Wiley Interdisciplinary Reviews: Computational Statistics

WILEY

WIREs COMPUTATIONAL STATISTICS

ADVANCED REVIEW OPEN ACCESS

A Comparison of Monte Carlo Based Marginal Likelihood Estimators

Aolan Li[1] | Pang-Yu Liu[1] | Yu-Bo Wang[2] | Analisa Milkey[3] | Paul O. Lewis[3] | Ming-Hui Chen[1]

[1]Department of Statistics, University of Connecticut, Storrs, Connecticut, USA | [2]School of Mathematical and Statistical Sciences, Clemson University, Clemson, South Carolina, USA | [3]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut, USA

Correspondence: Yu-Bo Wang (yubow@clemson.edu)

Received: 1 October 2025 | **Revised:** 3 December 2025 | **Accepted:** 2 January 2026

Commissioning Editor: Henry Horng-Shing Lu | **Review Editor:** David W Scott

Keywords: Bayes factor | marginal likelihood | Monte Carlo methods

ABSTRACT

Marginal likelihood plays a central role in Bayesian model comparison and hypothesis testing, but its computation is often challenging in practice. This article reviews recent Monte Carlo methods that rely on the availability of Markov chain Monte Carlo (MCMC) samples from the posterior and prior distributions along with the corresponding unnormalized kernels that can be evaluated numerically. Within this scope, we summarize the strengths, limitations, differences, and connections of different methods. Two in-depth applications are presented to illustrate their relative performance.

This article is categorized under:

Statistical and Graphical Methods of Data Analysis > Monte Carlo Methods

Statistical Models > Model Selection

Statistical Models > Bayesian Models

1 | Introduction

Marginal likelihood plays a central role in modern Bayesian statistics, providing a principled foundation for model comparison, hypothesis testing, and the evaluation of competing statistical frameworks. Unlike posterior summaries that focus solely on parameter estimation within a model, the marginal likelihood integrates over the entire parameter space, capturing the overall model fit while inherently balancing goodness-of-fit with model complexity. This property makes it a natural criterion for model selection and the cornerstone of Bayes factors.

Despite its importance, marginal likelihood estimation is notoriously challenging in practice. The required integration over high-dimensional parameter space is rarely available in closed form, particularly in complex models with intricate prior-likelihood structures. As a result, a wide range of computational techniques, ranging from Laplace approximations and bridge sampling to recent advances such as variational inference and sequential Monte Carlo (MC) methods, have been developed to approximate this integral (Fourment et al. 2020). Each method involves trade-offs between computational efficiency, stability, and accuracy, making the choice of estimation strategy highly context-dependent.

Recently, posterior-sample–based estimators for marginal likelihoods have been applied across a range of scientific domains. For example, they have been used in item response theory to compare logistic models (Liu et al. 2019) and have seen rapid adoption in phylogenetics for model selection and evolutionary inference. Early methodological reviews highlighted such estimators as promising posterior-sample–only approaches for normalizing constant estimation (Oaks et al. 2019), and they have since been incorporated into practical guidelines for phylogenetic software workflows (Barido-Sottani et al. 2024).

Alongside these field-specific applications, software development has also accelerated: the R packages `bridgesampling` (Gronau et al. 2020), `BayesPPD` (Shen et al. 2021) and `lorad` (Milkey et al. 2023) now provide implementations of the bridge sampling (Meng and Wong 1996), PWK (Wang et al. 2018), and LoRaD (Wang et al. 2023) methods, respectively, for computing marginal likelihoods, further lowering the barrier for applied researchers to adopt these methods.

In this article, we focus primarily on the MC methods that rely on just two minimal assumptions: the availability of Markov chain Monte Carlo (MCMC) samples from the posterior and prior distributions, and their unnormalized kernels can be numerically evaluated. This restriction is motivated by both practical and theoretical considerations. First, these minimal assumptions are met in most Bayesian analyses, making such methods broadly applicable across a wide range of models without requiring model-specific structures or additional sampling schemes. Second, this class of methods aligns closely with common practice, where posterior MCMC samples are routinely obtained from standard Bayesian software packages (e.g., Stan, JAGS, and NIMBLE) and the posterior kernel is readily accessible. Third, by restricting attention to methods that rely on commonly available inputs, we ensure a fair comparison across estimators without the confounding influence of context-specific tuning or auxiliary model information. Finally, these methods are typically easier to implement and more computationally scalable, especially in high-dimensional settings where advanced techniques, such as path sampling or nested sampling, can become prohibitively complex. To this end, we include and compare the harmonic mean (HM; Newton and Raftery 1994), the inflated density ratio (IDR; Petris and Tardella 2003), the partitioned weighted kernel (PWK; Wang et al. 2018), and the lowest radial distance (LoRaD; Wang et al. 2023) methods in this review. Additionally, we include importance sampling (IS) due to its special application in the power prior setting. Specifically, the preceding terms in the posterior and prior kernels naturally serve as ideal references for the current ones, thereby enhancing efficiency without incurring additional computational burden when exploring levels of information borrowing.

This review thus complements previous reviews (Fourment et al. 2020; Oaks et al. 2019), which covered a much wider array of approaches by focusing attention on a narrow class of methods that require only quantities that all Bayesian statistical software can provide, namely a single sample from the joint posterior distribution and the log-likelihood and log-prior of each sampled point. The newest of these methods (i.e., Wang et al. 2023) was not available to previous reviewers.

The remainder of the paper is organized as follows: In Section 2, we present the general setting of computational problems, the detailed descriptions of Monte Carlo methods, and the formulation for estimating Monte Carlo errors. Section 3.1 presents a detailed analysis of the lead-exposed children (TLC) trial data to assess ordered-variance patterns. Section 3.2 presents an important application in quantifying the degree of information borrowing within the power prior framework for the benchmark dose (BMD) data set. In these two case studies, the methods discussed in Section 2.2 are implemented and compared. Brief discussion is provided in Section 4.

2 | Monte Carlo Methods

2.1 | The General Setting

We consider a general setting for estimating a ratio of two normalizing constants. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)'$ denote a vector of parameters of interest. Assume that $q_0(\boldsymbol{\theta})$ is an unnormalized proper prior kernel up to a constant c_0 and $q_1(\boldsymbol{\theta})$ is an unnormalized proper posterior kernel up to a constant c_1 so that

$$c_0 = \int q_0(\boldsymbol{\theta}) d\boldsymbol{\theta} < \infty \text{ and } c_1 = \int q_1(\boldsymbol{\theta}) d\boldsymbol{\theta} < \infty \tag{1}$$

We are interested in estimating a ratio of two normalizing constants as follows:

$$R = \frac{c_1}{c_0} \tag{2}$$

Suppose that D denotes the current data set of interest and $L(\boldsymbol{\theta}|D)$ is the likelihood function. Taking $q_1(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|D) q_0(\boldsymbol{\theta})$, (2) reduces to the marginal likelihood given by

$$m(D) = \frac{c_1}{c_0} = \frac{\int L(\boldsymbol{\theta}|D) q_0(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int q_0(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{3}$$

We further assume that a MCMC sample of size T_1, $\{\boldsymbol{\theta}_{11}, \dots, \boldsymbol{\theta}_{1T_1}\}$, and another MCMC sample of size T_0, $\{\boldsymbol{\theta}_{01}, \dots, \boldsymbol{\theta}_{0T_0}\}$, from the posterior, which is proportional to $q_1(\boldsymbol{\theta})$, and the prior, which is proportional to $q_0(\boldsymbol{\theta})$, are readily available. In the subsequent subsection, our review focuses primarily on Monte Carlo methods that can be implemented only using one or both of these two MCMC samples from the posterior and prior distributions and under the additional assumption that $q_1(\boldsymbol{\theta})$ and $q_0(\boldsymbol{\theta})$ can be evaluated numerically.

2.2 | The Existing Methods

2.2.1 | Importance Sampling

Using the MCMC sample, $\{\boldsymbol{\theta}_{01}, \dots, \boldsymbol{\theta}_{0T_0}\}$, from the prior, the IS method is readily applied to estimate R in (2) or $m(D)$ in (3) as follows:

$$\hat{R}_{\mathrm{IS}} = \frac{1}{T_0} \sum_{t=1}^{T_0} \frac{q_1(\boldsymbol{\theta}_{0t})}{q_0(\boldsymbol{\theta}_{0t})} \tag{4}$$

when $q_1(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|D) q_0(\boldsymbol{\theta})$, $\hat{R}_{\mathrm{IS}}$ reduces to an estimate of $m(D)$ given by

$$\hat{m}_{\mathrm{IS}}(D) = \frac{1}{T_0} \sum_{t=1}^{T_0} L(\boldsymbol{\theta}_{0t}|D) \tag{5}$$

We note that Newton and Raftery (1994) discussed the IS estimate in (5) of the marginal likelihood. It is easy to show that $\hat{R}_{\text{IS}}$ is a consistent estimator of R and $\hat{m}_{\text{IS}}(D)$ is a consistent estimator of $m(D)$ as $T_0 \to \infty$ under some ergodic conditions.

2.2.2 | The Harmonic Mean (HM) Method

Similarly to the IS method, the harmonic mean (HM) method uses only the MCMC sample, $\{\boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{1T_1}\}$, from the posterior. The form of the HM estimate of R is given by

$$\hat{R}_{\text{HM}} = \left[\frac{1}{T_1}\sum_{t=1}^{T_1}\frac{q_0(\boldsymbol{\theta}_{1t})}{q_1(\boldsymbol{\theta}_{1t})}\right]^{-1} \tag{6}$$

when $q_1(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|D)q_0(\boldsymbol{\theta})$, $\hat{R}_{\text{HM}}$ reduces to an estimate of $m(D)$ given by

$$\hat{m}_{\text{HM}}(D) = \left[\frac{1}{T_1}\sum_{t=1}^{T_1}\frac{1}{L(\boldsymbol{\theta}_{1t}|D)}\right]^{-1} \tag{7}$$

Following Newton and Raftery (1994), we can show that (i) $\hat{R}_{\text{HM}}$ is a consistent estimator of R and (ii) $\hat{m}_{\text{HM}}(D)$ is a consistent estimator of $m(D)$ as $T_1 \to \infty$ under some ergodic conditions. As noted in Xie et al. (2011), the HM estimator does not guarantee a finite variance and tends to overestimate the marginal likelihood.

2.2.3 | The Inflated Density Ratio (IDR) Estimator

Unlike IS and HM, the IDR method estimates either c_1 from posterior draws $\{\boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{1T_1}\}$ or c_0 from prior draws $\{\boldsymbol{\theta}_{01}, \ldots, \boldsymbol{\theta}_{0T_0}\}$ separately. For clarity, we present the formulation for c_1, while the estimation of c_0 follows the same principle with prior samples. Let $\boldsymbol{\theta}_{\text{mode}}$ denote the mode of $q_1(\boldsymbol{\theta})$. Petris and Tardella (2003) proposed the IDR estimator with a perturbed posterior density.

$$q_{1r}(\boldsymbol{\theta}) = \begin{cases} q_1(\boldsymbol{\theta}_{\text{mode}}), & \text{if } [t]\|\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{mode}}\| \le r \\ q_1(w(\boldsymbol{\theta})), & \text{otherwise} \end{cases}$$

and of the form

$$\hat{c}_{1,\text{IDR}} = \frac{q_1(\boldsymbol{\theta}_{\text{mode}})b_r}{\frac{1}{T_1}\sum_{t=1}^{T_1}\frac{q_{1r}(\boldsymbol{\theta}_{1t})}{q_1(\boldsymbol{\theta}_{1t})} - 1} \tag{8}$$

where r is the pre-specified radius, $w(\boldsymbol{\theta}) = (\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{mode}})\left(1 - r^p/\|\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{mode}}\|^p\right)^{1/p}$, and b_r is the volume of the ball $\int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{mode}}\|\le r} d\boldsymbol{\theta} = \pi^{p/2}r^p/\Gamma(p/2+1)$. The form of the IDR estimate of R is given by

$$\hat{R}_{\text{IDR}} = \frac{\hat{c}_{1,\text{IDR}}}{\hat{c}_{0,\text{IDR}}}, \tag{9}$$

where $\hat{c}_{0,\text{IDR}}$ is defined analogously using prior samples $\{\boldsymbol{\theta}_{01}, \ldots, \boldsymbol{\theta}_{0T_0}\}$ and the perturbed prior density. When the prior is proper, $\hat{R}_{\text{IDR}}$ reduces to an estimate of $m(D)$ given by

$$\hat{m}_{\text{IDR}}(D) = \frac{q_1(\boldsymbol{\theta}_{\text{mode}})b_r}{\frac{1}{T_1}\sum_{t=1}^{T_1}\frac{q_{1r}(\boldsymbol{\theta}_{1t})}{q_1(\boldsymbol{\theta}_{1t})} - 1} \tag{10}$$

Note that (8) is valid due to

$$\begin{aligned}\int q_{1r}(\boldsymbol{\theta})d\boldsymbol{\theta} &= \int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{mode}}\|\le r} q_{1r}(\boldsymbol{\theta})d\boldsymbol{\theta} + \int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{mode}}\|> r} q_{1r}(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= q_1(\boldsymbol{\theta}_{\text{mode}})b_r + c_1\end{aligned} \tag{11}$$

As noted in Wang et al. (2018), although the IDR estimator enjoys the property of finite variance under some mild conditions, it can be sensitive to the choice of radius and mode specification. Additionally, to improve IDR performance, Wang et al. (2018) suggested working with the standardized MCMC sample instead of the one on the original scale.

2.2.4 | The Partitioned Weighted Kernel (PWK) Method

Motivated by Newton and Raftery (1994) and Petris and Tardella (2003), Wang et al. (2018) proposed the partition weighted kernel (PWK) estimator, which is constructed from posterior draws based on a new identity over a working parameter space $\boldsymbol{\Omega}$ (a subset of the full parameter space):

$$\frac{\Delta_1}{c_1} = \int_{\boldsymbol{\Omega}} \frac{h_1(\boldsymbol{\theta})}{q_1(\boldsymbol{\theta})}\frac{q_1(\boldsymbol{\theta})}{c_1}d\boldsymbol{\theta}$$

where $h_1(\boldsymbol{\theta}) = \sum_{k=1}^{K} q_1(\boldsymbol{\theta}_{1k}^*)1\{\boldsymbol{\theta} \in A_{1k}\}$ is a step function, $\Delta_1 = \int h_1(\boldsymbol{\theta})d\boldsymbol{\theta}$, $\{A_{1k}: k = 1, 2, \ldots, K\}$ denotes a partition of $\boldsymbol{\Omega}$, and $q_1(\boldsymbol{\theta}_{1k}^*)$ the representative point for each subset A_{1k} (e.g., the average kernel value of the MCMC samples in A_{1k}). Accordingly, Δ_1 has the closed form $\sum_{k=1}^{K} q_1(\boldsymbol{\theta}_{1k}^*)V(A_{1k})$, and the PWK estimator is of the form.

$$\frac{1}{\hat{c}_{1,\text{PWK}}} = \frac{\frac{1}{T_1}\sum_{t=1}^{T_1}\sum_{k=1}^{K}\frac{q_1(\boldsymbol{\theta}_{1k}^*)}{q_1(\boldsymbol{\theta}_{1t})}1\{\boldsymbol{\theta}_{1t} \in A_{1k}\}}{\sum_{k=1}^{K} q_1(\boldsymbol{\theta}_{1k}^*)V(A_{1k})} \tag{12}$$

The form of the PWK estimate of R is given by

$$\hat{R}_{\text{PWK}} = \frac{\hat{c}_{1,\text{PWK}}}{\hat{c}_{0,\text{PWK}}} \tag{13}$$

where $\hat{c}_{0,\text{PWK}}$ can be obtained analogously to (12). When the prior is proper, (13) reduces to the marginal likelihood. Under certain ergodic conditions, (12) has been proven to be consistent for $1/c_1$ and to have finite variance. Moreover, the optimal result can be theoretically achieved by continuously increasing the number of partitions K, thereby ensuring greater homogeneity of $q_1(\boldsymbol{\theta})$ over each region A_{1k}. To facilitate the partitioning

TABLE 1 | Comparison of marginal likelihood estimation methods.

Method	Stability and property	Computational efficiency	Dimensional scalability
IS	The variance of $\widehat{R}_{\text{IS}}$ in (4) is small if $q_0(\boldsymbol{\theta})$ in (4) is a good proposal density of $q_1(\boldsymbol{\theta})$.	Requires less than 1% of the PWK time.	Performance decreases as dimensionality increases; performance largely depends on how well $q_0(\boldsymbol{\theta})$ approximates $q_1(\boldsymbol{\theta})$.
HM	The estimator in (6) can have a large variance when $q_1(\boldsymbol{\theta})$ has a lighter tail than $q_0(\boldsymbol{\theta})$.	Similar computing time as IS; both require less than 1% of the PWK time.	Scales badly in higher dimensions; instability becomes more severe.
IDR	Uses a bounded ratio $q_{1r}(\boldsymbol{\theta})/q_1(\boldsymbol{\theta})$ in (8), so the estimator typically has finite variance.	Requires about three times the PWK time.	Monte Carlo error in (8) increases with dimension due to the shrinking volume of the perturbed region and finding the mode is not dimension-scalable.
PWK	Compared with all other reviewed methods, PWK is uniquely black-boxable.	Used as the reference baseline (100% computing time).	It is dimension-scalable. Constructing an optimal working parameter space remains difficult in the high dimensional case.
LoRaD	Excellent stability when local normality is achieved.	Reduces computing time by about 45%–50% compared with PWK.	Highly dimension-scalable.

Note: Relative computing times are based on the ECOG data application.

procedure and evaluation of the corresponding Δ_1, Wang et al. (2018) proposed the partitioning ring and slice approach and demonstrated its effectiveness in several challenging problems in the literature. It is worth mentioning that a recently developed method, the Truncated HArmonic Mean EStimator (THAMES), by Metodiev et al. (2025), is simply a special case of (12) when specifying $K = 1$ and $q_1(\boldsymbol{\theta}^*_{1k}) = 1$.

2.2.5 | The Lowest Radial Distance (LoRaD) Method

While PWK has appealing properties, its use in high dimensions is hindered by sparsity and the difficulty of identifying representative points, leading Wang et al. (2023) to propose the LoRaD method:

$$\frac{1}{\widehat{c}_{1,\text{LoRaD}}} = \frac{\frac{1}{T_1}\sum_{t=1}^{T_1}\frac{\widetilde{q}_1(\widetilde{\boldsymbol{\theta}}_{1t})}{q_1(\boldsymbol{\theta}_{1t})|\boldsymbol{\Sigma}_0^{1/2}|}\mathbf{1}\left\{\|\widetilde{\boldsymbol{\theta}}_{1t}\|\le r\right\}}{\int_{\|\widetilde{\boldsymbol{\theta}}\|\le r}\widetilde{q}_1(\widetilde{\boldsymbol{\theta}})d\widetilde{\boldsymbol{\theta}}} = \frac{\frac{1}{T_1}\sum_{t=1}^{T_1}\frac{\widetilde{q}_1(\widetilde{\boldsymbol{\theta}}_{1t})}{q_1(\boldsymbol{\theta}_{1t})|\boldsymbol{\Sigma}_0^{1/2}|}\mathbf{1}\left\{\|\widetilde{\boldsymbol{\theta}}_{1t}\|\le r\right\}}{\left(1-\Gamma(p/2, r^2/2)\right)/\Gamma(p/2)} \tag{14}$$

where $\widetilde{\boldsymbol{\theta}}_{1t} = \boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{\theta}_{1t} - \boldsymbol{\mu}_0)$ is the standardization of $\boldsymbol{\theta}_{1t}$ based on posterior means $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$, the reference density.

$$\widetilde{q}_1(\widetilde{\boldsymbol{\theta}}_{1t}) = (2\pi)^{-p/2}\exp\left(-\widetilde{\boldsymbol{\theta}}_{1t}'\widetilde{\boldsymbol{\theta}}_{1t}/2\right)$$

is the standard multivariate normal distribution, and $\Gamma(s,x)$ is the upper incomplete gamma function, which is defined as $\Gamma(s,x) = \int_x^\infty t^{s-1}e^{-t}dt$. Similarly, $\widehat{c}_{0,\text{LoRaD}}$ can be obtained as in (14). The ratio, $\widehat{c}_{1,\text{LoRaD}}/\widehat{c}_{0,\text{LoRaD}}$, reduces to the marginal likelihood when the prior is proper. Through standardization, the posterior distribution can be approximated by a standard multivariate normal distribution, enabling the estimator to construct an effective reference density, particularly in high-dimensional settings.

Table 1 summarizes the main practical features of the marginal likelihood estimators considered here, in terms of stability, computational efficiency, and dimensional scalability. LoRaD achieves a more favorable balance between stability and efficiency, combining good stability with reduced computing time relative to PWK and IDR.

2.3 | Monte Carlo Error Estimation

The Overlapping Batch Statistics (OBS) method (Schmeiser et al. 1990) is used to obtain an estimated Monte Carlo standard error (eMCSE), which is given by

$$\text{eMCSE}(\widehat{\eta}) = \sqrt{\widehat{\text{Var}(\widehat{\eta})}} = \left\{\left[\frac{B}{T-B}\right]\frac{\sum_{b=1}^{T-B+1}\left(\widehat{\eta}_b - \overline{\eta}\right)^2}{T-B+1}\right\}^{\frac{1}{2}} \tag{15}$$

where $\widehat{\eta}_b$ denotes the estimate of the normalizing constant (in log scale) obtained from the b-th batch $\{\boldsymbol{\theta}_t : t = b, b+1, \ldots, b+B-1\}$ with $B < T$, T is the total MCMC sample size, and $\overline{\eta} = \frac{1}{T-B+1}\sum_{b=1}^{T-B+1}\widehat{\eta}_b$ is the average of the batch estimates. Here, η can be defined as $\eta = \log c_1$ when estimating the posterior normalizing constant, or as $\eta = \log c_1 - \log c_0$ when estimating their ratio. Assuming that the posterior and prior samples are generated independently, it follows that $\text{eMCSE}(\hat{\eta}) = \sqrt{\widehat{\text{Var}}(\hat{\eta})} = \sqrt{\text{Var}\left(\widehat{\log c_1}\right) + \text{Var}\left(\widehat{\log c_0}\right)}$.

3 | Case Studies

3.1 | Analysis of TLC Data

In this section, we show how to calculate the marginal likelihood from MCMC samples using the methods outlined in the previous section. We consider a two-group setting in which repeated outcomes are observed for each subject. Data come from the Treatment of Lead-Exposed Children (TLC) trial (Rogan et al. 2000), where 100 children were randomized equally to placebo or succimer ($n_1 = n_2 = 50$). Let $\boldsymbol{y}_{ij}$ denote the three-dimensional vector of change-from-baseline blood lead levels at weeks 1, 4, and 6 for subject j in group i. Suppose

$$\boldsymbol{y}_{ij} \sim \mathcal{N}_3(\boldsymbol{\mu}_i, \Sigma_i), \qquad j = 1, \dots, n_i, \tag{16}$$

with $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \mu_{i3})'$ denoting the mean vector and Σ_i a positive-definite covariance matrix. Let $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the collection $\{(\boldsymbol{\mu}_i, \Sigma_i) : i = 1, 2\}$. We compare two models using marginal likelihood: (i) an unconstrained model $\mathcal{M}_0$ and (ii) a model $\mathcal{M}_1$ that satisfies an ordered-variance constraint, with parameter spaces

$$\mathcal{P}_0 = \left\{ (\boldsymbol{\mu}_i, \Sigma_i)_{i=1}^2 : -\infty < \mu_{ij} < \infty, j = 1, 2, 3, \;\; \Sigma_i \succ 0 \right\} \tag{17}$$

and

$$\begin{aligned} \mathcal{P}_1 = \{ &(\boldsymbol{\mu}_i, \Sigma_i)_{i=1}^2 : -\infty < \mu_{ij} < \infty, j = 1, 2, 3 \\ &\Sigma_i \succ 0, \Sigma_{i,11} < \cdots < \Sigma_{i,33} \text{ for } i = 1, 2 \} \end{aligned} \tag{18}$$

where $\Sigma_i \succ 0$ denotes that Σ_i is positive definite and $\Sigma_{i,11}, \Sigma_{i,22}, \Sigma_{i,33}$ denotes the diagonal elements. As shown in Table 2, both groups exhibit an ordered diagonal variance pattern.

Let D denote the observed data. The likelihood function for $\mathcal{M}_1$ is given by

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid D) = &\prod_{i=1}^{2} \prod_{j=1}^{n_i} |2\pi\Sigma_i|^{-\frac{1}{2}} \\ \times &\exp\left\{ -\frac{1}{2} (\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i)' \Sigma_i^{-1} (\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i) \right\} \mathbf{1}\{\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\} \in \mathcal{P}_1\} \end{aligned} \tag{19}$$

The prior is specified as a Normal–Inverse–Wishart distribution restricted to the constrained parameter space, given by

$$\begin{aligned} q_0(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = &\prod_{i=1}^{I} \frac{(\kappa_{0i})^{p/2} |\Psi_{0i}|^{\nu_{0i}/2}}{2^{\nu_{0i} p/2} (2\pi)^{p/2} \Gamma_3(\nu_{0i}/2)} |\Sigma_i|^{-(\nu_{0i}+p+2)/2} \\ \times &\exp\left\{ -\frac{1}{2} \text{tr}(\Psi_{0i}\Sigma_i^{-1}) - \frac{\kappa_{0i}}{2} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_{0i})' \Sigma_i^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_{0i}) \right\} \mathbf{1}\{\boldsymbol{\mu}, \boldsymbol{\Sigma} \in \mathcal{P}_1\} \end{aligned} \tag{20}$$

where $\boldsymbol{\mu}_{0i}$, κ_{0i}, ν_{01} and Ψ_{0i} for $i = 1, 2$ are the hyperparameters. Transformations are applied to map the parameters to an unconstrained space. For $\mathcal{M}_1$, we apply the decomposition of the covariance matrix.

$$\Sigma_i = \Lambda_i C_i \Lambda_i \tag{21}$$

where $\Lambda_i = \text{diag}(\lambda_{i,1}, \dots, \lambda_{i,3})$, $\lambda_{i,l}$ are the standard deviations, and C_i is the correlation matrix. The log transformation

$$\tilde{\lambda}_{i,l} = \log(\lambda_{i,l}), \qquad l = 1, \dots, 3 \tag{22}$$

The LKJ transformation (Lewandowski et al. 2009) allows one to express the entries of the correlation matrix $\rho_{i,jk}$ in terms of partial correlations $\eta_{i,jk}$. The Fisher transformation

$$\tilde{\eta}_{i,jk} = \frac{1}{2} \log\left(\frac{1 + \eta_{i,jk}}{1 - \eta_{i,jk}} \right) \tag{23}$$

and the ordered transformation

$$\xi_{i1} = \tilde{\sigma}_{i,1} = \log(\sigma_{i,11}), \qquad \tilde{\sigma}_{i,k} = \tilde{\sigma}_{i,k-1} + \exp(\xi_{ik}), \quad k = 2, 3 \tag{24}$$

For $\mathcal{M}_0$, we apply the Cholesky decomposition.

$$\Sigma_i = L_i L_i^\top \tag{25}$$

where L_i is a lower-triangular matrix with positive diagonal entries. The log transformation of diagonal elements.

$$\tilde{l}_{i,kk} = \log(l_{i,kk}), \qquad k = 1, \dots, 3 \tag{26}$$

where $l_{i,jj}$ denotes the j-th diagonal entry of L_i. Denote by $\boldsymbol{\theta}$ the vector of unconstrained parameters and write $T_l(\boldsymbol{\theta}) = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the above transformation. Then the posterior kernel is given by

TABLE 2 | Covariance and mean of data.

	Placebo			Succimer		
2–4 (lr)5–7 statistics	**Week 1**	**Week 4**	**Week 6**	**Week 1**	**Week 4**	**Week 6**
Covariance	9.57	5.27	4.46	53.15	38.64	22.72
	5.27	9.82	7.78	38.64	56.59	20.11
	4.46	7.78	14.21	22.72	20.11	64.73
Mean	−1.61	−2.20	−2.63	−13.02	−11.03	−5.78

$$q_1(\theta|\mathcal{M}_l) = L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|D)\, q_0(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{M}_l) \left| \frac{\partial(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \theta'} \right| \tag{27}$$

The marginal likelihood for $\mathcal{M}_1$ is given by

$$m(D|\mathcal{M}_1) = \frac{c_{1,\mathcal{M}_1}}{c_{0,\mathcal{M}_1}} = \frac{\int q_1(\theta|\mathcal{M}_1)d\theta}{\int q_0(\theta|\mathcal{M}_1)d\theta} \tag{28}$$

The marginal likelihood of $\mathcal{M}_0$ can be expressed in closed form as

$$m(D|\,\mathcal{M}_0) = \prod_{i=1}^{2} \pi^{-\frac{3n_i}{2}} \left(\frac{\kappa_{0i}}{\kappa_{0i}+n_i} \right)^{\frac{3}{2}} \frac{|\Psi_{0i}|^{\frac{\nu_{0i}}{2}}}{|\Psi_{ni}|^{\frac{\nu_{0i}+n_i}{2}}} \frac{\Gamma_3\left(\frac{\nu_{0i}+n_i}{2}\right)}{\Gamma_3\left(\frac{\nu_{0i}}{2}\right)} \tag{29}$$

where $\Psi_{ni} = \Psi_{0i} + \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_i)(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_i)' + \frac{n_i \kappa_{0i}}{n_i + \kappa_{0i}} (\bar{\boldsymbol{y}}_i - \boldsymbol{\mu}_{0i})(\bar{\boldsymbol{y}}_i - \boldsymbol{\mu}_{0i})'$. The marginal likelihood $m(D|\mathcal{M}_0)$ provides a reference value against which the performance of different estimation methods can be evaluated. We obtain 20,000 posterior MCMC samples via Stan (Carpenter et al. 2017) for both models under the prior hyperparameters $\boldsymbol{\mu}_{0i} = (0,0,0)'$, $\kappa_{0i} = 0.01$, $\nu_{0i} = 7$, and $\Psi_{0i} = 10I_3$ for $i = 1,2$. Stan sampling was performed with four chains. For the posterior, each chain used 1000 warmup iterations followed by 5000 post-warmup iterations, yielding a total of 20,000 posterior draws. For the prior, each chain used 10,000 warmup iterations followed by 5000 post-warmup iterations, also yielding 20,000 prior draws. Figures 1 and 2 present trace plots and autocorrelation (ACF) plots for diagonal elements, $\xi 11, \xi_{12}, \xi_{13}, \xi_{21}, \xi_{22}$, and ξ_{23} in (24), for both groups. The effective sample sizes (ESS) for these parameters were computed using the `effectiveSize` function from the `coda` package. These diagnostics confirm good mixing behavior across the selected parameters.

Under this setting, the normalizing constant for the ordered diagonal variance constraint can be determined.

$$\hat{c}_{0,\mathcal{M}_1} = \left(\frac{1}{3!}\right)^2 \tag{30}$$

This value is used consistently across all comparisons, while $c_{1,\mathcal{M}_1}$ is estimated using posterior samples. The eMCSE is obtained using (15). We set the batch size to $B = 2000$ for methods that do not require a training sample, and to $B = 1000$ for methods that require training, where an additional 10,000 samples are reserved exclusively for training purposes. Under this structure, HM directly estimates the marginal likelihood without requiring a proper prior. Methods (PWK, LoRaD) that require a working parameter space use an estimate based on the radius derived from the standardized unconstrained parameters of the MCMC samples. Specifically, the radius is determined from the quantile of the MCMC draws: for example, a 10% setting corresponds to defining $r_{\max}$ as the distance from the training sample's mode ($r = 0$) to the 10th percentile of the MCMC sample distribution. The working space (or truncated subset) is then defined within the range $r \in [0, r_{\max}]$. For IDR, the in-sample mode is used to maintain a reasonable $q(0)$. Table 3 shows the estimates and the corresponding eMCSEs. The subscripts indicate parameter settings. For example, IDR_{10} and IDR_{20} denote IDR with 10% and 20%, respectively. $\text{PWK}_{10,100}$ denotes PWK with radius defined by the 10% quantile and partition number $K = 100$. LoRaD_{20} denotes LoRaD with radius defined by the 20% quantile.

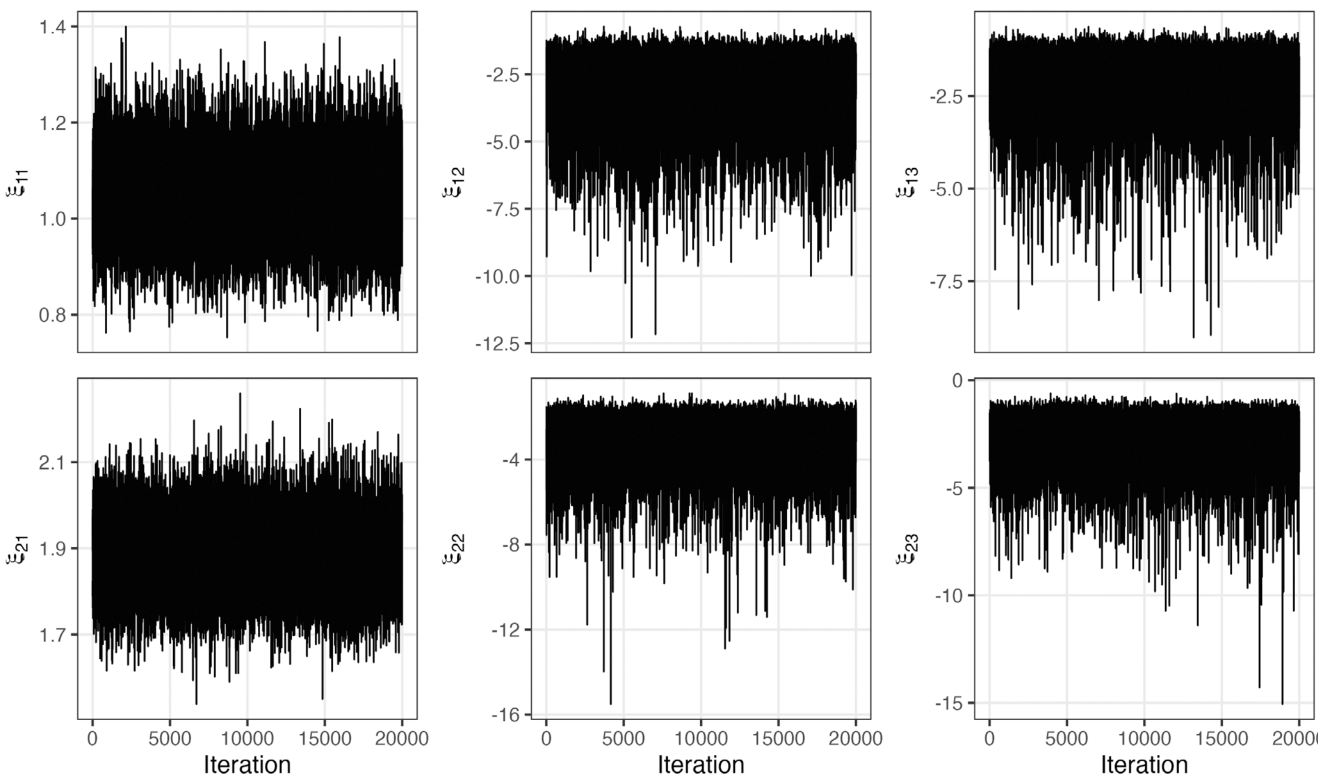

FIGURE 1 | Trace plots for six representative parameters from the TLC application.

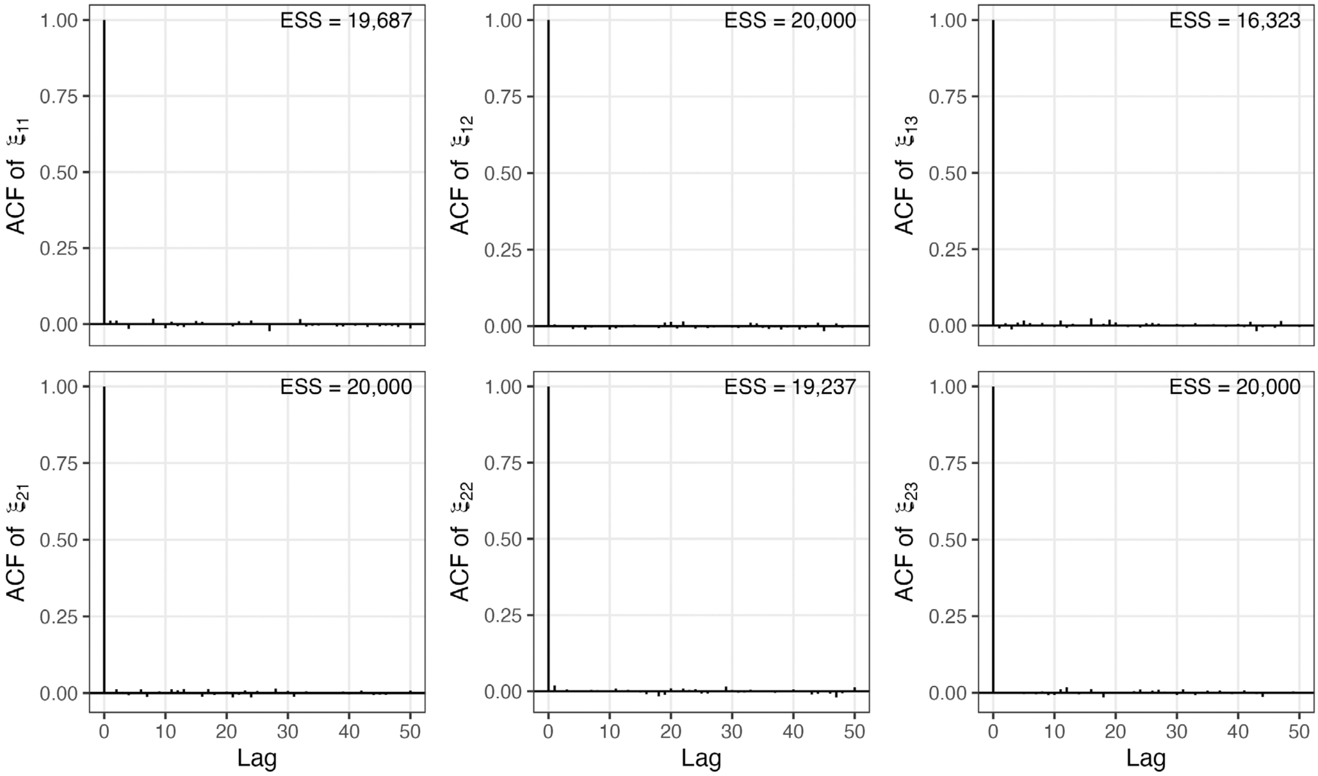

FIGURE 2 | Autocorrelation (ACF) plots for the six representative parameters from the TLC application.

TABLE 3 | Estimates in log scale with eMCSE for different methods.

	$\log \hat{m}(D\mid \mathcal{M}_1)$	$\log \hat{m}(D\mid \mathcal{M}_0)$
True value	—	−936.297
HM	−883.329 (0.749)	−880.656 (0.431)
IDR_{10}	−933.907 (0.143)	−936.283 (0.107)
IDR_{20}	−934.203 (0.168)	−936.171 (0.068)
$\text{PWK}_{10,100}$	−934.112 (0.205)	−936.322 (0.068)
$\text{PWK}_{20,100}$	−934.109 (0.199)	−936.240 (0.048)
$\text{PWK}_{10,50}$	−934.104 (0.208)	−936.322 (0.070)
$\text{PWK}_{20,50}$	−934.121 (0.196)	−936.237 (0.049)
LoRaD_{20}	−934.133 (0.017)	−936.226 (0.011)

Note: The subscripts indicate parameter settings. For example, IDR_{10} and IDR_{20} denote IDR with 10% and 20%, respectively. $\text{PWK}_{10,100}$ denotes PWK with radius defined by the 10% quantile and partition number $K = 100$. LoRaD_{20} denotes LoRaD with radius defined by the 20% quantile.

The HM estimator tends to overestimate marginal likelihoods, as the prior is relatively flat compared with the posterior. IDR, PWK, and LoRaD yield numerically similar estimates. All three methods favor $\mathcal{M}_1$ as providing a better fit to the data, while HM yields the opposite conclusion.

3.2 | Analysis of ECOG Data

In this subsection, we revisit the example of a power prior application presented in Wang et al. (2018), and conduct comprehensive comparisons of the methods reviewed in Section 2.2. To ensure fairness, all methods are restricted to using the same 21 MCMC samples (each of size 10,000 for every power-posterior and power-prior kernel) originally generated for the PWK calculation in Wang et al. (2018). We also include the IS method in our analysis, given its natural suitability for the power prior setting. This method does not require additional MCMC sampling but offers a more efficient alternative to determine the degree of information borrowing when complete exclusion of borrowing is not an option.

The motivating data arise from two clinical trials (E1684 and E1690) conducted by the Eastern Cooperative Oncology Group (ECOG) to evaluate the effectiveness of Interferon Alpha-2b (IFN) as an immunotherapy for melanoma patients. Trial E1684 is a phase III study with $n_0 = 286$ patients randomized to high-dose IFN or placebo. It was found that the IFN arm demonstrated a significantly better survival curve compared to the placebo group. To further assess IFN efficacy and explore potential benefits of a lower dosage, ECOG initiated a subsequent trial, E1690, which included three arms: high-dose IFN, low-dose IFN, and placebo. Given the chronological order and comparability of these trials, we treat E1684 as the historical data set to inform the subsequent trial E1690 within the power prior framework, and restrict our analysis to the high-dose IFN and placebo arms, which results in $n = 427$ patients in this subset.

For the i-th patient, let y_i denote the relapse-free survival time, ν_i the censoring indicator (1 if y_i is an event time, 0 otherwise), and $\boldsymbol{x}_i = \left(1, \text{trt}_i\right)'$ the covariate vector, where $\text{trt}_i = 1$ if the patient received high-dose IFN, and 0 otherwise. Let $D = \{n, \boldsymbol{y}, \boldsymbol{\nu}, \boldsymbol{X}\}$ and $D_0 = \left\{n_0, \boldsymbol{y}_0, \boldsymbol{\nu}_0, \boldsymbol{X}_0\right\}$ represent the current data and the historical

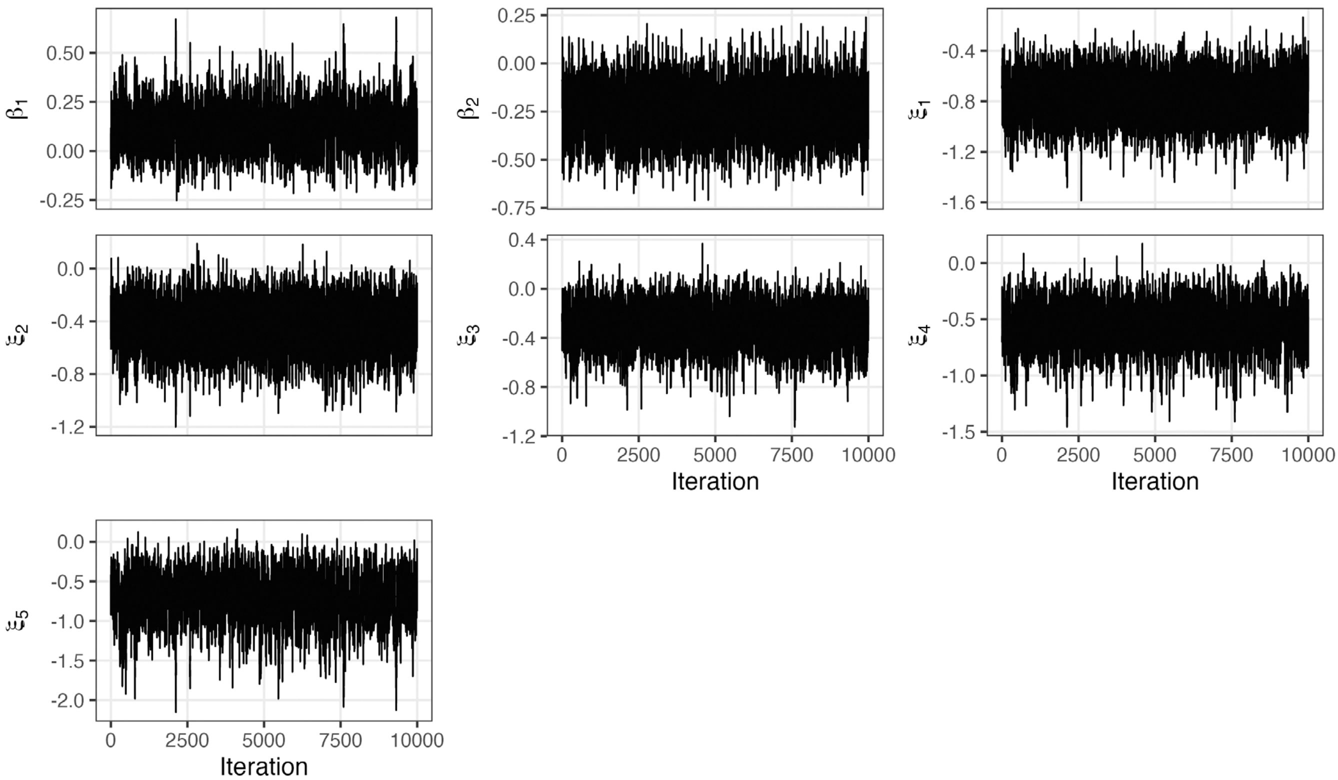

FIGURE 3 | Trace plots for all seven parameters in the ECOG application.

data, respectively. Following the same piecewise exponential model of Chen et al. (1999), the likelihood for the current data is

$$L(\boldsymbol{\theta}|\, D)=\prod_{i=1}^{n}\left\{\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)f(y_i|\,\xi)\right\}^{\nu_i}\exp\left\{-\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)F(y_i|\,\xi)\right\} \tag{31}$$

with $F(y|\,\xi)=1-\exp\left\{-\exp\left(\xi_j\right)\left(y-s_{j-1}\right)-\sum_{g=1}^{j-1}\exp\left(\xi_g\right)\left(s_g-s_{g-1}\right)\right\}$ and $f(y|\,\xi)=\frac{dF(y|\xi)}{dy}$, where $s_{j-1}\le y<s_j$, $\left(s_0,s_1,s_2,s_3,s_4,s_5\right)=(0,0.241,0.479,0.887,1.704,\infty)$, and $\boldsymbol{\theta}=\left(\boldsymbol{\beta}',\boldsymbol{\xi}'\right)=\left(\beta_0,\beta_1,\xi_1,\ \ldots,\xi_5\right)'$.

Let $L\left(\boldsymbol{\theta}|\,D_0\right)$ denote the likelihood for the historical data, which shares the same functional form as $L(\boldsymbol{\theta}|\,D)$, differing only in the data inputs. The power prior (Chen et al. 2025; Ibrahim and Chen 2000; Ibrahim et al. 2015) is then defined as

$$\pi\left(\boldsymbol{\theta}|\,D_0,a_0\right)=\frac{L\left(\boldsymbol{\theta}|\,D_0\right)^{a_0}\pi_0(\boldsymbol{\theta})}{\int L\left(\boldsymbol{\theta}|\,D_0\right)^{a_0}\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}}=\frac{q_0\left(\boldsymbol{\theta},a_0\right)}{c_0\left(a_0\right)} \tag{32}$$

where $0\le a_0\le 1$, $\pi_0(\boldsymbol{\theta})$ is a proper initial prior, which is specified as $\beta_0,\beta_1\overset{\text{i.i.d.}}{\sim}N(0,100)$ and $\xi_1,\ \ldots,\xi_5\overset{\text{i.i.d.}}{\sim}\pi_0(\xi)=\frac{1}{100}\exp\left\{-\frac{1}{100}\exp(\xi)\right\}\exp(\xi)$, where $-\infty<\xi<\infty$, and $c_0\left(a_0\right)$ is the prior normalizing constant. Using (32), the posterior density is given by

$$\pi\left(\boldsymbol{\theta}|\,D,D_0,a_0\right)=\frac{L(\boldsymbol{\theta}|\,D)L\left(\boldsymbol{\theta}|\,D_0\right)^{a_0}\pi_0(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta}|\,D)L\left(\boldsymbol{\theta}|\,D_0\right)^{a_0}\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}}=\frac{q_1\left(\boldsymbol{\theta},a_0\right)}{c_1\left(a_0\right)} \tag{33}$$

where $c_1\left(a_0\right)$ is the posterior normalizing constant. Consequently, the marginal likelihood is defined as

$$m\left(D|\,D_0,a_0\right)=\int L(\boldsymbol{\theta}|\,D)\left\{\frac{L\left(\boldsymbol{\theta}|\,D_0\right)^{a_0}\pi_0(\boldsymbol{\theta})}{c_0\left(a_0\right)}\right\}d\boldsymbol{\theta}=\frac{c_1\left(a_0\right)}{c_0\left(a_0\right)} \tag{34}$$

We consider 11 values of a_0, denoted by $a_{0,\ell}=\ell/10$, for $\ell=0,1,\ \ldots,10$. We are interested in determining which $a_{0,\ell}$ results in the largest value of $m\left(D|\,D_0,a_0\right)$ using 11 independent MCMC samples of size $T_1=10,000$, $\left\{\boldsymbol{\theta}_{1t}^{(\ell)},t=1,\ \ldots,T_1\right\}$, from the posterior $\pi\left(\boldsymbol{\theta}|\,D,D_0,a_{0,\ell}\right)$ in (33) for $\ell=0,1,\ \ldots,10$ and 10 additional independent MCMC samples of size $T_0=10,000$, $\left\{\boldsymbol{\theta}_{0t}^{(\ell)},t=1,\ \ldots,T_0\right\}$, from the prior $\pi\left(\boldsymbol{\theta}|\,D_0,a_{0,\ell}\right)$ in (32) for $\ell=1,2,\ \ldots,10$. We apply the methods discussed in Section 2.2 to carry out the computations. Figures 3 and 4 display trace plots and autocorrelation (ACF) plots for all the seven parameters in the ECOG application, corresponding to $\boldsymbol{\theta}=\left(\boldsymbol{\beta}',\boldsymbol{\xi}'\right)=\left(\beta_0,\beta_1,\xi_1,\ \ldots,\xi_5\right)'$, shown in this order from left to right and top to bottom. The effective sample sizes (ESS) for these parameters were computed using the `effectiveSize` function from the `coda` package.

For the IS method, we cannot calculate $m\left(D|\,D_0,a_{0,1}\right)/m\left(D|\,D_0,a_0=0\right)$ as no MCMC samples are available from the prior $\pi\left(\boldsymbol{\theta}|\,D_0,a_{0,0}\right)$. Observe that for $\ell=2,3,\ \ldots,10$,

$$R_{m,\ell,\ell-1}=\frac{m\left(D|\,D_0,a_{0,\ell}\right)}{m\left(D|\,D_0,a_{0,\ell-1}\right)}=\frac{c_1\left(a_{0,\ell}\right)}{c_1\left(a_{0,\ell-1}\right)}\times\left(\frac{c_0\left(a_{0,\ell}\right)}{c_0\left(a_{0,\ell-1}\right)}\right)^{-1} \tag{35}$$

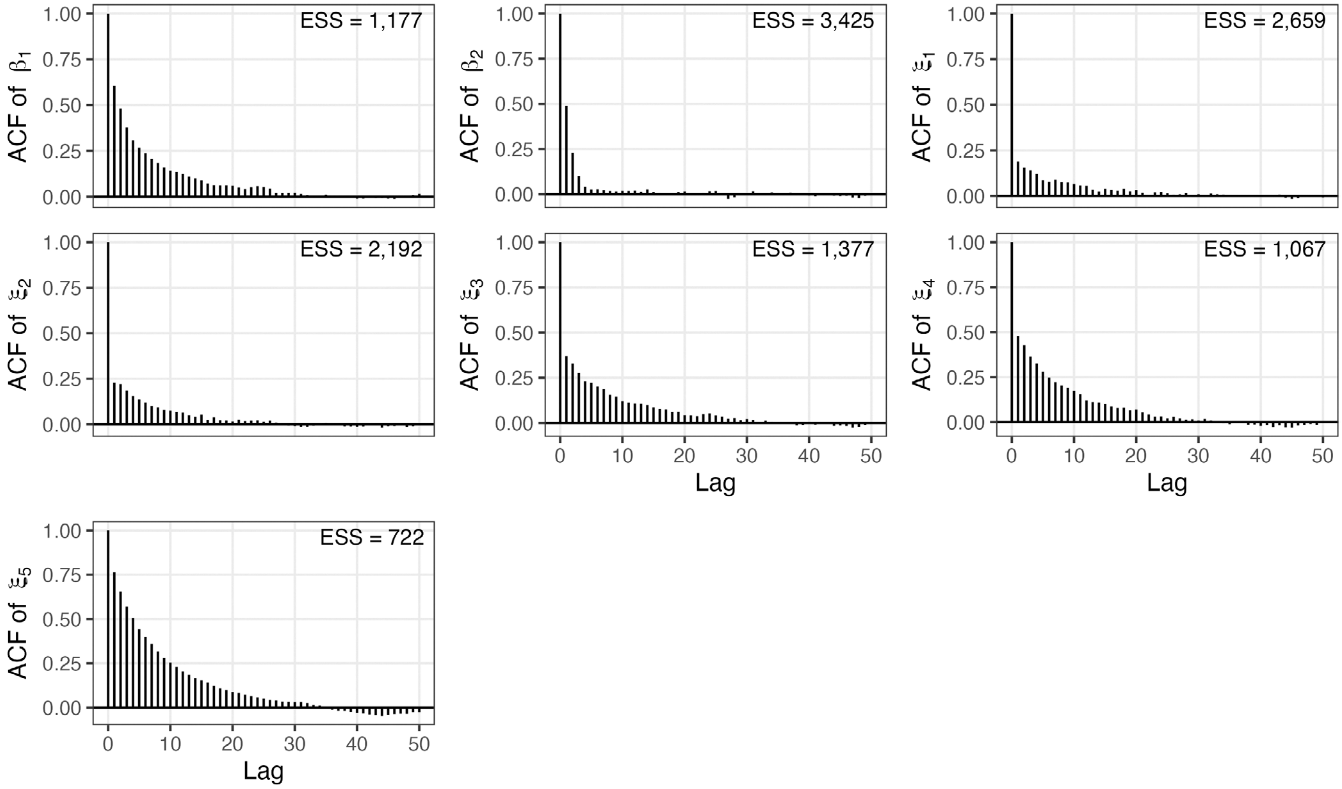

FIGURE 4 | Autocorrelation (ACF) plots for all seven parameters in the ECOG application.

$$R_{1,\ell,\ell-1} = \frac{c_1(a_{0,\ell})}{c_1(a_{0,\ell-1})} = \frac{\int L(\boldsymbol{\theta}|D)L(\boldsymbol{\theta}|D_0)^{a_{0,\ell}}\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int L(\boldsymbol{\theta}|D)L(\boldsymbol{\theta}|D_0)^{a_{0,\ell-1}}\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{36}$$

and

$$R_{0,\ell,\ell-1} = \frac{c_0(a_{0,\ell})}{c_0(a_{0,\ell-1})} = \frac{\int L(\boldsymbol{\theta}|D_0)^{a_{0,\ell}}\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int L(\boldsymbol{\theta}|D_0)^{a_{0,\ell-1}}\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{37}$$

For estimating $R_{1,\ell,\ell-1}$ in (36), using $\left\{\boldsymbol{\theta}_{1t}^{(\ell-1)}, t = 1, \ldots, T_1\right\}$ and (4) with $q_0(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|D)L(\boldsymbol{\theta}|D_0)^{a_{0,\ell-1}}\pi_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|D)L(\boldsymbol{\theta}|D_0)^{a_{0,\ell}}\pi_0(\boldsymbol{\theta})$, an IS estimate of $R_{1,\ell,\ell-1}$ is given by

$$\hat{R}_{1,\ell,\ell-1,\mathrm{IS}} = \frac{1}{T_1}\sum_{t=1}^{T_1} L\left(\boldsymbol{\theta}_{1t}^{(\ell-1)}|D_0\right)^{a_{0,\ell}-a_{0,\ell-1}} \tag{38}$$

Similarly, an IS estimate of $R_{0,\ell,\ell-1}$ is obtained as

$$\hat{R}_{0,\ell,\ell-1,\mathrm{IS}} = \frac{1}{T_0}\sum_{t=1}^{T_0} L\left(\boldsymbol{\theta}_{0t}^{(\ell-1)}|D_0\right)^{a_{0,\ell}-a_{0,\ell-1}} \tag{39}$$

Consequently, an IS estimate of $R_{m,\ell,\ell-1}$ is given by

$$\hat{R}_{m,\ell,\ell-1,\mathrm{IS}} = \hat{R}_{1,\ell,\ell-1,\mathrm{IS}} / \hat{R}_{0,\ell,\ell-1,\mathrm{IS}}$$

If the largest marginal likelihood value is not $m(D|D_0, a_{0,0})$ and $\hat{R}_{m,\ell,\ell-1,\mathrm{IS}}$ is a strictly decreasing function of ℓ, the largest value of $m(D|D_0, a_{0,\ell})$ is obtained at ℓ^* when $\ell^* > 1$, $\log\left[\hat{R}_{m,\ell^*,\ell^*-1,\mathrm{IS}}\right] > 0$, and $\log\left[\hat{R}_{m,\ell^*+1,\ell^*,\mathrm{IS}}\right] < 0$ or at $\ell^* = 1$ when $\log\left[\hat{R}_{m,2,1,\mathrm{IS}}\right] < 0$.

For the HM method, directly using $\left\{\boldsymbol{\theta}_{1t}^{(\ell)}, t = 1, \ldots, T_1\right\}$, an estimate of $m(D|D_0, a_{0,\ell})$ is given by

$$\hat{m}_{\mathrm{HM}}(D|D_0, a_{0,\ell}) = \frac{1}{\frac{1}{T_1}\sum_{t=1}^{T_1} L\left(\boldsymbol{\theta}_{1t}^{(\ell)}|D\right)^{-1}}$$

for $\ell = 0, 1, \ldots, 10$. For the IDR, PWK, and LoRaD methods, we need to use all 21 MCMC samples to estimate $c_1(a_{0,\ell})$ for $\ell = 0, 1, \ldots, 10$ and $c_0(a_{0,\ell})$ for $\ell = 1, \ldots, 10$. Note that $c_0(a_0 = 0) = 1$ as $\pi_0(\boldsymbol{\theta})$ is a normalized proper initial prior. For these methods, we use the first 50% of the MCMC sample as the training set and the remaining 50% of the MCMC sample to estimate these normalization constants.

We summarize all results in Table 4, including the eMCSEs computed by (15). Some comments are warranted. First, rather than estimating the marginal likelihood in each a_0, the IS method provides Bayes factor estimates relative to the preceding term in the sequence of a_0. As a result, it greatly benefits from the similarity between serial kernels and yields small eMCSEs. Second, when information borrowing is required, IS exhibits monotone results with respect to the degree of information borrowing. Third, although IS cannot obviate the no-borrowing case, this option has already been excluded by the other competing methods considered here. Fourth, in terms of eMCSE in marginal likelihood estimation, PWK and LoRaD provide comparable results, both producing smaller eMCSEs than HM and IDR. Finally, IS, PWK, and LoRaD

TABLE 4 | Log Bayes factor ($\log R_{m,\ell,\ell-1}$) or log marginal likelihood ($\log m(D|D_0, a_{0,\ell})$) estimates with eMCSEs (in parentheses) for different methods.

		$\log \hat{R}_{m,\ell,\ell-1}$	$\log \hat{m}(D\mid D_0, a_{0,\ell})$			
3–3 (lr)4–7 ℓ	$a_{0,\ell}$	**IS**	**HM**	**IDR**	**PWK**	**LoRaD$_{20}$**
0	0.0	—	−522.937 (0.484)	−556.958 (0.668)	−552.667 (0.031)	−552.553 (0.030)
1	0.1	—	−521.429 (0.405)	−520.653 (1.167)	−523.660 (0.067)	−523.552 (0.052)
2	0.2	1.641 (0.026)	−522.298 (0.489)	−519.555 (0.867)	−522.131 (0.058)	−522.332 (0.067)
3	0.3	0.623 (0.011)	−521.365 (0.278)	−519.918 (0.618)	−521.483 (0.063)	−521.795 (0.044)
4	0.4	0.247 (0.006)	−522.869 (0.482)	−515.430 (2.113)	−521.356 (0.061)	−521.690 (0.060)
5	0.5	0.099 (0.005)	−522.180 (0.317)	−517.717 (0.912)	−521.172 (0.101)	−521.966 (0.042)
6	0.6	0.029 (0.006)	−523.850 (0.525)	−520.102 (0.722)	−521.170 (0.058)	−521.655 (0.054)
7	0.7	−0.016 (0.005)	−522.122 (0.259)	−519.566 (0.591)	−521.396 (0.071)	−522.499 (0.049)
8	0.8	−0.045 (0.005)	−524.930 (0.573)	−519.766 (1.253)	−521.563 (0.109)	−522.126 (0.058)
9	0.9	−0.053 (0.004)	−524.355 (0.411)	−523.188 (1.193)	−521.643 (0.062)	−521.887 (0.047)
10	1.0	−0.052 (0.004)	−523.905 (0.168)	−520.955 (0.770)	−521.761 (0.075)	−522.026 (0.053)

Note: HM, IDR, PWK, and LoRaD estimate the log marginal likelihood, while IS reports log Bayes factor estimates. The radius $r = \sqrt{\chi^2_{7,0.95}}$ is the square-root of the 95th percentile of the Chi-square distribution with $p = \dim(\theta) = 7$ degrees of freedom, which is derived by computing the norm of p independent standard normal distributions. PWK additionally employs $K = 100$ spherical shells.

identify $a_0 = 0.6$ as the optimal degree of information borrowing, while HM and IDR favor $a_0 = 0.3$ and 0.4, respectively, although their corresponding estimates exhibit substantial differences. All computations in the analysis of the ECOG data were performed on a laptop equipped with a 12th Gen Intel(R) Core(TM) i7 processor and 16 GB RAM, running Windows 11. For each method, we record the total computing time required to evaluate the marginal likelihood in all values of $a_0 \in \{0, 0.1, \ldots, 1\}$. The reported times include the evaluation of both components $c_1(a_0)$ and $c_0(a_0)$, as well as the calculation of the marginal likelihood $\log m = \log c_1 - \log c_0$ and its associated Monte Carlo standard error obtained using the OBS method. Overall, the IDR estimator required a total of 1087.69 s, the PWK estimator required 755.78 s, and the LoRaD estimator required 499.92 s.

For LoRaD, we define the cutoff $r_{\max}$ via a coverage fraction ϕ as in Wang et al. (2023) and treat ϕ as a tuning parameter. In our numerical experiments, coverage fractions such as $\phi = 0.10$ and $\phi = 0.20$ produced very similar log marginal likelihood estimates, with differences well within the estimated MCSE, indicating that the method is not overly sensitive to moderate changes in ϕ. To provide a simple and reproducible default for practitioners, we therefore report results using $\phi = 0.20$ (i.e., using the 20% smallest radial distances to define $r_{\max}$), while noting that in more demanding applications ϕ can be tuned using the OBS MCSE strategy of Wang et al. (2023).

In the absence of a universal rule for choosing a reference or cutoff, for the TLC data, we report IDR and PWK estimates using coverage fractions of 10% and 20% for comparison. For the ECOG data, we follow Wang et al. (2018) and set the cutoff using the reference value $\sqrt{\chi^2_{7,0.95}}$. By analogy with the LoRaD implementation of Wang et al. (2023), practitioners can treat the coverage fraction or cutoff as a tuning parameter and prefer choices for which the log marginal likelihood estimates are stable and the associated MCSE is small.

4 | Discussion

We review Monte Carlo-based estimators that can be implemented using only one or both of two MCMC samples (drawn from the posterior and prior distribution), under the additional assumption that the posterior and prior kernels can be evaluated numerically. Through two illustrative examples first, the Treatment of Lead-Exposed Children (TLC) trial under the ordered-variance modeling framework, and second, the benchmark dose (BMD) dataset under the power prior setting we empirically compare different estimators. In the TLC case study, we model the data using a multivariate normal distribution with or without an ordered-variance constraint. IDR, PWK, and LoRaD, produce similar estimates and favor the ordered-variance model ($\mathcal{M}_1$), while HM not only presents higher variability but also selects the unconstrained model ($\mathcal{M}_0$). These findings underscore the instability of HM in settings where priors are relatively flat compared with posteriors, and highlight the greater reliability of IDR, PWK, and LoRaD. In the BMD power prior analysis, our focus shifts to quantifying the degree of information borrowing. Beyond the four main estimators (HM, IDR, PWK, and LoRaD), we also develop a sequential strategy based on IS, which leverages similarity between neighboring kernels rather than treating and solving each a_0 individually. This serial property yields substantial efficiency gains, particularly when information borrowing is needed. Consistent with PWK and LoRaD, IS identifies $a_0 = 0.6$ as the optimal level of borrowing, while HM and IDR selected smaller values ($a_0 = 0.3$ and $a_0 = 0.4$, respectively), but with greater variability and less agreement in terms of estimated values.

Variational inference (VI) has also been used for approximate Bayesian model selection. One line of work uses the evidence lower bound (ELBO) or variational free energy as a surrogate model score, ranking models by this quantity (see, e.g., Chérief-Abdellatif and Alquier (2018), Zhang and Yang (2024)). A second line combines VI with importance sampling, using the fitted variational posterior as a proposal distribution to obtain fast but approximate estimates of marginal likelihoods (Gunapati et al. 2022). Although such VI-based procedures can be orders of magnitude faster than MCMC or nested sampling, they introduce additional approximation bias, and their accuracy depends on how well the chosen variational family captures the true posterior. In contrast, the estimators reviewed here assume access to posterior samples and target the true marginal likelihood and are therefore best viewed as complementary to, rather than competing with, variational inference based methods.

Mixture model selection presents additional challenges due to multimodality. As long as the mode with the highest posterior density is identified, IDR, PWK, and LoRaD will be generally applicable given that a good posterior MCMC sample from the mixture model is readily available. Wang et al. (2018) empirically demonstrated that PWK works well through a simulation study involving a mixture model.

In summary, HM generally does not work well since the posterior usually has a lighter tail than the prior as noted in Table 1. HM is recommended when an informative prior is specified. IDR can be efficient if we can find the mode, and recent work shows that its efficiency deteriorates as the dimension increases (Oaks et al. 2019; Wang et al. 2020). IS is attractive when a good proposal distribution can be constructed. Similarly, generalized harmonic mean (GHM) type estimators rely on carefully chosen proposal distributions. In contrast, both the PWK estimator and LoRaD are particularly appealing when good proposal distributions are not readily available. In particular, PWK is black-boxable and LoRaD can be efficient if local normality is achieved.

Author Contributions

Aolan Li: formal analysis (equal), investigation (equal), software (equal), visualization (equal), writing – original draft (equal). **Pang-Yu Liu:** formal analysis (equal), investigation (equal), software (equal), writing – original draft (supporting). **Yu-Bo Wang:** formal analysis (equal), investigation (equal), methodology (equal), software (equal), supervision (supporting), writing – original draft (equal). **Analisa Milkey:** investigation (equal), methodology (supporting), writing – original draft (supporting), writing – review and editing (supporting). **Paul O. Lewis:** conceptualization (equal), resources (equal), supervision (equal), writing – review and editing (supporting). **Ming-Hui Chen:** conceptualization (equal), formal analysis (equal), investigation (equal), methodology (equal), resources (equal), supervision (equal), writing – original draft (equal), writing – review and editing (equal).

Acknowledgments

We sincerely thank the Editor and the anonymous reviewers for their careful reading of our manuscript and for their insightful comments and constructive suggestions, which have greatly improved the quality and clarity of this work.

Funding

A.M. was supported by the National Science Foundation Graduate Research Fellowship Program (Grant No. DGE 2136520). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Related WIREs Articles

A New Monte Carlo Method for Estimating Marginal Likelihoods.

LoRaD: Marginal Likelihood Estimation with Haste (But No Waste).

References

Barido-Sottani, J., O. Schwery, R. Warnock, C. Zhang, and A. Wright. 2024. "Practical Guidelines for Bayesian Phylogenetic Inference Using Markov Chain Monte Carlo (MCMC)." *Open Research Europe* 3: 204.

Carpenter, B., A. Gelman, M. D. Hoffman, et al. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76, no. 1: 1–32.

Chen, M.-H., Z. Guan, M. Lin, and M. Sun. 2025. "Power Priors for Leveraging Historical Data: Looking Back and Looking Forward." *Journal of Data Science* 23, no. 1: 1–30.

Chen, M.-H., J. G. Ibrahim, and C. Yiannoutsos. 1999. "Prior Elicitation, Variable Selection and Bayesian Computation for Logistic Regression Models." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 61: 223–242.

Chérief-Abdellatif, B.-E., and P. Alquier. 2018. "Consistency of Variational Bayes Inference for Estimation and Model Selection in Mixtures." *Electronic Journal of Statistics* 12, no. 2: 2995–3035.

Fourment, M., A. F. Magee, C. Whidden, F. A. M. Bilge IV, and V. N. Minin. 2020. "19 Dubious Ways to Compute the Marginal Likelihood of a Phylogenetic Tree Topology." *Systematic Biology* 69, no. 2: 209–220.

Gronau, Q. F., H. Singmann, and E.-J. Wagenmakers. 2020. "Bridgesampling: An R Package for Estimating Normalizing Constants." *Journal of Statistical Software* 92, no. 10: 1–29.

Gunapati, G., A. Jain, P. Srijith, and S. Desai. 2022. "Variational Inference as an Alternative to MCMC for Parameter Estimation and Model Selection." *Publications of the Astronomical Society of Australia* 39: e001.

Ibrahim, J. G., and M.-H. Chen. 2000. "Power Prior Distributions for Regression Models." *Statistical Science* 15, no. 1: 46–60.

Ibrahim, J. G., M.-H. Chen, Y. Gwon, and F. Chen. 2015. "The Power Prior: Theory and Applications." *Statistics in Medicine* 34, no. 28: 3724–3749.

Lewandowski, D., D. Kurowicka, and H. Joe. 2009. "Generating Random Correlation Matrices Based on Vines and Extended Onion Method." *Journal of Multivariate Analysis* 100, no. 9: 1989–2001.

Liu, Y., G. Hu, L. Cao, X. Wang, and M.-H. Chen. 2019. "A Comparison of Monte Carlo Methods for Computing Marginal Likelihoods of Item Response Theory Models." *Journal of the Korean Statistical Society* 48, no. 4: 503–512.

Meng, X.-L., and W. H. Wong. 1996. "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration." *Statistica Sinica* 6: 831–860.

Metodiev, M., M. Perrot-Dockès, S. Ouadah, N. J. Irons, P. Latouche, and A. E. Raftery. 2025. "Easily Computed Marginal Likelihoods From Posterior Simulation Using the THAMES Estimator." *Bayesian Analysis* 20, no. 3: 1003–1030.

Milkey, A., E. Korte, and P. O. Lewis. 2023. "LoRaD: Lowest Radial Distance Method of Marginal Likelihood Estimation." R package version 0.0.1.0. https://CRAN.R-project.org/package=lorad.

Newton, M. A., and A. E. Raftery. 1994. "Approximate Bayesian Inference With the Weighted Likelihood Bootstrap." *Journal of the Royal Statistical Society: Series B: Methodological* 56, no. 1: 3–48.

Oaks, J., K. Cobb, V. Minin, and A. Leaché. 2019. "Marginal Likelihoods in Phylogenetics: A Review of Methods and Applications." *Systematic Biology* 68, no. 5: 681–697.

Petris, G., and L. Tardella. 2003. "A Geometric Approach to Transdimensional Markov Chain Monte Carlo." *Canadian Journal of Statistics* 31, no. 4: 469–482.

Rogan, W., R. Bornschein, J. Chisolm, et al. 2000. "Safety and Efficacy of Succimer in Toddlers With Blood Lead Levels of 20–44 μ g/dL." *Pediatric Research* 48, no. 5: 593–599.

Schmeiser, B. W., T. N. Avramidis, and S. Hashem. 1990. "Overlapping Batch Statistics." In *Proceedings of the 22nd Conference on Winter Simulation*, 395–398. IEEE Press. https://www.personal.soton.ac.uk/aa1w07/wsc90b.pdf.

Shen, Y., M. A. Psioda, and J. G. Ibrahim. 2021. "BayesPPD: An R Package for Bayesian Sample Size Determination Using the Power and Normalized Power Prior for Generalized Linear Models." arXiv Preprint arXiv:2112.14616.

Wang, Y.-B., M.-H. Chen, L. Kuo, and P. O. Lewis. 2018. "A New Monte Carlo Method for Estimating Marginal Likelihoods." *Bayesian Analysis* 13, no. 2: 311–333.

Wang, Y.-B., M.-H. Chen, W. Shi, P. Lewis, and L. Kuo. 2020. "Inflated Density Ratio and Its Variation and Generalization for Computing Marginal Likelihoods." *Journal of the Korean Statistical Society* 49, no. 1: 244–263.

Wang, Y.-B., A. Milkey, A. Li, M.-H. Chen, L. Kuo, and P. O. Lewis. 2023. "LoRaD: Marginal Likelihood Estimation With Haste (But No Waste)." *Systematic Biology* 72, no. 3: 639–648.

Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. 2011. "Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection." *Systematic Biology* 60, no. 2: 150–160.

Zhang, Y., and Y. Yang. 2024. "Bayesian Model Selection via Mean-Field Variational Approximation." *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 86, no. 3: 742–770.