+1})+\widehat{\mathcal{U}}_{LWY}(b_s)-\mathcal{U}(b_s)}{2}\\
&= \sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})-\widehat{\mathcal{U}}_w(b_{s+1})+\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\\
&\quad + \sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_s)-\widehat{\mathcal{U}}_w(b_s)+\widehat{\mathcal{U}}_w(b_s)-\mathcal{U}(b_s)}{2}\\
&= \sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})-\widehat{\mathcal{U}}_w(b_{s+1})+\widehat{\mathcal{U}}_{LWY}(b_s)-\widehat{\mathcal{U}}_w(b_s)}{2}\\
&\quad + \sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})+\widehat{\mathcal{U}}_w(b_s)-\mathcal{U}(b_s)}{2}.
\end{aligned} \tag{47}$$

Hence, to prove (45), we need to prove

$$\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})-\widehat{\mathcal{U}}_w(b_{s+1})+\widehat{\mathcal{U}}_{LWY}(b_s)-\widehat{\mathcal{U}}_w(b_s)}{2} = o_p(1), \tag{48}$$

and

$$\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})+\widehat{\mathcal{U}}_w(b_s)-\mathcal{U}(b_s)}{2} = o_p(1). \tag{49}$$

It is easy to show that

$$\begin{aligned}&\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})-\widehat{\mathcal{U}}_w(b_{s+1})+\widehat{\mathcal{U}}_{LWY}(b_s)-\widehat{\mathcal{U}}_w(b_s)}{2}\\ &=\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_s)-\widehat{\mathcal{U}}_w(b_s)}{2}+\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})-\widehat{\mathcal{U}}_w(b_{s+1})}{2}.\end{aligned}$$

Hence, we only need to prove that

$$\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_s)-\widehat{\mathcal{U}}_w(b_s)}{2}=o_p(1),\ \sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})-\widehat{\mathcal{U}}_w(b_{s+1})}{2}=o_p(1).$$

Let $\bar{w}_{b_s}=\frac{1}{J}\sum_{j=1}^{J}w_{b_s}\left(\boldsymbol{\theta}^{tr}_{b_s,(j)}\right)$ when $b_s\in(1/n,1]$ or $\bar{w}_{b_s}=\frac{1}{J_0}\sum_{j=1}^{J}w_{0b_s}\left(\boldsymbol{\theta}_{0,(j)}\right)$ when $b_s\in[0,1/n]$. Note that $\widehat{\mathcal{U}}_{LWY}(b_{s+1})=\frac{1}{\bar{w}_{b_{s+1}}}\widehat{\mathcal{U}}_w(b_{s+1})$ and $E[\bar{w}_{b_s}]=1$ where E is the expectation with respect to $p_A(\boldsymbol{\theta}_{b_s}|\mathbf{y},b_s)$ when $b\in(\frac{1}{n},1]$ and $p(\boldsymbol{\theta})$ when $b\in[0,\frac{1}{n}]$.

Hence, we can get

$$\begin{aligned}&\left|\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})-\widehat{\mathcal{U}}_w(b_{s+1})}{2}\right|^2=\left|\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\frac{1}{\bar{w}_{b_s}}\widehat{\mathcal{U}}_w(b_{s+1})-\widehat{\mathcal{U}}_w(b_{s+1})}{2}\right|^2\\ &=\left|\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\left(\frac{1}{\bar{w}_{b_s}}-1\right)\widehat{\mathcal{U}}_w(b_{s+1})}{2}\right|^2=\left|\sum_{s=0}^{S-1}\frac{\sqrt{b_{s+1}-b_s}\left(\frac{1}{\bar{w}_{b_s}}-1\right)\sqrt{b_{s+1}-b_s}\widehat{\mathcal{U}}_w(b_{s+1})}{2}\right|^2\\ &\leq\left[\sum_{s=0}^{S-1}(b_{s+1}-b_s)\left(\frac{1}{\bar{w}_{b_s}}-1\right)^2\right]\left[\sum_{s=0}^{S-1}(b_{s+1}-b_s)\left(\frac{\widehat{\mathcal{U}}_w(b_{s+1})}{2}\right)^2\right].\end{aligned}\tag{50}$$

Since $b_s=(s/S)^c$ with $c\geq 1$, we have $b_{s+1}-b_s=\left(\frac{s+1}{S}\right)^c-\left(\frac{s}{S}\right)^c$ with $\min_s\{b_{s+1}-b_s\}=b_1-b_0=S^{-c}$ and $\max_s\{b_{s+1}-b_s\}=O(S^{-1}), s=1,2\cdots,S$. Furthermore, by the law of large numbers, $\bar{w}_{b_s}=1+o_p(1)$ and $\widehat{\mathcal{U}}_w(b_{s+1})=\mathcal{U}(b_{s+1})+o_p(1)=O_p(1)$. Hence, we get

$$\begin{aligned}&\sum_{s=0}^{S-1}(b_{s+1}-b_s)\left(\frac{1}{\bar{w}_{b_s}}-1\right)^2\leq\max_s\{b_{s+1}-b_s\}\sum_{s=0}^{S-1}\left(\frac{1}{\bar{w}_{b_s}}-1\right)^2=O(S^{-1})So_p(1)=o_p(1),\\ &\sum_{s=0}^{S-1}(b_{s+1}-b_s)\left(\frac{\widehat{\mathcal{U}}_w(b_{s+1})}{2}\right)^2\leq\max_s\{b_{s+1}-b_s\}\sum_{s=0}^{S-1}O_p(1)=O(S^{-1})SO_p(1)=O_p(1).\end{aligned}$$

Based on (50), we have

$$\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_{s+1})-\widehat{\mathcal{U}}_w(b_{s+1})}{2}=o_p(1).$$

Similarly, we can show that

$$\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_{LWY}(b_s)-\widehat{\mathcal{U}}_w(b_s)}{2}=o_p(1).$$

This completes the proof of Eq. (48).

We now prove (49). Similar to Theorem 4.2, for any $b_s\in[0,1]$, let

$$\begin{aligned}\widehat{M}_{w1}&=\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})}{2},\\ \widehat{M}_{w2}&=\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_w(b_s)-\mathcal{U}(b_s)}{2},\\ \widehat{M}_w&=\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})+\widehat{\mathcal{U}}(b_s)-\mathcal{U}(b_s)}{2}=\widehat{M}_{w1}+\widehat{M}_{w2}.\end{aligned}$$

Note that, since $E\left(\widehat{M}_w\right)=0$, we have

$$Var\left(\widehat{M}_w\right)=E\left(\widehat{M}_{w1}+\widehat{M}_{w2}\right)^2\le 2E\left(\widehat{M}_{w1}^2\right)+2E\left(\widehat{M}_{w2}^2\right). \tag{51}$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\widehat{M}_{w1}^2 &= \left[\sum_{s=0}^{S-1}(b_{s+1}-b_s)\frac{\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right]^2\\
&= \left[\sum_{s=0}^{S-1}\left(\sqrt{b_{s+1}-b_s}\frac{\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right)\left(\sqrt{b_{s+1}-b_s}\right)\right]^2\\
&\le \left[\sum_{s=0}^{S-1}\left(\sqrt{(b_{s+1}-b_s)}\frac{\widehat{\mathcal{U}}(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right)^2\right]\sum_{s=0}^{S-1}(b_{s+1}-b_s)\\
&= \sum_{s=0}^{S-1}(b_{s+1}-b_s)\left(\frac{\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right)^2.
\end{aligned}$$

Hence, we can get

$$\begin{aligned}
E\left(\widehat{M}_{w1}^2\right) &\le E\left[\sum_{s=0}^{S-1}(b_{s+1}-b_s)\left(\frac{\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right)^2\right]\\
&= \sum_{s=0}^{S_0}(b_{s+1}-b_s)E\left(\frac{\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right)^2+\sum_{s=S_0+1}^{S-1}(b_{s+1}-b_s)E\left(\frac{\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right)^2\\
&\le \sup_{b_{s+1}\in[0,\frac{1}{n}]}E\left(\frac{\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right)^2\sum_{s=0}^{S_0}(b_{s+1}-b_s)\\
&\quad+\sup_{b_{s+1}\in(\frac{1}{n},1]}E\left(\frac{\widehat{\mathcal{U}}_w(b_{s+1})-\mathcal{U}(b_{s+1})}{2}\right)^2\sum_{s=S_0+1}^{S-1}(b_{s+1}-b_s)\\
&= O(J_0^{-1})+O(J^{-1}).
\end{aligned}$$

Similarly, we can show that $E\left(\widehat{M}_{w2}^2\right)=O(J_0^{-1})+O(J^{-1})$. Based on (51), we have $Var\left(\widehat{M}_w\right)=E\left(\widehat{M}_w^2\right)=O(J_0^{-1})+O(J^{-1})$ $=o(1)$. Since $E\left(\widehat{M}_w\right)=0$, as $J_0,J\to\infty$, $\widehat{M}_w\xrightarrow{L^2}0$, which implies $\widehat{M}_w=o_p(1)$. This completes the proof of (49). Hence, Theorem 4.3 is proved.

Appendix B. Supplementary data

Supplementary material related to this article, such as the detailed proof of Theorem 4.1 and Lemmas 7.1–7.4, can be found online.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2021.11.009.

References

Annis, J., Evans, N.J., Miller, B.J., Palmeri, T.J., 2019. Thermodynamic integration and steppingstone sampling methods for estimating Bayes factors: A tutorial. J. Math. Psych. 89, 67–86.
Brodeur, A., Cook, N., Heyes, A.G., 2020. Methods matter: *p* -hacking and causal inference in economics. Amer. Econ. Rev. 110, 3634–3660.
Chen, C., 1985. On asymptotic normality of limiting density functions with Bayesian implications. J. R. Stat. Soc. Ser. B Stat. Methodol. 47, 540–546.
DiCiccio, T.J., Kass, R.E., Raftery, A., Wasserman, L., 1997. Computing Bayes factors by combining simulation and asymptotic approximations. J. Amer. Statist. Assoc. 92 (439), 903–915.
Friel, N., Pettitt, A.N., 2008. Marginal likelihood estimation via power posteriors. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (3), 589–607.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis, second ed. Chapman & Hall/CRC..
Han, C., Carlin, B.P., 2001. Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. J. Amer. Statist. Assoc. 96 (455), 1122–1132.
Harvey, C.R., 2017. Presidential address: The scientific outlook in financial economics. J. Finance 72 (4), 1399–1440.
Herbst, E.P., Schorfheide, F., 2015. Bayesian Estimation of DSGE Models. Princeton University Press..
Hoehna, S., Landis, M.J., Huelsenbeck, J.P., 2017. Parallel power posterior analyses for fast computation of marginal likelihoods in phylogenetics. Bioinformatics (forthcoming).
Hurn, S., Martin, V., Phillips, P.C.B., Yu, J., 2020. Financial Econometric Modeling. Oxford University Press..
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90, 773–795.

Kass, R.E., Tierney, L., Kadane, J.B., 1990. The validity of posterior expansions based on Laplace method. In: Geisser, S., Hodges, J.S., Press, S.J., Zellner, A. (Eds.), Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George a. Barnard, Vol. 7. Elsevier Science Publishers B.V., North-Holland, pp. 473–488.
Kleijn, B.J.K., van der Vaart, A.W., 2012. The Bernstein-von-Mises theorem under missepcification. Electron. J. Stat. 6, 354–381.
Koop, G., 2003. Bayesian Econometrics. Wiley-Interscience..
Li, Y., Yu, J., Zeng, T., 2020. Deviance information criterion for latent variable models and misspecified models. J. Econometrics 216 (2), 450–493.
Liu, X.B., Li, Y., Yu, J., Zeng, T., 2021. A posterior-based wald-type statistic for hypothesis testing. J. Econometrics http://dx.doi.org/10.1016/j.jeconom.2021.11.003, (forthcoming).
Sala-i Martin, X., Doppelhofer, G., Miller, R.I., 2004. Determinants of long-term growth: A Bayesian averaging of classical estimates approach. Amer. Econ. Rev. 94 (4), 813–835.
Moral-Benito, E., 2013. Model averaging in economics: An overview. J. Econ. Surv. 29 (1), 46–75.
Müller, U.K., 2013. Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. Econometrica 81 (5), 1805–1849.
Schervish, M.J., 2012. Theory of Statistics. Springer Science & Business Media..
Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., 2003. WinBUGS User Manual. MRC Biostatistics Unit., Cambridge,UK.
van der Vaart, A.W., 2000. Asymptotic Statistics. Cambridge University Press..
Waggoner, D.F., Wu, H., Zha, T., 2016. Striated Metropolis–Hastings sampler for high-dimensional models. J. Econometrics 192 (2), 406–420.
Xie, W., Lewis, P.O., Fan, Y., Kuo, L., Chen, M.H., 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60 (2), 150–160.
Young, K.D.S., Pettit, L.I., 1996. On priors and Bayes factors. J. Econometrics 75 (1), 113–119.