Marginal Likelihood Estimation in Item Response Theory Models

A Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Alex Nguyen, B.S.

Graduate Program in Statistics

The Ohio State University

2025

Master's Examination Committee:

Sally Paganin, Advisor

Christopher Hans

Abstract

This dissertation investigates the use of marginal likelihood estimation for Bayesian model comparison in Item Response Theory (IRT), focusing on 2-Parameter Logistic (2PL) models and their finite mixture extensions. Bayesian model comparison uses the Bayes factor, which is defined as the ratio of the marginal likelihood of the competing models. These are typically estimated via Monte Carlo methods, with bridge sampling being a popular and general purpose approach. However, it can become in-efficient when models are high dimensional.

The study applies bridge sampling to both standard 2PL models and finite mixture 2PL models, which allow the latent ability distribution to follow a flexible mixture of Gaussians. To improve estimation, we introduce a marginalization strategy that integrates out the latent abilities using a grid-based approximation. This approach avoids direct sampling of discrete cluster assignments and reduces the dimensionality of the posterior samples, resulting in faster and more stable computation.

Using simulated data under unimodal, bimodal, and multimodal ability distributions, we evaluate the effectiveness of bridge sampling in recovering calibrated Bayes factors. Results show that while finite mixture models are more flexible, the standard 2PL model can outperform them in cases where the true ability distribution is unimodal. This work provides practical insights into the application of bridge sampling

for psychometric data analysis and demonstrates its potential to enhance Bayesian model evaluation in high-dimensional settings.

Vita

April 9, 2001 ..... Born - Moscow, Russia

2023 ..... B.S. Applied Mathematics,
University of Cincinnati

2023-present ..... Graduate Teaching Associate,
The Ohio State University.

Fields of Study

Major Field: Statistics

Studies in Bayesian Model Comparison: Dr. Sally Paganin

Table of Contents

List of Figures

Figure **Page**

Chapter 1: Introduction

Item Response Theory (IRT) models are widely used in psychometric and educational testing, particularly for analyzing the relationship between unobservable latent traits and item responses. These traits, such as ability or proficiency, are often inferred from item responses such as answers to questions. For example, in an exam setting, we analyze students' responses to measure their abilities and assess characteristics of the test items. These models have seen extensive applications but they rely on the assumption that latent traits of the subjects follow a standard normal distribution and as such, often fall short in capturing the complex structures of real-world data, where the population might consist of distinct sub-populations with varying trait distributions. As a result, traditional parametric IRT models can fail to account for latent heterogeneity, leading to poor model fit, biased parameter estimates, or incorrect conclusions. This limitation has motivated the development of more flexible approaches for the latent trait distribution, including mixture models or Bayesian Non-parametric (BNP) approaches. These models allow the number of latent clusters to adapt to the data, thus accommodating a wider range of latent trait distributions and providing a more nuanced understanding of varying characteristics within diverse populations.

With these models in mind, the next logical step would be model selection to see which model best fits the observed data. However, this proves to be a major challenge. Often in Bayesian Statistics, the Bayes factor criterion, which quantifies the relative evidence in favor of one model over another, is used for model comparison. It relies heavily on accurately estimating the marginal likelihood (or the normalizing constant) of each model. Unfortunately, this is computationally challenging, particularly for complex models like BNP IRT models, which involve a high number of parameters, since they usually assume at least one parameter for each individual and one or more relative to the items. In such cases, the marginal likelihood is not analytically available and we need to rely on advance computational methods for estimation. While there has been considerable effort in developing methods to compare traditional IRT models with different numbers of item parameters [Liu et al., 2019], little research has focused on comparing models with varying assumptions about the distribution of latent traits (ability). Thus, this paper aims to fill that literature gap.

To estimate marginal likelihoods, several computational methods have been proposed, including importance sampling [Tokdar and Kass, 2010], harmonic mean estimators [Raftery et al., 2007], and bridge sampling [Gronau et al., 2017]. Among these, bridge sampling stands out as a particularly effective method for estimating the marginal likelihood of complex posterior distributions. Bridge sampling works by constructing a "bridge" between the posterior distribution and a simpler proposal distribution (most often a multivariate normal distribution), allowing for a more accurate and stable estimation of the marginal likelihood. This approach is especially well-suited for high-dimensional models much like BNP IRT models, where other estimators might suffer from high variance or convergence issues. Thus, this thesis

focuses on using bridge sampling to estimate the marginal likelihoods of parametric and non-parametric IRT models, investigating the efficacy and limitations of this method when applied to models with varying levels of complexity. By applying bridge sampling to both parametric and BNP IRT models, we aim to provide insights into its suitability for comparing models that differ not only in item parameters but also in their assumptions about latent trait distributions.

The remainder of this dissertation is organized as follows. Chapter 2 introduces the fundamental concepts of Bayesian model comparison, with an emphasis on marginal likelihoods and Bayes factors. It also provides a detailed discussion of bridge sampling and prior calibration methods used to improve the stability of model comparison. Chapter 3 presents the Item Response Theory (IRT) models considered in this study, including the standard parametric 2PL model, finite mixture IRT models, and semi-parametric extensions. Chapter 4 describes the application of bridge sampling to both the 2PL IRT model and its finite mixture extension. This chapter details the simulation setup, prior calibration procedures, and computational strategies implemented to address challenges in high-dimensional posterior estimation. Chapter 5 presents the results of the model comparison study, evaluating the performance of bridge sampling in selecting between unimodal and bimodal latent ability structures, along with a discussion of computational efficiency and limitations. Finally, Chapter 6 summarizes the key findings and contributions of this dissertation and outlines possible directions for future research in Bayesian model comparison for IRT and other complex hierarchical models.

Chapter 2: Model Comparison in Bayesian Statistics

2.1 Bayesian Statistics

In Bayesian statistics, the fundamental idea is to update our beliefs about parameter vector $\boldsymbol{\theta}$ after observing data vector $\boldsymbol{Y}$. This approach allows for the combination of prior knowledge (through the prior distribution) and new information (from the data) in a coherent way. The prior distribution $\pi(\boldsymbol{\theta})$ represents our initial beliefs or assumptions about the parameter before any data is observed. The likelihood $p(\boldsymbol{Y}|\boldsymbol{\theta})$, on the other hand, reflects the probability of the observed data given the parameter. After observing the data, the prior is updated using Bayes' theorem, resulting in the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{Y})$, which combines both the prior information and the likelihood derived from the data:

$$p(\theta|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\theta)\pi(\boldsymbol{\theta})}{m(\boldsymbol{Y})}.$$

The posterior distribution represents our updated belief about the parameter $\boldsymbol{\theta}$ after considering the evidence provided by the data. The denominator $m(\boldsymbol{Y})$, known as the marginal likelihood or evidence, is a normalizing constant that ensures the posterior distribution integrates to one. Bayesian inference revolves around this process of updating beliefs, allowing for a dynamic and flexible approach to statistical modeling, where prior assumptions and new data are continuously incorporated to

refine estimates. Overall, Bayesian statistics offers a flexible, intuitive, and robust framework for inference, particularly in settings where prior knowledge is valuable, uncertainty needs to be fully quantified, or complex models must be estimated.

2.2 Overview of Marginal Likelihood

The marginal likelihood is a critical component in Bayesian model comparison. It represents the probability of the observed data under a specific model, averaging over all possible values of the model parameters, weighted by their prior distributions. Mathematically, for a given model M with parameter vector $\boldsymbol{\theta}$ and data $\boldsymbol{Y}$, the marginal likelihood is defined as:

$$p(\boldsymbol{Y}|M) = \int\!\!\int\!\!\Big(p(\boldsymbol{Y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)\, d\boldsymbol{\theta},$$

where $p(\boldsymbol{Y}|M)$ is the likelihood function indicating how likely the observed data $\boldsymbol{Y}$ are, given the parameters vector $\boldsymbol{\theta}$ and the model M, and $p(\boldsymbol{\theta}|M)$, or $\pi(\boldsymbol{\theta})$, is the prior distribution of the parameters under model M.

The marginal likelihood (also known as the Bayesian evidence) serves a dual purpose. First, it acts as a normalizing constant in Bayesian inference, ensuring that the posterior distribution integrates to one. Second, it quantifies the overall fit of the model to the data, balancing the likelihood of the observed data against the complexity of the model. In Bayesian inference, more complex models, which typically involve a larger number of parameters, have broader parameter spaces to integrate over. This often results in a lower marginal likelihood unless the additional complexity is strongly supported by the data. As mentioned before, in cases with high dimensional parameters like the IRT models, computing this integral analytically is not feasible and we must rely on computational methods.

2.3 Methods for Marginal Likelihood Estimation

In most cases, the marginal likelihood does not have a closed form or is too complex to derive. Thus, it is commonly computed using Monte Carlo integration, a computational technique that relies on repeated random sampling to approximate numerical results. Generally, this method expresses the integral of interest as an expected value with respect to a probability distribution $p(x)$ of a random variable X. If we can simulate from this distribution, we can approximate the expectation of $m(x)$ using the sample mean. For example, the method estimates an expectation of $E_p[m(x)]$ with the empirical average:

$$\int\!\!\!\int p(x)m(x)dx = E_p[m(x)] \approx \frac{1}{N}\sum_{i=1}^{N} m(x_i), \quad x_i \sim p(x),$$

where x_i are i.i.d samples from the distribution of X and N is the number of samples. There are many different implementations of Monte Carlo integration to estimate the marginal likelihood because different methods vary in efficiency depending on the distribution of the likelihood relative to the prior. The following MC methods are some of them.

2.3.1 Naive Monte Carlo

The Naive Monte Carlo method is one of the simplest method of estimating the marginal likelihood. The idea is to rewrite the marginal likelihood as the expectation under the prior distribution and estimate it with the average:

$$\begin{aligned} p(\boldsymbol{Y}|M) &= \int\!\!\!\int p(\boldsymbol{Y}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= E_\pi\left[p(\boldsymbol{Y}|\boldsymbol{\theta}, M)\right] \\ &\approx \frac{1}{N}\sum_{i=1}^{N} p(\boldsymbol{Y}|\tilde{\boldsymbol{\theta}}_i, M), \quad \tilde{\boldsymbol{\theta}}_i \sim \pi(\boldsymbol{\theta}), \end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_i$ denotes the i-th sample from the prior distribution $\pi(\boldsymbol{\theta})$. Though this approach is simple, naive Monte Carlo is generally inefficient. Typically, the likelihood is concentrated in a smaller region of the parameter space compared to the prior. As a result, most values sampled from the prior yield likelihood values close to zero, which increases the variance of the estimator and can require a vast number of samples to accurately estimate the marginal likelihood.

2.3.2 Importance Sampling

To remedy the limitation of Naive Monte Carlo, Tokdar and Kass [2010] introduces a proposal based on importance sampling, which emphasizes sampling from a high density region of the parameter space. Say we are able to sample from a distribution Q with density $q(x)$, then:

$$\begin{aligned} E_p[m(x)] &= \int p(x)m(x)dx \\ &= \int q(x) \left[\frac{p(x)}{q(x)} m(x)\right] dx \\ &= E_q \left[\frac{p(x)}{q(x)} m(x)\right] \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \frac{p(x_i)}{q(x_i)} m(x_i), \quad x_i \sim q(x). \end{aligned}$$

In this case, we are sampling from $q(x)$ and not $p(x)$. This is extremely useful when $p(x)$ is difficult to sample from but we can still evaluate the un-normalized density. Depending on the choice of $q(x)$, this method improve the variance of the estimate's distribution and thus makes the Monte Carlo trace converges faster. To

estimate the marginal likelihood $p(\boldsymbol{Y})$, we would rewrite the equation similarly:

$$\begin{aligned} p(\boldsymbol{Y}|M) &= \int\Big(p(\boldsymbol{Y}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= E_q\left[\frac{p(\boldsymbol{Y}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta}|M)}\right]\Big(\\ &\approx \frac{1}{N}\sum_{i=1}^{N}\frac{p(\boldsymbol{Y}|\boldsymbol{\theta}_i, M)\pi(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i|M)}, \quad \boldsymbol{\theta}_i \sim q(\boldsymbol{\theta}|M), \end{aligned}$$

where $q(\boldsymbol{\theta}|M)$ is called the importance density. According to Gronau et al. [2017], a good importance density should be easy to evaluate, share the same domain as the posterior distribution, resemble the posterior distribution closely, and have slightly fatter tails. Even though this approach is much more efficient than the Naive Monte Carlo procedure, the heavy reliance on the choice of importance density makes this approach less flexible in high dimensional cases. If the importance density does not adequately satisfy the conditions above, the algorithm performs poorly and result in high variance in estimates.

2.3.3 Harmonic Mean Estimator

Instead of sampling from the importance distribution, we can estimate the marginal likelihood by sampling from the posterior distribution with the harmonic mean estimator [Raftery et al., 2007]. In the Bayesian marginal likelihood estimation context, the harmonic mean is used in a more specialized way. We use the identity:

$$\begin{aligned} \frac{1}{p(\boldsymbol{Y}|M)} &= \int\Big(\frac{1}{p(\boldsymbol{Y}|M)}q(\boldsymbol{\theta}|M)\, d\boldsymbol{\theta} = E_{post}\left[\frac{q(\boldsymbol{\theta}|M)}{p(\boldsymbol{Y}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)}\right]\Big(\\ p(\boldsymbol{Y}|M) &\approx \left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{w(\tilde{\boldsymbol{\theta}}_i)}\right)^{-1} \\ &\approx \left(\frac{1}{N}\sum_{i=1}^{N}\frac{q(\tilde{\boldsymbol{\theta}}_i|M)}{p(\boldsymbol{Y}|\tilde{\boldsymbol{\theta}}_i, M)\pi(\tilde{\boldsymbol{\theta}}_i|M)}\right)^{-1}, \quad \tilde{\boldsymbol{\theta}}_i \sim \pi(\tilde{\boldsymbol{\theta}}|\boldsymbol{Y}, M), \end{aligned}$$

where $E_{post}[.]$ means that we are taking expectation with respect to the posterior density. Different from importance sampling, we are sampling $\tilde{\boldsymbol{\theta}}_{\boldsymbol{i}}$ from the posterior distribution and not the importance density. An appropriate importance density in this case should exhibit thinner tails compared to the posterior distribution [Gronau et al., 2017] and share same other conditions as the importance sampling approach. Thus, this approach also suffers in high dimensions and a poor choice of importance density similar to importance sampling. Moreover, since the harmonic mean involves the reciprocal of likelihoods, if any likelihood values approach zero, the estimator can become extremely large or even unbounded. The estimator is also sensitive to outliers and tends to produce poor results in high-dimensional spaces or models with complex likelihood functions. To reduce the potential high variance in high dimensional cases, bridge sampling was introduced.

2.3.4 Bridge Sampling

The bridge sampling method was introduced in the late 90s [Meng and Wong, 1996], and a R-package *bridgsampling* was later popularized by Gronau et al. [2017]. The idea is as the name suggested, constructing a bridge between the two probability distributions: the posterior distribution of the model parameters (which is typically complex and difficult to integrate) and a simpler, well-known distribution (often a multivariate normal distribution) referred to as the proposal distribution. By evaluating the relative densities of these two distributions at strategically chosen points, bridge sampling allows for efficient and accurate estimation of the marginal likelihood

$p(\boldsymbol{Y}|M)$:

$$
\begin{aligned}
p(\boldsymbol{Y} \mid M) &= \frac{\int p(\boldsymbol{Y} \mid \boldsymbol{\theta}, M)\pi(\boldsymbol{\theta} \mid M)h(\boldsymbol{\theta} \mid M)q(\boldsymbol{\theta} \mid M)d\boldsymbol{\theta}}{\int \frac{p(\boldsymbol{Y}|\boldsymbol{\theta},M)\pi(\boldsymbol{\theta}|M)}{p(\boldsymbol{Y}|M)} h(\boldsymbol{\theta} \mid M)q(\boldsymbol{\theta} \mid M)d\boldsymbol{\theta}} \\
&= \frac{E_{q(\theta)}[p(\boldsymbol{Y} \mid \boldsymbol{\theta}, M)\pi(\boldsymbol{\theta} \mid M)h(\boldsymbol{\theta} \mid M)]}{E_{\pi(\theta|Y,M)}[h(\boldsymbol{\theta} \mid M)q(\boldsymbol{\theta} \mid M)]} \\
&\approx \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} p(\boldsymbol{Y} \mid \boldsymbol{\theta}_i^*, M)\pi(\boldsymbol{\theta}_i^* \mid M)h(\boldsymbol{\theta}_i^* \mid M)}{\frac{1}{N_1}\sum_{j=1}^{N_1} h(\tilde{\boldsymbol{\theta}}_j \mid M)q(\tilde{\boldsymbol{\theta}}_j \mid M)],}
\end{aligned}
$$

where $\tilde{\boldsymbol{\theta}}_j \sim \pi(\boldsymbol{\theta}|Y, M)$, the posterior distribution and $\boldsymbol{\theta}_i^* \sim q(\boldsymbol{\theta}|M)$, the proposal distribution. And $h(\boldsymbol{\theta}|M)$ is the bridge function defined as:

$$
h(\boldsymbol{\theta}|M) = C \cdot \frac{1}{s_1 p(\boldsymbol{Y}|\boldsymbol{\theta}, M) + s_2 p(\boldsymbol{Y}|M)q(\boldsymbol{\theta}|M)},
$$

where $s_1 = \frac{N_1}{N_1+N_2}$, $s_2 = \frac{N_2}{N_1+N_2}$, and C is a constant.

Bridge Sampling Framework

The implementation of bridge sampling begins with drawing $2N_1$ posterior samples $\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_{2N_1}$ from the posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}, M)$, where $\boldsymbol{\theta}$ represents model parameters. These samples are typically obtained using Markov Chain Monte Carlo (MCMC) methods. The posterior samples are divided into two equally sized batches of N_1 samples each and transformed onto the real line $\tilde{\boldsymbol{\phi}}_1, \tilde{\boldsymbol{\phi}}_2, \ldots, \tilde{\boldsymbol{\phi}}_{2N_1}$, if necessary.

Next, a proposal distribution $q(\boldsymbol{\phi} \mid M)$ is selected. Typically, a multivariate normal distribution is used as the proposal, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ estimated from the first batch of transformed posterior samples $q(\boldsymbol{\phi} \mid M) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\phi}$ represents the parameters transformed onto the real line. N_2 samples $\boldsymbol{\phi}_1^*, \boldsymbol{\phi}_2^*, \ldots, \boldsymbol{\phi}_{N_2}^*$ are then drawn from this proposal distribution and inverse transformed to the original space $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots, \boldsymbol{\theta}_{N_2}^*$, if needed.

Using the N_2 proposal samples, we evaluate the un-normalized posterior density as $q_{21,i} = p(\boldsymbol{Y} \mid \boldsymbol{\theta}_i^*, M)\pi(\boldsymbol{\theta}_i^*)$, and the un-normalized proposal density as $q_{22,i} = q(\boldsymbol{\phi}_i^* \mid$

M). We define $l_{2,i} = q_{21,i}/q_{22,i}$ to represent the density ratio for the i-th proposal samples.

Similarly, we use the second batch of N_1 posterior samples to evaluate the un-normalized log posterior density and the un-normalized log proposal density. Specifically, for each sample in this batch, we compute the un-normalized posterior density, denoted as $q_{11,j} = p\left(\boldsymbol{Y} \mid \tilde{\boldsymbol{\theta}}_j, M\right) \pi\left(\tilde{\boldsymbol{\theta}}_j\right)$, and the un-normalized proposal density, denoted as $q_{12,j} = q\left(\tilde{\boldsymbol{\phi}}_j \mid M\right)$. We then define the density ratio for the j-th posterior samples as $l_{1,j} = q_{11,j}/q_{12,j}$.

Iterative Scheme for Marginal Likelihood Estimation

Once we have obtained the density ratio for the posterior samples and proposal samples, we iteratively estimate the marginal likelihood $p(\boldsymbol{Y} \mid M)$ using the update proposed by Meng and Wong [1996, p. 837]. The updated value at iteration $t+1$ is computed as:

$$p(\boldsymbol{Y} \mid M)^{(t+1)} = \frac{1}{N_2} \sum_{i=1}^{N_2} \frac{l_{2,i}}{s_1 l_{2,i} + s_2 p(\boldsymbol{Y} \mid M)^{(t)}} \Bigg/ \frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 l_{1,j} + s_2 p(\boldsymbol{Y} \mid M)^{(t)}},$$

where $p(\boldsymbol{Y}|M)^{(0)}$ is usually set as 1, and N_1 is set to be equal to N_2, and thus, $s_1 = s_2 = 1/2$.

This update is repeated until convergence, typically determined by the relative change in the marginal likelihood estimate falling below a small threshold such as 10^{-6}. The iterative nature of this scheme helps stabilize the estimate and reduces the variance inherent in direct importance sampling approaches. By leveraging both posterior and proposal samples, bridge sampling provides a robust and efficient means of computing marginal likelihoods, which are essential for Bayesian model comparison.

2.4 Standard Bayes Factors and Calibrated Bayes Factors

2.4.1 Standard Bayes Factors

The Bayes factor uses the marginal likelihood to compare two models directly. It is defined as a ratio of the marginal likelihoods of two competing models, M_1 and M_2,

$$BF_{12} = \frac{p(\boldsymbol{Y}|M_1)}{p(\boldsymbol{Y}|M_2)}.$$

This ratio provides a quantitative measure of the relative evidence in favor of one model over another. A Bayes factor greater than 1 indicates that model M_1 is more strongly favored by the observed data as compared to M_2, while a value less than 1 favors M_2. The strength of evidence can be interpreted using conventional thresholds; for example, a Bayes factor between 3 and 10 indicates moderate evidence, while values above 10 suggest strong evidence for one model over the other [Kass and Raftery, 1995] (see also [Jeffreys, 1961]). More details regarding the marginal likelihood and Bayes factor can be found at Gelman et al. [2013].

2.4.2 Calibrated Bayes Factors

One common issue when using Bayes Factor is that Bayesian inference typically starts with non-informative or weakly informative priors, allowing the data to drive parameter estimation without strong subjective assumptions. While this flexibility is advantageous, it can lead to computational inefficiencies, particularly in high-dimensional models where diffuse priors increase variance and slow Markov chain Monte Carlo (MCMC) convergence.

To mitigate this issue, Xu et al. [2019] propose using training samples to calibrate prior distributions, ensuring they provide sufficient information for stable inference without unduly influencing posterior estimates. We adopt a similar approach,

leveraging a subset of the data to construct weakly informative priors that improve estimation efficiency.

The calibrated Bayes factor modifies the standard Bayes factor by incorporating information from a calibration sample. Suppose that we select a subset of the data, called the training sample, and use it to update the prior distributions for both models. The calibrated Bayes factor is then defined as:

$$\log CB_{12}(\boldsymbol{Y}) = \log BF_{12}(\boldsymbol{Y}) - \frac{1}{H}\sum_{h=1}^{H}\log BF_{12}(\boldsymbol{Y}^{(h)}),$$

where $\boldsymbol{Y}^{(h)}$ denotes the h-th randomly selected subset of the data used for calibration, and H is the total number of such subsets. Essentially, the CBF adjusts the original Bayes factor by removing the average contribution from the calibration samples.

2.5 Prior Calibration Using Training Samples

To implement this, we first select a representative subset $\boldsymbol{Y}_{sub}$ from the full dataset $\boldsymbol{Y}$. This subset should be large enough to capture key characteristics of the full data while remaining computationally feasible. We fit the model to this subset and extract summary statistics such as posterior means and variances from this subset are then used to inform the prior specification for the model.

This strategy improves the stability of posterior inference in the model, reducing sensitivity to initial parameter choices and minimizing variance in marginal likelihood estimates. By balancing non-informative priors with data-driven weakly informative priors, this method enhances computational efficiency while preserving the flexibility of Bayesian inference.

Chapter 3: Item Response Theory (IRT) Models

Item Response Theory (IRT) is a family of models widely used in educational testing and psychometrics to scale binary responses into continuous latent constructs, describing characteristics of individuals and items. IRT models can include different numbers of parameters that describe the characteristics of the items. The standard IRT models include the 1-Parameter Logistic (1PL) model, the 2-Parameter Logistic (2PL) model, and the 3-Parameter Logistic (3PL) model, where the name derives from the number of parameters associated to each item.

In this section, we introduce model notation in the context of educational assessment, where typically data are students' responses to exam items, with the latent trait interpreted as their ability. Throughout the paper, we will focus on the 2PL model.

3.1 Parametric IRT Models

The observed data in Item Response Theory (IRT) models are typically binary responses, denoted as y_{ip}, where $y_{ip} = 1$ indicates that the individual p answers item i correctly and $y_{ip} = 0$ otherwise. The probability of a correct response, denoted by π_{ip}, is modeled conditionally on a set of latent parameters encoding characteristics of the individuals and items. In the 2PL model, each item has a difficulty parameter β_i,

which reflects how challenging the item is, and a discrimination parameter λ_i, which measures how well the item differentiates between individuals of different ability levels. The individual's latent ability η_p represents their underlying skill, with higher values corresponding to higher ability.

In the 2PL model, the probability of a correct answer is modeled as:

$$\pi_{ip} = P(y_{ip} = 1|\eta_p, \beta_i, \lambda_i) = \frac{1}{1 + \exp(-\lambda_i(\eta_p - \beta_i))},$$

$$i = 1, \ldots, I, \quad p = 1, \ldots, P,$$

where I is number of items and P is number of individuals.

A high value difficulty parameter β_i makes the item harder to be answered correctly, requiring individuals with higher abilities η_p to achieve a 50% chance of answering correctly. When $\eta_p = \beta_i$, the probability of a correct response is 50%. If $\eta_p > \beta_i$, we have that $\pi_{ip} > 0.5$, indicating a higher likelihood of a correct response. If $\eta_p < \beta_i$, then $\pi_{ip} < 0.5$ and the item is less likely to be answered correctly. We assume that the discriminations λ_i are positive and the latent traits η_p follow a standard normal distribution.

A 1PL model assumes that all items have the same discrimination, typically by setting $\lambda_i = 1$, simplifying the logistic function. This makes the model suitable when the items are of similar quality in distinguishing between individuals of varying abilities. A 3PL model adds a guessing parameter for each item to the 2PL (typically denoted as c_i), which accounts for the possibility that individuals may answer correctly by guessing.

In Bayesian settings, maintaining the normality assumption for the latent trait, we include priors for the parameters. The complete 2PL IRT model specification is

as follows:

$$y_{pi} \sim \text{Bernoulli}\left(\left(\frac{1}{1+\exp(-\lambda_i(\eta_p - \beta_i))}\right)\right($$

$$\beta_i \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2), \quad \log(\lambda_i) \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2), \quad \eta_p \sim \mathcal{N}(0,1),$$

$$i = 1, ..., I, \quad p = 1, ..., P,$$

assuming conditional independence of responses across individuals and items. This model, as well as the 1-PL and 3-PL IRT models, are parametric, meaning that they assume all individuals are drawn from a population with a single, well-defined latent trait distribution. While these models are computationally efficient and widely used, they struggle to capture the complexity of real-world data where latent traits may be heterogeneous across sub-populations. This limitation is particularly noticeable when individuals come from different cultural or educational backgrounds, which might require more flexible models to accurately capture the variability in their latent traits. To address this limitations, one can relax the normality assumption and model the ability using a mixture of Gaussian distribution either using finite or infinite mixtures.

3.2 Finite Mixture IRT Models

The finite mixture IRT model extends the standard 2PL model by allowing the latent trait distribution to consist of a finite mixture of normal components. This approach accommodates heterogeneity in the population's latent traits by assuming that individuals belong to one of K latent subgroups, each with its own distribution of abilities.

Let $\boldsymbol{z} = (z_1, ..., z_P)$ be a vector denoting the subgroup allocation for each individual, with $z_p \in 1, \ldots, K$, the model is specified as:

$$y_{pi} \sim \text{Bernoulli}\left(\frac{1}{1+\exp(-\lambda_i(\eta_p - \beta_i))}\right)\Bigg($$

$$\beta_i \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2), \quad \log(\lambda_i) \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2), \quad z_p \sim \text{Categorical}(w_1, \ldots, w_K),$$

$$\eta_p \mid z_p = k \sim \mathcal{N}(\mu_k, \sigma_k^2), \quad k = 1, \ldots, K,$$

$$i = 1, ..., I, \quad p = 1, ..., P.$$

Similar to the standard 2PL model, we assume that item parameters are independently distributed across items and independent of the person-level parameters. Given their cluster assignments z_p, the latent abilities η_p are conditionally independent across individuals. The prior over z_p induces dependence in η_p marginally, but conditional independence is preserved within clusters. The mixture weights $\boldsymbol{w} = (w_1, ..., w_k)$ are modeled using a Dirichlet prior:

$$\boldsymbol{w} \sim \text{Dirichlet}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right)\Bigg($$

The cluster means and variances are typically assigned hierarchical priors:

$$\mu_k \sim \mathcal{N}(0, \sigma_k^2/\gamma), \quad \sigma_k^2 \sim \text{Inverse-Gamma}(a, b),$$

where γ is a constant scale factor. This structure allows the model to identify subgroups within the population that share similar latent traits, improving model flexibility and fit. Unlike the standard 2PL model, which assumes a single normal distribution for abilities, the finite mixture model captures multimodal or skewed distributions in the population. This makes it particularly useful when the data suggest the existence of distinct subpopulations, such as students from different educational backgrounds. Alternatively, one can introduce infinite mixture models or semi-parametric

IRT models, which extend this approach by allowing the number of latent subgroups to be potentially infinite or determined by the data. These models remove the need to pre-specify the number of clusters K and provide greater flexibility in capturing complex latent trait distributions.

3.3 Semi-Parametric IRT Models

An alternative to the finite mixture approach is the use of semi-parametric or Bayesian nonparametric (BNP) models. These models further relax the assumption of a fixed number of latent subgroups by allowing the number of clusters to grow with the data. A commonly used BNP method is the Chinese Restaurant Process (CRP), which assigns individuals to clusters in a probabilistic manner. The process encourages the formation of new clusters as more data are observed, regulated by a concentration parameter α. An accessible overview of Bayesian nonparametric models and the CRP can be found in Li et al. [2019].

The semi-parametric IRT model can be expressed as:

$$y_{pi} \sim \text{Bernoulli}\left(\left(\frac{1}{1+\exp(-\lambda_i(\eta_p-\beta_i))}\right)\right($$

$$\beta_i \sim \mathcal{N}(0,\sigma_\beta^2), \quad \log(\lambda_i) \sim \mathcal{N}(0,\sigma_\lambda^2),$$

$$\eta_p \mid z_p = k \sim \mathcal{N}(\mu_k,\sigma_k^2), \quad \boldsymbol{z} \sim \text{CRP}(\alpha),$$

$$\mu_k \sim \mathcal{N}(0,\sigma_k^2/\gamma), \quad \sigma_k^2 \sim \text{Inverse-Gamma}(a,b),$$

$$k = 1,\ldots,K, \quad i = 1,...,I, \quad p = 1,...,P.$$

Here, the independence assumption is conceptually the same as the finite mixture model and the cluster assignment vector $\boldsymbol{z}$ is determined by the CRP, and the cluster parameters μ_k and σ_k^2 have their own priors. This model is highly flexible, allowing the

data to determine both the number and structure of clusters without pre-specifying K.

While the semi-parametric approach offers substantial flexibility, it is computationally intensive and can be less interpretable than finite mixture models. Therefore, in this chapter, we primarily focus on the parametric and finite mixture models, which provide a good balance between flexibility, interpretability, and computational feasibility.

3.4 Identifiability

A key aspect of IRT models is the issue of identifiability—whether the model parameters can be uniquely determined from the observed data. In the standard parametric IRT models, the assumption of a normal distribution for the latent ability plays an important role in ensuring identifiability. This assumption allows the model parameters to be well-defined and separable from one another.

However, when the normality assumption is relaxed, as in finite mixture or semi-parametric IRT models, identifiability becomes more challenging. In these models, additional constraints are typically required to avoid issues such as label switching or unidentifiable location and scale parameters. For example, constraints on the ordering of cluster means or fixing certain parameters may be imposed to guarantee identifiability.

Since the primary focus of this study is on the estimation of the marginal likelihood for model comparison purposes, and not on the detailed inference of model parameters, we do not delve into the technical aspects of identifiability here. For a

thorough discussion of identifiability conditions in mixture IRT models and related models, see Section 2.3 of Paganin et al. [2023].

Chapter 4: Marginal Likelihood Estimation for IRT models

This chapter discusses our experiments in estimating the marginal likelihood for IRT models. Among the approaches presented in Chapter 2, we focused on using bridge sampling. However, we found that direct application of bridge sampling proved difficult for the IRT models described in Chapter 3, due to the large parameter space implied by these models, where we have one parameter for each individual and at least one per item.

In addition, when using Finite Mixture IRT Models, the discrete nature of the cluster assignments presents additional obstacles. Specifically, it increases the dimensionality of the parameter space by adding P additional parameters—one for each individual. Since bridge sampling requires parameters to reside in a continuous, real-valued space, including the categorical cluster labels necessitates a transformation. However, there is no valid or principled way to transform these discrete variables to the real line and invert that transformation without distorting the target distribution. Even if a transformation were forced, it would introduce instability in the iterative bridge sampling scheme, leading to numerical issues and unreliable estimates. One can integrate the discrete parameters by marginalizing all the cluster assignments to use bridge sampling with only continuous parameters. However, this does not significantly reduce the dimensions of the posterior space.

As a solution, we propose an "hybrid" approach. The idea is to calculate the marginal likelihood by considering one subset of the parameters at a time. For the first subset (latent abilities), we would like to use numerical integration, while for the rest (item parameters), we rely on bridge sampling.

4.1 Optimized Log Likelihood Computation for marginal likelihood estimation

To improve computational efficiency while maintaining accuracy in estimating the marginal likelihood, we derive an optimized formulation for the log-likelihood for 1PL and 2PL IRT models. The goal of marginal likelihood estimation is to evaluate the integral of the joint density of the observed data and all model parameters over the parameter space:

$$m(\boldsymbol{Y}) = \int\!\!\int\!\!\int \pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\eta})\, p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\eta})\, d\boldsymbol{\beta}\, d\boldsymbol{\lambda}\, d\boldsymbol{\eta}.$$

This integral is typically approximated using Monte Carlo methods, such as bridge sampling, which require sampling from the posterior distribution of all parameters. In this section, we propose an alternative strategy to simplify this computation, taking advantage of the structure of IRT models, where item parameters and person parameters are typically assumed mutually independent. We rewrite the multidimensional integral above as:

$$\begin{aligned}
&\int\!\!\int\!\!\int \pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\eta})\, p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\eta})\, d\boldsymbol{\beta}\, d\boldsymbol{\lambda}\, d\boldsymbol{\eta} \\
&\qquad = \int_{\mathcal{L}} \int_{\mathcal{B}} \left[\int_{\mathcal{H}} \int\!\!\int p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\eta}) \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} \right] \pi(\boldsymbol{\beta}) \pi(\boldsymbol{\lambda}) d\boldsymbol{\beta} d\boldsymbol{\lambda} \\
&\qquad = \int_{\mathcal{L}} \int_{\mathcal{B}} \int\!\!\int p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) \pi(\boldsymbol{\beta}) \pi(\boldsymbol{\lambda}) d\boldsymbol{\beta} d\boldsymbol{\lambda},
\end{aligned}$$

where $p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda})$ is sometimes referred to as an integrated likelihood and:

- $\mathcal{L} = (\mathbb{R}^+)^I$ is the space of discrimination parameters $\boldsymbol{\lambda}$,
- $\mathcal{B} = \mathbb{R}^I$ is the space of difficulty parameters $\boldsymbol{\beta}$,
- $\mathcal{H} = \mathbb{R}^P$ is the space of latent abilities $\boldsymbol{\eta}$.

We propose to approximate the integral over the latent abilities $\boldsymbol{\eta}$ analytically or numerically, reducing the dimensionality of the integral to be estimated by bridge sampling. Specifically, since individuals' abilities are *i.i.d.*, we precompute the integral over η_p for each individual, so that the remaining marginal likelihood computation only involves the item parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ and mixture parameters in the case of finite mixture models. This strategy simplifies the computation and improves the numerical stability of the marginal likelihood estimation.

Considering an individual p and an item i, the likelihood function in a two-parameter logistic (2PL) item response theory (IRT) model is given by:

$$P(y_{pi} \mid \eta_p, \beta_i, \lambda_i) = \frac{\exp(y_{pi}\lambda_i(\eta_p - \beta_i))}{1 + \exp(\lambda_i(\eta_p - \beta_i))},$$

where y_{pi} represents the binary response of individual p to item i, λ_i is the discrimination parameter, and β_i is the difficulty parameter.

In both the standard 2PL and finite mixture 2PL model, the integrated likelihood of the observed data conditional on the model parameters is expressed as:

$$p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) = \prod_{i=1}^{I} \prod_{p=1}^{P} \left(\int_{\mathcal{H}} P(y_{pi} \mid \eta_p, \beta_i, \lambda_i) f(\eta_p)\, d\eta_p, \right.$$

and the integrated log-likelihood function is therefore:

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{i=1}^{I} \sum_{p=1}^{P} \left(\log \int_{\mathcal{H}} P(y_{pi} \mid \eta_p, \beta_i, \lambda_i) f(\eta_p)\, d\eta_p, \right.$$

where $f(\eta_p)$ denotes the density of the latent ability distribution, which may be standard normal in the standard 2PL model or a mixture of normal distributions in the finite mixture 2PL model.

Using numerical approximation, we can estimate the integral by approximating the continuous function $f(x)$ as discrete distribution over R points:

$$\int_{\eta} P(y_{pi} \mid \eta_p, \beta_i, \lambda_i) f(\eta_p) d\eta_p \approx \sum_{r=1}^{R} \tilde{f}(\eta^{(r)}) \frac{\exp(y_{pi}\lambda_i(\eta^{(r)} - \beta_i))}{1 + \exp(\lambda_i(\eta^{(r)} - \beta_i))},$$

where

$$\tilde{f}(\eta^{(r)}) = \frac{f(\eta^{(r)})}{\sum_{s=1}^{R} f(\eta^{(s)})}.$$

For computational efficiency, we precompute $f(\eta^{(r)})$ for a grid of η and normalize it with $\tilde{f}(\eta^{(r)})$. We found that typically using $R = 101$ grid points is sufficient.

Given the independence of item responses and using the approximation above, finally, the integrated log-likelihood can be approximated as:

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) \approx \sum_{i=1}^{I} \quad n_{0i} \log \sum_{r=1}^{R} \tilde{f}(\eta^{(r)}) \frac{1}{1 + \exp(\lambda_i(\eta^{(r)} - \beta_i))} + n_{1i} \log \sum_{r=1}^{R} \tilde{f}(\eta^{(r)}) \frac{\exp(\lambda_i(\eta^{(r)} - \beta_i))}{1 + \exp(\lambda_i(\eta^{(r)} - \beta_i))},$$

where n_i^0 and n_i^1 represent the number of individuals who responded 0 or 1 to item i, respectively. This formulation allows for a direct computation of the marginal likelihood without explicitly sampling η_p.

4.2 Results for the Finite Mixture IRT Models

In our finite mixture model, we assume that the latent trait distribution follows a mixture of Gaussian components:

$$f(\eta_p \mid \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\eta_p | \mu_k, \sigma_k^2),$$

where the mixture weights follow a symmetric Dirichlet distribution:

$$\boldsymbol{w} \sim \text{Dirichlet}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right)($$

The prior structure for the cluster-specific parameters is given by:

$$\mu_k|\sigma_k^2 \sim \mathcal{N}\left(\mu_0, \frac{\sigma_k^2}{\gamma}\right)(\quad \sigma_k^2 \sim \text{Inv-Gamma}(\alpha, \beta),$$

where γ is a scale parameter, that we consider fixed. To derive the marginal likelihood, we integrate out the latent cluster assignments and the mixture component parameters:

$$f(\eta_p \mid \boldsymbol{w}) = \sum_{k=1}^{K}(w_k \int\int(p(\eta_p|\mu_k, \sigma_k^2)p\left(\mu_k|\mu_0, \frac{\sigma_k^2}{\gamma}\right) p(\sigma_k^2|\alpha, \beta)\, d\mu_k\, d\sigma_k^2.$$

Using the completing-the-square technique and the probability density function (PDF) transformation trick for normal and inverse-gamma distributions, integrating out μ_k and σ_k^2 results in a non-central t-distribution for η_p within each cluster:

$$\eta_p \mid k \sim t_{2\alpha}\left(\mu_0, \frac{(1+\gamma)\beta}{\gamma\alpha}\right)($$

where 2α is the degrees of freedom, μ_0 is the location parameter, and the scale parameter is given by $\frac{(1+\gamma)\beta}{\gamma\alpha}$.

To obtain the marginal distribution of η_p, we average over the mixture weights:

$$f(\eta_p \mid \boldsymbol{w}) = \sum_{k=1}^{K}(w_k \cdot t_{2\alpha}\left(\mu_0, \frac{(1+\gamma)\beta}{\gamma\alpha}\right).$$

This marginalization approach offers computational advantages. By integrating out the cluster-specific parameters and directly evaluating the marginal distribution of η_p as a mixture of t-distributions, we avoid the need to explicitly sample or iterate over all P individuals' latent abilities in each step of the computation. This is particularly

beneficial when the sample size P is large, as it reduces the computational burden and accelerates the evaluation of the marginal likelihood. Consequently, this strategy leads to faster convergence and improved scalability of the bridge sampling procedure in high-dimensional settings.

4.3 Prior Calibration in the 2PL Finite Mixture IRT Model

As discussed in Section 2.4, we consider a prior calibration approach to mitigate the effect of vague priors when computing Bayes factors. In particular, we fit the model using a subset of the data. Using the posterior samples, we estimate the hyper-parameters of item parameters ($\boldsymbol{\beta}$, $\boldsymbol{\lambda}$) using moment matching, as well as parameters of the mixture components for the Finite Mixture IRT Models.

Thus, we fit the finite mixture model outlined in Section 3.2 to a subset of the data, consisting of 200 randomly selected individuals out of 1000 (more will be discussed in Chapter 5). The resulting posterior samples are then used to estimate the hyper-parameters for prior calibration in the full data analysis, via moment matching.

Our 2PL IRT Model with finite mixture after calibration would be:

$$y_{pi} \sim \text{Bernoulli}\left(\left(\frac{1}{1+\exp(-\lambda_i(\eta_p-\beta_i))}\right)\left(\right.\right.$$

$$\beta_i \sim \mathcal{N}(\hat{\mu}_\beta, \hat{\sigma}^2_\beta), \quad \log(\lambda_i) \sim \mathcal{N}(\hat{\mu}_\lambda, \hat{\sigma}^2_\lambda), \quad \eta_p \mid z_p = k \sim \mathcal{N}(\mu_k, \sigma^2_k),$$

$$z_p \sim \text{Categorical}(w_1, \ldots, w_K), \quad \boldsymbol{w} \sim \text{Dirichlet}(\hat{\boldsymbol{\alpha}}_{\boldsymbol{w}}),$$

$$\mu_k \sim \mathcal{N}(\hat{\mu}, \sigma^2_k/\gamma), \quad \sigma^2_k \sim \text{Inv-Gamma}(\hat{\alpha}, \hat{\beta}),$$

$$k = 1, \ldots, K, \quad i = 1, ..., I, \quad p = 1, ..., P,$$

where the estimated hyperparameters used for prior calibration are:

$$\hat{\mu}_\beta, \hat{\sigma}^2_\beta, \hat{\mu}_\lambda, \hat{\sigma}^2_\lambda, \hat{\mu}, \hat{\gamma}, \hat{\alpha}, \hat{\beta}, \hat{\boldsymbol{\alpha}}_{\boldsymbol{w}}.$$

These were obtained via moment matching based on posterior samples from the subset of 200 individuals. This calibrated model allows for more efficient posterior inference and facilitates more stable marginal likelihood estimation using bridge sampling.

4.4 Prior Calibration in the 2PL Basic IRT Model

For the standard 2PL IRT model, we also consider the possibility of applying prior calibration as discussed in Section 2.4. Our 2PL IRT Model after calibration would be:

$$y_{pi} \sim \text{Bernoulli}\left(\left(\frac{1}{1+\exp(-\lambda_i(\eta_p-\beta_i))}\right)\right($$

$$\beta_i \sim \mathcal{N}(\hat{\mu}_\beta, \hat{\sigma}^2_\beta), \quad \log(\lambda_i) \sim \mathcal{N}(\hat{\mu}_\lambda, \hat{\sigma}^2_\lambda), \quad \eta_p \sim \mathcal{N}(0,1),$$

$$i = 1, ..., I, \quad p = 1, ..., P.$$

Note that for the traditional IRT model the standard normal distribution for the latent trait is considered as an assumption on the population rather than a prior, so we do not update that distribution. Then, we use these estimated hyper-parameters when fitting the model on the rest of the data.

Chapter 5: Simulations

5.1 Simulated Data

The primary goal of this study is to assess whether Bayes factors obtained via bridge sampling appropriately favor the true underlying model. Specifically, we aim to evaluate whether the Bayes factor supports the standard 2PL IRT model when the data-generating process is unimodal; and whether the Bayes factor favors the finite mixture IRT model when the data-generating process is bimodal.

In this study, we simulate synthetic data following the approach of Paganin et al. [2023]. We specify three different scenarios for the distribution of latent abilities: unimodal, bimodal, and multimodal. For each scenario, responses are generated for $P = 1000$ individuals across $I = 10$ binary items. The item parameters include discrimination parameters $\{\lambda_i\}_{i=1}^{10}$, sampled from a uniform distribution $\mathrm{U}(0.5, 1.5)$, and difficulty parameters $\{\beta_i\}_{i=1}^{10}$, equally spaced between -3 and 3.

For the unimodal scenario, the latent abilities η_p are generated from a normal distribution with mean 0 and variance 1.25^2. In the bimodal scenario, abilities are generated from a mixture of two normal distributions with equal weights, specifically $\mathcal{N}(-2, 1.25^2)$ and $\mathcal{N}(2, 1.25^2)$. For the multimodal scenario, abilities are drawn from a mixture of three components: two normal distributions $\mathcal{N}(-2, 1)$ and $\mathcal{N}(0, 0.5)$, and

one skew-normal distribution with location parameter 3, scale parameter 1, and shape parameter -3. The mixture weights for the multimodal case are set to $(0.2, 0.4, 0.4)$.

5.2 Results

5.2.1 Calibrated Bayes Factor Analysis

To compare the standard 2PL model with the finite mixture 2PL model, we use the Calibrated Bayes Factor (CBF) as defined in Section 2.4.2. The CBF accounts for prior calibration by subtracting the average log Bayes factor computed on a set of training subsets from the log Bayes factor on the full dataset. Specifically, for each simulation, we compute

$$\log CB_{12}(\boldsymbol{Y}) = \log BF_{12}(\boldsymbol{Y}) - \frac{1}{H}\sum_{h=1}^{H} \log BF_{12}(\boldsymbol{Y}^{(h)}),$$

where $\log BF_{12}(\boldsymbol{Y}) = \log p(\boldsymbol{Y} \mid \mathcal{M}_1) - \log p(\boldsymbol{Y} \mid \mathcal{M}_2)$ compares the finite mixture model ($\mathcal{M}_1$) to the standard 2PL model ($\mathcal{M}_2$) on the full dataset $\boldsymbol{Y}$, and $\log BF_{12}(\boldsymbol{Y}^{(h)})$ is the log Bayes factor computed from the h-th subset $\boldsymbol{Y}^{(h)}$. In our study, we set $H = 10$ and use the same 10 subsets for both prior calibration and computing the calibration term in the CBF. This averaging is done separately within each dataset.

To evaluate how the CBF behaves under repeated data simulation and estimation, we consider the following setup. Under each latent trait scenario (unimodal, bimodal, and multimodal), we simulate a single set of 50 datasets. For each of these datasets, we fit both models, obtain posterior samples via MCMC, and estimate the marginal likelihoods for the full dataset and each of the 10 calibration subsets using bridge sampling. To account for variability in the bridge sampling step, we repeat the procedure 10 times per dataset, each time drawing a new set of proposal samples

from the fitted proposal distribution. In each data simulation, the posterior samples and calibration subsets are held fixed; only the proposal samples vary across these repetitions.

A negative log CBF indicates that the standard 2PL model generalizes better than the mixture model, while a positive CBF favors the finite mixture model.

5.2.2 Unimodal Scenario Results

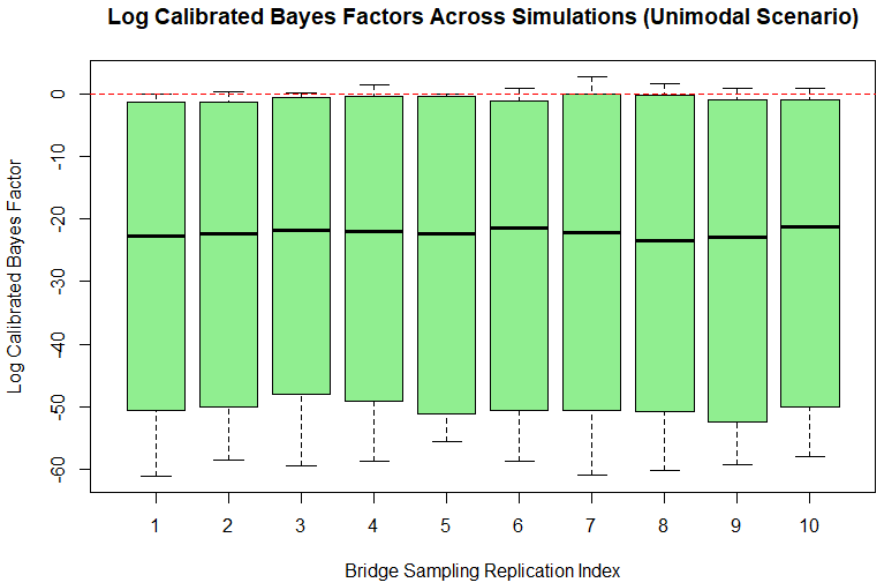

Figure 5.1: Log Calibrated Bayes Factors across simulations in the unimodal scenario. Each box represents variability across datasets for a fixed bridge sampling replication.

Figure 5.1 displays the distribution of log Calibrated Bayes Factors (CBFs) across 10 bridge sampling replications, each evaluated on a set of independently simulated datasets under the unimodal scenario. Each boxplot corresponds to a single

bridge sampling replication, with the spread reflecting variation across different data-generating simulations.

Overall, the CBF values are predominantly negative, indicating consistent preference for the standard 2PL model over the finite mixture model. This aligns with expectations, as the data were generated from a unimodal latent trait distribution, which the standard 2PL is designed to capture. While a few replications yield log CBFs that are close to or slightly above zero, suggesting occasional support for the mixture model, these cases are rare and do not overturn the broader trend. Additionally, the narrow spread of log CBF values within each replication highlights the numerical stability of bridge sampling, suggesting that it yields consistent marginal likelihood estimates even under repeated runs.

5.2.3 Bimodal Scenario Results

Figure 5.2 shows the distribution of log Calibrated Bayes Factors (CBFs) under the bimodal scenario, across 10 bridge sampling replications. Each box represents the variation in CBF values across independently generated datasets for a fixed bridge sampling repetition.

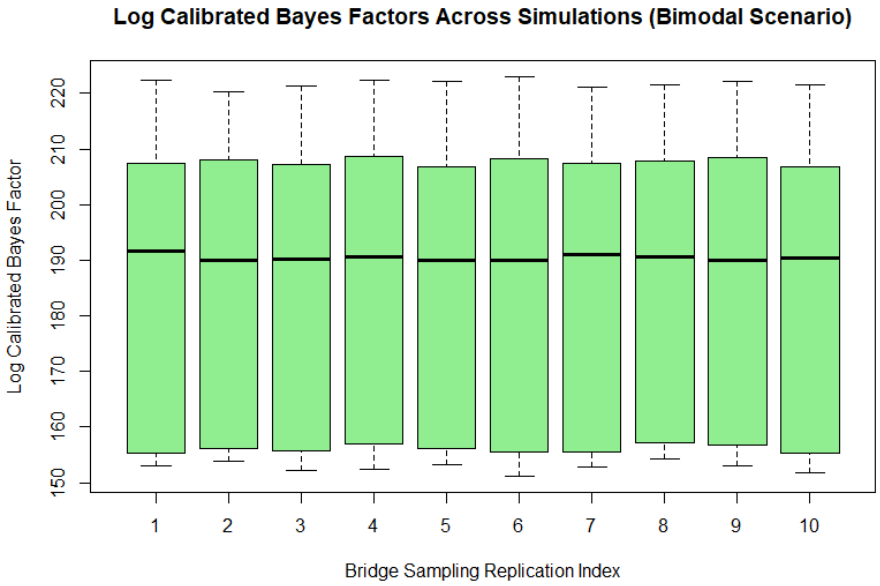

Figure 5.2: Log Calibrated Bayes Factors across simulations in the bimodal scenario. Each box represents variability across datasets for a fixed bridge sampling replication.

In contrast to the unimodal case, all CBF values here are strongly positive, consistently favoring the finite mixture model over the standard 2PL. This result aligns with the data-generating process, which involves a bimodal latent trait distribution that the mixture model is better equipped to capture. Moreover, the similarity in CBF magnitude and variability across replications reinforces both the stability of bridge sampling and the robustness of the model comparison conclusion. Under bimodal structure, the finite mixture model is clearly preferred.

5.2.4 Multimodal Scenario Results

Figure 5.3 displays the log Calibrated Bayes Factors (CBFs) across simulations in the multimodal scenario. As with the previous plots, each box represents the distribution of CBF values across datasets for a fixed bridge sampling replication.

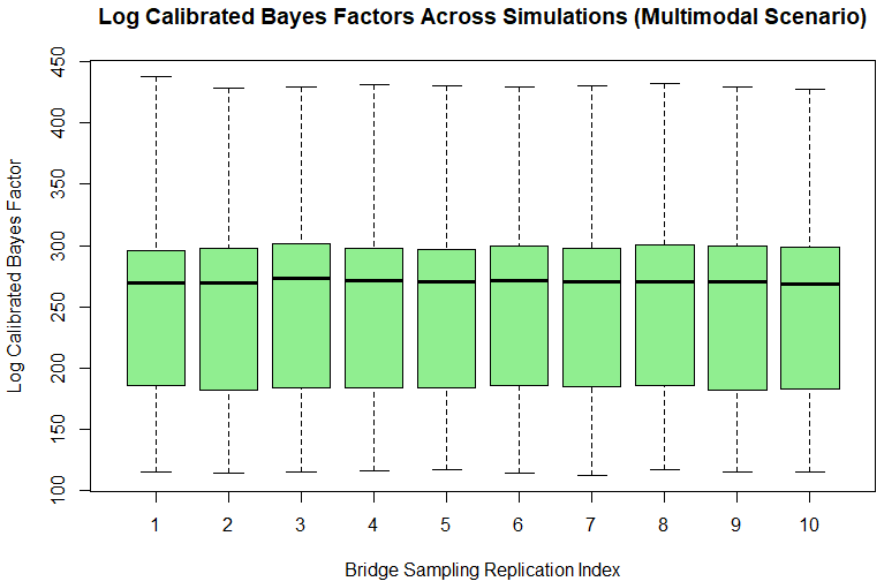

Figure 5.3: Log Calibrated Bayes Factors across simulations in the multimodal scenario. Each box represents variability across datasets for a fixed bridge sampling replication.

Compared to the bimodal case, the CBF values in the multimodal scenario are not only consistently positive, but substantially larger—reaching values as high as 450 in some simulations. This indicates an even stronger preference for the finite mixture model over the standard 2PL. The larger magnitude reflects the increased mismatch between the true data-generating process and the unimodal constraint imposed by the

standard model, which becomes more pronounced in the presence of multiple latent subpopulations.

Despite the increased complexity of the data, the CBF distributions remain relatively stable across bridge sampling replications. This further supports the reliability of bridge sampling in high-dimensional, nonstandard settings and affirms the mixture model's ability to flexibly adapt to heterogeneity in the latent trait distribution.

Chapter 6: Conclusions

6.1 Summary Findings

This thesis explored the use of marginal likelihood estimation for Bayesian model comparison in 2PL Item Response Theory (IRT) models, focusing on both the standard and finite mixture formulations. Through simulations under unimodal, bimodal, and multimodal latent ability distributions, we assessed model performance, generalizability, and sensitivity to prior calibration using calibrated Bayes factors (CBFs).

A key methodological contribution was the use of an integrated log posterior formulation. By numerically integrating over the latent ability parameters $\boldsymbol{\eta}$, we reduced the dimensionality of the posterior density and improved numerical stability in marginal likelihood estimation. This integration also removed dependence on latent class indicators $\boldsymbol{z}$, which are discrete parameters. Since bridge sampling requires continuous densities to function properly, marginalizing over these latent classes enabled the application of bridge sampling to finite mixture models, which would otherwise violate this assumption due to their use of discrete latent assignments.

Across all three scenarios, the marginal likelihood estimates obtained via bridge sampling exhibited strong numerical stability. For each simulated dataset, repeated bridge sampling runs produced tightly clustered estimates. This consistency confirms

the robustness of bridge sampling, even in models with hierarchical structures and latent mixture components.

Model comparison was conducted using the log calibrated Bayes factor (CBF), which evaluates the change in model evidence between subset and full data. Positive CBF values indicate support for the finite mixture model, while negative values favor the standard 2PL. In the unimodal scenario, CBF values ranged from approximately 0 to -60, reflecting consistent preference for the simpler 2PL model and suggesting overfitting by the mixture model. In contrast, both the bimodal and multimodal scenarios yielded strongly positive CBFs, ranging from 160 to 220 in the bimodal case, and up to 450 in the multimodal case, highlighting the mixture model's advantage in capturing latent heterogeneity when the true ability distribution is more complex.

Taken together, these findings support the use of finite mixture models when there is evidence of multimodality in the latent structure, while also emphasizing the value of CBFs as a tool for validating model generalization. The integration of the log posterior not only improves the theoretical validity of bridge sampling for mixture models but also enhances computational efficiency by drastically reducing dimensionality.

6.2 Limitations

While the use of integrated log posterior and bridge sampling provided accurate and efficient marginal likelihood estimation, several limitations remain. First, the current approach relies on numerical integration over a fixed grid for $\boldsymbol{\eta}$, which may not scale well for high-dimensional or adaptive models.

Another limitation lies in the fixed number of mixture components K, which was manually tuned to match the underlying data-generating process. Future work could explore the BNP models that allow the number of components to adaptively grow with data complexity, although this would require further methodological extensions to enable marginal likelihood computation.

Finally, while the calibrated Bayes factor is useful for assessing generalization, it assumes that the subset data is sufficiently representative and that the calibration process does not introduce bias. More sophisticated resampling or cross-validation strategies could be considered to better assess model robustness and reduce variability in marginal likelihood estimates across partitions.

Bibliography

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 3rd edition, 2013.

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

Harold Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, UK, 3rd edition, 1961.

Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. doi: 10.1080/01621459.1995.10476572.

Y. Li, E. Schofield, and M. Gönen. A tutorial on dirichlet process mixture modeling. *Journal of Mathematical Psychology*, 91:128–144, 2019. doi: 10.1016/j.jmp.2019.04.004.

Y. Liu, G. Hu, L. Cao, X. Wang, and M.-H. Chen. A comparison of monte carlo methods for computing marginal likelihoods of item response theory models. *Journal of the Korean Statistical Society*, 48(4):503–512, 2019. doi: 10.1016/j.jkss.2019.04.001.

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

S. Paganin, C. J. Paciorek, C. Wehrhahn, A. Rodríguez, S. Rabe-Hesketh, and P. de Valpine. Computational strategies and estimation performance with bayesian semiparametric item response theory models. *Journal of Mathematical Psychology*, 91:128–144, 2023. doi: 10.1016/j.jmp.2019.04.004.

A. E. Raftery, M. A. Newton, J. M. Satagopan, and P. N. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In *Bayesian Statistics*, volume 8, pages 371–416. Oxford University Press, 2007.

S. T. Tokdar and R. E. Kass. Importance sampling: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

X. Xu, P. Lu, S. N. MacEachern, and R. Xu. Calibrated bayes factors for model comparison. *Journal of Statistical Computation and Simulation*, 89(4):591–614, 2019. doi: 10.1080/00949655.2018.1563091.