Psychological Methods

Bayesian Evaluation for Latent Variable Models: A Tutorial on Computing Information Criteria and Bayes Factors With the R Package bleval

Xiaohui Luo, Jieyuan Dong, Hongyun Liu, Yang Liu, and Edgar C. Merkle

Online First Publication, June 29, 2026. https://dx.doi.org/10.1037/met0000852

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

https://doi.org/10.1037/met0000852

Bayesian Evaluation for Latent Variable Models: A Tutorial on Computing Information Criteria and Bayes Factors With the R Package bleval

Xiaohui Luo[1], Jieyuan Dong[1], Hongyun Liu[1, 2], Yang Liu[3], and Edgar C. Merkle[4]

[1] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Faculty of Psychology, Beijing Normal University

[2] Research Center for Capacity Building in Educational Assessment and Evaluation (Beijing Higher Education Innovation Center for Philosophy and Social Sciences), Beijing Normal University

[3] Department of Human Development and Quantitative Methodology, University of Maryland

[4] Department of Psychological Sciences, University of Missouri

Abstract

Model evaluation is crucial in psychological methodology, particularly in the context of Bayesian latent variable models. Bayesian evaluation methods require integrating out latent variables to compute marginal likelihoods for information criteria and integrating both latent variables and model parameters to compute fully marginal likelihoods for Bayes factors. These processes can be computationally demanding, as likelihoods in many latent variable models are intractable. Moreover, the role of latent variables in model evaluation has been insufficiently addressed in applied research, likely due to a lack of practical guidance and user-friendly software. This tutorial fills these gaps by (a) offering step-by-step instructions for approximating marginal likelihoods using an efficient numerical method (i.e., adaptive Gauss–Hermite quadrature), (b) developing an R package, bleval, designed specifically for computing information criteria and fully marginal likelihoods in Bayesian latent variable models, and (c) demonstrating the application of this package to various empirical scenarios, including structural equation models, item response theory models, and multilevel models. The empirical examples also investigate practical issues such as the sensitivity of model evaluation results to the number of quadrature nodes, the performance of the numerical quadrature method in high-dimensional latent variable models, and the handling of latent class variables in Bayesian mixture models.

Translational Abstract

Latent variable models, including structural equation models, item response theory (IRT) models, and multilevel models, are widely used in psychological research. Within the Bayesian framework, researchers often compute information criteria and Bayes factors to evaluate and compare competing models. However, popular Bayesian software packages compute information criteria based on *conditional* likelihoods by default. These conditional metrics treat latent variables as fixed effects specific to the observed sample, whereas information criteria based on *marginal* likelihoods (where latent variables are integrated out) are generally considered more appropriate for generalizing inferences to a broader population. In addition, computing fully marginal likelihoods for Bayes factors via bridge sampling can be challenging. Direct approximation approaches (e.g., defaults in the brms package) often fail to converge due to the difficulty of integrating over high-dimensional latent variables and model parameters simultaneously. To address these challenges, this tutorial presents a practical workflow that employs adaptive Gauss–Hermite quadrature to approximate marginal likelihoods and introduces an R package, bleval, designed to facilitate the computation of marginal likelihood-based information criteria and fully marginal likelihoods. By numerically integrating out latent variables, this approach makes the approximation of fully marginal likelihoods for Bayes factors more feasible. We demonstrate this workflow using four examples: a Gaussian linear mixed model, a latent moderation structural equation model, a five-dimensional IRT model, and a mixture IRT model. Designed as a downstream toolkit compatible with common software (e.g., JAGS, Stan, and brms), bleval offers applied researchers an accessible pathway to implement these rigorous evaluation methods.

Keywords: Bayesian evaluation, latent variable models, Bayesian information criteria, marginal likelihood, Bayes factor

Supplemental materials: https://doi.org/10.1037/met0000852.supp

Andreas M. Brandmaier served as action editor.

Hongyun Liu https://orcid.org/0000-0002-3472-9102

This study was not preregistered. The data and code have been uploaded to the R package bleval hosted on GitHub (https://github.com/luoxh3/bleval). A preliminary version of this work was presented orally at the

continued

Model evaluation plays a crucial role in psychological research. Within the Bayesian framework, evaluating the goodness of fit of models is the basis for model comparison, selection, and averaging (Hoeting et al., 1999; Vehtari & Ojanen, 2012; Wasserman, 2000). A common approach to evaluating a Bayesian model is through estimating out-of-sample predictive accuracy (Gelman et al., 2014). Specifically, several information criteria, including the deviance information criterion (DIC; Spiegelhalter et al., 2002), the Watanabe–Akaike information criterion (WAIC; Watanabe, 2010), and the leave-one-out (cross-validation) information criterion (LOOIC; Vehtari et al., 2017), are commonly used to approximate expected predictive accuracy using available data. These criteria can be computed based on model likelihoods and are readily accessible in popular Bayesian software such as BUGS (Lunn et al., 2013; Spiegelhalter et al., 1996), JAGS (Plummer, 2003), and Stan (Stan Development Team, 2023). Another traditional approach for evaluating and comparing models is computing Bayes factors, defined as the ratio of the fully marginal likelihoods of two models (Jeffreys, 1961; Kass & Raftery, 1995; see later for a formal definition of fully marginal likelihood). Bayes factors can be used to assess the relative predictive accuracy of one model over another (i.e., how well one model predicts the data compared with another), or to quantify the relative evidence in favor of one model over another (Rouder & Morey, 2019). This measure can also be obtained using Bayesian software and packages (Garofalo et al., 2024; Hoijtink et al., 2019; Veenman et al., 2024).

However, when latent variable models are the target of evaluation, special attention is warranted regarding the treatment of latent variables (e.g., random vs. fixed effects) and the associated challenges in likelihood computation (Li et al., 2016; Merkle et al., 2019). Common latent variable models used in psychological research include structural equation models (SEMs), item response theory (IRT) models, and multilevel models (MLMs). In SEMs and IRT models, latent variables typically represent individuals' latent characteristics, whereas in MLMs, they represent random effects (e.g., random intercepts and slopes) of clusters (e.g., schools in data with students nested within schools) or persons (in data with repeated measures nested within individuals). Mixture modeling adds latent class variables that define discrete groups (i.e., latent class membership) of units. In the following discussion, the term "cluster" will refer to the entities represented by latent variables, whereas "unit" will denote the directly observed entities grouped by these clusters.

Marginal Likelihood-Based Information Criteria and Their Computational Challenges

For Bayesian models with latent variables, an easy implementation of information criteria is to treat the latent variables as model parameters and directly sample them from the MCMC (Markov chain Monte Carlo) posterior draws. This approach corresponds to the data augmentation technique (Tanner & Wong, 1987), where latent variables are introduced into the sampling process to facilitate posterior computation. When data augmentation is used for posterior sampling, the default information criteria obtained from generic MCMC software (e.g., BUGS, JAGS, and Stan) are typically based on *conditional likelihoods*, that is, likelihoods conditional on the latent variables. In contrast, a more computationally demanding implementation of information criteria is based on *marginal likelihoods*, which treats latent variables separately from model parameters and integrates out the latent variables.

The distinction between conditional and marginal likelihood-based information criteria has been discussed in previous research (Gelman et al., 2014; Merkle et al., 2019; Millar, 2018; Spiegelhalter et al., 2002; Zhang et al., 2019). The choice between conditional and marginal likelihood can be framed in terms of whether the latent variables are treated as fixed or random effects when predicting new data. Specifically, conditional likelihood treats the latent variables as fixed effects (i.e., specific to the observed clusters), which is appropriate for making inferences within the same observed clusters. In contrast, marginal likelihood views the latent variables as random effects (i.e., assumed to be drawn from a broader population), which can be used when generalizing inference beyond the observed sample is desired. Given that psychological modeling typically aims for generalization beyond specific observed clusters (i.e., treating latent variables as random effects), information criteria based on marginal likelihoods are generally more appropriate and recommended (Merkle et al., 2019; Millar, 2018; Zhang et al., 2019).

When the model is simple and its marginal likelihood is tractable (e.g., mean and covariance structure models), data augmentation is not necessary for Bayesian estimation, and information criteria can be straightforwardly computed (as implemented in the blavaan package; Merkle et al., 2021). In more complex latent variable models, however, it is typically impossible to express the marginal likelihood in closed form. Specifically, for latent variable models involving nonnormal data types (e.g., binary or ordinal responses in IRT models), nonlinear structures (e.g., latent variable interactions) or mixture components, data augmentation with

International Meeting of the Psychometric Society in July 2025. Xiaohui Luo, Jieyuan Dong, and Hongyun Liu were supported by the National Natural Science Foundation of China (Grants 32471145 and 32300938), and the work of Hongyun Liu was also supported by the Beijing Higher Education Innovation Center for Philosophy and Social Sciences. Xiaohui Luo and Jieyuan Dong contributed equally to this article, and both should be considered first authors.

Xiaohui Luo served as lead for visualization and writing–original draft and contributed equally to conceptualization, formal analysis, and methodology. Jieyuan Dong contributed equally to formal analysis and methodology and served in a supporting role for writing–original draft. Hongyun Liu served as lead for funding acquisition and contributed equally to conceptualization, methodology, supervision, and writing–review and editing. Yang Liu contributed equally to conceptualization, methodology, supervision, and writing–review and editing and served in a supporting role for formal analysis. Edgar C. Merkle contributed equally to conceptualization, methodology, supervision, and writing–review and editing and served in a supporting role for formal analysis.

Correspondence concerning this article should be addressed to Hongyun Liu, Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Faculty of Psychology, Beijing Normal University, No. 19, Xin Jie Kou Wai Street, Hai Dian District, Beijing 100875, PR China, or to Yang Liu, Department of Human Development and Quantitative Methodology, University of Maryland, 3304R Benjamin Building, 3942 Campus Drive, College Park, MD 20742, United States. Email: hyliu@bnu.edu.cn or yliu87@umd.edu

MCMC is typically required for parameter estimation. In these cases, numerical integration methods are needed to integrate out latent variables and approximate the marginal likelihoods used to compute information criteria.

Bayes Factors and Their Computational Challenges

Compared with information criteria, Bayes factors are computationally more demanding, as computing them requires integrating out both latent variables and model parameters. When Bayes factors are computed based on marginal likelihoods, only the model parameters need to be integrated out, as the latent variables have already been marginalized. In contrast, computing Bayes factors based on conditional likelihoods requires integrating out both the latent variables and the model parameters, making the computation more demanding. To minimize ambiguity in terminology, we use the term "marginal likelihoods" to refer to the likelihoods integrating out latent variables from the product of a conditional likelihood and a latent variable density, and the term "fully marginal likelihoods" for those that integrate out both latent variables and model parameters from the product of a conditional likelihood, a latent variable density, and a prior density. In this tutorial, we compute information criteria using marginal likelihoods. For Bayes factor computation, we obtain fully marginal likelihoods by further integrating out the model parameters from the marginal likelihoods.

Several methods have been proposed to approximate fully marginal likelihoods for competing models to compute Bayes factors. In simple cases where researchers are interested in comparing nested models that differ in only a single parameter, the Savage–Dickey density ratio (Dickey & Lientz, 1970) can be used as a feasible approximation of Bayes factors. In more general cases, Monte Carlo sampling methods are required to approximate the fully marginal likelihood, with bridge sampling (Bennett, 1976; Meng & Wong, 1996) being a common choice due to its advantages over other sampling-based methods (Gronau et al., 2017) and its relatively easy implementation in the R package bridgesampling (Gronau et al., 2020). The key idea of bridge sampling is to introduce an optimal bridge function connecting the posterior distribution to a proposal distribution, which yields accurate and robust estimates even when the proposal does not match the posterior tails.

Although there are methods and tools to approximate fully marginal likelihoods for Bayesian models, their applications to latent variable models are often challenging, as the total number of dimensions to be integrated out—including both latent variables and model parameters—can easily become overwhelming and unmanageable. For instance, consider an SEM with three factors (i.e., three latent variables) and a sample size of 300 (i.e., 300 clusters). This results in 900 quantities from the latent variables, plus model parameters, to be integrated out when computing fully marginal likelihoods in Bayes factors. In such cases, it may be helpful to compute the marginal likelihoods before using bridge sampling to approximate the fully marginal likelihood. The marginal likelihood computation can be accomplished using numerical integration, as discussed in the computation of information criteria. This adjustment, specifically for evaluating latent variable models, can effectively reduce the number of quantities (i.e., the number of dimensions that must be integrated over) involved in bridge sampling, making it more practically applicable.

The Current Tutorial

Taken together, the aforementioned discussion highlights the importance of distinguishing latent variables from model parameters and integrating out latent variables when evaluating the model's likelihood. For information criteria, computing marginal likelihoods is more appropriate for predictive objectives in most psychological studies and aligns better with traditional likelihood-based methods (Merkle et al., 2019). For Bayes factors, computing marginal likelihoods before using bridge sampling can effectively reduce the computation burden of fully marginal likelihoods.

Despite the critical role of latent variables in model evaluation, they have received insufficient attention and have not been adequately addressed in applied research, which we believe may be due to a lack of relevant guidance and user-friendly software from an applied perspective. Although there are some general introductions and tutorials on Bayesian model evaluation using information criteria (Gelman et al., 2014; Vehtari et al., 2017) and Bayes factors (Garofalo et al., 2024; Hoijtink et al., 2019), practical tutorials on computing information criteria and Bayes factors specifically focused on Bayesian latent variable models estimated via data augmentation are still lacking.

Furthermore, although existing packages offer some relevant functionality, important limitations remain for the evaluation of data-augmented models. For example, the blavaan package (Merkle et al., 2021), commonly used for fitting Bayesian SEMs, can compute marginal likelihood-based information criteria, but its default marginal approach is limited to multivariate normal SEMs and cannot readily handle more general models (e.g., those with non-normal data, latent interactions, or mixture components). It also does not directly support the computation of fully marginal likelihoods for Bayes factors. In addition, the brms package (Bürkner, 2017), which is a powerful tool for MLMs and IRT models, computes information criteria based on the *conditional* likelihood (an issue demonstrated later in the Comparison and Integration With Existing Packages section). Moreover, when brms is used to compute fully marginal likelihoods via bridge sampling, it treats latent variables as model parameters and passes both sets of quantities to the sampler. This requires integration over a very large number of dimensions (often hundreds or thousands), which frequently leads to convergence difficulties.

Therefore, the present work aims to fill these gaps by (a) offering practical guidance on approximating marginal likelihoods using efficient numerical methods (i.e., adaptive quadrature) when the model parameters are estimated by data augmentation, (b) developing an R package bleval to compute information criteria and fully marginal likelihoods for these models, and (c) demonstrating application of the package in various empirical scenarios using different latent variable models (i.e., SEMs, IRT models, and MLMs).[1]

Note that there are other model evaluation measures for Bayesian latent variable models, such as approximate fit indices adapted for Bayesian SEM (e.g., the Bayesian root mean square error of approximation, BRMSEA; Garnier-Villarreal & Jorgensen, 2020) and calibrated posterior predictive p values (van Kollenburg et al., 2017). The current tutorial does not examine these

[1] The R package bleval can be downloaded from GitHub at https://github.com/luoxh3/bleval.

measures but focuses on information criteria and Bayes factors. Readers interested in these alternative evaluation measures are referred to Depaoli et al. (2024), Garnier-Villarreal and Jorgensen (2025), Levy (2011), and Winter and Depaoli (2022).

The article is organized as follows. We begin by introducing the bleval package, presenting its workflow and describing the numerical integration methods on which it relies. We also clarify the distinct role of bleval relative to other popular packages (i.e., blavaan and brms) and demonstrate how it can be integrated with brms to enhance practical utility. Next, we present a comprehensive worked example using a Gaussian linear mixed model (a model with a tractable marginal likelihood) to explain the core concepts (i.e., marginal likelihood, information criteria, fully marginal likelihoods, and Bayes factors) and to demonstrate step by step how the corresponding bleval functions are applied. This example also serves to validate the numerical approximations against an analytical solution. Following this foundation, we further demonstrate the broader applications of the developed package for models with intractable marginal likelihoods using three empirical examples. Example 1 shows the evaluation of latent moderation SEMs, which are commonly analyzed nonlinear models in SEMs. In this example, we also provide a preliminary demonstration of the sensitivity of model evaluation results to different numbers of quadrature nodes per latent variable. Example 2 presents a five-dimensional IRT model, which is used to explore how the numerical quadrature method in the developed package performs for models with a relatively large number of latent variables. Example 3 uses a mixture IRT model to illustrate how latent class variables in Bayesian mixture models can be effectively addressed using the developed package (presented in Supplemental Material S1). The tutorial concludes with a discussion of the contributions, practical recommendations and limitations.

The bleval Package: A Downstream Toolkit for Evaluating Bayesian Latent Variable Models

In this section, we introduce the bleval package, which provides a set of functions for computing marginal-likelihood-based information criteria and fully marginal likelihoods (for Bayes factors) from posterior samples of Bayesian latent variable models. We first describe the overall workflow of bleval, focusing on the three functions that users call directly. We then turn to the numerical integration methods that underlie the package, along with the two internal functions that implement them. Finally, we compare bleval with two widely used packages (i.e., blavaan and brms) to clarify its niche and demonstrate how it can be integrated with the brms package.

Workflow of the bleval Package

The bleval package is designed as a "downstream" tool. It does not perform the initial Bayesian estimation; instead, it takes the MCMC posterior samples generated by other software (e.g., JAGS and Stan) and performs the necessary integration to compute model evaluation metrics. The core idea is to take posterior samples of the model parameters and the latent variables to compute the marginal likelihood for each cluster using adaptive Gauss–Hermite quadrature (described in detail in the next subsection). From these marginal likelihoods, the package then calculates information criteria and approximates the fully marginal likelihood needed for Bayes factors.

Figure 1 outlines the workflow of the bleval package. The package includes five functions. Two of these (i.e., `get_quadrature` and `log_marglik_i`) are internal helpers that handle the numerical integration and will be discussed in the next subsection. Here we focus on the three user-facing functions that are the primary tools for model evaluation. To compute information criteria (i.e., DIC, WAIC, and LOOIC), users need to provide some basic arguments and call the `log_marglik` function to obtain the pointwise log marginal likelihood and the log marginal likelihood based on point estimates of model parameters. With these likelihood approximations, users can compute information criteria using the `calc_IC` function. To approximate the log fully marginal likelihood for Bayes factor computation, users can provide additional inputs and use the `log_fmarglik` function.

The `calc_IC` and `log_fmarglik` functions leverage powerful functions from existing packages (i.e., the `waic` and `loo` functions in the loo package; Vehtari et al., 2017, and the `bridge_sampler` function in the bridgesampling package; Gronau et al., 2020). Detailed introduction and instructions are available in the bleval package hosted on GitHub (https://github.com/luoxh3/bleval).

Numerical Integration Methods and Internal Functions

To provide background for how `log_marglik` computes the marginal likelihood, we now review the numerical integration methods that make this possible, starting from basic quadrature principles and building up to the adaptive approach used in bleval. We then describe the two internal functions that implement this method.

From a numerical perspective, the marginal likelihood can be estimated using numerical quadrature, which approximates deterministic integrals through a finite weighted sum of evaluations of the integrand (i.e., the function to be integrated):

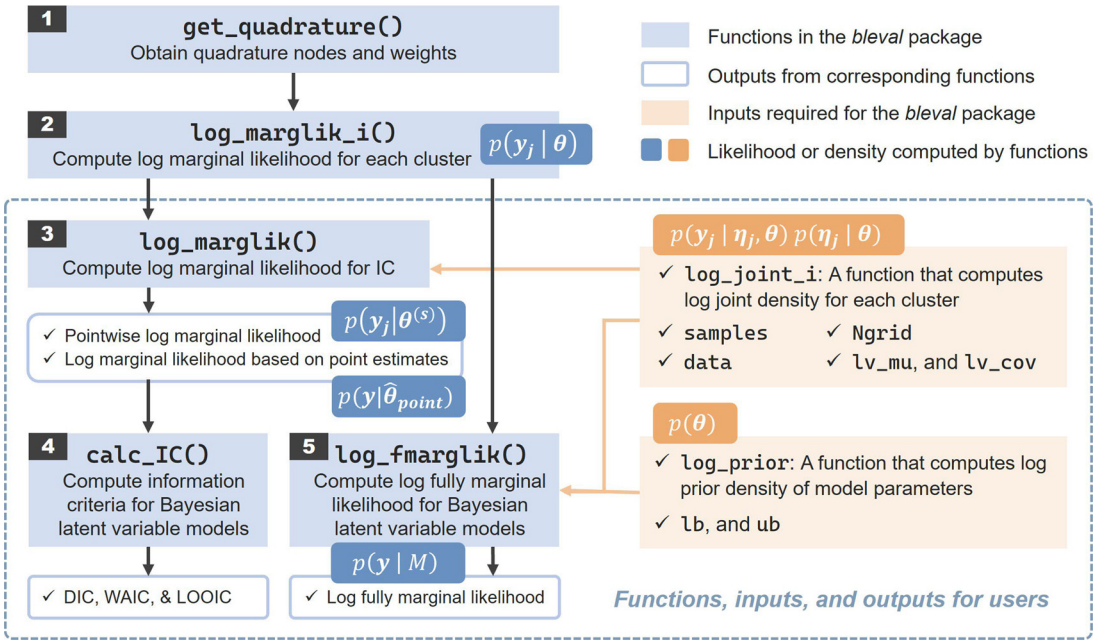
$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) \approx \sum_{q=1}^{Q} w_q \, p(\boldsymbol{y} \mid \boldsymbol{\eta} = \boldsymbol{z_q}, \boldsymbol{\theta}). \tag{1}$$

Here, $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the marginal likelihood, $p(\boldsymbol{y} \mid \boldsymbol{\eta}, \boldsymbol{\theta})$ is the conditional likelihood, Q is the total number of quadrature points, and w_q and $\boldsymbol{z_q}$ are the qth quadrature weight and node ($q = 1, 2, \ldots, Q$), respectively.

One of the most commonly used numerical quadrature approaches is Gaussian quadrature (Davis & Polonsky, 1972). When integrating over a normal random variable, a specific variant known as Gauss–Hermite quadrature is typically employed, which provides accurate approximations with a relatively small number of quadrature points. Integrating over multivariate normal variables typically relies on an outer-product Gauss–Hermite quadrature. However, the Gauss–Hermite quadrature employs a fixed quadrature rule, in which the nodes (i.e., the locations where the function is evaluated) and weights (i.e., the multipliers for each function evaluation) are determined given the number of quadrature points. When the integrand is concentrated in a region far from the origin, or when its shape differs markedly from a standard normal density, this fixed rule can become inefficient or inaccurate (Cagnone & Monari, 2013; Rabe-Hesketh et al., 2005).

Figure 1

Workflow for Using the bleval Package



Note. The first two functions, located outside the dashed box, are used internally within the package to compute the marginal likelihood for each cluster. The other three functions, highlighted within the dashed box, are user-facing functions designed for computing model evaluation measures. DIC = deviance information criterion; WAIC = Watanabe–Akaike information criterion; LOOIC = leave-one-out (cross-validation) information criterion. See the online article for the color version of this figure.

To overcome this limitation, adaptive Gauss–Hermite quadrature has been developed (Naylor & Smith, 1982) and widely used in Bayesian analyses (Bilodeau et al., 2024). It shifts and scales quadrature nodes and weights based on the specific integrand, thereby placing greater weight on evaluations near the peak of the integrand (Rabe-Hesketh et al., 2002, 2005). Specifically, the nodes and weights can be adapted using (a) the mode and curvature (i.e., the inverse of the information matrix evaluated at the mode) of the integrand (Liu & Pierce, 1994; Pinheiro & Bates, 1995), or (b) the means and the covariance matrices of the integrand (Merkle et al., 2019; Rabe-Hesketh et al., 2002, 2005). In this tutorial, we assume that posterior samples of latent variables are available. Therefore, we adopted the second method (i.e., adapting nodes and weights using their posterior mean and covariance), as it is more straightforward and computationally efficient in this context.

The original version of the second method computes the means and the covariance matrices of the *conditional* posterior distribution of latent variables given each posterior sample of model parameters (Rabe-Hesketh et al., 2002, 2005). Recently, a modified version of this method has been proposed and applied to Bayesian latent variable models, which computes the means and covariance matrices of the *unconditional* posterior distribution of latent variables (Merkle et al., 2019). Although the quadrature nodes and weights obtained using this modified version may not be as precisely targeted as those from the original method, the modified method offers greater computational efficiency and is better suited for models with an increasing number of latent variables. Therefore, this modified version of the adaptive Gauss–Hermite quadrature is used in this tutorial and within the bleval package. Detailed computation formulas for adaptive Gauss–Hermite quadrature are provided in Supplemental Material S2.

This adaptive quadrature procedure is implemented by two internal functions. The `get_quadrature` function first generates a set of quadrature nodes and weights for each latent variable and produces the outer products for all node and weight sets, respectively. These standard nodes and weights are then passed to the `log_marglik_i` function, along with the posterior means and covariance matrices of the latent variables (supplied by the user when calling `log_marglik`). The `log_marglik_i` function adapts the nodes and weights using these moments, evaluates the log joint density at each adapted node for a given cluster and posterior draw, and returns the log marginal likelihood for that cluster. This process is repeated across all clusters and all posterior draws. The results are then aggregated by `log_marglik` to produce the pointwise log marginal likelihood matrix and the log marginal likelihood vector based on posterior means, which are used for computing information criteria and fully marginal likelihoods.

Comparison and Integration with Existing Packages

To clarify the role of bleval, Table 1 summarizes key differences between bleval and two commonly used packages for Bayesian latent variable modeling (i.e., blavaan; Merkle & Rosseel, 2018, and brms; Bürkner, 2017). In short, bleval is a downstream toolkit that accepts MCMC output and focuses on computing marginal likelihood-based information criteria and fully marginal likelihoods

Table 1
Comparison of the bleval Package With Commonly Used Packages for Bayesian Latent Variable Modeling

Analysis	bleval	blavaan[a]	brms
Parameter estimation	No; downstream toolkit for posterior samples from other tools (e.g., JAGS, Stan).	Yes, for multivariate normal SEMs (using a marginal approach).	Yes, for MLMs and some IRT models.
Information criteria	Marginal likelihood-based DIC, WAIC, and LOOIC	Marginal likelihood-based WAIC and LOOIC	Conditional likelihood-based WAIC and LOOIC
Fully marginal likelihood (for Bayes factors)	Integrates out latent variables first, then passes model parameters to the bridge sampler.	Not directly supported.	Passes both latent variables and model parameters to the bridge sampler.

Note. There is also the potential to compute improved fully marginal likelihoods and Bayes factors for blavaan models via the bridgesampling package, although this functionality is not seamless. DIC = deviance information criterion; WAIC = Watanabe–Akaike information criterion; LOOIC = leave-one-out (cross-validation) information criterion; SEM = structural equation model; IRT = item response theory.
[a] In addition to the default marginal likelihood approach, blavaan offers functionality for computing conditional likelihoods that involve latent variables (see Merkle, 2022) and for approximating fully marginal likelihoods via a Laplace approximation.

for data-augmented latent variable models; blavaan implements a marginal approach suitable for multivariate normal SEMs (Merkle et al., 2021); brms provides general Bayesian estimation (especially for MLMs) but computes information criteria based on conditional likelihood. Overall, bleval complements these packages by making numerical marginalization (and subsequent computation for information criteria and fully marginal likelihoods) more accessible for data-augmented latent variable models.

We conducted supplementary analyses to empirically illustrate these comparisons. Supplemental Material S3 reports a confirmatory factor analysis example where WAIC and LOOIC estimated by bleval and blavaan are highly consistent, validating bleval's accuracy for models with tractable marginal likelihoods. Supplemental Material S4 reports an analysis of the illustrative mixed-effects model (see the next section) using brms. It shows that WAIC and LOOIC obtained from bleval (based on marginal likelihood) differ from those reported by brms (based on conditional likelihood). Furthermore, the bridge sampler in brms failed to converge when attempting to compute the fully marginal likelihood by passing all latent variables and model parameters (totaling 1,006 quantities) to the sampler, highlighting the necessity of the marginalization approach implemented in bleval for stable Bayes factor computation.

A key characteristic distinguishing bleval from blavaan and brms is its aim to serve as a *general-purpose tool* for computing information criteria and fully marginal likelihoods across a wide range of latent variable models. This generality is particularly valuable for models with intractable marginal likelihoods that cannot be readily handled by existing packages' default workflows. This design goal, however, can entail a trade-off with user-friendliness, as defining model-specific components (e.g., the `log_joint_i` function) requires some user input.

To lower this barrier and enhance practical utility, we explored an integration between the bleval package and the brms package. Specifically, we developed a helper function `bleval_with_brmsfit`, which automates the computation of marginal likelihood-based information criteria and fully marginal likelihoods for certain models fitted using brms. This function accepts a fitted model object from brms (created using `brm`) and internally handles the key steps required by bleval: It extracts the model structure and posterior samples, constructs the necessary `log_joint_i` and `log_prior` functions internally, and then calls the core bleval routines to compute the evaluation metrics. The current implementation is designed for two-level models with a univariate response variable and supports Gaussian, Bernoulli, and Poisson families. Although it does not cover all model classes in brms, this function provides a substantially more accessible workflow for common cases and demonstrates the potential for making these advanced evaluation methods more readily applicable to applied research. Supplemental Material S5 provides six detailed examples demonstrating the application of this helper function to models with varying random-effects structures and response distributions.

A Worked Multilevel Linear-Model Example: Concepts and bleval Usage

Gaussian linear mixed models are classic examples of latent variable models with tractable marginal likelihoods. In this section, we use a Gaussian linear mixed model with two random effects (i.e., two latent variables) to introduce the key concepts of marginal likelihood, information criteria, and fully marginal likelihood, and to show how the corresponding bleval functions are used in practice. Since the marginal likelihood of this model is tractable, this example allows us to validate the numerical approximations produced by bleval against an exact analytical solution.

Data Generation and Parameter Estimation

The data were generated based on a mixed-effects model with two random effects (i.e., a random intercept and a random slope):

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{u}_j + \boldsymbol{\varepsilon}_j, \tag{2}$$

where $\boldsymbol{y}_j$ is an $n_j \times 1$ vector of the Level 1 units in Level 2 cluster j, $\boldsymbol{X}_j$ is an $n_j \times 2$ design matrix for the fixed effects in cluster j (with 1 in the first column representing intercepts, and the second column representing a within-cluster predictor), $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ is a 2×1 vector of fixed effects, $\boldsymbol{Z}_j$ is a $n_j \times 2$ design matrix for the random effects in cluster j, $\boldsymbol{u}_j$ is a 2×1 vector of random effects for cluster j (which follows a bivariate normal distribution $\boldsymbol{u}_j \sim \boldsymbol{N}\left(0, \boldsymbol{G} = \begin{bmatrix} \sigma^2_{u0} & \sigma_{u0}r\sigma_{u1} \\ \sigma_{u0}r\sigma_{u1} & \sigma^2_{u1} \end{bmatrix}\right)$), and $\boldsymbol{\varepsilon}_j$ is an $n_j \times 1$ vector of residuals for the units in cluster j (which follows a normal distribution $\boldsymbol{\varepsilon}_j \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{R}_j = \sigma^2\boldsymbol{I})$; $\boldsymbol{I}$ is a $n_j \times n_j$ identity matrix).

A multilevel data set with 500 clusters ($N = 500$), each containing 50 units ($n_j = 50$ for all j), was generated in R. We first set the covariance matrix for random effects as $\boldsymbol{G} = \begin{bmatrix} 1 & 0.09 \\ 0.09 & 0.09 \end{bmatrix}$ and the fixed effects as $\boldsymbol{\beta} = \begin{bmatrix} 0 \\ 0.3 \end{bmatrix}$ to generate the random effects for each cluster. The $\boldsymbol{G}$ matrix implies a correlation of 0.3 between random effects, which reflects typical dependencies between random effects in psychological studies. Then, the within-cluster predictor was generated from a normal distribution with mean zero and variance of one, and the variance of residuals σ^2 was set as 0.01 to generate the within-cluster units $\boldsymbol{y_j}$.

The data-generating model was fitted to the simulated data. We placed weakly informative priors over the fixed effects (β_0 and β_1). For the elements in the between-cluster covariance matrix $\boldsymbol{G}$ and the within-cluster covariance matrix $\boldsymbol{R_j}$, we specified Gamma priors for the inverse of variances (σ_{u0}^{-2}, σ_{u1}^{-2}, and σ^{-2}) and a uniform prior between -1 and 1 for the correlation (r).

The model parameters were estimated using JAGS (Plummer, 2003) through the rjags package (Plummer et al., 2024) in R. Four independent Markov chains were generated, with 2,500 iterations from each chain (that is, 10,000 iterations in total) obtained after burning in the first 5,000 iterations and applying a thinning interval of 10.

Computing Marginal Likelihood

Theoretical Background: Conditional and Marginal Likelihoods of Latent Variable Models

Suppose we fit latent variable models to observed data y_{ij} (i.e., the unit i for cluster j, where $i = 1, 2, \ldots, n_j$, and n_j is the number of units in cluster j; $j = 1, 2, \ldots, N$, and N is the number of clusters). For each cluster j, we construct latent variables $\boldsymbol{\eta}_j$. Specifically, in SEMs and IRT models, y_{ij} is the observed value of item i for person j, and $\boldsymbol{\eta}_j$ represents the latent characteristics for person j. In MLMs, y_{ij} denotes the unit i (e.g., students or time) in cluster j (e.g., schools or persons), and $\boldsymbol{\eta}_j$ represents the cluster-specific effects (i.e., random effects for cluster j). The observed data y_{ij} and latent variables $\boldsymbol{\eta}_j$ are modeled with certain distributions parametrized by $\boldsymbol{\theta}$. The marginal likelihood for all units in the data ($\boldsymbol{y}$) is given by

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{j=1}^{N} p(\boldsymbol{y_j} \mid \boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{y_j}$ is the vector of all units in cluster j, and $\boldsymbol{y}$ is the vector of all units (i.e., $\boldsymbol{y_j}$ stacked across clusters). The marginal likelihood for cluster j, $p(\boldsymbol{y_j} \mid \boldsymbol{\theta})$, is defined by integrating over latent variables:

$$p(\boldsymbol{y_j} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{y_j} \mid \boldsymbol{\eta_j}, \boldsymbol{\theta})\, p(\boldsymbol{\eta_j} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\eta_j}. \tag{4}$$

Here, $p(\boldsymbol{y_j} \mid \boldsymbol{\eta_j}, \boldsymbol{\theta})$ is the conditional likelihood of the units for cluster j, and $p(\boldsymbol{\eta_j} \mid \boldsymbol{\theta})$ is the density for the latent variables.

In many commonly used latent variable models, the conditional independence between units within each cluster is assumed. As a result, the conditional likelihood of the units for cluster j can be further factorized as

$$p(\boldsymbol{y_j} \mid \boldsymbol{\eta_j}, \boldsymbol{\theta}) = \prod_{i=1}^{n_j} p(y_{ij} \mid \boldsymbol{\eta_j}, \boldsymbol{\theta}), \tag{5}$$

where $p(y_{ij} \mid \boldsymbol{\eta_j}, \boldsymbol{\theta})$ is the conditional likelihood of the unit i in cluster j given the latent variables and model parameters (often also involving covariates, which are omitted for simplicity).

Computation in bleval

To compute the (log) marginal likelihood for each cluster in the mixed-effects model using bleval, we can call the `log_marglik` function (see Figure 2). There are six basic required arguments for this function: (a) `samples`, a matrix or data frame containing the posterior samples of model parameters (obtained using the rjags package in this example); (b) `data`, a list that must contain an element N indicating the number of clusters (here $N = 500$); beyond this requirement, the structure of data is flexible and can be adapted to the specific model (in this example, it also includes the vectors of the observed responses $\boldsymbol{y}$ and the covariate $\boldsymbol{x}$, but users can define additional components as needed); (c) `Ngrid`, the number of quadrature nodes for each latent variable; (d) `lv_mu`, a list of posterior means of latent variables for each cluster; (e) `lv_cov`, a list of posterior covariance matrices of latent variables for each cluster; and (f) `log_joint_i`, a function that computes the log joint density for each cluster.

Among these, the `log_joint_i` function is crucial and can be relatively challenging to specify.[2] For the mixed-effects model, the main components of this function are shown in Figure 2. This function takes a named vector with parameter values in a posterior sample (`samples_s`), a data object (`data`), an index of the cluster (`i`), a numeric number of quadrature nodes (`Ngrid`), and a matrix of latent variable values transformed from the quadrature nodes (`nodes`; output of the `get_quadrature` function) as inputs, and returns the log joint density for each cluster (i.e., the log of the integrand in Equation 4).

The `log_marglik` function then returns a list containing two objects: a matrix of log marginal likelihoods for each data point (`log_marglik_point`), and a vector of log marginal likelihoods given the point estimates (i.e., posterior means) of the model parameters (`log_marglik_postmean`).

Computing Information Criteria

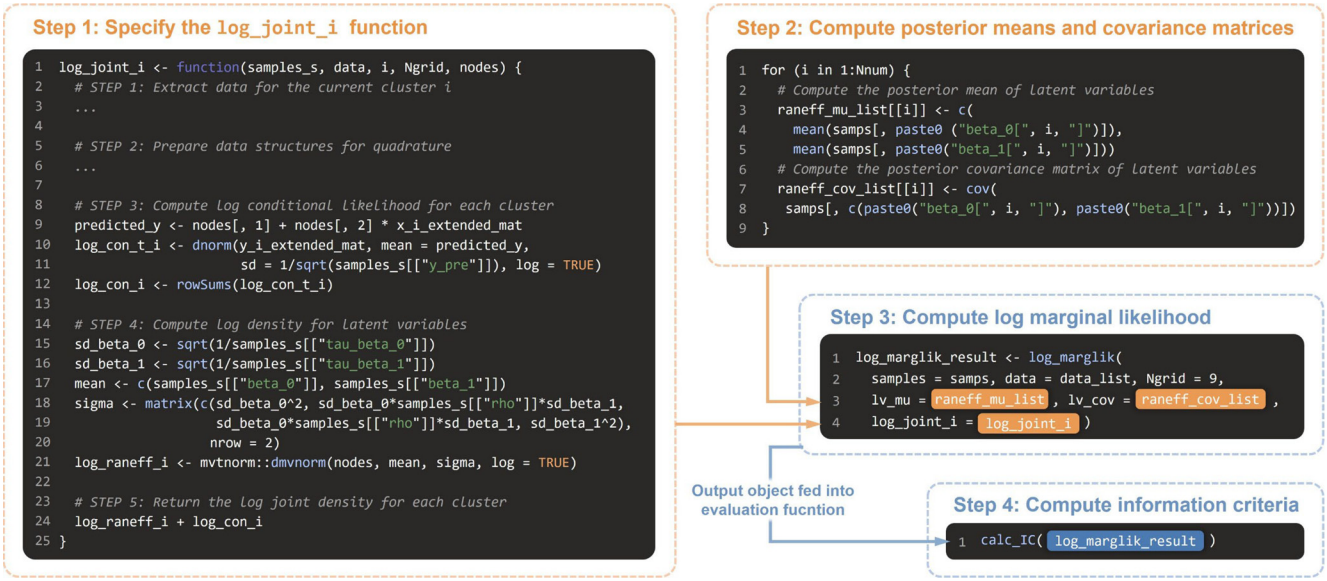
Theoretical Background: Information Criteria Based on Marginal Likelihood

For latent variable models, information criteria can be computed based on marginal likelihoods. Before diving into the definition and computation formulas for information criteria, we first introduce the expected log predictive density as a theoretical measure of the predictive performance of models. Then, we demonstrate how to practically compute information criteria to approximate the

[2] Please see Supplemental Material S6 for more detailed instructions for specifying `log_joint_i` and `log_prior` functions.

Figure 2

Computing Log Marginal Likelihood and Information Criteria in bleval



Note. Complete code and step-by-step guidance are available at https://github.com/luoxh3/bleval. See the online article for the color version of this figure.

expected log predictive density and assess the out-of-sample predictive accuracy of latent variable models.

Expected Log Predictive Density. Information criteria are used as measures of out-of-sample predictive accuracy. Theoretically, the out-of-sample predictive accuracy for a new data set $\{\tilde{y}_k\}$ $(k = 1, 2, \ldots, K)$ is defined as

$$\text{elpd} = \sum_{k=1}^{K} \int \log p(\tilde{y}_k \,|\, \boldsymbol{y}) p_{\text{true}}(\tilde{y}_k) \mathrm{d}\tilde{y}_k, \tag{6}$$

where elpd denotes the expected log (posterior) predictive density. $p(\tilde{y}_k \,|\, \boldsymbol{y}) = \int p(\tilde{y}_k \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}$ is the posterior predictive density for $\tilde{y}_k$, and $p_{\text{true}}(\tilde{y}_k)$ is the density based on the true data-generating process for $\tilde{y}_k$.

Practical Computations of Information Criteria. For historical reasons, information criteria are defined as elpd multiplied by -2 (Gelman et al., 2014). However, the true model-based density $p_{\text{true}}(\tilde{y}_k)$ in Equation 6 is typically unknown. Therefore, several practical methods have been proposed to approximate elpd and compute information criteria to evaluate out-of-sample predictive accuracy. These methods can be categorized into two types (Stone, 1977): those based on bias correction (e.g., DIC and WAIC) and those based on cross-validation (i.e., LOOIC).

The first category of information criteria, based on bias correction, starts by estimating log predictive density (e.g., lpd_{DIC}, and lpd_{WAIC}) and then subtracting a correction term (e.g., p_{DIC} and p_{WAIC}, which represent the effective number of parameters and can be used as a measure of model complexity).

The DIC (Spiegelhalter et al., 2002) is defined as

$$\text{DIC} = -2\,\text{elpd}_{\text{DIC}} = -2\,(\text{lpd}_{\text{DIC}} - p_{\text{DIC}}). \tag{7}$$

Based on posterior samples $\boldsymbol{\theta}^{(s)}$ $(s = 1, 2, \ldots, S$, where S is the total number of posterior samples), lpd_{DIC} is estimated using point estimates (typically posterior means) of parameters $\hat{\boldsymbol{\theta}}_{\text{point}}$,

$$\widehat{\text{lpd}}_{\text{DIC}} = \log p(\boldsymbol{y} \,|\, \hat{\boldsymbol{\theta}}_{\text{point}}), \tag{8}$$

and p_{DIC} has two computed versions:

$$\hat{p}_{\text{DIC}_1} = 2\left[\log p(\boldsymbol{y} \,|\, \hat{\boldsymbol{\theta}}_{\text{point}}) - \frac{1}{S}\sum_{s=1}^{S} \log p\left(\boldsymbol{y} \,|\, \boldsymbol{\theta}^{(s)}\right)\right], \tag{9}$$

$$\hat{p}_{\text{DIC}_2} = 2\,\widehat{\text{Var}}[\log p(\boldsymbol{y} \,|\, \boldsymbol{\theta}^{(s)})], \tag{10}$$

where $\widehat{\text{Var}}$ represents the sample variance. The mean-based $\hat{p}_{\text{DIC}_1}$ is numerically more stable than the variance-based $\hat{p}_{\text{DIC}_2}$ and is more commonly used (Gelman et al., 2014). The marginal likelihood for all units in the data, given the point estimates of model parameters $p(\boldsymbol{y} \,|\, \hat{\boldsymbol{\theta}}_{\text{point}})$, or given their estimates in each posterior sample $p\left(\boldsymbol{y} \,|\, \boldsymbol{\theta}^{(s)}\right)$, can be calculated according to Equation 3.

The WAIC (Watanabe, 2010) provides a more fully Bayesian approach than criteria like DIC that rely on point estimates. Although DIC computes predictive accuracy using the posterior mean (a plug-in estimator), WAIC accounts for posterior uncertainty by utilizing the entire posterior distribution and computing pointwise posterior predictive density. Specifically, WAIC is defined as

$$\text{WAIC} = -2\,\text{elpd}_{\text{WAIC}} = -2\,(\text{lpd}_{\text{WAIC}} - p_{\text{WAIC}}), \tag{11}$$

where lpd_{WAIC} can be computed as

$$\widehat{\text{lpd}}_{\text{WAIC}} = \sum_{j=1}^{N} \log\left[\frac{1}{S}\sum_{s=1}^{S} p\left(\boldsymbol{y}_j \,|\, \boldsymbol{\theta}^{(s)}\right)\right]. \tag{12}$$

Similar to that of DIC, the effective number of parameters in WAIC can also be computed based on mean ($\hat{p}_{\text{WAIC}_1}$) and variance ($\hat{p}_{\text{WAIC}_2}$).

$$\hat{p}_{\text{WAIC}_1} = 2\sum_{j=1}^{N}\left\{\log\left[\frac{1}{S}\sum_{s=1}^{S}p\left(\boldsymbol{y}_j\,|\,\boldsymbol{\theta}^{(s)}\right)\right] - \frac{1}{S}\sum_{s=1}^{S}\log p\left(\boldsymbol{y}_j\,|\,\boldsymbol{\theta}^{(s)}\right)\right\}, \tag{13}$$

$$\hat{p}_{\text{WAIC}_2} = \sum_{j=1}^{N}\widehat{\text{Var}}\left[\log p\left(\boldsymbol{y}_j\,|\,\boldsymbol{\theta}^{(s)}\right)\right]. \tag{14}$$

The variance-based $\hat{p}_{\text{WAIC}_2}$ is more stable than the mean-based $\hat{p}_{\text{WAIC}_1}$, as it sums over the variance computed for each data point. This variance-based version of p_{WAIC} is also used in the loo package for computing WAIC (Vehtari et al., 2017). The marginal likelihood $p\left(\boldsymbol{y}_j\,|\,\boldsymbol{\theta}^{(s)}\right)$ for cluster j given estimates of model parameters in each posterior sample can be calculated according to Equation 4.

The other category of information criteria is based on cross-validation. Cross-validation is a natural way to estimate out-of-sample prediction accuracy (Vehtari & Lampinen, 2002). Information criteria based on bias correction (e.g., DIC and WAIC) can be viewed as approximations to different versions of cross-validation (Stone, 1977). Information criteria based on cross-validation to estimate out-of-sample predictive accuracy typically use leave-one-out cross-validation and compute the LOOIC. In latent variable models, the elpd based on leave-one-out cross-validation is defined as

$$\text{elpd}_{\text{LOO}} = \sum_{j=1}^{N}\log p(\boldsymbol{y}_j\,|\,\boldsymbol{y}_{-j}), \tag{15}$$

where $p(\boldsymbol{y}_j\,|\,\boldsymbol{y}_{-j}) = \int\log p(\boldsymbol{y}_j\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta}\,|\,\boldsymbol{y}_{-j})\text{d}\boldsymbol{\theta}$ is the leave-one-out (leave-one-*cluster*-out) predictive density for units in cluster j, given the units in all other clusters. The LOOIC is defined as

$$\text{LOOIC} = -2\,\text{elpd}_{\text{LOOIC}}. \tag{16}$$

Theoretically, we could omit units in cluster j to generate N sets of posterior samples and estimate the elpd for each validation cluster. However, this process can be computationally demanding. Researchers are therefore more interested in approximating elpd_{LOO} using all observed clusters and posterior samples derived from the full data set. The introduced WAIC is one such approximation method, whereas others are based on importance sampling (IS) or its refinements (Li et al., 2016; Millar, 2018; Vehtari et al., 2017). Among the IS-based estimates for $\text{elpd}_{\text{LOOIC}}$, the method using Pareto smoothed IS (Vehtari et al., 2017) has higher accuracy, reliability, and robustness compared with other methods. This Pareto smoothed IS-based approach and the corresponding LOOIC can be easily implemented in the R package loo (Vehtari et al., 2017) with MCMC posterior samples of the pointwise log likelihood as input. Therefore, the LOOIC estimate appropriate for Bayesian latent variable models can be obtained using the loo package, after computing the log marginal likelihood for each data point given model parameter estimates in each posterior sample according to Equation 4.

Computation in bleval

Given the output of `log_marglik()`, computing the information criteria is straightforward with the `calc_IC()` function (see Figure 2). This function returns a list containing DIC, WAIC, and LOOIC estimates, along with the corresponding effective parameter counts and expected log predictive densities. It internally calls the `waic()` and `loo()` functions from the loo package.

Computing Fully Marginal Likelihood and Bayes Factors

Theoretical Background: Fully Marginal Likelihood and Bayes Factors

Another way to evaluate and compare models is by examining the fully marginal likelihood of each model and the Bayes factors between competing models. Fully marginal likelihood refers to the probability of the observed data given the model. In contrast to the marginal likelihood that is used in the computation of information criteria, fully marginal likelihood requires further integrating out the model parameters in latent variable models. For a latent variable model (M), the fully marginal likelihood is defined as

$$p(\boldsymbol{y}\,|\,M) = \int_{\boldsymbol{\theta}}\left[\int_{\boldsymbol{\eta}_j}p(\mathbf{y}\;|\;\boldsymbol{\eta}_j,\boldsymbol{\theta},M)\,p(\boldsymbol{\eta}_j\;|\;\boldsymbol{\theta},M)\text{d}\boldsymbol{\eta}_j\right]p(\boldsymbol{\theta}\;|\;M)\,\text{d}\boldsymbol{\theta}, \tag{17}$$

where the term in the square bracket represents the marginal likelihood, and $p(\boldsymbol{\theta}\;|\;M)$ is the prior density of model parameters given model M. The fully marginal likelihood in Equation 17 can also be interpreted as a weighted average of the likelihood over latent variables given specific values of parameters, where the weights are the density of the corresponding parameter values.

In model comparison, the fully marginal likelihood of a model describes its strength of evidence or predictive accuracy (Gronau et al., 2017; Kass & Raftery, 1995; Rouder & Morey, 2019). The ratio of fully marginal likelihoods between two models, known as the Bayes factor, can be used to evaluate the relative evidence or relative predictive accuracy between competing models (Garofalo et al., 2024; Rouder & Morey, 2019; Veenman et al., 2024).

In practice, however, the fully marginal likelihood for most models cannot be computed analytically. The R package bridgesampling (Gronau et al., 2020) based on the bridge sampling method (Bennett, 1976; Meng & Wong, 1996) can be used to approximate fully marginal likelihood. To estimate this likelihood, the package requires the following input: (a) MCMC posterior samples, (b) a function to compute the log posterior density for a set of model parameters, (c) the observed data (and other relevant quantities for the provided function), and (d) the lower and upper bounds for model parameters (Gronau et al., 2020). Notably, when using bridge sampling, it is essential to include all constant terms in the likelihood and prior densities to ensure a valid approximation of the fully marginal likelihood.

For latent variable models, two methods can be used to approximate the fully marginal likelihood with the bridgesampling package, both of which ultimately integrate out both the latent

variables and the model parameters. The first method treats latent variables as part of the parameter space and draws MCMC samples from their joint posterior with the model parameters; the fully marginal likelihood is then approximated by integrating over all latent variables and parameters simultaneously. This can result in a large number of quantities (i.e., the number of dimensions that must be integrated over; sometimes hundreds or thousands) to be addressed in the sampling process. The second method first integrates out latent variables and computes marginal likelihoods using numeric integration before passing only the MCMC samples of model parameters to the bridge sampler. Both methods are theoretically feasible; however, as will be shown in this example, the first method may be impractical for common latent variable models with moderate sample sizes, since the bridge sampling estimator can encounter convergence issues when handling too many quantities. Therefore, the second method that distinguishes latent variables and model parameters is recommended and used in this tutorial and the developed R package bleval.

Computation in bleval

To approximate the log fully marginal likelihood, we call the `log_fmarglik` function (see Figure 3). The first six arguments are the same as those for the `log_marglik` function, and the additional three required arguments are a function that computes the log prior density for model parameters (`log_prior`), and two named vectors with the lower and upper bounds of the parameters (`lb` and `ub`). For the mixed-effects model in this example, the specification of the `log_prior` function is shown in Figure 3. It takes a named vector with parameter values from a posterior sample (`samples_s`) as input and returns the log prior density for each cluster. The `log_fmarglik` function then returns an object containing the value of the log fully marginal likelihood and the number of iterations for the bridge sampling estimator to converge. This object from one model can be used to compute the Bayes factor between two models.

Validation Against the Analytical Solution

Because the marginal likelihood for the Gaussian linear mixed model is analytically tractable, we can compare the numerical approximations obtained with bleval to the exact values. Under the fitted model, the log marginal likelihood of $\boldsymbol{y}_j$ can be expressed in closed form as

$$\ell(\boldsymbol{\beta}, \boldsymbol{V}_j) = -\frac{1}{2}\{n_j \log(2\pi) + \log|\boldsymbol{V}_j| + (\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta})^T \boldsymbol{V}_j^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta})\}, \tag{18}$$

where $\boldsymbol{V}_j = \mathbf{cov}(\boldsymbol{y}_j) = \boldsymbol{Z}_j\boldsymbol{G}\boldsymbol{Z}_j^T + \boldsymbol{R}_j$. This marginal likelihood integrates out the random effects and includes only the data and model parameters, which can be directly used to compute the information criteria. The fully marginal likelihood was obtained by integrating the analytical marginal likelihood over the parameters using bridge sampling (with the same `log_prior` and parameter bounds).

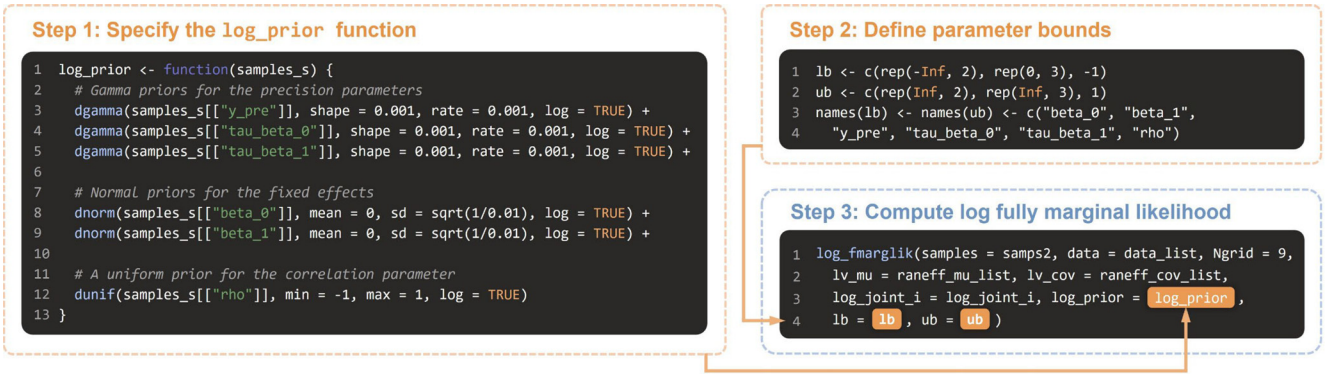
The numerical results from bleval (using nine quadrature nodes per latent variable) matched the analytical values to at least three decimal places for DIC, WAIC, and LOOIC and to the second decimal place for the log fully marginal likelihood. The R code and detailed results are provided in Supplemental Material S7.

We also reanalyzed the same model using brms and our integration helper function `bleval_with_brmsfit()`. This numerical approach yielded results virtually identical to the analytical solution and those from the standard bleval workflow. These results demonstrate the effectiveness of the numerical integration method (i.e., adaptive Gauss–Hermite quadrature) in approximating the marginal likelihood used for evaluating the mixed-effects model.

In addition, we attempted to treat random effects as model parameters and provided posterior samples of 1,006 quantities (i.e., 500 × 2 random effects plus 6 model parameters) to the `bridge_sampler` function in the bridgesampling package (Gronau et al., 2020) to approximate the fully marginal likelihood. As expected, the bridge sampling estimator failed to converge within 1,000 iterations. This occurred with both the default method (which uses a multivariate normal distribution as the proposal distribution) and the Warp-III method (an improved bridge sampling technique that accounts for potential skewness in the posterior distribution). We also increased the number of posterior samples (from 10,000 to 50,000) and the maximum number

Figure 3

Computing Log Fully Marginal Likelihood in bleval



Note. Complete code and step-by-step guidance are available at https://github.com/luoxh3/bleval. See the online article for the color version of this figure.

of iterations (from 1,000 to 5,000). Despite these adjustments, the bridge sampler still failed to converge.[3]

Although the numerical results confirm the accuracy of the bleval package, it is important to recognize that model comparison criteria, especially the log fully marginal likelihood and Bayes factors, can be sensitive to the choice of prior distributions (e.g., Kass & Raftery, 1995; van Erp et al., 2018). To illustrate this issue, we conducted a sensitivity analysis for the mixed-effects model by varying the priors on its scale parameters (see Supplemental Material S9 for details). As a comprehensive prior sensitivity analysis is beyond the scope of this tutorial, the subsequent empirical examples focus primarily on computing the evaluation metrics for each model individually. We return to the practical implications of prior choice in the Discussion section.

Empirical Examples

Up to this point, we have demonstrated the bleval workflow on an MLM with a tractable marginal likelihood; we now apply the package to two more complex latent variable models where the marginal likelihood is intractable: a latent moderation SEM and a five-dimensional IRT model. A third example using a mixture Rasch model is presented in Supplemental Material S1. The first example uses JAGS for parameter estimation, whereas the latter two use Stan. All empirical data and code are available in the R package bleval hosted on GitHub (https://github.com/luoxh3/bleval).

Example 1: Latent Moderation Structural Equation Model

Data

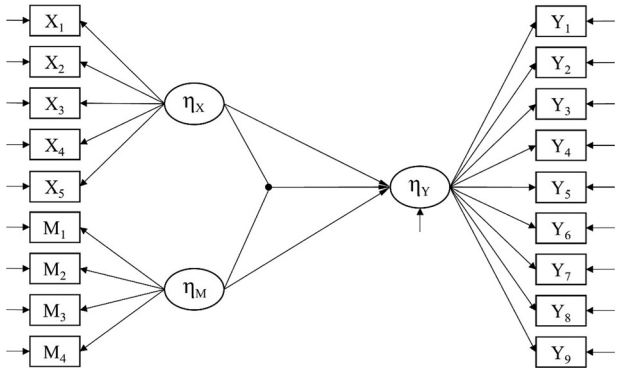
The data were collected from 356 Chinese college students (75.84% female), with a mean age of 20.658 years (ranging from 17 to 25, $SD = 1.642$). Participants completed measures of their ruminative response styles, emotion regulation strategies, and depression symptoms. The brooding subscale in the Ruminative Responses Scale (Nolen-Hoeksema & Morrow, 1991) was used to assess participants' general tendency of brooding. The scale consists of five items (e.g., "think 'What am I doing to deserve this?'"), which are evaluated on a 4-point Likert scale from 1 (*almost never*) to 4 (*almost always*). The expressive suppression subscale in the emotion regulation scale (Gross & John, 2003) assessed individuals' use of response-focused emotion regulation strategies. The scale consists of four items (e.g., "I control my emotions by not expressing them"), rated on a 7-point Likert scale, from 1 (*strongly disagree*) to 7 (*strongly agree*). Depression symptoms were evaluated using the Patient Health Questionnaire-9 (Kroenke et al., 2001), which consists of nine items. Participants reported how often they experienced depressive symptoms (e.g., "feeling down, depressed, or hopeless") during the past week on a 4-point Likert scale ranging from 1 (*never*) to 4 (almost every day).

Analysis

In this example, we examined expressive suppression (η_M) as a potential moderator of the relation between brooding (η_X) and depressive symptoms (η_Y) using a Bayesian latent moderation SEM (see Figure 4). The first loading of each latent factor (i.e., latent variable) was fixed at one for model specification. Regarding

Figure 4

Latent Moderation Structural Equation Model in Example 1



prior specifications, the regression coefficients as well as the factor loadings for the second and subsequent indicators of each latent factor were assigned normal priors with a mean of 0 and a precision of 0.01. The precision parameters (inverse variances) for the observed variables and the latent variables were assigned Gamma priors with shape and rate parameters of 0.001.

Model parameters were estimated using JAGS (Plummer, 2003) via the rjags package (Plummer et al., 2024) in R. We ran four independent Markov chains and obtained 2,500 iterations from each (i.e., 10,000 iterations in total), after burning in the first 5,000 iterations and applying a thinning interval of 10. The model was considered convergent if the potential scale reduction factor (PSRF or $\hat{R}$; Brooks & Gelman, 1998) for all parameters was less than 1.1. The sampling efficiency was monitored via the effective sample size (ESS).

To evaluate the model using the bleval package, we first extracted the posterior samples for 39 model parameters and computed the posterior means and covariance matrices of the three latent variables for each participant (to adapt quadrature nodes and weights). Then, we specified the log joint density function (`log_joint_i`) and the log prior density function (`log_prior`). Finally, we used the `log_marglik` and the `calc_IC` function to compute DIC, WAIC, and LOOIC and used the `log_fmarglik` function to approximate the log fully marginal likelihood of the model. In addition, we set the number of quadrature nodes per latent variable to 5, 9, and 15 to preliminarily examine the sensitivity of the model evaluation results to different numbers of quadrature nodes.

Results

The model converged with all PSRF values below 1.1. The sampling efficiency was high, as evidenced by large ESSs across all parameters (minimum ESS = 4,561). The results showed a positive moderating effect ($B = .159$, 95% credible interval [CrI; .024, .304]) of expressive suppression on the relation between brooding and depressive symptoms. This suggests that individuals

[3] We provide a lower dimensional example (a Gaussian linear mixed model with a random intercept and 50 units) where the bridge sampler did converge. The resulting fully marginal likelihood estimate agrees with our bleval estimates to the second decimal place. Please see Supplemental Material S8 for more details.

with a higher tendency for expressive suppression experience a stronger negative impact of brooding on depressive symptoms.

The Bayesian model evaluation results are presented in Table 2. As shown in the table, the estimates of DIC, WAIC, LOOIC, and log fully marginal likelihood demonstrate negligible variation across different numbers of quadrature nodes. Specifically, for the DIC, WAIC, and LOOIC estimates, the values remain unchanged to three decimal places when increasing the number of nodes from 9 to 15. For the log fully marginal likelihood, the differences are minimal, with only the third decimal place changing across all three node settings (5, 9, and 15). These results suggest that the evaluation metrics for the current model are relatively robust to changes in the number of quadrature nodes.

We also applied the `bridge_sampler` function by treating latent variables as model parameters, totaling 1,107 quantities (i.e., 356 × 3 latent variables plus 39 model parameters). As with the illustrative example, the bridge sampling estimator failed to converge within 1,000 iterations, using both the default method and the Warp-III method.

Example 2: Multidimensional Item Response Model

Data

A random subset of 1,000 participants was drawn from a publicly available Big Five Inventory data set (the full data set can be found online at https://www.kaggle.com/tunguz/big-five-personality-test). Each participant provided responses to the 20-item Mini International Personality Item Pool (Donnellan et al., 2006) scale. The scale consists of five subscales measuring distinct personality traits (extraversion, agreeableness, conscientiousness, neuroticism, and intellect/imagination), with four items per subscale. All items were rated on a 5-point Likert scale.

Table 2

Bayesian Model Evaluation Results of the Latent Moderation Structural Equation Model (Example 1) Under Different Numbers of Quadrature Nodes per Latent Variable

Bayesian model evaluation measures	$N_{node} = 5$	$N_{node} = 9$	$N_{node} = 15$
DIC			
$\hat{p}_{DIC}$	39.110	39.103	39.103
$\widehat{elpd}_{DIC}$	−7,769.708	−7,769.701	−7,769.701
$\widehat{DIC}$	15,539.415	15,539.402	15,539.402
WAIC			
$\hat{p}_{WAIC}$	44.381	44.381	44.381
$\widehat{elpd}_{DIC}$	−7,772.701	−7,772.698	−7,772.698
$\widehat{WAIC}$	15,545.403	15,545.396	15,545.396
LOOIC			
$\hat{p}_{LOO}$	44.418	44.417	44.417
$\widehat{elpd}_{LOO}$	−7,772.736	−7,772.734	−7,772.734
$\widehat{LOOIC}$	15,545.475	15,545.468	15,545.468
Log fully marginal likelihood	−7,998.337	−7,998.334	−7,998.331

Note. $\hat{p}_{DIC}$, $\hat{p}_{WAIC}$, and $\hat{p}_{LOO}$ denote the estimated effective number of parameters. N_{node} denotes the number of quadrature nodes per latent variable. elpd = expected log predictive density; DIC = deviance information criterion; WAIC = Watanabe–Akaike information criterion; LOOIC = leave-one-out (cross-validation) information criterion.

Analysis

A generalized partial credit model (Muraki, 1992) was applied to the data (see Figure 5). Each item had a discrimination parameter (α_j) and four step difficulty parameters (β_{js}) estimated ($j = 1, 2, \ldots, 20; s = 1, 2, 3, 4$). The variances of the traits (i.e., latent variables) were constrained to 1.00, whereas the correlations among traits $\mathbf{\Omega}$ were freely estimated. To handle the reverse-worded items, we modified the discrimination parameter from α_j to $w_j\alpha_j$. Here, w_j represents an item-wording weight, where $w_j = -1$ for reverse-worded items and $w_j = 1$ for nonreverse-worded items, with α_j correspondingly constrained to be positive. For item parameters, the prior distributions were set as $\beta_{js} \sim N(0, 5^2)$ and $\alpha_j \sim \text{Log} - \text{normal}(0, 1^2)$. For person parameters, five traits were assumed to follow a multivariate normal distribution with mean vector $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = \mathbf{\Omega}$. The correlation matrix $\mathbf{\Omega}$ was specified with an LKJ (Lewandowski–Kurowicka–Joe) correlation prior distribution (Lewandowski et al., 2009) with shape 1.

Model parameters were estimated using Stan Version 2.33.1 (Stan Development Team, 2023) via the cmdstanr package Version 0.6.1 (Gabry et al., 2023) in R. Four chains were run, each with 2,500 warm-up iterations and 2,500 postwarm-up iterations. Model convergence was assessed by the PSRF, with its values below 1.1 for all parameters considered as satisfactory.

Since the number of quadrature points grows exponentially with dimensionality, higher node settings become computationally intensive in this moderately high-dimensional scenario. We therefore explored quadrature nodes of 3, 5, and 7 per dimension, striking a practical balance between potential accuracy gains and computational feasibility. After defining the log joint density function (`log_joint_i`) and the log prior density function (`log_prior`) for the generalized partial credit model, we computed DIC, WAIC, LOOIC, and log fully marginal likelihood using the bleval package.

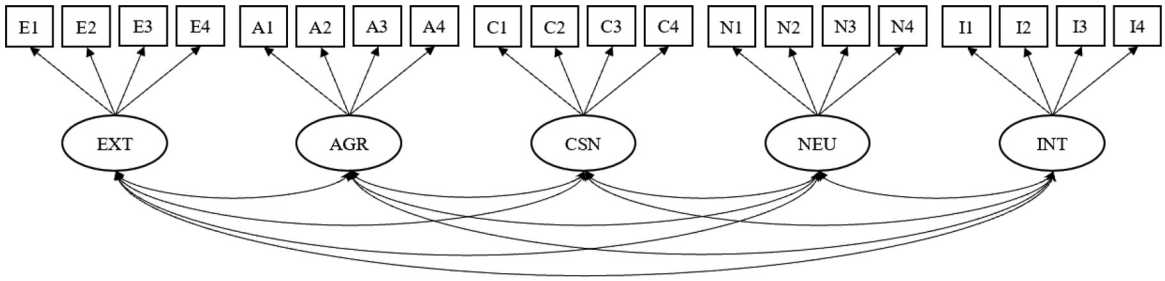
Results

The model demonstrated strong convergence (all PSRF values below 1.1) and high sampling efficiency (minimum ESS = 1,347). The discrimination of the item "I sympathize with others' feelings" in the agreeableness subscale was estimated to be the strongest (2.247, 95% CrI [1.787, 2.827]), whereas the item "I seldom feel blue" in the neuroticism subscale turned out to have the weakest discrimination (.446, 95% CrI [.358, .540]). As for the correlations among traits, sufficient evidence supported that neuroticism was negatively correlated with both extraversion (−.124, 95% CrI [−.202, −.046]) and conscientiousness (−.365, 95% CrI = [−.444, −.282]), and the correlation between extraversion and conscientiousness was positive (.097, 95% CrI [.014, .176]). Besides, agreeableness exhibited positive correlations with extraversion (.241, 95% CrI [.165, .313]), openness (.197, 95% CrI [.115, .277]), and neuroticism (.130, 95% CrI [.050, .207]).

The Bayesian model evaluation results are presented in Table 3. The DIC estimate exhibited a substantial change when quadrature nodes per latent variable increased from 3 to 5, whereas the change markedly diminished between five and seven quadrature nodes. This convergence pattern was also consistent across estimates of WAIC, LOOIC, and log fully marginal likelihood. Close

Figure 5

Five-Dimensional Generalized Partial Credit Model in Example 2



Note. EXT = extraversion; AGR = agreeableness; CSN = conscientiousness; NEU = neuroticism; INT = intellect/imagination.

agreement between results from five and seven quadrature nodes per latent variable indicates that five quadrature nodes may achieve a sufficiently accurate and computationally manageable approximation. Although seven quadrature nodes yield marginal improvement, they incur considerably higher computational cost ($7^5 = 16{,}807$ vs. $5^5 = 3{,}125$ points).

Similar to the exploration in Example 1, we applied the `bridge_sampler` function by treating latent variables as model parameters, totaling 5,110 quantities (i.e., $1{,}000 \times 5$ latent variables plus 110 model parameters). The bridge sampling estimator could not reach convergence within 1,000 iterations, using both the default method and the Warp-III method.

Discussion

Model evaluation is an important component of Bayesian modeling, where information criteria, including DIC, WAIC, and LOO, along with fully marginal likelihoods, have been widely employed as evaluation metrics. In this tutorial, we have emphasized the importance of distinguishing conditional likelihoods, marginal likelihoods (integrating out latent variables), and fully marginal likelihoods (integrating out both latent variables and parameters) when evaluating Bayesian latent variable models. We have introduced the R package bleval, which implements an adaptive Gauss–Hermite quadrature method to approximate marginal likelihoods from MCMC posterior samples and then uses these approximations to compute information criteria (DIC, WAIC, LOOIC) and fully marginal likelihoods (for Bayes factors) via existing packages loo and bridgesampling. We demonstrated the package's workflow through a worked multilevel linear-model example, validated the numerical approximations against an analytical solution, and applied the package to more complex models with intractable marginal likelihoods (nonlinear SEMs and IRT models). These examples collectively demonstrate the broad applicability of the bleval package across diverse latent variable models.

Table 3

Bayesian Model Evaluation Results of the Five-Dimensional Generalized Partial Credit Model Under Different Numbers of Quadrature Nodes per Latent Variable

Bayesian model evaluation measures	$N_{node} = 3$	$N_{node} = 5$	$N_{node} = 7$
DIC			
$\hat{p}_{DIC}$	111.143	110.462	110.267
$elpd_{DIC}$	−27,205.855	−27,187.561	−27,186.846
$\widehat{DIC}$	54,411.710	54,375.122	54,373.692
WAIC			
$\hat{p}_{WAIC}$	122.434	122.468	122.490
$\widehat{elpd_{DIC}}$	−27,211.804	−27,193.878	−27,193.250
$\widehat{WAIC}$	54,423.607	54,387.715	54,386.501
LOOIC			
$\hat{p}_{LOO}$	122.497	122.531	122.552
$\widehat{elpd_{LOO}}$	−27,211.866	−27,193.920	−27,193.313
$\widehat{LOOIC}$	54,423.733	54,387.841	54,386.627
Log fully marginal likelihood	−27,525.440	−27,507.550	−27,506.940

Note. $\hat{p}_{DIC}$, $\hat{p}_{WAIC}$, and $\hat{p}_{LOO}$ denote the estimated effective number of parameters. N_{node} denotes the number of quadrature nodes per latent variable. elpd = expected log predictive density; DIC = deviance information criterion; WAIC = Watanabe–Akaike information criterion; LOOIC = leave-one-out (cross-validation) information criterion.

Contributions

This tutorial emphasizes the importance of differentiating three key concepts in the context of Bayesian latent variable modeling: (a) the likelihoods conditional on model parameters and latent variables (referred to as "*conditional* likelihoods"), (b) the likelihoods with latent variables integrated out (referred to as "marginal likelihoods"), and (c) the likelihoods involving integration over both latent variables and model parameters (referred to as "fully marginal likelihoods" in this article). Although the difference between conditional likelihoods and marginal likelihoods has been discussed in previous research (Li et al., 2016; Merkle et al., 2019; Millar, 2018; Spiegelhalter et al., 2002), information criteria based on marginal likelihoods rather than conditional likelihoods have still received inadequate attention, even in methodological studies from Bayesian multilevel modeling (e.g., Liu et al., 2025) to Bayesian IRT modeling (e.g., Ulitzsch et al., 2022). Critically, marginal likelihoods align with the canonical definition of likelihood, and information criteria based on marginal likelihoods show better generalizability to new clusters (Merkle et al., 2019; Zhang et al., 2019). As Bayesian latent variable models become increasingly popular, explicitly distinguishing these types of likelihoods can enhance applied researchers' comprehension of these concepts, thereby preventing inappropriate calculation of model evaluation indices.

Building on these theoretical clarifications, we propose a practical workflow to approximate marginal likelihoods, based on which

the information criteria and the fully marginal likelihoods are computed. Inspired by Rabe-Hesketh et al. (2002) and Merkle et al. (2019), we implemented a post hoc approximation of marginal likelihoods using adaptive Gauss–Hermite quadrature based on posterior samples, which offers a computationally efficient solution for real-world applications. We also addressed a practical issue related to bridge sampling: its computational bottleneck when applied to latent variable models. Although bridge sampling is recommended for estimating fully marginal likelihoods (Gronau et al., 2017), its direct application to latent variable models, where both latent variables and model parameters are treated as parameters to be sampled, often leads to convergence failures (as shown in our illustrative and empirical examples). In this article, we addressed this problem by preintegrating the latent variables via adaptive Gauss–Hermite quadrature, thereby reducing the number of quantities in the posterior samples passed to the bridgesampling package. This strategy echoes the recommendations in high-dimensional Bayesian inference (Bilodeau et al., 2024) and improves the computational feasibility of fully marginal likelihoods.

To support researchers in implementing the proposed workflow, we developed the bleval package, a practical tool for evaluating Bayesian latent variable models. The package combines adaptive Gauss–Hermite quadrature with existing R packages (loo, bridgesampling), enabling convenient computation of both information criteria and fully marginal likelihoods for Bayesian latent variable models. Diverse scenarios in practice can be accommodated by the bleval package. For instance, it can handle common nonlinear structures of latent variables, such as the product of latent variables involved in Example 1. Example 3 also demonstrates its feasibility when models contain latent class variables. Moreover, through posterior sample inputs, the bleval package can serve as a downstream analysis toolkit compatible with mainstream Bayesian estimation platforms, including but not limited to JAGS and Stan. Taken together, the bleval package contributes to the proper and practical application of Bayesian evaluation for latent variable models in future research.

Practical Recommendations

Despite the distinction between conditional likelihoods (given latent variables) and marginal likelihoods (with latent variables integrated out), researchers usually rely on the default settings of software, without being fully aware of how information criteria were calculated. Therefore, when evaluating Bayesian latent variable models, users should be cautious about the information criteria automatically calculated by default functions in estimation toolkits, as they are probably based on conditional likelihoods rather than marginal likelihoods. For more appropriate evaluation of Bayesian latent variable models, we highly suggest adopting our bleval package, which features integrating out latent variables before calculating the indices.

When utilizing the bleval package, some recommendations merit attention. First, it is crucial to carefully specify the functions `log_joint_i` and `log_prior` to ensure the integrand aligns with the target model. Taking the mixed-effects model as a template, we offered step-by-step instructions on how to specify these user-defined functions and compute evaluation indices via the bleval package. Second, adequate posterior samples are needed for ensuring the stability of evaluation results, particularly in the approximation of fully marginal likelihoods. Despite being a solution to the issue of latent variables, bridge sampling may still struggle with handling hundreds of model parameters (e.g., large-scale IRT models). In such cases, increasing MCMC iterations can improve convergence by reducing Monte Carlo errors, albeit necessitating compromises in computational efficiency. Third, users should be aware of the influence of the number of quadrature nodes on result stability and computational efficiency. In Examples 1 and 3, we set the number of nodes per latent variable as 5, 9, and 15 and found that the differences between the results of two neighboring iterations tended to be stably small. To explore larger numbers of nodes per dimension, users may follow Furr's (2017) approach beginning with an odd number (e.g., 5 or 7) and increasing by approximately 50% while remaining odd so that a quadrature point is at the posterior mean. Stopping criteria should be tailored to model-specific approximation performance. For high-dimensional models, as demonstrated in Example 2, practical trade-offs between accuracy and computational demands are necessary.

Beyond these considerations, attention should also be given to verifying the accuracy of user-defined likelihood and prior functions. For models with tractable marginal likelihoods, results from bleval can be validated against known analytical solutions or established software like blavaan. For more complex models where such benchmarks are unavailable, our experience suggests the utility of a stepwise approach: (a) during the writing phase, one may employ the "$\sim$" heuristic to ensure all model components are accounted for, while also verifying that the parameterization of R's density functions (e.g., `dnorm` using mean and standard deviation) matches that of their Bayesian software (e.g., JAGS's `dnorm` using mean and precision); (b) after writing them, a basic test of the functions with a single posterior draw can confirm whether they execute and return a plausible log-density value; and (c) after obtaining the results, it is advisable to examine whether the estimated effective number of parameters aligns reasonably with the actual number of model parameters (excluding latent variables), as a notable divergence may signal potential issues in computation. It is also crucial to use properly normalized density functions (both likelihood and prior), meaning they include all necessary constant terms. Omitting these constants will bias the approximations from numerical integration. These checks can strengthen confidence in both the custom implementation and the resulting model evaluation metrics.

In addition to the computation process, the report and interpretation of model evaluation indices are noteworthy as well. When reporting the evaluation results from the bleval package, users should clearly describe the MCMC sampling settings, the integration method, the rule for selecting integration nodes, the final number of integration nodes, and the calculation formulas of evaluation indices. When interpreting the evaluation results, it is important to note that Bayesian information criteria, including DIC, WAIC, and LOOIC, as well as the fully marginal likelihood, are all *relative* measures of predictive accuracy or evidence, which means their practical utility lies in their comparison across competing models. In practice, we could compare these metrics to analogous metrics from other models to weight models or select the best model.

Furthermore, although the bleval package offers a stable numerical approximation of the log fully marginal likelihood, several important interpretative caveats surrounding Bayes factors warrant emphasis. A key concern well-documented in the literature is their sensitivity to prior specification. As Tendeiro and Kiers (2019) note, Bayes factors can vary substantially with different prior

choices. Thus, we would like to highlight that it is generally recommended to accompany applications of fully marginal likelihoods and Bayes factors with a prior sensitivity analysis (van Erp et al., 2018). If results remain consistent across a reasonable range of priors, conclusions are more robust. However, if results prove highly sensitive to prior choice, this uncertainty should be transparently reported and discussed. The computational stability of Bayes factor estimates should also be verified, a concern highlighted by Schad et al. (2023), who recommend incorporating stability checks and sensitivity analyses into the workflow. Prior predictive checks (Winter & Depaoli, 2023) are also suggested to ensure that priors are appropriately aligned with substantive knowledge. In summary, although valuable for model comparison, log fully marginal likelihoods and Bayes factors offer evidence conditional on the full model specification, including prior choices. We therefore advocate their use as one component within a comprehensive workflow, complemented by parameter estimation and posterior predictive checks to form a more robust and complete inferential picture. For deeper discussions, readers are referred to Tendeiro and Kiers (2019), Schad et al. (2023), and van Erp et al. (2018).

Limitations and Future Directions

Although adaptive Gauss–Hermite quadrature based on posterior means and covariance matrices performed robustly in different kinds of latent variable models, we acknowledge that its accuracy for severely skewed or multimodal posterior distributions remains untested. Future work could explore the limits of the current adaptive quadrature method and investigate alternative strategies (e.g., mode-curvature adjustments; Pinheiro & Bates, 1995).

In addition, Example 2 empirically revealed the application boundary of latent variable dimensions for the bleval package, suggesting that our package shares the same challenges faced by existing numerical integration methods. Specifically, the excessive number of quadrature nodes required in high-dimensional scenarios leads to prohibitive computational costs. Although this issue extends beyond the scope of model evaluation discussed in this tutorial, it warrants further investigation, as resolving high-dimensional integration challenges could unlock breakthroughs in evaluating more complex latent variable models with more random effects or latent constructs. Here, the bleval package represents a preliminary yet important attempt, paving the way for transforming the Bayesian evaluation of latent variable models from a technical challenge into a routine practice.

References

Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, *22*(2), 245–268.

Bilodeau, B., Stringer, A., & Tang, Y. (2024). Stochastic convergence rates and applications of adaptive quadrature in Bayesian inference. *Journal of the American Statistical Association*, *119*(545), 690–700.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), Article 28.

Cagnone, S., & Monari, P. (2013). Latent variable models for ordinal data by using the adaptive quadrature approximation. *Computational Statistics*, *28*(2), 597–619.

Davis, P. J., & Polonsky, I. (1972). Numerical interpolation, differentiation, and integration. In M. Abramowitz & I. A. Stegun (Eds.), *Handbook of mathematical functions* (pp. 875–924). Dover.

Depaoli, S., Winter, S. D., & Liu, H. (2024). Under-fitting and over-fitting: The performance of Bayesian model selection and fit indices in sem. *Structural Equation Modeling*, *31*(4), 604–625.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *Annals of Mathematical Statistics*, *41*(1), 214–226.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP Scales: Tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment*, *18*(2), 192–203. https://doi.org/10.1037/1040-3590.18.2.192

Furr, D. C. (2017). *Bayesian and frequentist cross-validation methods for explanatory item response models* [Unpublished doctoral dissertation]. University of California.

Gabry, J., Češnovar, R., & Johnson, A. (2023). *Cmdstanr: R interface to CmdStan*. https://mc-stan.org/cmdstanr/

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*, *25*(1), 46–70. https://doi.org/10.1037/met0000224

Garnier-Villarreal, M., & Jorgensen, T. D. (2025). Evaluating local model misspecification with modification indices in Bayesian structural equation modeling. *Structural Equation Modeling*, *32*(2), 304–318.

Garofalo, S., Finotti, G., Orsoni, M., Giovagnoli, S., & Benassi, M. (2024). Testing Bayesian informative hypotheses in five steps with JASP and R. *Advances in Methods and Practices in Psychological Science*, *7*(4), 1–23.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, *92*(10), 1–29.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, *85*(2), 348–362. https://doi.org/10.1037/0022-3514.85.2.348

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–417.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*(5), 539–556. https://doi.org/10.1037/met0000201

Jeffreys, H. (1961). *The theory of probability*. Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling*, *18*(4), 663–685.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Li, L., Qiu, S., Zhang, B., & Feng, C. X. (2016). Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. *Statistics and Computing*, *26*(4), 881–897.

Liu, Q., & Pierce, D. A. (1994). A note on Gauss–Hermite quadrature. *Biometrika*, *81*(3), 624–629.

Liu, Y., Fang, F., & Liu, H. (2025). Model selection for mixed-effects location-scale models with confidence interval for LOO or WAIC difference. *Multivariate Behavioral Research*, *60*(4), 678–694. https://doi.org/10.1080/00273171.2025.2462033

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book. A practical introduction to Bayesian analysis*. Chapman Hall.

Meng, X. L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*(4), 831–860.

Merkle, E. C. (2022). *Adventures in ordinal model likelihoods*. https://ecmerkle.org/cs/ord_ic.html

Merkle, E. C., Fitzsimmons, E., Uanhoro, J., & Goodrich, B. (2021). Efficient Bayesian structural equation modeling in Stan. *Journal of Statistical Software*, *100*(6), 1–22.

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, *84*(3), 802–829. https://doi.org/10.1007/s11336-019-09679-0

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30.

Millar, R. B. (2018). Conditional vs marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation. *Statistics and Computing*, *28*(2), 375–385.

Muraki, E. (1992). A Generalized Partial Credit Model:Application of an EM Algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Naylor, J. C., & Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society*, *31*(3), 214–225.

Nolen-Hoeksema, S., & Morrow, J. (1991). A prospective study of depression and posttraumatic stress symptoms after a natural disaster: The 1989 Loma Prieta earthquake. *Journal of Personality and Social Psychology*, *61*(1), 115–121. https://doi.org/10.1037//0022-3514.61.1.115

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*(1), 12–35.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing. https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Plummer, M., Stukalov, A., & Denwood, M. (2024). *rjags: Bayesian graphical models using MCMC* (version 4-16). R package. https://CRAN.R-project.org/package=rjags

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal*, *2*(1), 1–21.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*(2), 301–323.

Rouder, J. N., & Morey, R. D. (2019). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *American Statistician*, *73*(2), 186–190.

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2023). Workflow techniques for the robust use of Bayes factors. *Psychological Methods*, *28*(6), 1404–1426. https://doi.org/10.1037/met0000472

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, *64*(4), 583–639.

Spiegelhalter, D. J., Thomas, A., Best, N., & Gilks, W. (1996). *BUGS 0.5: Bayesian inference using Gibbs sampling manual* (Version ii). MRC Biostatistics Unit, Institute of Public Health.

Stan Development Team. (2023). *Stan modeling language users guide and reference manual* (Version 2.33). https://mcstan.org

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society*, *39*(1), 44–47.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540.

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795. https://doi.org/10.1037/met0000221

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *British Journal of Mathematical and Statistical Psychology*, *75*(3), 668–698. https://doi.org/10.1111/bmsp.12272

van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods*, *23*(2), 363–388. https://doi.org/10.1037/met0000162

van Kollenburg, G. H., Mulder, J., & Vermunt, J. K. (2017). Posterior calibration of posterior predictive p values. *Psychological Methods*, *22*(2), 382–396. https://doi.org/10.1037/met0000142

Veenman, M., Stefan, A. M., & Haaf, J. M. (2024). Bayesian hierarchical modeling: An introduction and reassessment. *Behavior Research Methods*, *56*(5), 4600–4631. https://doi.org/10.3758/s13428-023-02204-3

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Vehtari, A., & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, *14*(10), 2439–2468. https://doi.org/10.1162/08997660260293292

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107. https://doi.org/10.1006/jmps.1999.1278

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(12), 3571–3594.

Winter, S. D., & Depaoli, S. (2022). Performance of model fit and selection indices for Bayesian structural equation modeling with missing data. *Structural Equation Modeling*, *29*(4), 531–549.

Winter, S. D., & Depaoli, S. (2023). Illustrating the value of prior predictive checking for Bayesian structural equation modeling. *Structural Equation Modeling*, *30*(6), 1000–1021.

Zhang, X., Tao, J., Wang, C., & Shi, N. (2019). Bayesian model selection methods for multilevel IRT models: A comparison of five DIC-based indices. *Journal of Educational Measurement*, *56*(1), 3–27.

Received June 18, 2025
Revision received March 17, 2026
Accepted May 7, 2026 ■