Proceedings of the 4th International Conference on Statistics: Theory and Applications (ICSTA'22)
Prague, Czech Republic – July 28- 30, 2022
Paper No. 140
DOI: 10.11159/icsta22.140

Approximate Bayesian Inference for Educational Attainment Models

Shuhrah ALghamdi[1], Nema Dean[2], Ludger Evers[3]
[1] School of Mathematics and Statistics, University of Glasgow
Glasgow G12 8SQ, UK
s.alghamdi.3@research.gla.ac.uk; Nema.Dean@glasgow.ac.uk
[2] Princess Nourah Bint Abdul Rahman University
Riyadh, Saudi Arabi

Abstract **-** The rapidly expanding volume of educational testing data from online assessments has posed a problem for researchers in modern education. Their main goal is to utilise this information in a timely and adaptive manner to infer skills mastery, improving learning facilities and adapting them to individual learners. Over the past few years, a number of static statistical models have been proposed for extracting knowledge about skills mastery from item response data. However, realistic models typically lead to complex, computationally expensive fitting methods such as MCMC. So these methods will not tend to scale well for streaming data and large-scale real-time systems. The main objective of this paper is to develop approximate Bayesian inference based on the Laplace approximation method (LA), which allows faster inference. The LA estimation method's performance for the one-parameter logistic item response theory (IRT) model has been compared with the MCMC method in a simulation study. Based on the results of several comparison criterion methods such as bias, RMSE and Kendell's $\boldsymbol{\tau}$ measurement distance, the performance of the LA is very good in small, moderate, and relatively large sample size settings. The LA approximately estimated abilities results are very close to the actual values and sometimes even better than the estimated abilities resulting from MCMC. In addition, LA resulted in between a 120 to 900 times speedup over MCMC, making it a more practical alternative for large educational testing datasets.

Keywords**:** MCMC, Laplace approximation, Item Response Theory, Bayesian Inference, Streaming Data, Online Inference

1. Introduction

Item Response models are a common tool in educational research for estimating students' skill mastery based on test results. There have been many estimation techniques in Item Response Theory (IRT) models approaches explored in the literature, both Bayesian and frequentist. Researchers such as [1] suggested that Bayesian estimations methods can be useful for complex IRT models and for small data sets because of their ability to utilise expert prior information. With the help of modern computer techniques, Bayesian estimation methods have been widely used for IRT models via Markov chain Monte Carlo (MCMC) (see e.g. [2]). The challenge arises when the data arrives in real-time, and the parameters need to be estimated online. MCMC techniques may be unsuitable as they need to generate a different chain run for each posterior as new data arrives, and hence, are too computationally expensive and slow for streaming or large volumes of data. The Laplace approximation (LA) [3] is a mathematically simple and computationally cheap approximation method for Bayesian inference. In contrast to MCMC, LA only has to find the mode of the posterior rather than having to explore the whole space of posterior distribution, leading to a substantial speedup. In order to assess LA's usability, this paper aims to compare the performance of the LA method of estimation for the one-parameter logistic IRT model with the MCMC method in a simulation study.

2. Item Response Theory Model (IRT)

Item response theory (IRT) models demonstrate the relationship between the ability or attitude and an item response (e.g., questions, survey). These models can be categorized based on different factors such as the dimensionality of the ability, type of questions or the number of item parameters. The focus of the present paper is on the unidimensional ability one-parameter logistic (1PL) model (also called the Rasch model) [4], where all items measure one common ability, e.g., overall attainment of a subject. The model contains one item parameter which is the difficulty parameter:

The data of student responses to items will be represented as a matrix $\boldsymbol{X}$ where;

$$X_{ij} = \begin{cases} 1 & \text{if examinee } i \text{ answer item j correctly} \\ 0 & \text{if examinee i answer item j incorrectly,} \end{cases}$$

with $i = 1, 2, \ldots, n$ (number of rows) and $j = 1, 2, \ldots, m$ (number of columns). The 1PL probability of a correct response to an item j by examinee i can be written as:

$$p(X_{ij} = 1) = \frac{\exp(\theta_i - b_j)}{\left(1 + \exp(\theta_i - b_j)\right)}, \theta_i \text{ and } b_j \in \mathbb{R} \quad (1)$$

The parameter $\boldsymbol{b}$ represents the difficulty of the questions, where high (low) values of b_j means hard (easy) questions. The parameter $\boldsymbol{\theta}$ represents examinees' abilities, where high (low) values of θ_i mean high (low) levels of examinee skill. In this setting, the questions or items measure the same skill (unidimensional ability), and the examinees are assumed to answer all questions. Therefore, $X_{ij} \sim Bernoulli(\pi_{ij})$, giving the likelihood of this model as; $L(\boldsymbol{x} \mid \theta_i, b_j) = \prod_{i=1}^{n} \prod_{j=1}^{m} \pi_{ij}^{x_{ij}} (1 - \pi_{ij})^{1 - x_{ij}}$, where $\pi_{ij} = \frac{\exp(\theta_i - b_j)}{\left(1 + \exp(\theta_i - b_j)\right)}$, and $\text{logit}(\pi_{ij}) = (\theta_i - b_j)$.

3. Laplace Approximation Method

The Laplace approximation (LA) [3] is an analytical approximation method that aims to find a Gaussian approximation to a continuous target distribution. The idea behind the Laplace approximation is using the second-order Taylor expansion of the log-posterior of interest $p(\boldsymbol{\theta}|X) = \log p(X \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$ around its maximum $\widehat{\boldsymbol{\theta}}$, which corresponds to a Gaussian approximation at the mode. Formally, we have: $p(\boldsymbol{\theta} \mid X) \approx \mathcal{N}(\widehat{\boldsymbol{\theta}}, \mathbf{H}^{-1})$, where

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\theta} p(\boldsymbol{\theta} \mid X) = \arg\max_{\theta} p(X, \boldsymbol{\theta}) = \arg\max_{\theta} p(X \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) = \arg\max_{\theta} [\log p(X \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})], \quad (2)$$

and **H** is the Hessian matrix of the negative log-posterior at the mode

$$\mathbf{H} = -\nabla^2 \log p(\boldsymbol{\theta} \mid \mathrm{X})|_{\theta=\widehat{\theta}} = -\nabla^2 \log p(\mathrm{X}, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\nabla^2 [\log p(\mathrm{X} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})]|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \quad (3)$$

Therefore, at the first stage, we need to find the maximum points of the log posterior distributions for each parameter. For example, for the 1PL IRT model, we need to get the maximum point for each individual's ability θ_i and the maximum point for each question's difficulty b_j. In this study, the Broden-Fletcher-Goldfarb-Shanno (BFGS) iterative optimisation algorithm will be used (via the optim function in R [5]) to find the maximum points [6].

4. Comparison Study

The main objective of this comparison study is to evaluate the performance of the Laplace approximation method compared to MCMC method.

4.1. Comparison Criterion

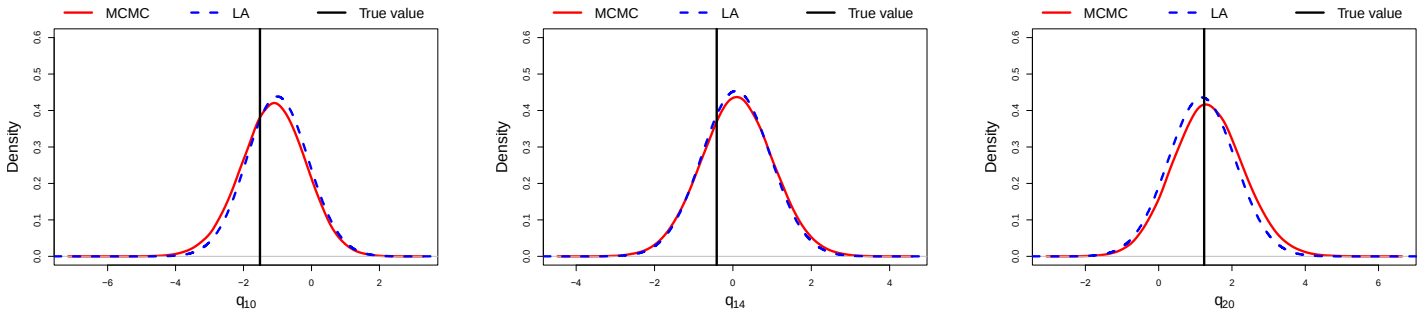
The performance of LA will be investigated in terms of comparison of distributions by comparing the shape of two posterior densities and measuring the distance using Jensen-Shannon divergence (JSD) [7]. Bias, RMSE and Kendell's $\boldsymbol{\tau}$ [8] will be used to assess and compare the accuracy of point estimates resulting from both methods. The run times will be recorded to determine the speedup offered by LA.

4.2. Simulated Data

Data is simulated from the1PL model with binary responses and unidimensional ability. The $\theta_i's$ range uniformly from -4 to 4, $b_j's$ range uniformly from -2 to 2 (ranges suggested by [9]). The comparison study will investigate the performance of the LA in small, moderate and relatively large sample sizes of students; n=30, 300, and 600. It will also consider a variety of numbers of items/questions; m= 10, 30, 50, 70 and 100. For each setting, the simulated dataset is repeated 20 times and the previously mentioned numerical measurements; bias, RMSE and Kendell's $\boldsymbol{\tau}$ are calculated and averaged for point estimates (mean after burn-in in MCMC, mode for LA).

4.3. Comparison Results

Figure 1 presents three different levels of randomly selected students' abilities. We can see that LA provides similar estimates of the ability's parameter $\boldsymbol{\theta}$, with the LA posterior mode closely matching MCMC for all three ability levels. However, because the posterior densities resulting from MCMC are not entirely symmetric, the mode of the two densities is not precisely in the same place. On the other hand, when the density of the posterior resulting from MCMC is almost symmetric, such as θ_{14}, the two posteriors' modes become almost identical. Moreover, the shape and the highest of the densities are nearly the same, suggesting that the approximate posterior distributions generated from LA explore the proper parameters space well and similarly to MCMC. The JSD divergences method is used to measure the dissimilarity between the resulting posterior distributions obtained from each method. The average JSD values in these experiments range between 0.005 to 0.24. However, this maximum value of 0.24 is relatively small since the original JSD ranges between 0 and 1, where 0 means the two distributions are identical, and 1 means strongly different. That indicates that the two resulting posterior distributions are similar. Although the posterior distribution densities presented here are for one specific simulated data, n=30 and m=10, the performance, and the relationship between the two methods in other simulated datasets are similar to this example. The main difference is that as we increase the sample sizes of students to 300 and 600, the maximum divergence between the two methods (JSD) rises to 0.4. However, about only 10 % of the resulting posterior densities of ability parameters are higher than 0.25. The largest difference (JSD values) between the two posterior distributions appears for very high/low ability students (i.e., at the extremes of the range).



(a) Number of Correct Answers= 3 (b) Number of Correct Answers= 5 (c) Number of Correct Answers= 7

Fig. 2: Posterior density plots for MCMC and LA methods of estimating selected examinees' abilities with different numbers of correct answers for sample size n=30 and m=10.

Table 1: Average bias, RMSE, and Kendall's $\boldsymbol{\tau}$ values between the point estimates and the true values for the ability parameter.

Sample size	**Item size**	**Bias**		**RMSE**		**Kendell's τ**	
		MCMC	**LA**	**MCMC**	**LA**	**MCMC**	**LA**
n=30	10	-0.005	-0.004	1.04	0.85	0.80	0.86
	30	-0.006	-0.007	0.75	0.58	0.88	0.90
	50	0.015	0.013	0.63	0.50	0.92	0.93
	70	-0.03	-0.02	0.57	0.44	0.93	0.94
	100	0.017	0.015	0.48	0.37	0.95	0.95
n=300	10	-0.0004	0	1.014	0.85	0.77	0.83
	30	-0.0004	-0.0002	0.71	0.59	0.86	0.88
	50	0	-0.0005	0.56	0.48	0.89	0.90
	70	-0.002	-0.003	0.48	0.42	0.90	0.91
	100	-0.0004	0	0.40	0.36	0.90	0.90
n=600	10	-0.0005	-0.0002	1.02	0.86	0.76	0.82
	30	0.0008	0.0012	0.70	0.58	0.86	0.88
	50	0.0004	0.002	0.54	0.47	0.89	0.90

	70	-0.0003	-0.001	0.46	0.40	0.90	0.91
	100	0	-0.0012	0.39	0.35	0.91	0.92

Regarding the point estimates of ability parameters, the results presented in Table 1 show that the biases for the LA were generally smaller than those from MCMC. The only exceptions to this result occurred for 50 questions and a sample size of 300 and for 600 sample size except test of length 10, in which cases the MCMC approach yielded the lowest biases. In addition, the RMSE for the LA estimates were lower than those of the MCMC estimator across all sample sizes and the number of questions, with the most noticeable differences in smaller sample sizes and shorter test lengths. Kendall's $\boldsymbol{\tau}$, which evaluates the degree of similarity between the order of actual abilities set and the order of estimated abilities set resulting from both methods, were generally larger for the LA method, with the most marked differences occurring with smaller sample sizes and shorter test lengths. The original value of $\boldsymbol{\tau}$ ranges from -1 to 1, where 1 means the two rankings are identical, and -1 means one is opposite of the other. These large tau values indicate that the rank order resulting from abilities estimates is very close to the actual abilities' order. The LA method has slightly better ordered abilities than the MCMC. However, Kendall's $\boldsymbol{\tau}$ values became almost identical for longer tests under both methods. In terms of the computational cost, the average computational time for maximum and minimum each sample size experiment was recorded in seconds and presented in Table 2. The MCMC took approximately 120 to 3729 seconds (2 to 62 minutes) to produce the results, while the LA took only 0.01 to 4 seconds. Even for large, simulated data, n=600 and m=100, the LA provided the results in only seconds.

Table 2: Comparison of the computation time between MCMC method and LA method.

Sample size	**Item**	**Time (in seconds)**	
		MCMC	LA
n=30	10	120	0.01
	100	442	0.07
n=300	10	871	0.33
	100	1874	0.89
n=600	10	1691	0.53
	100	3729	4.00

5. Conclusion

A comparison study was carried out for the 1PL model with binary responses and unidimensional ability for three levels of sample sizes; small, moderate and relatively large, and a variety of test lengths from short (m=10) to long (m=100). The main goal of the simulation study is to compare the performance of LA with MCMC. From the results of these comparison studies, we found that the LA method provides very accurate approximations computationally cheaply in this case. The approximately estimated abilities results are very close to the actual values and sometimes even better than the estimated abilities resulting from MCMC. Therefore, the LA method seems to be a useful tool for researchers interested in obtaining estimates of students' abilities in real-time. Although the focus is on one-parameter IRT models, it would be straightforward to extend the LA method similarly to two-parameter or three-parameter IRT models.

References

[1] F. Baker, *Item response theory*. New York: Marcel Dekker, 2004.

[2] Kim and D. Bolt, "Estimating Item Response Theory Models Using Markov Chain Monte Carlo Methods", *Educational Measurement: Issues and Practice*, vol. 26, no. 4, pp. 38-51, 2007.

[3] L. Tierney and J. Kadane, "Accurate Approximations for Posterior Moments and Marginal Densities", Journal *of the American Statistical Association*, vol. 81, no. 393, pp. 82-86, 1986.

[4] G. H. Fischer en I. W. Molenaar, Rasch models: Foundations, recent developments, and applications. Springer Science & Business Media, 2012.

[5] "R: The R Project for Statistical Computing", *R-project.org*, 2022. [Online]. Available: https://www.r-project.org/. [Accessed: 07- Apr- 2022].

[6] Y. Yuan, "A Modified BFGS Algorithm for Unconstrained Optimization", *IMA Journal of Numerical Analysis*, vol. 11. no. 3, pp. 325-332, 1991.

[7] F. Nielsen, "On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means", *Entropy*, vol. 21, no. 5, p. 485, 2019.

[8] H. Abdi, "The Kendall rank correlation coefficient", *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pp. 508--510, 2007
[9] C. DeMars, *Item Response Theory*. Oxford University Press, USA, 2010.