Handling Missing Data In Variational Autoencoder based Item Response Theory

July 9, 2024

Abstract

Recently Variational Autoencoders (VAEs) have been proposed as a method to estimate high dimensional Item Response Theory (IRT) models on large datasets. Although these improve the efficiency of estimation drastically compared to traditional methods, they have no natural way to deal with missing values. In this paper, we adapt three existing methods from the VAE literature to the IRT setting and propose one new method. We compare the performance of the different VAE-based methods to each other and to marginal maximum likelihood estimation for increasing levels of missing data in a simulation study for both three- and ten-dimensional IRT models. Additionally, we demonstrate the use of the VAE-based models on an existing algebra test dataset. Results confirm that VAE-based methods are a time-efficient alternative to marginal maximum likelihood, but that a larger number of importance-weighted samples are needed when the proportion of missing values is large.

Keywords: Multidimensional Item Response Theory, Variational Autoencoders, Missing Data

Data availability statement

All data and code required to reproduce our experiments are available on our GitHub: `https://anonymous.4open.science/r/VAE-MIRT-Missing-CE67/`.

1 Introduction

Item response theory (IRT) is a popular latent variable framework in educational testing and psychology (Lord & Novick, 1968). It is used to estimate scores on latent variables based on a set of categorical item responses. This can be either a single latent variable in traditional IRT or multiple latent variables in multidimenisonal item response theory (MIRT; Reckase, 2009) . Applications of these models include large-scale online educational testing (Klinkenberg et al., 2011; Meyer & Zhu, 2013) and clinical assessment (Thomas, 2019), and often involve large datasets and high dimensional latent variables.

The classical way to estimate MIRT models is with marginal maximum likelihood (MML) using the expectation maximisation (EM) algorithm (Bock & Aitkin, 1981). This involves maximizing the marginal likelihood, in which the latent variables are integrated out. The dimensionality of this integral depends on the number of latent variables in the IRT model, making traditional Gauss-Hermite quadrature approximations computationally expensive for higher dimensional IRT models (e.g., Wood et al., 2002). Using adaptive quadrature can reduce the number of quadrature points needed per dimension (Schilling & Bock, 2005), but the number of quadrature points still grows exponentially with the number of dimensions.

Contemporary research has focused on various ways to approximate the integral in the marginal likelihood efficiently, such as Markov chain Monte Carlo methods (e.g., Edwards, 2010) or stochastic approximation (e.g., Cai, 2010; von Davier & Sinharay, 2010). Although these methods substantially improve the efficiency of estimation of multi-dimensional models, the exponential growth of complexity as the latent dimensionality increases still limits the use of these methods in high-dimensional applications.

Yet another approach that avoids the computation of the marginal likelihood is the use of variational inference (Blei et al., 2017). Instead of maximizing the marginal likelihood, these methods approximate the posterior distribution of the latent variables with a different distribution and maximize the resulting lower bound on the marginal log-likelihood. Recently, Cho et al. (2021) have successfully used these methods to estimate up to 7 dimensional IRT models. Ma et al. (2023) showed that such variational estimation of MIRT parameters might lead to slightly biased discrimination estimates, but show that taking multiple importance weighted samples from the approximate posterior can mitigate this issue, at the cost of slightly higher computational time. The computational efficiency of these methods can be further improved using amortised variational inference (AVI) (Zhang et al., 2018). Rather than optimizing posterior parameters directly, in AVI a function is estimated that predicts the posterior parameters from the observed data by maximizing the same lower bound. By equating the parameters of this function across all observations, AVI drastically improves efficiency for large datasets. The most canonical form of AVI is the variational autoencoder (VAE), which uses a feedforward neural network (FNN) (Svozil et al., 1997) as the function estimating the posterior parameters. This is a class of models used in the machine learning community that serve as universal function approximators (Hornik et al., 1989). Curi et al. (2019) use this VAE approach to estimate high dimensional MIRT models on large datasets and show that the estimation is on par with traditional MML while being up to 40 times faster in estimation. Later research expands on this idea, showing that a generalization of the lower bound using importance sampling can approximate the true marginal likelihood arbitrarily well (Urban & Bauer, 2021; Burda et al., 2015). Urban & Bauer (2021) use this adapted lower bound to estimate MIRT models with up to ten dimensions. Moreover, research has generalized this approach to various applications, such as correlated posteriors (Converse et al., 2021), and three- and four-parameter logistic models (Liu et al., 2022).

Despite being a promising framework for estimating high dimensional MIRT models, a drawback of the VAE approach is that dealing with missing data is not straightforward. Where traditional methods draw strength from being full information maximum likelihood procedures in which missing data can easily be dealt with, a FNN can not handle missing input values naturally. There are various approaches to missing data in the general literature on VAEs, but only two pragmatic and potentially suboptimal methods have been adapted to the MIRT context (Liu et al., 2022; Montecino, 2023). In this study, we adapt three missing data techniques from the VAE literature to VAE-based MIRT modeling, and propose one new technique. In a simulation study in which we vary the degree of missing data, the performances of the four approaches are systematically compared to each other and to full information MML using Metropolis-Hastings Robbins-Monro (MHRM) estimation, which is a state-of-the-art MML estimation procedure (Cai, 2010).

In this paper we focus on estimation based on (a lower bound to) the marginal likelihood, as arguably, marginal likelihood based estimation is currently the most popular in IRT. However we note that there are other estimation procedures that are not based on the (marginal) likelihood which can give a fast alternative to high-dimensional MML estimation, such as constrained (Chen et al., 2019) and regularized (Bergner et al., 2022) joint-maximum likelihood, and least squares estimation (Browne, 1974; Muthén, 1984).

The outline of this paper is as follows: First, we briefly discuss FNNs, as these are important tools used in VAE-based MIRT. We then formally introduce MIRT and show the challenges associated with MML for high-dimensional IRT models. Next, we discuss variational inference and VAEs for MIRT models and derive different approaches to dealing with missing data in this paradigm. We present a thorough parameter recovery simulation study comparing the different missing data techniques to each other and to MML, and a short second simulation to demonstrate that the proposed models can deal with correlated latent variables. Finally, we demonstrate the use of the methods on a real dataset pertaining to a large scale algebra test, and end with our general conclusion and recommendations on how to deal with missing data in VAE based MIRT modeling.

2 Feed forward neural networks

FNNs are a popular class of methods in the field of machine learning. They map an p dimensional vector of predictor variables $\mathbf{x}$ to an o dimensional vector of outcome variables $\mathbf{y}$. As the name suggests, a feedforward neural network consists of multiple layers, where the values of variables in the next layer depend on the values of the previous layer:

$$\mathbf{h}_{i+1} = f_i(\mathbf{W}_i\mathbf{h}_i + \mathbf{b}_i), i = 0, 1, \ldots I, \tag{1}$$

where $\mathbf{W}_i$ is a matrix of weight parameters for layer i, $\mathbf{b}_i$ is a vector of bias parameters for layer i, and f_i is a nonlinear function called an activation function. $\mathbf{h_i}$ denotes the n_i dimensional representation vector in layer i, where n_i is the dimensionality of the layer. Note that for convenience $\mathbf{h}_0$ denotes the N dimensional vector of original observations $\mathbf{x}$, and $\mathbf{h}_I$ denotes the outcome predictions $\hat{\mathbf{y}}$. The intermediate representations $\mathbf{h}_i, 0 < i < I$, are generally referred to as hidden layer activations, and the individual elements in $\mathbf{h}_i$ are referred to as hidden nodes. Conceptually, a FNN consists of multiple linear transformations, each followed by a nonlinear function. Hornik et al. (1989) has shown that theoretically, FNNs with just a single hidden layer can approximate any Borel measurable function to any degree of accuracy, given enough hidden nodes, making FNNs a form of universal function approximators.

The parameters of an FNN are generally estimated by numerical optimization of a loss function $\Lambda(\mathbf{y}, \hat{\mathbf{y}})$, such as the log-likelihood, using gradient descent. Modern implementations use Stochastic gradient descent (SGD) (Amari, 1993). The core idea of SGD is to iteratively update parameter values by partitioning the dataset into random disjoint subsets $\{\mathbf{x}_b\}_{b=1}^{B}$ (called mini-batches), computing the predicted values $\{\hat{\mathbf{y}}_b\}_{b=1}^{B}$, and updating the parameters by adding the gradients of the loss function with respect to the weights and biases multiplied by a learning rate, which is a hyperparameter that determines the size of the updates. This process is repeated until some stopping criterion is reached. The partitioning of the dataset allows the estimation process to be faster, while also making it less sensitive to local minima. The specific algorithm used in this study is the AMSGrad algorithm (Reddi et al., 2019), which is currently one of the most popular adaptations of SGD for FNNs. It makes use of an adaptive learning rate, allowing for bigger updates when gradients are small accelerating convergence, and smaller updates when gradients are big, preventing overshooting. We refer the reader to Reddi et al. (2019) for a detailed discussion of AMSGrad.

3 MIRT

The Multidimensional 2PL (M2PL) model is the most common MIRT model (McKinley & Reckase, 1983) due to its relation to item factor analysis (Wirth & Edwards, 2007; Takane & De Leeuw, 1987). In this model, the probability of a correct response is modeled as

$$P(X_{ij} = 1|\boldsymbol{\theta}_j) = \frac{1}{1 + \exp(-\mathbf{a}_j^T \boldsymbol{\theta}_i - b_j)}, \tag{2}$$

where X_{ij} is a random variable representing the response of participant i on item j, $\boldsymbol{\theta}_i$ is the latent variable vector for person i, $\mathbf{a}_j$ is the item slope vector for item j, and b_j is the item intercept for item j. The latent variable parameter vectors represent scores on the multidimensional latent ability constructs, whereas the item parameters represent the degree to which an item discriminates between different levels of each latent construct, as well as how likely an item is to receive a positive response generally. The marginal log-likelihood for person i is given by

$$\log P(X_i = \mathbf{x}_i|\Omega) = \log \int \cdots \int \prod_{j=1}^{J} P(X_{ij} = x_{ij}|\boldsymbol{\theta};\Omega_j)p(\boldsymbol{\theta})d\theta_1 d\theta_2 \cdots d\theta_D, \tag{3}$$

where X_i is a vector of random variables representing the binary item responses for subject i with elements X_{ij}, $\Omega_j = \{\mathbf{a}_j, b_j\} \in \Omega$ is the set of item parameters for item j, $p(\boldsymbol{\theta})$ is the prior probability of the ability estimate $\boldsymbol{\theta}$, and D is the number of latent dimensions. Some elements of the item discrimination parameters $\mathbf{a}_j, j = 1, \ldots, J$, have to be fixed to zero in order to avoid rotational indeterminacy. The complete marginal log-likelihood is simply the sum over all participants.

3.1 Variational inference

As discussed in the introduction, variational inference provides an alternative way to estimate MIRT parameters, avoiding the need for this computationally expensive integral that needs to be approximated in MML. In variational inference, rather than optimizing the marginal likelihood directly, the posterior density $p(\boldsymbol{\theta}_i|\mathbf{x}_i)$, for which there is generally no closed expression, is approximated by another density $\hat{q}(\boldsymbol{\theta}_i|\mathbf{x}_i)$ such that the Kullback-Leibler (KL) divergence is minimized over a (parameterized) family $\mathcal{F}$ of proposal distributions:

$$\hat{q}(\boldsymbol{\theta}_i \mid \mathbf{x}_i) = \underset{q(\boldsymbol{\theta}|\mathbf{x}) \in \mathcal{F}}{\operatorname{argmin}} \, \mathrm{KL}[q(\boldsymbol{\theta}_i|\mathbf{x}_i)||p(\boldsymbol{\theta_i}|\mathbf{x}_i)] \tag{4}$$

(Blei et al., 2017). Since this KL divergence in this expression requires the true posterior, we can not optimize it directly. However, it can be shown that minimizing this KL divergence is equivalent to maximizing the evidence lower bound (ELBO), which is defined as

$$\mathrm{ELBO} = \mathbb{E}_{\hat{q}}[\log p(\mathbf{x}_i|\boldsymbol{\theta}_i)] - \mathrm{KL}[q(\boldsymbol{\theta}_i|\mathbf{x}_i)||p(\boldsymbol{\theta}_i)], \tag{5}$$

where $p(\boldsymbol{\theta}_i)$ is the prior density over $\boldsymbol{\theta}_i$ (i.e., the population density of the latent ability scores). The ELBO derives its name from the fact that it is a lower bound to the marginal log-likelihood. Specifically, the marginal log-likelihood can be expressed as

$$\log p(\mathbf{x}_i) = \mathrm{KL}[\hat{q}(\boldsymbol{\theta}_i|\mathbf{x}_i)||p(\boldsymbol{\theta}_i|\mathbf{x}_i)] + \mathrm{ELBO}. \tag{6}$$

This indicates that the ELBO approaches the marginal log-likelihood as the approximation distribution approaches the true posterior. Choosing a flexible family $\mathcal{F}$ for q allows this divergence to be minimized as much as possible. We will address the choice for q below. Variational inference has successfully been used by as method of estimating MIRT parameters in the past (Cho et al., 2021; Ma et al., 2023).

3.2 Amortised variational inference

Where traditional variational inference optimizes the parameters of the approximate posterior directly, amortized variational inference estimates a mapping from the observed data to the posterior parameters. This means that instead of estimating a separate posterior distribution for the latent ability of each respondent, we estimate the parameters of a so-called inference model that predicts the posterior distribution of the latent ability parameters from a pattern of item responses. This function is generally chosen to be approximated with a FNN (Curi et al., 2019; Urban & Bauer, 2021; Converse et al., 2021; Liu et al., 2022; Wu et al., 2020), and the parameters are shared across observations. As a result, if the posterior distribution of the latent variables is assumed to be a multivariate normal distribution, the latent variables $\boldsymbol{\theta}_i$ are modeled by:

$$\boldsymbol{\theta}_i \sim MVN(\boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i)), \text{ with } (\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) = \mathrm{FNN}_\phi(\mathbf{x}_i). \tag{7}$$

where ϕ are the parameters of the inference model. The full model is a VAE (Kingma & Welling, 2013) in which the parameters are estimated iteratively by taking a sample $\hat{\boldsymbol{\theta}}_i$ from the multivariate normal (cf. Equation (7)), with the FNN evaluated at the current best estimates of $\hat{\phi}$. Then, $\hat{\Omega}$ and $\hat{\phi}$ are jointly updated to maximize the ELBO with respect to these parameters. That is, the new estimates are

$$\hat{\phi}, \hat{\Omega} = \underset{\phi,\Omega}{\mathrm{argmax}}\, \mathbb{E}[\log p(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_i; \Omega)] - KL[q(\boldsymbol{\theta}_i|\mathbf{x}_i; \phi)||p(\boldsymbol{\theta}_i)], \tag{8}$$

where the approximate posterior $q(.)$ is given by Equation (7). In the simplest case, the population density $p(\boldsymbol{\theta}_i)$ is fixed to a standard multivariate normal density, but it is straightforward to accommodate correlated latent abilities in this prior. Below we discuss how this assumption of a normal approximate posterior can be relaxed.

3.3 Importance weighted amortized variational inference

As discussed above, VAEs generally assume that the posterior distribution of $\boldsymbol{\theta}_i$ is normal and highly factorized (e.g., multivariate normal with diagonal covariance matrix). These constraints on the posterior can cause the ELBO to severely underestimate the true marginal log-likelihood. Recent research has focused on various ways of tightening this lower bound (Rezende & Mohamed, 2015; Burda et al., 2015). One of these approaches, proposed by Burda et al. (2015), is the importance-weighted variational autoencoder (IWVAE). This model has the same structure as a regular VAE, but instead of taking a single sample of the approximate posterior, multiple samples are taken, in order to calculate an importance-weighted estimate of the log-likelihood:

$$\text{IW-ELBO} = \mathbb{E}_{1:k}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_i^{(k)})}{q(\hat{\boldsymbol{\theta}}_i^{(k)}|\mathbf{x}_i)}\right], \tag{9}$$

where K is the number of importance samples and the term inside the sum are the unnormalized importance samples for the joint distribution. When $K = 1$ the importance-weighted evidence lower bound (IW-ELBO) is equivalent to the regular ELBO in Equation (5). Using multiple importance-weighted samples from the approximate posterior tightens the lower bound to the marginal log-likelihood. Burda et al. (2015) show that the IW-ELBO approaches $\log p(\mathbf{x}_i)$ as K goes to infinity, making equation (9) a MML estimator. In a later paper Cremer et al. (2017) showed that by taking these multiple importance-weighted samples from the approximate posterior, you are effectively using a more flexible mixture distribution as your approximate posterior, relaxing the assumption that the posterior must be a multivariate normal. They also demonstrate that you can compute expected a posteriori (EAP) ability estimates from this more flexible posterior by using importance sampling. Specifically, you can sample from the approximate posterior by taking K samples from the multivariate normal distribution, and taking a sample from these K samples using the importance weights. By repeating this process M times and taking the mean a Monte Carlo estimate of the EAP ability can be obtained. The IWVAE was first used in the context of MIRT by Urban & Bauer (2021) who show empirically that the IWVAE can provide accurate MIRT parameter estimates for up to 10 dimensions while being much faster than traditional EM algorithms.

3.4 The challenge of missing data

Although the amortization step allows IWVAEs to efficiently estimate complex latent variable models, it introduces a problem regarding missing data. Since the model now includes an inference model that predicts the parameters of the latent variable distribution based on an observed response pattern (equation (7)), this function needs to be capable of dealing with missing data. A standard FNN does not meet this requirement, since missing values make it impossible to calculate gradients with respect to the network parameters. Simply neglecting the missing values in $\boldsymbol{x}_i$ will cause $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ in Equation (7) to be on a different scale for respondents with different missing value patterns which will greatly bias the results. Furthermore, using list-wise deletion, pair-wise deletion, or mean-imputation are not recommended in general for loss of statistical efficiency. Liu et al. (2022) employ an intuitive trick to handle missing data in VAEs. However, as will be discussed below, this approach is suboptimal and potentially biases item parameter estimates, especially in cases where the majority of the data is missing. Thus, more work is needed to adapt VAE-based IRT models to missing data. Below, we derive four explicit approaches to handle missing data in IWVAE and provide a systematic comparison.

4 Handling missing data

4.1 Input drop-out

An intuitive way of dealing with missing data is the so-called *input drop-out trick* (Nazabal et al., 2020; Liu et al., 2022). This trick consists of replacing the missing values in the input pattern with zeros, and only calculating the loss of the VAE over the non-missing datapoints. The new loss function for the VAE using the input drop-out trick is

$$ELBO_{ID} = \mathbb{E}_{q(\boldsymbol{\theta}_i|\tilde{\mathbf{x}}_i;\phi)}[\log p(x_{ij}, j \in O_i|\boldsymbol{\theta}_i;\Omega)] - KL[q(\boldsymbol{\theta}_i|\tilde{\mathbf{x}}_i;\phi)||p(\boldsymbol{\theta}_i)], \quad (10)$$

where $\tilde{\mathbf{x}}_i$ is the response pattern with missing values replaced by zero, and O_i is the set of indices of observed item responses. This is the method that was used in the context of MIRT by Liu et al. (2022).

A problem with the input-dropout trick is that although the reconstruction loss is only calculated over the observed data points, the inference model has no way to distinguish between missing values and responses that are truly zero, making it harder to estimate the latent ability variables. This may especially be suboptimal in MIRT applications, as zero responses are frequent in binary response data.

4.2 Conditional VAEs

To mitigate the problem above, Collier et al. (2020) introduced the *conditional variational autoencoder (CVAE).* In this approach, the loss function is equivalent to the input drop-out loss $ELBO_{ID}$ in equation (10), but the inference network now takes both the response pattern and the mask of missing data as input:

$$\boldsymbol{\theta}_i \sim N_d(\boldsymbol{\theta};\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i), \text{ with } (\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i) = \text{FNN}_\phi(\mathbf{x}_i,\mathbf{m}_i). \quad (11)$$

Here $\mathbf{m}_i$ is a binary *mask* vector that indicates which values are missing in $\mathbf{x}_i$ and which are not—i.e., $m_{ij} = 1$ if $j \in O_i$ and $m_{ij} = 0$ if $j \notin O_j$. This additional input to the inference model allows it to differentiate between missing values and true zero responses.

This model has not been considered yet in the context of MIRT, and we consider this model as an alternative approach to handling missing data in this context.

4.3 Partial VAEs

A third approach to missing data consists of *partial variational autoencoders (PVAE)* (Ma et al., 2018). PVAEs were originally developed for the application of recommender systems, in which large fractions of missing data are common. They are based on neural networks for point net classification[1] that can deal with variable input sizes (Qi et al., 2017). Although FNNs generally require a fixed input size for each respondent, PVAEs bypass this requirement by estimating a so-called *embedding* vector for each item. An embedding vector in this context is an estimated vector of parameters. By combining the embeddings from all items with an observed response using a permutation invariant aggregation operation, unobserved items can be simply ignored in the hope that the estimated embeddings will still encode essential information about $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$.

$$\boldsymbol{\theta}_i \sim N(\boldsymbol{\theta}; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i), \text{ with } (\boldsymbol{\mu}_{,}\boldsymbol{\sigma}_i) = \mathrm{FNN}_\phi(g(\{\mathbf{e}_j : j \in O_i\})). \tag{12}$$

Here $\mathbf{e}_j$ denotes the embedding for item j, and g is a well-defined permutation invariant function that maps a set of points in embedding space to a single point in that same space, for example, the mean vector, or the vector of coordinate-wise maxima. This function g allows for a variable amount of embeddings and returns an output vector of a fixed length. The FNN takes the output from g and computes the variational parameters as in a regular VAE. Alternatively, g can be extended to return a vector of descriptive statistics of the embeddings, such as the mean, standard deviation, or distribution quantiles, to yield a more detailed representation of the distribution of the embedding vectors.

4.4 Imputation VAE

Finally, we propose a new approach which we will refer to as the *imputation variational autoencoder (IMVAE)*. In this model, we use the same input dropout loss function $ELBO_{ID}$ in equation (10), but rather than imputing the input with zeros, we impute the missing values with the expected values for the responses given the current IRT model parameter estimates. Specifically, we set

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i \odot \mathbf{m}_i + p(\mathbf{x}_i | \hat{\boldsymbol{\theta}}_i; \hat{\Omega}) \odot (\mathbf{1} - \mathbf{m}_i), \tag{13}$$

where $\odot$ denotes the Hadamard products, $\mathbf{1}$ is a commensurate vector of ones, and $\hat{\boldsymbol{\theta}}_i$ again is the vector of latent variable estimates for subject i based on the current estimates $\hat{\phi}$ of the inference model parameters. In the first iteration, $\hat{\boldsymbol{\theta}}_i$ is set to the expected value of the prior. This adaptation still calculates the loss using only the observed data patterns but allows the inference model to make better estimates of the latent ability variables by using the current model parameters to estimate the most likely values of the missing data. Note that our proposed method is similar to an approach proposed by Montecino (2023, Chapter 3).[2]. However, our approach differs in two ways: First, we

[1] This concerns the classification of objects based on unordered observed points in 3D space.

[2] We thank the anonymous reviewer that brought this dissertation to our attention

use the actual predicted probabilities while Montecino (2023) rounded these probabilities to zero and one which distorts the stochastic properties of the responses. Additionally, Montecino (2023) used a regular VAE while we embed our approach in an IWVAE which -as discussed above- will provide a closer approximation to the true posterior.

5 Simulation study

We simulated datasets of 10.000 respondents from three and ten-dimensional MIRT models. Latent variable values were sampled from a standard normal $\boldsymbol{\theta} \sim N_d(0, \mathbf{I}_d)$. For simplicity, latent variables were uncorrelated. However, note that the flexibility of the approximate posterior introduced by taking multiple importance samples does actually allow for correlations between factors, which we will demonstrate in Section 7. Item slope parameters were sampled from a uniform distribution $a_j \sim U(.5, 2), j = 1, \dots, J$, and intercept parameters b_j were set to equally spaced values between -2 and 2. Several slope parameters were set to zero. For the three-dimensional model, we simulate 28 items and fix slope parameters to zero based on a factor configuration from the literature (da Silva et al., 2019; Curi et al., 2019). For the ten-dimensional model, we simulate 110 items based on our own configuration where each item depends on one or two latent factors. Both configurations are available on the GitHub page,[3] along with all of the code necessary to reproduce our experiments or to apply our models to new data.

Missing values were introduced at random by deleting a proportion of observed responses. The proportion of missing data was varied to ten equally spaced values between 0 and .75, inclusive. We ran 50 replications for each combination of dimensionality and missingness, resulting in $1,000$ unique datasets. We fit the input dropout variational autoencoder (IDVAE), CVAE, PVAE, and IMVAE to each dataset and estimate each model using 1, 5, and 25 importance weight samples, resulting in a 2 by 10 by 4 by 3 factorial design. We compare results to MML using state-of-the-art MHRM (Cai, 2010) estimation as implemented in the the R package *mirt* (Chalmers, 2012). We used the EAP ability estimates for the MML models, which were estimated using quasi=Monte-Carlo estimating in the ten-dimensional case. For the VAE-based models we use the ability estimation procedure outlined in Section 3.2, which produces a Monte-Carlo estimate of the expectation of the approximate posterior.

5.1 Implementation details

All our models are implemented in PyTorch (Paszke et al., 2019), and all code is publicly available on GitHub. Models were estimated using the AMSgrad algorithm using a learning rate of .005 and a batch size of 32. We stopped estimation when the loss did not decrease by more than 1e-7 over 10 iterations.

[3] https://anonymous.4open.science/r/VAE-MIRT-Missing-CE67/

We used a single hidden layer of 20 nodes for the IDVAE, CVAE, and the IMVAE. In the PVAE, an embedding of length 12 was learned for each item. these embeddings were transformed to a vector of length 24 using a single layer FNN. We transform these item vectors to a fixed-length person vector using a permutation invariant operation. Specifically, we compute the mean, median, standard deviation, and the 25th and 75th quartiles of each person's item distribution, and concatenate the results, resulting in a vector of length 120 for each person. The resulting person vectors are used as input for a regular VAE encoder with a single hidden layer of 24 nodes. We have also experimented with using a single permutation invariant function, but using a concatenation has provided better results in our experience.

6 Results

Figure 1 presents the average mean square error of the parameter estimates for the different models and different numbers of importance-weighted samples as the proportion of missing data increases. Parameter estimates using a single importance sample were very poor and were left out of the figure to prevent obscuring the image. Overall we see that the accuracy of VAE-based parameter estimates decreases rapidly as the proportion of missing data increases. However, as the amount of importance samples increases the variational methods approach the accuracy of MML. This is in line with expectations, as increasing the number of samples tightens the bound to the marginal log-likelihood. As more data is missing a larger number of samples is needed to attain accurate parameter estimates. However, as becomes clear from Table 1[4], using VAE based methods with a large number of IW samples is still computationally much faster than MML in high dimensional applications. The complete results for the three- and ten-dimensional models are available in table 2 and 5 respectively.

Table 1: Average runtime (minutes) of each model over 10 replications. IW is the number of importance weighted samples and d is the dimensionality of the model. All models were estimated on a single CPU core of the Apple MacBook Pro M3.

		IW=1		IW=5		IW=25	
miss	model	$d = 3$	$d = 10$	$d = 3$	$d = 10$	$d = 3$	$d = 10$
0.00	cvae	0.55	.46	.47	.41	.66	.90
	idvae	0.53	.42	.53	.52	.53	.96
	ivae	0.54	.42	.59	.65	.52	.86
	pvae	1.67	1.46	.98	3.88	4.26	3.84
	mirt	4.27	19.05	-	-	-	-
	cvae	0.45	.55	.50	.53	.61	.94
	idvae	0.51	.48	.47	.49	.55	1.00

[4] Since the main simulation study was run on a cluster computer, where different simulation runs are executed on different computational nodes, we performed a separate 10 iteration study locally solely for the purpose of fairly comparing runtimes.

(continued)

		IW=1		IW=5		IW=25	
miss	model	$d = 3$	$d = 10$	$d = 3$	$d = 10$	$d = 3$	$d = 10$
0.25	ivae	0.51	.49	.43	.50	.60	.95
	pvae	1.18	1.48	1.03	2.11	3.47	3.57
	mirt	4.40	20.34	-	-	-	-
0.75	cvae	0.39	.48	.45	.46	.73	.91
	idvae	0.40	.50	.54	.35	.54	1.09
	ivae	0.42	.48	.57	.42	.62	.89
	pvae	1.02	1.36	1.27	1.67	2.15	2.17
	mirt	4.49	11.54	-	-	-	-

The relative performance of variational methods is generally consistent across conditions. The CVAE parameter estimates have the lowest MSE. The CVAE and IMVAE are very close with respect to their item parameter estimates, but the extra information provided to the inference model appears to allow the CVAE to estimate ability parameters with more precision. Both the CVAE and IMVAE are consistently more accurate than the IDVAE, which is in line with expectations, as these two methods can be conceived of as improvements upon the IDVAE. The PVAE is clearly the least effective method to estimate MIRT parameters, performing substantially worse than all other methods across conditions. Somewhat surprisingly, the variational method outperform MML in terms of the ability estimates, especially so in the high dimensional case. This could potentially be due to the fact that we use quasi-Monte Carlo integration to approximate the posterior theta distribution for the MML model.

Table 2: Simulation study results for three-dimensional models. m denotes the proportion of missing data, p denotes the parameter that is being estimated, and σ^2 denotes the average variance of the parameter estimates. Note that MML is arbitrarily placed under $IW = 1$ to keep the table concise,but it does not make use of importance weighting. For the sake of brevity, the table only contains three levels of missing data. The complete table is available on GitHub.

			IW=1			IW=5			IW=25		
p	m	model	σ^2	$bias^2$	mse	σ^2	$bias^2$	mse	σ^2	$bias^2$	mse
	0.00	cvae	0.0036	0.1524	0.1560	0.0037	0.0591	0.0628	0.0001	0.0256	0.0258
		idvae	0.0129	0.2059	0.2188	0.0024	0.0730	0.0754	0.0001	0.0258	0.0259
		ivae	0.0231	0.2236	0.2467	0.0022	0.0737	0.0759	0.0001	0.0258	0.0260
		pvae	0.0071	0.4332	0.4403	0.0005	0.1326	0.1331	0.0002	0.0354	0.0356
		mirt	0.0009	0.0000	0.0009	-	-	-	-	-	-
	0.25	cvae	0.0140	0.2853	0.2993	0.0015	0.1108	0.1123	0.0021	0.0380	0.0401
		idvae	0.0150	0.3967	0.4117	0.0012	0.1134	0.1146	0.0019	0.0375	0.0395
		ivae	0.0176	0.3994	0.4170	0.0013	0.1116	0.1129	0.0014	0.0368	0.0381
		pvae	0.0013	0.6304	0.6317	0.0015	0.1540	0.1555	0.0007	0.0491	0.0498
		mirt	0.0023	0.0000	0.0024	-	-	-	-	-	-
		cvae	0.0008	0.6378	0.6386	0.0054	0.5154	0.5208	0.0062	0.2039	0.2102
		idvae	0.0007	0.6381	0.6389	0.0070	0.4640	0.4710	0.0074	0.1876	0.1950
		ivae	0.0007	0.6371	0.6379	0.0065	0.4748	0.4812	0.0040	0.1754	0.1795
		pvae	0.0008	0.6376	0.6384	0.0048	0.5062	0.5111	0.0058	0.1817	0.1875

a											
	0.75										
		mirt	0.0859	0.0690	0.1549	-	-	-	-	-	-
b	0.00	cvae	0.0001	0.0555	0.0556	0.0002	0.0435	0.0436	0.0002	0.0335	0.0336
		idvae	0.0004	0.0652	0.0656	0.0001	0.0462	0.0462	0.0001	0.0339	0.0341
		ivae	0.0008	0.0685	0.0693	0.0001	0.0463	0.0464	0.0001	0.0341	0.0342
		pvae	0.0005	0.0828	0.0833	0.0001	0.0491	0.0492	0.0001	0.0371	0.0372
		mirt	0.0010	0.0000	0.0011	-	-	-	-	-	-
	0.25	cvae	0.0005	0.0798	0.0803	0.0001	0.0564	0.0565	0.0002	0.0454	0.0455
		idvae	0.0004	0.0945	0.0949	0.0001	0.0580	0.0581	0.0002	0.0456	0.0458
		ivae	0.0003	0.0949	0.0952	0.0001	0.0578	0.0579	0.0001	0.0457	0.0459
		pvae	0.0000	0.1025	0.1025	0.0001	0.0613	0.0614	0.0001	0.0482	0.0483
		mirt	0.0022	0.0000	0.0022	-	-	-	-	-	-
	0.75	cvae	0.0000	0.1208	0.1208	0.0000	0.1196	0.1196	0.0001	0.1015	0.1016
		idvae	0.0000	0.1208	0.1208	0.0000	0.1184	0.1184	0.0001	0.0992	0.0993
		ivae	0.0000	0.1208	0.1208	0.0000	0.1186	0.1186	0.0001	0.0980	0.0981
		pvae	0.0000	0.1209	0.1209	0.0000	0.1194	0.1194	0.0001	0.0984	0.0985
		mirt	0.0341	0.0182	0.0523	-	-	-	-	-	-
$\boldsymbol{\theta}$	0.00	cvae	0.0147	0.4927	0.5074	0.0232	0.3610	0.3842	0.0083	0.3363	0.3446
		idvae	0.0456	0.5239	0.5696	0.0211	0.3903	0.4115	0.0084	0.3368	0.3453
		ivae	0.0645	0.5314	0.5959	0.0202	0.3901	0.4103	0.0085	0.3367	0.3452
		pvae	0.0619	0.7426	0.8045	0.0163	0.4897	0.5061	0.0099	0.3418	0.3518
		mirt	0.1990	0.1369	0.3359	-	-	-	-	-	-
	0.25	cvae	0.0432	0.6278	0.6709	0.0163	0.5048	0.5211	0.0144	0.4098	0.4243
		idvae	0.0516	0.7057	0.7573	0.0155	0.5145	0.5300	0.0139	0.4103	0.4242
		ivae	0.0574	0.6983	0.7557	0.0151	0.5125	0.5277	0.0123	0.4094	0.4217
		pvae	0.0110	0.9799	0.9908	0.0211	0.5465	0.5675	0.0123	0.4209	0.4331
		mirt	0.2222	0.2526	0.4747	-	-	-	-	-	-
	0.75	cvae	0.0060	0.9952	1.0012	0.0275	0.8946	0.9221	0.0224	0.7492	0.7717
		idvae	0.0021	0.9948	0.9968	0.0279	0.8636	0.8915	0.0225	0.7443	0.7668
		ivae	0.0024	0.9960	0.9984	0.0274	0.8738	0.9013	0.0196	0.7389	0.7585
		pvae	0.0055	0.9975	1.0030	0.0261	0.8924	0.9184	0.0211	0.7412	0.7623
		mirt	0.0770	0.8328	0.9098	-	-	-	-	-	-

Table 3: Simulation study results for ten-dimensional model. m denotes the proportion of missing data, p denotes the parameter that is being estimated, and σ^2 denotes the variance of the parameter estimates. Note that MML is arbitrarily placed under $IW = 1$ to keep the table concise, but it does not make use of importance weighting. For the sake of brevity, the table only contains three levels of missing data. The complete table is available on GitHub.

			IW=1			IW=5			IW=25		
p	m	model	σ^2	$bias^2$	mse	σ^2	$bias^2$	mse	σ^2	$bias^2$	mse
	0.00	cvae	0.0158	0.0647	0.0805	0.0016	0.0249	0.0266	0.0001	0.0165	0.0167
		idvae	0.0141	0.0711	0.0852	0.0018	0.0243	0.0261	0.0003	0.0170	0.0173
		ivae	0.0130	0.0724	0.0854	0.0021	0.0245	0.0266	0.0002	0.0170	0.0172
		pvae	0.0039	0.2680	0.2720	0.0024	0.1024	0.1049	0.0009	0.0559	0.0568
		mirt	0.0004	0.0001	0.0004	-	-	-	-	-	-
	0.25	cvae	0.0143	0.1006	0.1148	0.0043	0.0374	0.0418	0.0009	0.0213	0.0222
		idvae	0.0141	0.1409	0.1549	0.0053	0.0393	0.0445	0.0004	0.0212	0.0216
		ivae	0.0169	0.1454	0.1623	0.0041	0.0353	0.0394	0.0007	0.0202	0.0209
		pvae	0.0002	0.3095	0.3097	0.0038	0.1150	0.1188	0.0013	0.0593	0.0606
		mirt	0.0009	0.0000	0.0009	-	-	-	-	-	-
		cvae	0.0003	0.3100	0.3103	0.0051	0.1932	0.1983	0.0059	0.0770	0.0830
		idvae	0.0003	0.3098	0.3101	0.0074	0.1645	0.1719	0.0058	0.0594	0.0651

a	0.75	ivae	0.0003	0.3101	0.3104	0.0052	0.1923	0.1976	0.0068	0.0711	0.0779
		pvae	0.0003	0.3102	0.3105	0.0044	0.2057	0.2101	0.0055	0.0826	0.0881
		mirt	0.0393	0.0034	0.0427	-	-	-	-	-	-
b	0.00	cvae	0.0014	0.1665	0.1679	0.0003	0.2119	0.2122	0.0000	0.2193	0.2193
		idvae	0.0010	0.1707	0.1717	0.0001	0.2141	0.2142	0.0000	0.2194	0.2194
		ivae	0.0010	0.1677	0.1687	0.0002	0.2139	0.2141	0.0001	0.2187	0.2187
		pvae	0.0001	0.1451	0.1453	0.0002	0.1658	0.1660	0.0001	0.1966	0.1967
		mirt	0.0011	0.0003	0.0014	-	-	-	-	-	-
	0.25	cvae	0.0003	0.1342	0.1345	0.0004	0.1691	0.1695	0.0001	0.1860	0.1861
		idvae	0.0005	0.1421	0.1426	0.0002	0.1796	0.1798	0.0001	0.1911	0.1911
		ivae	0.0005	0.1429	0.1434	0.0002	0.1783	0.1785	0.0001	0.1907	0.1907
		pvae	0.0000	0.1528	0.1528	0.0002	0.1526	0.1528	0.0001	0.1766	0.1767
		mirt	0.0020	0.0009	0.0028	-	-	-	-	-	-
	0.75	cvae	0.0000	0.1713	0.1713	0.0000	0.1653	0.1654	0.0001	0.1495	0.1496
		idvae	0.0000	0.1713	0.1713	0.0001	0.1616	0.1617	0.0001	0.1491	0.1492
		ivae	0.0000	0.1713	0.1713	0.0000	0.1656	0.1657	0.0001	0.1503	0.1504
		pvae	0.0000	0.1713	0.1713	0.0000	0.1662	0.1663	0.0001	0.1509	0.1510
		mirt	0.0454	0.0035	0.0489	-	-	-	-	-	-
$\boldsymbol{\theta}$	0.00	cvae	0.1169	0.3778	0.4947	0.0206	0.3279	0.3485	0.0092	0.3184	0.3276
		idvae	0.1237	0.3992	0.5229	0.0278	0.3288	0.3566	0.0133	0.3194	0.3327
		ivae	0.1202	0.4083	0.5284	0.0292	0.3301	0.3593	0.0132	0.3195	0.3327
		pvae	0.1113	0.8111	0.9225	0.0654	0.4859	0.5513	0.0297	0.3939	0.4236
		mirt	0.1967	0.1809	0.3777	-	-	-	-	-	-
	0.25	cvae	0.1166	0.4746	0.5912	0.0349	0.3925	0.4274	0.0138	0.3700	0.3838
		idvae	0.1335	0.5466	0.6801	0.0450	0.4066	0.4516	0.0150	0.3720	0.3870
		ivae	0.1578	0.5336	0.6914	0.0382	0.3956	0.4338	0.0161	0.3698	0.3859
		pvae	0.0075	0.9932	1.0006	0.0648	0.5339	0.5986	0.0284	0.4383	0.4667
		mirt	0.2049	0.2304	0.4353	-	-	-	-	-	-
	0.75	cvae	0.0052	0.9959	1.0011	0.0419	0.7497	0.7916	0.0316	0.6201	0.6518
		idvae	0.0023	0.9950	0.9973	0.0483	0.7107	0.7590	0.0294	0.6080	0.6374
		ivae	0.0037	0.9958	0.9995	0.0409	0.7528	0.7937	0.0335	0.6160	0.6495
		pvae	0.0063	0.9966	1.0029	0.0362	0.7852	0.8214	0.0298	0.6349	0.6647
		mirt	0.1271	0.7221	0.8493	-	-	-	-	-	-

7 Correlated latent variables

Although we have so far only considered data obtained with uncorrelated latent factors in order to keep the simulation design simple, this is not a requirement for the VAE-based approach: Because the IWAE encoder can be used to sample from an approximation of the posterior that approaches the true posterior as $K \to \infty$ (Cremer et al., 2017), we can obtain an estimate of the covariance matrix of the latent factors from samples from this approximate posterior by means of the following expression

$$\widehat{\Sigma} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{m=1}^{M} \eta_{im}\eta_{im}', \quad \eta_{im} \sim f(\boldsymbol{\theta}|X_1 = x_{i1}, \ldots, X_J = x_{iJ}; \hat{\Omega}, \hat{\phi}).$$

Here the η_{im} are samples from the approximate posterior $f(\boldsymbol{\theta}|X_1, \ldots, X_J; \hat{\Omega}, \hat{\phi})$ for each of the N observations. In appendix A we show that this estimator approaches the marginal maximum likelihood estimator as $M \to \infty$ if $\hat{\Omega}$ approaches the marginal maximum likelihood estimator (i.e., if $K \to \infty$). Below we show that with $M = 1$ these estimates are reasonably accurate in a small

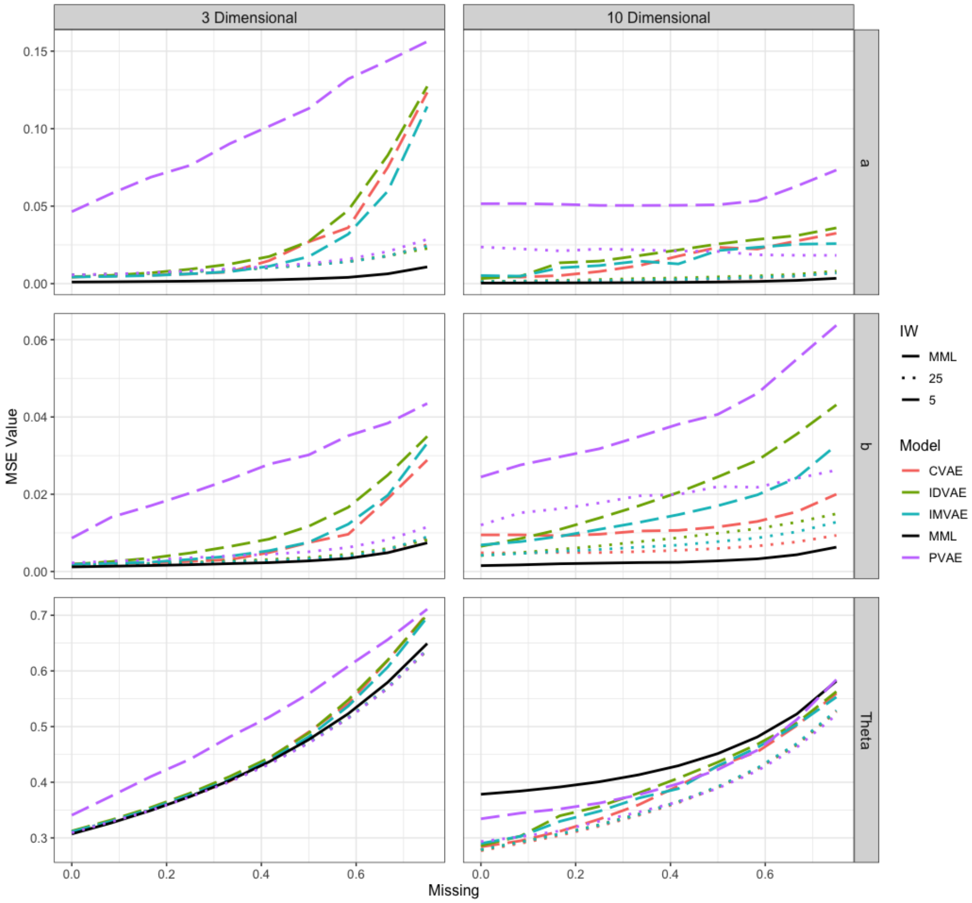

Figure 1: Mean square error of parameter estimates for increasing proportion of missing data. Note that results for models estimated using a single importance weight are omitted.

secondary simulation conducted for just this purpose. We also use this method to estimate factor correlations in the real data application.

We simulate 50 datasets of 10.000 observations from a ten-dimensional MIRT model, where half of the observations are missing. All conditions are kept equal to the simulations study in Section 5, except for the fact that the true ability values are drawn from a multivariate normal distribution where all correlations are set to .4. We estimate parameters using MML and using a CVAE with 25 importance weights. Based on pilot studies, we found that MHRM provides better slope parameters when factors are correlated whereas quasi-Monte Carlo EM (QMCEM) provides better intercept estimates. Therefore, we include both estimation methods in this simulation.

Table 4 provides the recovery statistics for the different parameters. First of all, results indicate that the CVAE provides the most accurate ability estimates, which is in line with the results of the uncorrelated simulation study. Surprisingly, MML using MHRM actually results in clearly worse intercept estimates than the other two methods, whereas QMCEM performs clearly worse on the slopes, indicating that both MML methods struggle to obtain accurate item parameters on high dimensional correlated latent factors, when a large proportion of data is missing. In terms of the factor correlation estimates the two MML methods are both more accurate than the CVAE. Overall the results show that VAE-based methods are also applicable in situations where factors are correlated, although the factor correlation estimates themselves might be more precise with MML.

Table 4: Mean square error of parameter estimates for correlated latent factors. Standard errors are reported between brackets. a denotes the slope parameters, b denotes the intercept parameters, and θ denotes the ability estimates. r denotes the factor correlations.

	a			b			θ			r		
model	σ^2	$bias^2$	mse	σ^2	$bias^2$	mse	σ^2	$bias^2$	mse	σ^2	$bias^2$	mse
CVAE	.0001	.0024	.0025	.0001	.0070	.0071	.0095	.3820	.392	<.0001	.0748	.0748
MHRM	.0051	..0006	.0056	.0061	.0128	.0189	.1970	.3470	.544	.0004	.0065	.0069
QMCEM	.0250	.0199	.0449	.0061	.0004	.0065	.1930	.3570	.550	.0161	.0277	.0438

8 Real data application

To demonstrate the practical viability of the different missing data approaches, we apply our models to a subset of the Bridge to Algebra (BTA) dataset, which was initially published in the context of the 2010 Association for Computing Machinery (ACM) Knowledge Discovery and Data-Mining (KDD) competition (Stamper & Pardos, 2016). The dataset is publicly available in the Pittsburgh Science of Learning Center Datashop (Koedinger et al., 2010) and was gathered from an online environment for learning mathematics called the cognitive tutor (Ritter et al., 2007). The complete BTA dataset consists of responses of $6,034$ students to over $50,000$ algebra problems. Importantly, each problem is divided into several steps, where each step is assumed to require one or more algebra skills. The full item set measures 60 different algebra skills. These skills consist of basic arithmetic operations, such as 'identifying number as a common factor' or 'calculating zero partial product'. In our study, we treat the individual steps as items. The skill requirements for each item make the data well suited for MIRT as the skills can be seen as latent dimensions and the dataset already describes which items require which skills, effectively describing a large

Q-matrix.

To demonstrate the use of our approach, we selected 75 items which measure 9 skills in total. Table 6 contains the descriptions of the latent dimensions as well as the number of items that load on each dimension. $1,289$ students completed this set of items. All students completed all 75 items, which allowed us to introduce missing data artificially and compare results between the complete and missing data datasets. The preprocessed dataset is available on GitHub.

Table 6: Number of items per latent dimension.

Latent dim.	Number of items
List consecutive multiples of a number	16
Calculate zero partial product	16
Identify number as common factor	12
Calculate partial product – carry out	16
Identify number as common multiple	16
List factor of large number	8
Calculate partial product – carry in	13
Calculate partial product – carry in and out	12
Calculate partial product – no carry	23

As a gold standard, we fit a 9-dimensional M2PL model to the complete dataset using MML. Slopes were fit to zero based on the skill requirement data from the BTA dataset. Most items load on a single latent dimension, but some load on up to 5 dimensions. Each latent dimension is measured by at least 8 items, and at most 23 items. The complete matrix of skill requirements is available on GitHub. To compare the approaches under missing data, we introduced 30% missing values at random. Parameters were estimated both using MML and using the CVAE. We used 25 IW samples. Other hyperparameters were kept equal to the simulation study.

Parameter estimation on the missing data dataset took 22 minutes using MML and just 16 seconds using the CVAE, highlighting the computational efficiency of variational methods. Figure 8 plots the parameter estimates on missing data against the parameter estimates on the entire dataset. Note that we only show the ability and discrimination parameters for a single dimension, which we think is representative of the other dimensions. The complete set of plots for all 9 dimensions is available on GitHub, and additionally, Table 7 reports the correlation coefficients between the missing data parameter estimates and the parameter estimates on the entire dataset for all dimensions. Overall, we see that MML estimates on the dataset with missing data are more similar to the full-data MML estimates. This is in line with expectations, as these parameters are estimated using the same method with slightly different data. However, CVAE estimates also correlate highly with the older standard.

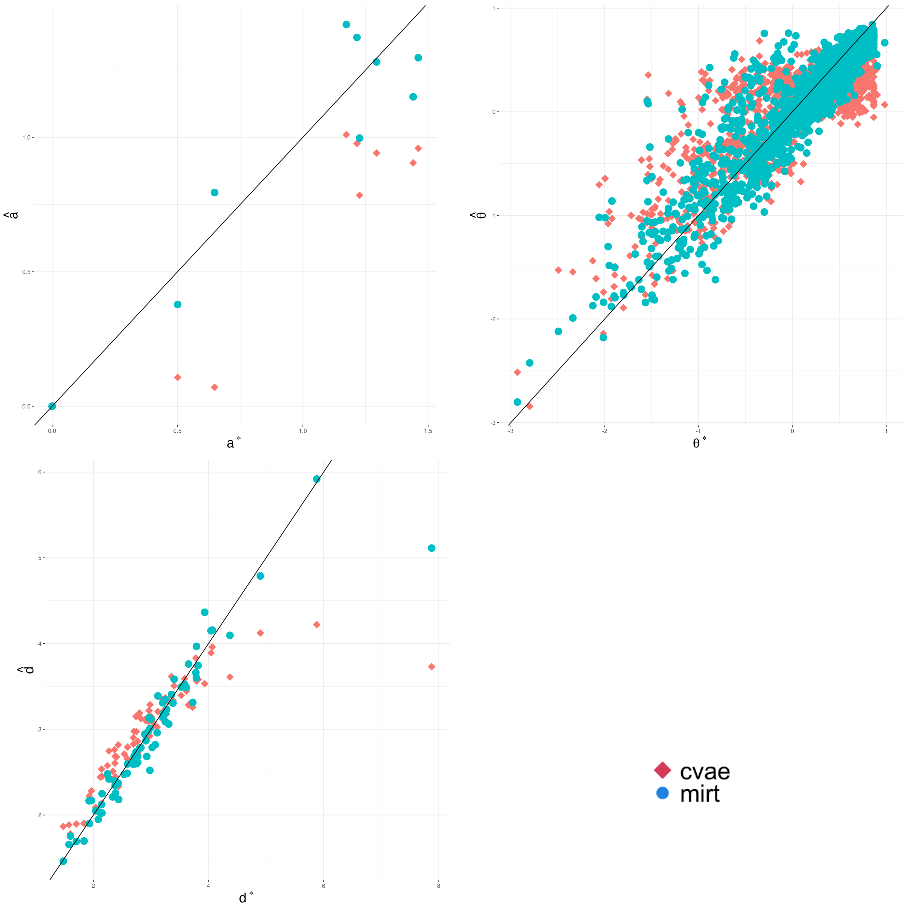

Figure 2: MIRT and CVAE parameters for 30% missing data, plotted against the mirt parameter estimates based on the complete dataset.

9 Discussion

We studied four different variational methods of estimating M2PL models in the presence of missing data. In a simulation study, we compared the different VAE methods to the performance of MML estimation. Results confirmed that variational methods are an efficient alternative to MML to estimate high dimensional MIRT models given that enough IW samples are used to attain performance up to par with MML. In situations where little to no data is missing, taking a small number of importance samples in the VAE-based models is already sufficient for a comparable performance to MML. However, when large parts of the data are not available, which might occur in adaptive testing settings or test equating environments for example, the ELBO appears to underestimate the marginal log-likelihood, and more IW samples are needed to

Table 7: Correlation between CVAE and MML parameter estimates with the gold standard for each dimension. a denotes the slopes, b denotes the intercepts and θ denotes the abilities. r denotes the factor correlations between each latent factor and the 8 other factors.

Latent dim.	1	2	3	4	5	6	7	8	9
a_{cvae}	.9291	.9734	.9567	.9151	.9696	.9714	.5003	.7368	.6028
a_{mml}	.9424	.9469	.9917	.9810	.9968	.9856	.8240	.8540	.8967
θ_{cvae}	.7496	.7465	.8618	.8000	.8356	.7828	.3680	.3792	.7685
θ_{mml}	.8310	.8375	.9319	.9150	.9149	.9000	.8171	.8373	.8904
r_{cvae}	.8296	.8245	.9603	.8090	.9493	.8993	.7583	.8838	.8676
r_{mml}	.9455	.9606	.9963	.9730	.9728	.9846	.8864	.9695	.9257
b_{cvae}	.8384	-	-	-	-	-	-	-	-
b_{mml}	.9356	-	-	-	-	-	-	-	-

accurately recover the model parameters. In terms of the different variational methods used, it is clear that the CVAE and IMVAE should be preferred to the straight-forward input dropout method, which has previously been used in the context of VAE based MIRT (Liu et al., 2022). Imputing missing values with 0 leads to suboptimal ability estimates, which in turn affects the estimation of the item parameters. This is most relevant if a substantial portion of data is missing. In terms of parameter recovery accuracy, the CVAE is the best model, as it provides the inference model with extra information regarding missing values, enabling the model to deal with missing values correctly. The IMVAE might be used as a simpler alternative, sacrificing some estimation accuracy for a less complex model. Finally, we have shown that the CVAE can provide accurate parameter estimates on real high dimensional datasets over 80 times faster than traditional MML.

One limitation of our study concerns the missing data model. In the current simulations, as well as the real data application, we assumed that all data is missing completely at random (MCAR). This simplifying assumption allowed us to make a general comparison between different approaches to handling missing data. Our model approaches full information marginal maximum likelihood as the number of samples increases, so we expect the present approach to be viable in the case of missing at random (MAR) as well. However, in future work, it is important to verify whether our proposed methods generalize to situations where data is MAR.

References

Amari, S.-i. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, *5*(4-5), 185–196.

Bergner, Y., Halpin, P., & Vie, J.-J. (2022). Multidimensional item response theory in the style of collaborative filtering. *psychometrika*, *87*(1), 266–288.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, *112*(518), 859–877.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*(4), 443–459.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, *8*(1), 1–24.

Burda, Y., Grosse, R., & Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a metropolis–hastings robbins–monro algorithm. *Psychometrika*, *75*, 33–57.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, *48*, 1–29.

Chen, Y., Li, X., & Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, *84*, 124–146.

Cho, A. E., Wang, C., Zhang, X., & Xu, G. (2021). Gaussian variational estimation for multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *74*, 52–85.

Collier, M., Nazabal, A., & Williams, C. K. (2020). Vaes in the presence of missing data. *arXiv preprint arXiv:2006.05301*.

Converse, G., Curi, M., Oliveira, S., & Templin, J. (2021). Estimation of multidimensional item response theory models with correlated latent variables using variational autoencoders. *Machine learning*, *110*(6), 1463–1480.

Cremer, C., Morris, Q., & Duvenaud, D. (2017). Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*.

Curi, M., Converse, G. A., Hajewski, J., & Oliveira, S. (2019). Interpretable variational autoencoders for cognitive models. In *2019 international joint conference on neural networks (ijcnn)* (pp. 1–8).

da Silva, M. A., Liu, R., Huggins-Manley, A. C., & Bazán, J. L. (2019). Incorporating the q-matrix into multidimensional item response theory models. *Educational and Psychological Measurement*, *79*(4), 665–687.

Edwards, M. C. (2010). A markov chain monte carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*(3), 474–497.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359–366.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813–1824.

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the edm community: The pslc datashop. *Handbook of educational data mining*, *43*, 43–56.

Liu, T., Wang, C., & Xu, G. (2022). Estimating three-and four-parameter mirt models with importance-weighted sampling enhanced variational autoencoder. *Frontiers in Psychology*, *13*, 4189.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* IAP.

Ma, C., Gong, W., Hernández-Lobato, J. M., Koenigstein, N., Nowozin, S., & Zhang, C. (2018). Partial vae for hybrid recommender system. In *Nips workshop on bayesian deep learning* (Vol. 2018).

Ma, C., Ouyang, J., Wang, C., & Xu, G. (2023). A note on improving variational estimation for multidimensional item response theory. *Psychometrika*, 1–33.

McKinley, R. L., & Reckase, M. D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space.* (Tech. Rep.). American Coll Testing Program Iowa City Ia Resident Programs Dept.

Meyer, J. P., & Zhu, S. (2013). Fair and equitable measurement of student learning in moocs: An introduction to item response theory, scale linking, and score equating. *Research & Practice in Assessment*, *8*, 26–39.

Montecino, C. E. E. (2023). *Using vae for incomplete educational data* (Unpublished doctoral dissertation). Universidade de São Paulo.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

Nazabal, A., Olmos, P. M., Ghahramani, Z., & Valera, I. (2020). Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, *107*, 107501.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-lear`

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 652–660).

Reckase, M. D. (2009). *Multidimensional item response theory models.* Springer.

Reddi, S. J., Kale, S., & Kumar, S. (2019). On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*.

Rezende, D., & Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning* (pp. 1530–1538).

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic bulletin & review*, *14*, 249–255.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555.

Stamper, J., & Pardos, Z. A. (2016). The 2010 kdd cup competition dataset: Engaging the machine learning community in predictive learning analytics. *Journal of Learning Analytics*, *3*(2), 312–316.

Svozil, D., Kvasnicka, V., & Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, *39*(1), 43–62.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.

Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. *Psychological assessment*, *31*(12), 1442.

Urban, C. J., & Bauer, D. J. (2021). A deep learning algorithm for high-dimensional exploratory item factor analysis. *psychometrika*, *86*(1), 1–29.

von Davier, M., & Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics*, *35*(2), 174–193.

Wirth, R., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological methods*, *12*(1), 58.

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. (2002). *Testfact: Test scoring, item statistics, and item factor analysis. chicago: Scientific software international.* Inc.

Wu, M., Davis, R. L., Domingue, B. W., Piech, C., & Goodman, N. (2020). Variational item response theory: Fast, accurate, and expressive. *arXiv preprint arXiv:2002.00276*.

Zhang, C., Bütepage, J., Kjellström, H., & Mandt, S. (2018). Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, *41*(8), 2008–2026.

A Appendix: Deriving the MML estimator of the covariance matrix

The marginal likelihood of a set of responses $\{x_{ij}\}_{j=1}^{J}$ for person i is

$$L_i = \int d\boldsymbol{\theta} f(\boldsymbol{\theta}; \boldsymbol{\Sigma}) \prod_{j=1}^{J} P(X_{ij} = x_{ij} | \Omega_j, \boldsymbol{\theta}),$$

where $f(\boldsymbol{\theta}; \boldsymbol{\Sigma})$ is the multivariate joint normal density, given a mean vector $\boldsymbol{\mu} = 0$ and a covariance matrix $\boldsymbol{\Sigma}$ which is $d \times d$:

$$f(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta}\right)$$

The derivative of $\log L_i$ with respect to $\boldsymbol{\Sigma}$ is

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\Sigma}} \log L_i &= \frac{1}{L_i} \frac{\partial}{\partial \boldsymbol{\Sigma}} L_i \\ &= \frac{1}{L_i} \int d\boldsymbol{\theta} \frac{\partial f}{\partial \boldsymbol{\Sigma}}(\boldsymbol{\theta}; \boldsymbol{\Sigma}) P(X_1, \ldots, X_J | \Omega, \boldsymbol{\theta}_i = \boldsymbol{\theta}) \end{aligned}$$

Logarithmic differentiation ($dg/dy = g(y)(d \log g/dy)$) gives

$$\begin{aligned} \frac{\partial f}{\partial \boldsymbol{\Sigma}} &= f(\boldsymbol{\theta}; \boldsymbol{\Sigma}) \frac{\partial}{\partial \boldsymbol{\Sigma}} [-\frac{d}{2} \log |\boldsymbol{\Sigma}| - \frac{d}{2} \boldsymbol{\theta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta}] \\ &= f(\boldsymbol{\theta}; \boldsymbol{\Sigma}) [-\frac{d}{2} \boldsymbol{\Sigma}^{-1} + \frac{d}{2} \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} \boldsymbol{\theta}' \boldsymbol{\Sigma}^{-1}]. \end{aligned}$$

Therefore

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} \log L_i = \frac{1}{L_i} \int d\boldsymbol{\theta} f(\boldsymbol{\theta}; \boldsymbol{\Sigma}) \frac{d}{2} [-\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} \boldsymbol{\theta}' \boldsymbol{\Sigma}^{-1}] P(X_1, \ldots, X_J | \Omega, \boldsymbol{\theta}_i = \boldsymbol{\theta}).$$

Note that

$$\int d\boldsymbol{\theta} f(\boldsymbol{\theta};\boldsymbol{\Sigma})[-\frac{d}{2}\boldsymbol{\Sigma}^{-1}]P(X_1,\ldots,X_J|\Omega,\boldsymbol{\theta}_i=\boldsymbol{\theta}) = -\frac{d}{2}\boldsymbol{\Sigma}^{-1}L_i$$

and

$$\begin{aligned}\int d\boldsymbol{\theta} f(\boldsymbol{\theta};\boldsymbol{\Sigma})[\frac{d}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}]P(X_1,\ldots,X_J|\Omega,\boldsymbol{\theta}_i=\boldsymbol{\theta})&\\ =\frac{d}{2}\boldsymbol{\Sigma}^{-1}\left[\int d\boldsymbol{\theta} f(;\boldsymbol{\Sigma})\boldsymbol{\theta}\boldsymbol{\theta}'P(X_1,\ldots,X_J|\Omega,\boldsymbol{\theta}_i=\boldsymbol{\theta})\right]\boldsymbol{\Sigma}^{-1}&\end{aligned}$$

Therefore

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\Sigma}}\log L_i =\\ &\frac{d}{2\,L_i}\left(-\boldsymbol{\Sigma}^{-1}L_i+\boldsymbol{\Sigma}^{-1}\left[\int d\boldsymbol{\theta} f(\boldsymbol{\theta};\boldsymbol{\Sigma})\boldsymbol{\theta}\boldsymbol{\theta}'P(X_1,\ldots,X_J|\Omega,\boldsymbol{\theta}_i=\boldsymbol{\theta})\right]\boldsymbol{\Sigma}^{-1}\right).\end{aligned}$$

To find the MML estimator of $\boldsymbol{\Sigma}$ we equate the derivative of the sample log-likelihood to zero: Summing over $i=1,\ldots,N$, we have

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\Sigma}}\log L = 0\\ &\iff N\boldsymbol{\Sigma}^{-1}=\boldsymbol{\Sigma}^{-1}\left[\sum_{i=1}^{N}\frac{1}{L_i}\int d\boldsymbol{\theta} f(\boldsymbol{\theta};\boldsymbol{\Sigma})\boldsymbol{\theta}\boldsymbol{\theta}'P(X_1,\ldots,X_J|\Omega,\boldsymbol{\theta}_i=\boldsymbol{\theta})\right]\boldsymbol{\Sigma}^{-1}\\ &\iff \boldsymbol{\Sigma}=\frac{1}{N}\sum_{i=1}^{n}\int\boldsymbol{\theta}\boldsymbol{\theta}'\frac{1}{L_i}f(\boldsymbol{\theta};\boldsymbol{\Sigma})P(X_1,\ldots,X_J|\Omega,\boldsymbol{\theta}_i=\boldsymbol{\theta})d\boldsymbol{\theta}.\end{aligned}$$

Consequently, using Bayes' theorem,

$$\therefore\boldsymbol{\Sigma}=\frac{1}{N}\sum_{i=1}^{N}\int\boldsymbol{\theta}\boldsymbol{\theta}'f(\boldsymbol{\theta}|X_1=x_{i1},\ldots,X_J=x_{iJ};\Omega,\boldsymbol{\Sigma})d\boldsymbol{\theta}.$$

That is, the ML estimator of the covariance matrix satisfies the equation

$$\boldsymbol{\Sigma}=\frac{1}{N}\sum_{i=1}^{N}E\{\boldsymbol{\theta}\boldsymbol{\theta}'|X_1=x_{i1},\ldots,X_J=x_{iJ};\Omega,\boldsymbol{\Sigma}\}.$$

Note that this is in fact a sample implementation of the equation $E\{\eta\eta'\}=E\{E\{\eta\eta'|\mathbf{x}\}\}$. In other words, the MML estimator of $\boldsymbol{\Sigma}$ solves a *method of moments* estimator equation.

Note that if we can sample from the posterior for each subject i, yielding $\eta_i, i=1,\ldots,n$, we can obtain an estimate of $\boldsymbol{\Sigma}$

$$\widehat{\boldsymbol{\Sigma}}=\frac{1}{N}\sum_{i=1}^{N}\eta_i\eta_i',\quad \eta_i\sim f(\boldsymbol{\theta}|X_1=x_{i1},\ldots,X_J=x_{iJ};\Omega,\boldsymbol{\Sigma}).$$

If we average many such estimates, repeatedly sampling from the posteriors, this estimate converges in probability to the MML estimate if $\Omega=\hat{\Omega}$ the MML estimate.