Article

Evaluating Stan's Variational Bayes Algorithm for Estimating Multidimensional IRT Models

Esther Ulitzsch ^{1,*} and Steffen Nestler ² ¹ Department of Educational Measurement, IPN–Leibniz Institute for Science and Mathematics Education, 24118 Kiel, Germany² Department of Statistics and Psychological Methods, University of Münster, 48149 Münster, Germany; steffen.nestler@uni-muenster.de

* Correspondence: ulitzsch@ipn.uni-kiel.de

Abstract: Bayesian estimation of multidimensional item response theory (IRT) models in large data sets may come with impractical computational burdens when general-purpose Markov chain Monte Carlo (MCMC) samplers are employed. Variational Bayes (VB)—a method for approximating the posterior distribution—poses a potential remedy. Stan's general-purpose VB algorithms have drastically improved the accessibility of VB methods for a wide psychometric audience. Using marginal maximum likelihood (MML) and MCMC as benchmarks, the present simulation study investigates the utility of Stan's built-in VB function for estimating multidimensional IRT models with between-item dimensionality. VB yielded a marked speed-up in comparison to MCMC, but did not generally outperform MML in terms of run time. VB estimates were trustworthy only for item difficulties, while bias in item discriminations depended on the model's dimensionality. Under realistic conditions of non-zero correlations between dimensions, VB correlation estimates were subject to severe bias. The practical relevance of performance differences is illustrated with data from PISA 2018. We conclude that in its current form, Stan's built-in VB algorithm does not pose a viable alternative for estimating multidimensional IRT models.

Citation: Ulitzsch, E.; Nestler, S. Evaluating Stan's Variational Bayes Algorithm for Estimating Multidimensional IRT Models. *Psych* **2022**, *4*, 73–88. <https://doi.org/10.3390/psych4010007>

Academic Editor: Philipp Doebler

Received: 5 November 2021

Accepted: 3 February 2022

Published: 5 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: Bayesian estimation; Markov chain Monte Carlo; Variational Bayes; multidimensional IRT; Stan

1. Introduction

Multidimensional item response theory (IRT) models are the method of choice for analyzing data from cognitive tests assessing multiple competencies (e.g., science, mathematical literacy, and reading). The vast majority of statistical software for estimating (multidimensional) IRT models draws on marginal maximum likelihood (MML) estimation (see e.g., [1–4]), but Bayesian IRT modeling has gained increasing attention in recent years (see [5,6] for introductions) due to its great flexibility in model specification and its capability of handling high-dimensional models well.

To estimate the parameters of Bayesian IRT models, researchers can use customized Markov chain Monte Carlo (MCMC) samplers that were developed for specific IRT models, such as the Gibbs sampler by [7] for the Rasch model, or they can use general-purpose software for Bayesian estimation such as JAGS [8] or Stan [9]. In our experience, applied researchers typically employ the latter software packages because they provide them with a high flexibility in model specification. This allows them to take the specifics of their data into account without the need to develop their own customized samplers (for which they may not have the time). Furthermore, general-purpose software packages quickly adopt new state-of-the-art Bayesian techniques, allowing applied researchers to easily consider these new techniques in their work. However, the last point may also be a disadvantage, at least when there is little research on the performance of these new techniques as implemented in the general-purpose software.

A case in point is the implementation of Variational Bayes (VB) in Stan. VB is a deterministic method for approximating the posterior distribution [10] that trades off the performance guarantees of MCMC-based Bayesian estimation against speed. VB may therefore be a real alternative for researchers aiming to draw on Bayesian techniques for estimating complex models in large data sets that could alleviate the potentially infeasible computation times of general-purpose MCMC samplers that are not customized to the specified models.

We believe that VB will receive increasing attention as an alternative estimation approach for multidimensional IRT models. First, research showed that the method performed well in estimating simpler IRT models [11]. Second, many applications of multidimensional IRT models involve large data sets. A prominent example is large-scale assessments (LSAs) such as the Programme for the International Assessment of Adult Competencies (PIAAC; [12]) or the Programme for International Student Assessment (PISA; [13]), which provide publicly available, rich data sets that support investigating multiple competencies in large, representative samples from multiple countries. For the analysis of such data, VB may pose a computationally more feasible alternative to general-purpose MCMC samplers. Third, and most importantly for the present context, general-purpose VB algorithms are now available in the statistical programming language Stan. This provides the opportunity to leverage VB methods to a wide audience while also maintaining flexibility in model specification. However, although Stan's VB algorithms have been shown to perform well in simple linear and hierarchical logistic regression analyses [14,15], it is currently unclear whether the results of Stan's general-purpose VB algorithms are also trustworthy for complex IRT models.

Focusing on multidimensional IRT models with between-item dimensionality, the aim of the present study is to investigate not only how well the general approach of Stan works for estimating complex IRT models, but also whether applied researchers can use the current implementation in their work. In doing so, we (a) add to an improved understanding of the performance of Stan's general-purpose VB function for estimating complex models with high-dimensional parameter spaces, (b) further investigate the utility of VB for psychometric applications, and (c) provide psychometric researchers with an evaluation as to whether drawing on Stan's VB algorithms may be a promising direction to take for implementing their newly developed IRT models. In the following, we first briefly review MML and MCMC-based Bayesian estimation of multidimensional IRT models. Thereafter, we give a short introduction to the fundamentals of VB methods and their implementation in Stan. We then report the results of a simulation study in which we examine the performance of VB as implemented in Stan under different conditions of model complexity and sample size. We continue with an illustration of the approach using a real-world example data set.

2. Conventional Estimation of IRT Parameters

The focus of the present study is a D -dimensional two-parameter logistic (2PL) model with between-item dimensionality (i.e., each item loads on one dimension only). We focus on multidimensional IRT models with between-item dimensionality because we believe (1) that they are more relevant for and more widely employed by applied researchers, because they correspond to the design of typical cognitive assessments and allow the factors to be interpreted more clearly. In typical LSA designs, for example, items are assumed to be reflective of one proficiency only (e.g., science, mathematical literacy, or reading). As we see LSA data as a potential use case for employing VB methods, we aim at investigating a model that is applied in this context. Furthermore, we (2) deem it promising to start with a relatively simple model to investigate the utility of Stan's VB algorithm for estimating multidimensional IRT models and better understand its performance. In the case that VB does not yield trustworthy results even for a simple multidimensional model, it may be worthwhile to first improve performance for this simpler model before moving to more

complex models such as models with within-item dimensionality (i.e., cross-loadings) or non-compensatory models with multiplicative terms.

In multidimensional IRT models with between-item dimensionality, the dichotomously scored item responses y_{ij} , containing examinee i 's, $i = 1, \dots, N$, score on item j , $j = 1, \dots, K$, measuring dimension d , $d = 1, \dots, D$, are modeled as

$$p(y_{ij} = 1) = \frac{\exp(a_j \eta_{id[d[j]]} + b_j)}{1 + \exp(a_j \eta_{id[d[j]]} + b_j)}, \quad (1)$$

where b_j and a_j denote items j 's difficulty and discrimination parameter, respectively. Person i 's proficiency on the d th dimension measured by item j is given by $\eta_{id[d[j]]}$. The function $d[j]$ maps the item's index j to the index d of the latent proficiency variable it is assumed to measure.

A joint multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ is assumed for person proficiencies $\boldsymbol{\eta}_i = (\eta_{i1}, \dots, \eta_{iD})$.

$$\boldsymbol{\eta}_i \sim \mathcal{N}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (2)$$

This yields the following likelihood

$$L(\mathbf{Y}|\boldsymbol{\gamma}) = \prod_{i=1}^N \prod_{j=1}^K p(y_{ij}|a_j, b_j, \eta_{id[d[j]]}) f(\boldsymbol{\eta}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (3)$$

where $\boldsymbol{\gamma}$ is a vector of the parameters to be estimated and $f(\cdot)$ denotes a multivariate normal density. For model identification, all elements in $\boldsymbol{\mu}$ can be set to 0 and variances (i.e., diagonal elements of $\boldsymbol{\Sigma}$) can be set to 1. Depending on the research questions, both incidental (i.e., item difficulties and discriminations) or structural parameters (e.g., correlations among the dimensions) may be of interest.

2.1. Marginal Maximum Likelihood Estimation

Standard statistical software for estimating (multidimensional) IRT models such as TAM [1] or mirt [2], but also software used in LSA reporting for item parameter calibration such as mdlm [4] currently employed in PIAAC and PISA, draw on MML estimation via the expectation–maximization (EM) algorithm. However, using MML is limited by the number of dimensions D , as maximizing the marginal log-likelihood involves solving high-dimensional integrals that are usually analytically intractable. Hence, numerical approximation methods to the integrals have to been used [16]. These methods are known to be computationally demanding under conditions with three or more dimensions, and may come with a computational burden that jeopardizes the applicability of higher dimensional IRT models in practical settings.

2.2. Bayesian Estimation via Markov Chain Monte Carlo

In contrast to the MML approach, Bayesian estimation combines the likelihood function of the data with prior distributions $\pi(\boldsymbol{\gamma})$ for the to-be estimated model parameters $\boldsymbol{\gamma}$. This allows the posterior distribution $p(\boldsymbol{\gamma}|\mathbf{Y})$ of the parameters to be obtained, given the data

$$p(\boldsymbol{\gamma}|\mathbf{Y}) = \frac{L(\mathbf{Y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})}{p(\mathbf{Y})}. \quad (4)$$

The posterior distribution is used to construct point and interval estimators of the model parameters by computing certain summary measures and quantiles, respectively, of the posterior distribution itself or the marginal distribution of a model parameter. However, for complex models with many parameters it is difficult to analytically derive these measures. Therefore, MCMC algorithms (see [17], for an introduction), such as the Gibbs

sampler or the no-U-turn sampler [18] are commonly used to approximate the posterior distribution by iteratively drawing random numbers from it. The sampled values constitute a Markov chain and once the chain has stabilized, one uses the random draws to compute summary statistics such as the mean or quantiles.

Bayesian estimation of (multidimensional) IRT models bears various advantages compared with MML. First, state-of-the-art MCMC sampling techniques (e.g., [18]) can deal well with high-dimensional, highly correlated parameter spaces, rendering them particularly useful for complex and high-dimensional IRT models. Thus, restrictions on the dimensionality under MML estimation do not apply to MCMC-based estimation techniques commonly employed for Bayesian estimation (e.g., [19]). Second, the Bayesian approach can be more easily adapted to new data and modeling situations and hence “supports in a natural way extensions of common item response models” (see [5], p. 2). These extensions are important to address challenges such as complex response behavior (e.g., guessing or switches in response processes), missingness and nonresponse, or complex sampling designs.

MCMC algorithms are important methods for Bayesian statistics and they are important for the (still) increasing adoption of the Bayesian approach in applied research. However, a problem of MCMC algorithms is that they can be very computationally demanding, especially when researchers draw on general-purpose samplers instead of samplers tailored to the specific model employed. Hence, while general-purpose samplers make Bayesian estimation techniques easily accessible and provide researchers with great flexibility in model specification, their computation times can become unbearable when sample sizes are large, rendering Bayesian estimation in such contexts infeasible. Tailoring samplers to a specific model can be a cumbersome endeavor, and developing well-scalable samplers is not equally straightforward for all classes of models. Therefore, alternative approximation methods of the posterior distribution have been investigated, which may also be used in the context of multidimensional IRT models.

3. Variational Bayesian Methods

In contrast to MCMC algorithms that approximate the posterior distribution by sampling from it, VB is a deterministic technique in which approximating the posterior is transferred into an optimization problem [20,21]. VB methods were originally developed in the context of machine learning, but are now used in various fields with large and complex data sets, including computational biology, computational neuroscience, or natural language processing and speech recognition [21]. VB methods only recently gained attention in the psychometric literature, where they have been applied to estimate the parameters of customary IRT models [11,22], cognitive diagnostic models [23–25], latent class analysis [26], and models from cognitive and mathematical psychology [27,28]. We note that variational methods can not only be used to approximate the posterior distribution, but also to approximate any (complex) probability density. For this reason, variational methods are also suitable for MML estimation where one is dealing with complex likelihood functions. Research on likelihood-based variational inference has covered complex IRT models such as models with random item effects [29] or multidimensional models [30], as well as the general case of generalized linear latent variable models [31,32].

The basic idea behind VB is to approximate the posterior distribution $p(\gamma|\mathbf{Y})$ with a second distribution over the model parameters $q(\gamma|\theta)$ [10]. The second distribution is controlled by the parameters θ that determine how far or close q is to the posterior distribution. In fact, the main goal of VB is to determine those parameter values θ for which the distance between q and the posterior distribution is minimal. Once this optimization problem is solved, one can use q instead of the posterior distribution to compute summary statistics such as the mean or quantiles.

In order to determine the variational parameters for which q is as close as possible to the posterior distribution, one must define a measure that maps the distance between the

two probability densities. The most commonly used measure is the Kullback–Leibler (KL) divergence that, applied to the present context, is given by:

$$\text{KL}(q(\gamma|\theta)||p(\gamma|\mathbf{Y})) = \int q(\gamma|\theta) \log \left[\frac{q(\gamma|\theta)}{p(\gamma|\mathbf{Y})} \right] d\gamma. \quad (5)$$

The formula for the KL divergence can be further transformed because it is an expectation with respect to $q(\gamma|\theta)$:

$$\begin{aligned} \text{KL}(q(\gamma|\theta)||p(\gamma|\mathbf{Y})) &= \mathbb{E}_q \left(\log \left[\frac{q(\gamma|\theta)}{p(\gamma|\mathbf{Y})} \right] \right) \\ &= \log[p(\mathbf{Y})] + \mathbb{E}_q(\log[q(\gamma|\theta)]) - \mathbb{E}_q(\log[L(\mathbf{Y}|\gamma)\pi(\gamma)]) \end{aligned} \quad (6)$$

where the last line follows from inserting Equation (4). This shows that it is actually not possible to minimize the KL divergence, because the marginalizing constant $p(\mathbf{Y})$ is unknown. However, as $p(\mathbf{Y})$ does not depend on q and hence θ , one can *maximize* [10,21,27]:

$$\Omega(\theta) = \mathbb{E}_q(\log[L(\mathbf{Y}|\gamma)\pi(\gamma)]) - \mathbb{E}_q(\log[q(\gamma|\theta)]) \quad (7)$$

with respect to θ to obtain a density q that is as close as possible to the posterior distribution. This function is also known as the evidence lower bound (ELBO), because it is a lower bound for the marginalizing constant, or ‘evidence’, $p(\mathbf{Y})$.

In addition to specifying the distance measure, it is also necessary to specify which distribution q is to be used as the approximating distribution. Of course, q should be chosen in a way that the function is easy to handle but also flexible enough to achieve a good approximation. One way to find such a distribution is to assume that q can be factorized as

$$q(\gamma) = \prod_{g=1}^G q(\gamma_g). \quad (8)$$

Equation (8) assumes complete independence of all individual parameters γ_g . This is sometimes referred to as ‘naive’ mean-field assumption. One can also assume that the elements in γ can be divided into M parameter groups and that q factorizes with respect to these groups [10].

Note that Equation (8) makes no assumption about the form or type of distribution (e.g., that it is a normal distribution). Nevertheless, the assumption is sufficient to derive an optimization algorithm (called coordinate ascent variational inference; CAVI) that can be used to find the approximation density for a parameter that maximizes the ELBO. In fact, the optimal solution is given by

$$q_g^*(\gamma_g) \propto \exp \left(\mathbb{E}_{-g} \left[\log p(\gamma_g|\gamma_{-g}, \mathbf{Y}) \right] \right), \quad (9)$$

where the expectation is taken with respect to the parameters that are currently not considered. Equation (9) shows that VB methods and Gibbs sampling are closely related because both use the conditional densities of the parameters. Furthermore, even if the form of q has not been specified, the use of Equation (9) leads to variational densities that are standard distributions and whose parameters are optimized during CAVI. For example, if one uses VB to estimate the expected value of a normally distributed variable, then $q^*(\mu)$ is a normal distribution whose mean and variance are updated during CAVI. In fact, the distribution of q is implicitly defined by the shape of the likelihood and the prior distributions, again showing the close connection between VB and Gibbs sampling.

Equations (9) and (10) can be used to derive a VB approach for a specific statistical model. However, since these derivations are often technically very complex and are tied to the specific model, an alternative general-purpose approach would be (a) to assume a specific distribution per se for all model parameters and all models and (b) to maximize

Equation (7) to determine the parameters of this distribution. For instance, one can assume that the approximating distributions are independent normal distributions:

$$q(\gamma|\theta) = \prod_{g=1}^G q(\gamma_g|\theta = (\mu_{\gamma_g}, \sigma_{\gamma_g}^2)). \quad (10)$$

In this case, Equation (7) is

$$\Omega(\theta) = \mathbb{E}_q(\log[L(\mathbf{Y}|\gamma)\pi(\gamma)]) - H[q(\gamma|\theta)] \quad (11)$$

where H is the entropy of the normal distribution. This approach is called mean-field Gaussian and, in fact, is the approach taken in Stan's built-in VB function under default settings (see [33]). It is also possible to specify a multivariate normal as the approximating distribution in Equation (10) and this is called full-rank Gaussian in Stan. The normal distribution assumption can also be exploited to approximate the expectation in Equation (11) and the gradient of Equation (11) via Monte Carlo integration by sampling from q . In Stan, the gradient is computed using automatic differentiation (hence the name automatic differentiation variational inference; ADVI; see [15]), and together with the normal distribution assumption, this is another reason why Stan's approach can be used for a wide variety of statistical models—such as multidimensional 2PL models—for which the derivation of a specific variational inference algorithm would be too complicated.

The price for Stan's default general-purpose VB algorithm is the rather restrictive mean-field Gaussian assumption that may be too restrictive in the context of complex models with high-dimensional, highly correlated parameter spaces. That is, the parameters may neither be independent of each other, nor may they be normally distributed. The full-rank algorithm implemented in Stan, in contrast, overcomes the assumption of parameters being independent of each other; it may, however, also be more challenged by complex models. It is therefore worthwhile to investigate Stan's built-in VB algorithms' ability to handle complex models, whether its default yields trustworthy results even under potential violations of the mean-field Gaussian assumption, and to what extent Stan's algorithms are challenged by the complexity and high-dimensional parameter spaces multidimensional IRT models entail.

4. Present Research

The aim of the present study is to investigate the utility of Stan's built-in VB function for estimating multidimensional 2PL models. In doing so, we add to the current literature on VB for psychometric applications in several ways. First, research on VB multidimensional IRT modeling is sparse, and it has so far only been evaluated with regard to the recovery of item and person parameter estimates in single simulated data sets [22]. Comprehensive evaluations of VB estimation of multidimensional IRT models are therefore still pending, although they will improve our understanding of the utility of VB for psychometric applications. Second, previous studies on VB in the context of psychometrics [22–25] have focused on customized VB algorithms, limiting the usability for applied researchers and hindering the ease of model adaptation. Stan's VB function, in turn, is easy to use and, as such, comes with great potential for making the advantages of VB widely accessible for psychometric applications. Although Stan's VB function has only recently been developed and is still at a beta version stage, first evaluations of its performance are very promising, finding Stan's VB algorithms to yield unbiased point estimates for a variety of model classes, including logistic regression [15]. Evaluations of their performance concerning the estimation of IRT models in general and under multidimensionality in particular will thus inform both applied and psychometric researchers as to whether drawing on Stan's VB algorithms may be a promising direction to take for implementing their IRT model of choice.

5. Simulation

5.1. Population Model and Simulation Conditions

We simulated data according to a multidimensional 2PL model with between-item dimensionality as given by Equation (1). For data generation, we aimed at mirroring the conditions encountered in LSAs, as we see the analysis of LSA data as a potential use case for the application of VB. Data-generating parameters were set to resemble those reported in PISA 2018 [13]. For item parameters, we considered 10 different values stemming from the sequences $\{0.40 + 0.10l\}_{l=1}^{10}$ for discriminations a and $\{-1.50 + 0.25l\}_{l=1}^{10}$ for difficulties b . We varied four factors in our simulation: The number of dimensions D was set to 2 or 4, representing a model of low or high dimensionality, respectively, and reflecting the typical number of domains investigated in major LSAs. PISA for instance, administers three core domains (reading, mathematical literacy, and science) and, in some countries, an additional innovative domain (e.g., global competence in 2018). The correlations between the dimensions ρ were defined to be either 0.00 or 0.75. The latter corresponds to a typical correlation among domains encountered in LSAs (see Table 18.6. in [12]). A zero correlation renders the mean-field assumption of uncorrelated parameters more plausible, at least for examinees' proficiency estimates. We therefore considered a zero correlation to investigate to what extent this facilitates VB estimation. Furthermore, the number of items per dimension K_d was set to 20 or 50, resulting in a total number of items of $\sum_{d=1}^D K_d$. Finally, the number of examinees N was set to 5000, 20,000, or 150,000. This corresponds to low and high country-level sample sizes as well as to total sample sizes encountered in typical mid-sized international LSA settings. In PIAAC 2012, for instance, a total of 164,711 examinees was assessed, with country-level and sub-national entity-level sample sizes, respectively, ranging from 3761 in Northern Ireland to 27,285 in Canada [12]. Due to computational constraints, $\rho = 0$ was only assessed under $N = 5000$ as well as under conditions with $N = 20,000$ and $K_d = 20$. For $\rho = 0.75$, the considered sample sizes, number of dimensions, and the number of items per dimension were fully crossed. This resulted in $3 \times 2 \times 2 + 2 \times 2 + 2 = 18$ data-generating conditions. In each condition, we generated 100 data sets.

5.2. Dependent Measures

We investigated performance in terms of the run time, bias, and variability of parameter estimates. To decrease the influence of extreme parameter estimates, we computed robust measures for bias (i.e., the median) and variability (i.e., the median absolute deviation). In line with [34], we considered relative bias below 10% to be acceptable. For evaluating the quality of the VB approximation to the full posterior distribution, we further investigated Pareto $\hat{k}$ diagnostic values. For these, values below 0.70 were considered acceptable (see guidelines in [35]).

5.3. Estimation

Bayesian estimation was implemented in Stan version 2.25.0 [9], using rstan [36]. We made use of Stan's built-in VB function. In preliminary analyses, we investigated both Stan's default, relying on the mean-field assumption, as well as its full-rank Gaussian algorithm, specifying a multivariate normal as the approximating distribution. Using the full-rank Gaussian algorithm, however, led to Stan stopping the estimation procedure, yielding the warning that the model may be ill-conditioned. Since we were not able to find out why this warning was released (and because the model was estimated with other procedures in Stan, thus ruling out misspecification as a potential cause), we only report results that refer to Stan's mean-field algorithm. Furthermore, we also probed importance resampling as a potential remedy to the low accuracy of VB parameter estimates. However, as this did not yield marked improvements, we report results for the mean-field algorithm that adhered to Stan's defaults.

For benchmarking, we also analyzed data employing MCMC-based and MML estimation. For MCMC-based Bayesian estimation, Stan employs the no-U-turn sampler [18],

an adaptive form of Hamiltonian Monte Carlo sampling. We acknowledge that samplers tailored to the estimation of IRT models are likely to outperform Stan's MCMC no-U-turn sampler in terms of speed; however, the aim was to use a general-purpose sampler as a benchmark for fair comparisons with the employed general-purpose VB algorithm. Note that Stan offers multiple options for speeding up MCMC-based estimation, ranging from reparameterization to within-chain parallelization via Stan's rectangular mapping functionality. MCMC run times may therefore be seen upper bounds, which can certainly be improved. We ran two chains with 1000 iterations each, with a warm-up period of 500. Due to its computational burden, only data sets with $N = 5000$, $K_d = 20$, and $\rho = 0.75$ were analyzed with MCMC. As under MCMC the bias and variability of parameter estimates can reasonably be assumed to decrease as the size of the data set increases (see [37] for an evaluation of Stan's parameter recovery of 2PL model parameters using MCMC-based estimation), MCMC results for this selected set of conditions can be seen as lower bounds in the present simulation set-up. We employed TAM [1] for MML-based estimation. Due to the high number of dimensions considered, we employed quasi Monte Carlo integration with 500 nodes. Again, to reduce computational burden, only conditions with $\rho = 0.75$ were analyzed with MML. Under $N = 150,000$, we were mainly interested in run time differences between MML and VB rather than differences in the accuracy of parameter estimates. Therefore, under these conditions, we only analyzed data from 10 replications with MML for the purpose of gauging run time differences. All analyses were implemented in R version 4.0.2 [38]. Exemplary R and Stan code is provided in the OSF repository accompanying this article.

5.4. Results

Table 1 compares the median and median absolute deviations of run time in seconds for VB against the considered MCMC and MML benchmarks. Note that these comparisons need to be interpreted with caution, as run times of MCMC and MML are heavily dependent on the software as well as the number of iterations and nodes employed, respectively. As was to be expected, MCMC was by far outperformed by both VB and MML in terms of run time. Under conditions with smaller data sets of $N = 5000$ and $K_d = 20$, MML was markedly faster than VB. Differences decreased with an increasing size of the data set. Nevertheless, VB was, on average, faster than MML only when $N = 150,000$ and $K_d = 50$, where it yielded a speed-up in a median run time of 34%. At the same time, across all conditions, VB's run time was much more variable than run time under MML, as indicated by higher median absolute deviations.

Table 1. Comparison of median and median absolute deviation (MAD) run time in seconds.

N	K_d	D	VB		MML		MCMC	
			Median	MAD	Median	MAD	Median	MAD
5000	20	2	220.51	92.32	47.39	1.74	3974.47	833.20
		4	575.68	166.63	190.23	20.81	7519.80	1711.84
	50	2	404.84	120.78	251.62	13.62		
		4	1067.82	225.25	734.38	231.20		
20,000	20	2	840.53	156.93	205.90	6.15		
		4	2155.94	499.51	859.58	29.07		
	50	2	1813.55	383.51	1169.36	28.81		
		4	4962.79	1831.03	3560.65	685.39		
150,000	20	2	6673.44	2274.47	2177.35	31.63		
	50	2	7884.13	2768.26	12013.06	185.32		

Table 2 compares bias for VB against the MCMC and MML benchmarks. To simplify the exposition, we averaged the results across data-generating values for item discriminations and difficulties. Bias for the benchmark methods MCMC and MML in absolute terms

was below 0.01 for all parameters and conditions. As such, relative bias under MCMC and MML remained markedly below 10% for all parameter types and conditions. Under VB, in all analyzed data sets, Stan issued warnings that VB approximations may not be reliable, as indicated by large Pareto $\hat{k}$ diagnostic values. Across all conditions, difficulties were well recovered. Item discriminations were downward biased and this bias was larger the greater the number of dimensions (i.e., $D = 4$). Nevertheless, when $D = 2$, relative bias remained below 10%. Note that in preliminary analyses, we also investigated bias in item discriminations in unidimensional models. The results were comparable to those under $D = 2$. We found correlations between latent dimensions to be the most sensitive to data-generating values and conditions. Zero correlations were recovered without bias. However, when the true correlation was 0.75, estimates were severely downward biased. This bias was smaller the higher the number of items per dimension, but it was aggravated with a high number of dimensions. Consequently, relative bias for $\rho = 0.75$ was just above 10% when $N = 5000$, $K_d = 50$ and $D = 2$, but as high as 28% when $K_d = 20$ and $D = 4$. We did not observe marked improvements with an increasing sample size.

Table 2. Bias under variational Bayesian inference (VB) compared against Markov chain Monte Carlo (MCMC) and marginal maximum likelihood (MML).

N	K_d	D	Method	ρ		a		b	
				$\rho = 0.00$	$\rho = 0.75$	$\rho = 0.00$	$\rho = 0.75$	$\rho = 0.00$	$\rho = 0.75$
5000	20	2	VB	0.005	−0.171	−0.065	−0.070	−0.001	0.001
			MCMC		0.000		0.002		−0.002
			ML		−0.003		−0.003		−0.002
		4	VB	0.000	−0.194	−0.112	−0.116	−0.001	−0.003
			MCMC		−0.001		0.005		0.000
			ML		0.004		−0.003		−0.004
	50	2	VB	−0.001	−0.077	−0.054	−0.077	0.003	0.001
			ML		−0.002		0.000		0.000
		4	VB	−0.001	−0.105	−0.084	−0.119	−0.008	−0.013
			ML		0.008		0.013		−0.010
20,000	20	2	VB	0.001	−0.170	−0.064	−0.068	0.001	0.003
			ML		−0.003		−0.001		0.001
		4	VB	0.000	−0.210	−0.097	−0.107	−0.001	−0.003
			ML		0.003		0.001		−0.001
	50	2	VB	−0.007	−0.077	−0.056	−0.081	0.002	0.002
			ML		−0.001		0.001		−0.001
		4	VB		−0.126		−0.122		−0.015
			ML		0.008		0.013		−0.010
150,000	20	2	VB		−0.172		−0.068		0.001
			ML		−0.002		0.000		0.000
	50	2	VB		−0.078		−0.078		0.003
			ML		−0.001		0.002		0.001

Table 3 compares median absolute deviations of parameter estimates retrieved from VB against the benchmarks MCMC and MML. For discriminations and difficulties, we present median absolute deviations of differences between estimates and data-generating parameters across all data-generating values. Similar to the results concerning the bias, the variability of the VB correlation estimates was highly contingent on the data-generating parameters. When the true correlation was zero, variability was lower the larger the size of the sample. By contrast, when $\rho = 0.75$, estimates were considerably more variable under $D = 4$ than under $D = 2$ (e.g., 0.019 as compared with 0.090 under $N = 5000$ and $K_d = 20$). For the benchmark methods MCMC and MML, median absolute deviations of correlation estimates remained below 0.020 even under conditions with smaller data sets, and, as such, below the best result observed for VB across all considered conditions. We observed comparable effects of the number of dimensions on variability for item discriminations a .

For instance, when $N = 5000$, $K_d = 20$, and $\rho = 0.75$, VB estimates of item discriminations showed a median absolute deviation of 0.057 when $D = 2$, but of 0.093 when $D = 4$. The variability of the VB difficulties was comparable to the variabilities obtained with MCMC and MML. For MCMC and MML, we found that the variability of estimates of discriminations and difficulties only depended on sample size, with median absolute deviations around 0.040 when $N = 5000$, around 0.020 when $N = 20,000$, and around 0.010 when $N = 150,000$ for both parameter types.

Table 3. Median absolute deviations under variational Bayesian inference (VB) compared against Markov chain Monte Carlo (MCMC) and marginal maximum likelihood (MML).

N	K_d	D	Method	ρ		a		b	
				$\rho = 0.00$	$\rho = 0.75$	$\rho = 0.00$	$\rho = 0.75$	$\rho = 0.00$	$\rho = 0.75$
5000	20	2	VB	0.027	0.019	0.060	0.057	0.049	0.047
			MCMC		0.009		0.045		0.037
			ML		0.010		0.045		0.037
		4	VB	0.031	0.090	0.123	0.093	0.052	0.048
			MCMC		0.010		0.044		0.035
			ML		0.016		0.044		0.036
	50	2	VB	0.026	0.016	0.053	0.056	0.047	0.048
			ML		0.006		0.042		0.036
		4	VB	0.022	0.049	0.079	0.075	0.044	0.046
			ML		0.009		0.042		0.041
20,000	20	2	VB	0.023	0.020	0.043	0.041	0.030	0.030
			ML		0.005		0.022		0.017
		4	VB	0.023	0.101	0.085	0.068	0.033	0.030
			ML		0.013		0.022		0.018
	50	2	VB	0.021	0.017	0.037	0.042	0.031	0.033
			ML		0.005		0.021		0.018
		4	VB		0.063		0.069		0.029
			ML		0.006		0.021		0.022
150,000	20	2	VB		0.020		0.037		0.026
			ML		0.002		0.008		0.007
	50	2	VB		0.013		0.038		0.026
			ML		0.002		0.008		0.007

6. Illustrative Data Example

We also examined differences between VB, MCMC and MML using data from PISA 2018 [13]. We focused on the Spanish sample, as this, with $N = 35,922$ examinees, comprised the largest country-level sample size. Examinees were administered a total of $K = 445$ items, measuring three domains. For the major domain, reading, 260 items were administered. Items measuring reading fluency were excluded, as these were scaled separately from the other reading items in PISA 2018. For the minor domains, science and mathematical literacy, 115 and 70 items were administered, respectively. Note that due to the incomplete block designs employed for the minor domains and the multistage adaptive test design employed for the major domain, each examinee was administered only a fraction of the items. The data set comprised a total of 1,984,724 item responses; item-level sample sizes ranged from 2055 to 17,815. For estimation, we employed the same set-up as in the simulation study.

As expected, we observed vast differences in both run time and parameter estimates. VB yielded the by far shortest run time of 16 min, followed by MML with 5 h and MCMC with 21 h. In line with results from the simulation study, there were pronounced differences in parameter estimates for correlations among the three domains (see Table 4) retrieved with VB as compared to MML and MCMC. While there was strong agreement between estimates retrieved from MML and MCMC, VB yielded markedly lower correlations. For instance, while MML and MCMC yielded correlations between science and mathematical

literacy of about 0.81, the correlation obtained with VB was 0.42. Hence, in the present application, VB would result in conclusions that are drastically different from those drawn from MML and MCMC results. While researchers investigating the relationship between science and mathematical literacy using MCMC or MML would conclude that these competencies are highly correlated, researchers drawing on VB would conclude that these competencies are related only to a moderate degree. MML and MCMC yielded comparable item parameter estimates, with correlations between estimates of discriminations and difficulties, respectively, above 0.99. As shown in Figure 1, VB yielded estimates comparable to those from MCMC for difficulties, but somewhat different results for discriminations. The correlation between estimates was 0.95 for the difficulties and 0.86 for the discriminations. Nevertheless, in line with the simulation study, differences in discriminations were much less pronounced than differences in correlations.

Table 4. Correlation between domains.

	MML		MCMC		VB	
	Science	Reading	Science	Reading	Science	Reading
Reading	0.88		0.88		0.60	
Mathematical Literacy	0.81	0.82	0.81	0.82	0.42	0.56

Figure 1. Comparisons of item parameters retrieved under MCMC and VB. (a) Discriminations a . (b) Difficulties b .

7. Discussion

The aim of the present study was to investigate Stan's VB function for estimating multidimensional 2PL models with between-item dimensionality under realistic research conditions. Stan provides general-purpose VB algorithms that strongly facilitate the implementation of variational inference for a wide array of models, rendering it a potentially powerful tool for fitting complex IRT models to large data sets. Stan provides algorithms either relying on the mean-field assumption, i.e., assuming the approximate distribution to factorize into independent normal distributions, or assuming a multivariate normal as the approximating distribution. In the given context, however, only Stan's default, implementing the rather restrictive mean-field Gaussian assumption, yielded results (see [39] for comparable results in the context of Poisson regression with random effects). A potential reason for the failure of the full-rank VB algorithm in the context of multidimensional 2PL models may be their (potentially highly correlated) high-dimensional parameter space and, as such, the high dimensionality of the approximating multivariate normal distribution as well as the associated increased computational costs.

Based on the results of the simulation study, we conclude that in its current form, Stan's built-in VB algorithm does not (yet) pose a viable alternative to MML and MCMC-based Bayesian estimation of multidimensional IRT models in large data sets. In general, no reliable mean-field VB approximations could be obtained, as indicated by consistently large Pareto $\hat{k}$ diagnostic values, the accompanying warnings issued by Stan, and biased

parameter estimates. Concerning parameters of the measurement model, mean-field VB yielded trustworthy results only for item difficulties, while estimates of item discriminations were downward biased and also slightly more variable compared with estimates of the MCMC and MML benchmarks. These results mirror those from previous studies. For a multilevel random intercept logistic regression model [40], for instance, reported mean-field VB yielded good results for fixed effects regression parameters, but poor performance for variance components. [41] reported comparable results for a Poisson regression model with random effects. Note that we fixed variances of the fitted multidimensional 2PL models to one for identification, such that issues related to the estimation of variances may ‘leak’ into the estimation of discrimination parameters. Furthermore, we found discrimination parameters to be especially prone to bias and high variability under conditions with a high number of dimensions, while under conditions with only two dimensions, bias remained below 10%. One may argue that researchers might be willing to trade off this amount of loss in accuracy of parameter estimates against speed of estimation. However, while mean-field VB yielded a marked speed-up in comparison to MCMC-based Bayesian estimation, it did not generally outperform MML in terms of run time, even with as many as 500 nodes employed for quasi Monte Carlo integration, rendering MML the better choice in terms of both run time and trustworthiness of parameter estimates.

Mean-field VB estimates of the considered structural parameters (i.e., correlations between dimensions) were unbiased and displayed low variability when the true population correlations were zero. Given typical competence assessment data, however, zero correlations pose an unrealistic scenario. Under more realistic conditions with highly correlated dimensions, challenging the mean-field assumption, estimates of correlations were severely downward biased, even more so when a high number of dimensions was measured with relatively few items. The poor performance of mean-field VB under highly correlated dimensions may indicate that the mean-field assumption is too restrictive in these settings. We found comparable differences in estimates of correlations between mean-field VB and the MML and MCMC benchmarks in the illustrative real-data example, indicating that performance differences between the estimation approaches are of real relevance under real-data conditions.

Limitations and Future Directions

Stan’s VB algorithm is applicable to a wide array of models and has not been tailored to the context of multidimensional IRT models. We found that in its current form, Stan’s built-in VB algorithm can not well handle multidimensional IRT models, especially when the number of dimensions is high and/or when the latent dimensions are highly correlated. While the default algorithm relying on the mean-field assumption was prone to bias, we were not able to retrieve results under Stan’s full-rank algorithm, which assumes a multivariate normal as the approximating distribution. Future research may investigate to what extent model reparameterizations may aid in retrieving results when drawing on Stan’s full-rank algorithm. It should, however, be kept in mind that Stan’s full-rank algorithm has been reported to actually provide worse approximations to the MCMC posterior means than the mean-field version for a variety of model classes [42], rendering it unlikely that—even when results can be obtained—Stan’s full-rank algorithm outperforms its mean-field algorithm in terms of the accuracy of parameter estimates of multidimensional 2PL models.

We would like to note that VB has only recently been implemented in Stan and is, at present, still at a beta version stage. Given the rapid development of both VB algorithms and Stan’s functionalities, we are optimistic that the applicability of general-purpose VB algorithms implemented in accessible software for estimating complex multidimensional IRT models will improve in the nearer future. Nevertheless, our research also suggests that the performance of these improved algorithms must be examined before they are used in applied contexts.

In the meantime, to draw on VB algorithms that are both scalable and trustworthy in the context of IRT analyses, we think that future research should examine the perfor-

mance of VB approaches customized to this specific context. Such developments may be especially useful in contexts where customized MCMC samplers may be challenging to develop, e.g., when no sufficiency of statistics can be assumed (see [7,43]). For instance, partially-factorized VB methods—avoiding a fully factorized approximation, and instead assuming a factorization only for a specific subset of parameters (see [44], for developments and evaluations in the context of Probit models)—weaken the mean-field assumption and may therefore pose a better alternative for settings with highly correlated latent dimensions, while at the same time being less challenged by complex models than full-rank Gaussian algorithms. Another alternative may be to build on VB approaches that have been developed in the context of binary outcomes that circumvent the assumption of independent normal approximate distributions. The authors of [45], for instance, achieved good performance by employing the Jaakkola–Jordan variational lower bound [46] in the context of semiparametric logistic regression. The authors of [47] further developed the approach suggested by [48] and achieved accurate results for logistic regression. Furthermore, ref. [22] recently proposed the variational item response theory lower bound (VIBO) as an optimization objective for VB tailored to the context of multidimensional IRT models. In their evaluations of VIBO for multidimensional IRT models, however, the authors considered a single data set and focused only on the recovery of proficiency estimates. While their results are promising, we think that a more thorough investigation of their approach is needed.

Due to computational constraints, we focused on a small set of conditions and considered only 100 replications per condition in the simulation study. To corroborate the conclusions of the present study, future research may aim at replicating results for some selected conditions with an increased number of replications. Furthermore, concerning correlations between latent dimensions, we only considered two rather extreme cases. While correlations of zero were well recovered, correlations among highly correlated dimensions were severely biased. Future research may aim at establishing boundary conditions of maximum latent correlations that still support retrieving trustworthy parameter estimates.

It should also be noted that conditions encountered in LSA data as a major potential use case for VB are commonly much more complex than those investigated in our simulation study. It may well be that Stan’s built-in VB algorithm is even more challenged by these more complex conditions, and we note that this also challenges the development of customized VB algorithms. For instance, LSA data commonly stem from complex sampling designs with multilevel structures that researchers may want to consider in their analyses. Further, LSA data commonly have high proportions of missing responses—both planned (e.g., due to incomplete block or multistage adaptive test designs) and unplanned (e.g., due to item omissions, time restrictions, or quitting behavior). This can result in rather small item-level sample sizes, even when the overall sample size is large—in the illustrative real-data example, for instance, even for an overall sample size of 35,922, the smallest item-level sample size was as low as 2055. While MML and MCMC can well handle such conditions (see [49,50]), VB may yield less accurate estimates for items with small item-level sample sizes. Investigating and developing VB algorithms for more complex models and data conditions is an important and challenging task for future research.

In the present study, we investigated models with between-item dimensionality that represent a rather simple special case of multidimensional IRT models and that mirror typical LSA designs. Stan’s built-in VB algorithm may be even further challenged under multidimensional IRT models with more complex structures, e.g., due to complex loading patterns under within-item dimensionality or multiplicative structures in non-compensatory multidimensional IRT models. At the same time, developing customized VB algorithms for more complex multidimensional IRT models poses a further pertinent topic for future research, as customary estimation of such models with MML or MCMC can be very computationally demanding, limiting their applicability [51].

Supplementary Materials: Exemplary R and Stan code is provided at <https://osf.io/swfqc/>.

Author Contributions: Conceptualization, E.U. and S.N.; methodology, E.U. and S.N.; software, E.U.; formal analysis, E.U.; investigation, E.U.; data curation, E.U.; writing—original draft preparation, E.U. and S.N.; writing—review and editing, S.N.; visualization, E.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Data employed in the empirical illustration are available at <https://www.oecd.org/pisa/data/2018database/> (accessed on 5 November 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADVI	Automatic differentiation variational inference
CAVI	Coordinate ascent variational inference
ELBO	Evidence lower bound
EM	Expectation–maximization
IRT	Item response theory
KL	Kullback–Leibler
LSA	Large-scale assessment
MCMC	Markov chain Monte Carlo
MML	Marginal maximum likelihood estimation
PIAAC	Programme for the International Assessment of Adult Competencies
PISA	Programme for International Student Assessment
VIBO	Variational item response theory lower bound
2PL model	Two-parameter logistic model

References

1. Robitzsch, A.; Kiefer, T.; Wu, M. TAM: Test Analysis Modules. 2021. R Package Version 3.6-45. Available online: <https://cran.r-project.org/web/packages/TAM/TAM.pdf> (accessed on 4 February 2020).
2. Chalmers, R.P. mirt: A multidimensional item response theory package for the R environment. *J. Stat. Softw.* **2012**, *48*, 1–29. [CrossRef]
3. Wu, M.; Adams, R.; Wilson, M. ConQuest: Generalised Item Response Modelling Software. 2020. Available online: <https://www.acer.org/au/conquest> (accessed on 4 February 2020).
4. Von Davier, M. A general diagnostic model applied to language testing data. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 287–307. [CrossRef] [PubMed]
5. Fox, J.P. *Bayesian Item Response Modeling: Theory and Applications*; Springer: New York, NY, USA, 2010.
6. Luo, Y.; Jiao, H. Using the Stan program for Bayesian item response theory. *Educ. Psychol. Meas.* **2018**, *78*, 384–408. [CrossRef] [PubMed]
7. Maris, G.; Bechger, T.; San Martin, E. A Gibbs sampler for the (extended) marginal Rasch model. *Psychometrika* **2015**, *80*, 859–879. [CrossRef] [PubMed]
8. Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria, 20–22 March 2003.
9. Stan Development Team. *Stan: A Probabilistic Programming Language*; Stan Modeling Language Users Guide and Reference Manual 2.25.0; Stan Development Team: Sydney, NSW, Australia, 2020.
10. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
11. Natesan, P.; Nandakumar, R.; Minka, T.; Rubright, J.D. Bayesian prior choice in IRT estimation using MCMC and variational Bayes. *Front. Psychol.* **2016**, *7*, 1422. [CrossRef] [PubMed]
12. OECD. *Technical Report of the Survey of Adult Skills (PIAAC)*; Technical Report; Organisation for Economic Co-operation and Development: Paris, France, 2013.
13. OECD. *PISA 2018 Technical Report*; Technical Report; Organisation for Economic Co-Operation and Development: Paris, France, 2020.
14. Kucukelbir, A.; Ranganath, R.; Gelman, A.; Blei, D. Fully Automatic Variational Inference of Differentiable Probability Models. NIPS Workshop on Probabilistic Programming. 2014. p. 69. Available online: https://statmodeling.stat.columbia.edu/wp-content/uploads/2014/12/pp_workshop_nips2014.pdf (accessed on 4 February 2020).

15. Kucukelbir, A.; Tran, D.; Ranganath, R.; Gelman, A.; Blei, D.M. Automatic differentiation variational inference. *J. Mach. Learn. Res.* **2017**, *18*, 430–474.
16. González, J.; Tuerlinckx, F.; De Boeck, P.; Cools, R. Numerical integration in logistic-normal models. *Comput. Stat. Data Anal.* **2006**, *51*, 1535–1548. [[CrossRef](#)]
17. Robert, C.; Casella, G. *Monte Carlo Stat. Methods*; Springer: New York, NY, USA, 2013.
18. Hoffman, M.D.; Gelman, A. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **2014**, *15*, 1593–1623.
19. Béguin, A.A.; Glas, C.A. MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika* **2001**, *66*, 541–561. [[CrossRef](#)]
20. Jordan, M.I.; Ghahramani, Z.; Jaakkola, T.S.; Saul, L.K. An introduction to variational methods for graphical models. *Mach. Learn.* **1999**, *37*, 183–233. [[CrossRef](#)]
21. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [[CrossRef](#)]
22. Wu, M.; Davis, R.L.; Domingue, B.W.; Piech, C.; Goodman, N.D. Variational item response theory: Fast, accurate, and expressive. *arXiv* **2020**, arXiv:abs/2002.00276.
23. Yamaguchi, K.; Okada, K. Variational Bayes inference for the DINA Model. *J. Educ. Behav. Stat.* **2020**, *45*, 569–597. [[CrossRef](#)]
24. Yamaguchi, K. Variational Bayesian inference for the multiple-choice DINA model. *Behaviormetrika* **2020**, *47*, 159–187. [[CrossRef](#)]
25. Yamaguchi, K.; Okada, K. Variational Bayes inference algorithm for the saturated diagnostic classification model. *Psychometrika* **2021**, *85*, 973–995. [[CrossRef](#)]
26. White, A.; Murphy, T.B. BayesLCA: An R package for Bayesian latent class analysis. *J. Stat. Softw.* **2014**, *61*, 1–28. [[CrossRef](#)]
27. Galdo, M.; Bahg, G.; Turner, B.M. Variational Bayesian methods for cognitive science. *Psychol. Methods* **2020**, *25*, 535–559. [[CrossRef](#)]
28. Ostwald, D.; Kirilina, E.; Starke, L.; Blankenburg, F. A tutorial on variational Bayes for latent linear stochastic time-series models. *J. Math. Psychol.* **2014**, *60*, 1–19. [[CrossRef](#)]
29. Rijmen, F.; Jeon, M. Fitting an item response theory model with random item effects across groups by a variational approximation method. *Ann. Oper. Res.* **2013**, *206*, 647–662. [[CrossRef](#)]
30. Cho, A. E.; Wang, C.; Zhang, X.; Xu, G. Gaussian variational estimation for multidimensional item response theory. *Br. J. Math. Stat. Psychol.* **2021**, *74*, 52–85. [[CrossRef](#)]
31. Jeon, M.; Rijmen, F.; Rabe-Hesketh, S. A variational maximization–maximization algorithm for generalized linear mixed models with crossed random effects. *Psychometrika* **2017**, *82*, 693–716. [[CrossRef](#)] [[PubMed](#)]
32. Hui, F.K.C.; Warton, D.I.; Ormerod, J.T.; Haapaniemi, V.; Taskinen, S. Variational approximations for generalized linear latent variable models. *J. Comput. Graph. Stat.* **2017**, *26*, 35–43. [[CrossRef](#)]
33. Kucukelbir, A.; Ranganath, R.; Gelman, A.; Blei, D.M. Automatic variational inference in Stan. *arXiv* **2015**, arXiv:1506.03431.
34. Muthén, L.K.; Muthén, B.O. How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equ. Model.* **2002**, *9*, 599–620. [[CrossRef](#)]
35. Yao, Y.; Vehtari, A.; Simpson, D.; Gelman, A. Yes, But Did It Work?: Evaluating Variational Inference. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018, pp. 5581–5590.
36. Stan Development Team. RStan: The R interface to Stan. 2020. R Package Version 2.19.3. Available online: <https://cran.r-project.org/web/packages/rstan/rstan.pdf> (accessed on 4 February 2020).
37. Chang, M.I.; Sheng, Y. A comparison of two MCMC algorithms for the 2PL IRT model. In *The Annual Meeting of the Psychometric Society*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 71–79.
38. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing; R Core Team: Vienna, Austria, 2020.
39. Chatzilena, A.; Van Leeuwen, E.; Ratmann, O.; Baguelin, M.; Demiris, N. Contemporary statistical inference for infectious disease models using Stan. *Epidemics* **2019**, *29*, 100367. [[CrossRef](#)]
40. Lee, C.Y.Y.; Wand, M.P. Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. *Biom. J.* **2016**, *58*, 868–895. [[CrossRef](#)]
41. Luts, J.; Wand, M.P. Variational inference for count response semiparametric regression. *Bayesian Anal.* **2015**, *10*, 991–1023. [[CrossRef](#)]
42. Giordano, R.; Broderick, T.; Jordan, M.I. Covariances, rbustness, and variational Bayes. *J. Mach. Learn. Res.* **2018**, *19*, 1–49.
43. Koops, J.; Bechger, T.; Maris, G. Bayesian inference for multistage and other incomplete designs. *arXiv* **2020**, arXiv:10.31234/osf.io/u9ekt.
44. Fasano, A.; Durante, D.; Zanella, G. Scalable and accurate variational Bayes for high-dimensional binary regression models. *arXiv* **2019**, arXiv:1911.06743.
45. Wand, M.P. Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *J. Am. Stat. Assoc.* **2017**, *112*, 137–168. [[CrossRef](#)]
46. Jaakkola, T.S.; Jordan, M.I. Bayesian parameter estimation via variational methods. *Stat. Comput.* **2000**, *10*, 25–37. [[CrossRef](#)]
47. Nolan, T.H.; Wand, M.P. Accurate logistic variational message passing: Algebraic and numerical details. *Stat* **2017**, *6*, 102–112. [[CrossRef](#)]

-
48. Knowles, D.; Minka, T. Non-conjugate variational message passing for multinomial and binary regression. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 1701–1709.
 49. Han, K.C.T.; Paek, I. A review of commercial software packages for multidimensional IRT modeling. *Appl. Psychol. Meas.* **2014**, *38*, 486–498. [[CrossRef](#)]
 50. Garnier-Villareal, M.; Merkle, E.C.; Magnus, B.E. Between-item multidimensional IRT: How far can the estimation methods go? *Psych* **2021**, *3*, 404–421. [[CrossRef](#)]
 51. Wang, C.; Nydick, S.W. Comparing two algorithms for calibrating the restricted non-compensatory multidimensional IRT model. *Appl. Psychol. Meas.* **2015**, *39*, 119–134. [[CrossRef](#)]