



Estimation of multidimensional item response theory models with correlated latent variables using variational autoencoders

Geoffrey Converse¹ · Mariana Curi² · Suely Oliveira¹ · Jonathan Templin¹

Received: 8 October 2020 / Revised: 1 May 2021 / Accepted: 20 May 2021 /

Published online: 2 June 2021

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2021

Abstract

Artificial neural networks with a specific autoencoding structure are capable of estimating parameters for the multidimensional logistic 2-parameter (ML2P) model in item response theory (Curi et al. in International joint conference on neural networks (IJCNN), 2019), but with limitations, such as uncorrelated latent traits. In this work, we extend variational autoencoders (VAE) to estimate item parameters and correlated latent abilities, and directly compare the ML2P-VAE method to more traditional parameter estimation methods, such as Monte Carlo expectation-maximization. The incorporation of a non-identity covariance matrix in a VAE requires a novel VAE architecture, which can be utilized in applications outside of education. In addition, we show that the ML2P-VAE method is capable of estimating parameters for models with a large number of latent variables with low computational cost, where traditional methods are infeasible for data with high-dimensional latent traits.

Keywords Item response theory · Interpretable neural networks · Parameter estimation · Educational measurement

Editors: Derek Greene.

✉ Geoffrey Converse
converseg@gmail.com

Mariana Curi
mcuri@icmc.usp.br

Suely Oliveira
suely-oliveira@uiowa.edu

Jonathan Templin
jonathan-templin@uiowa.edu

¹ University of Iowa, Iowa City, USA

² University of São Paulo, São Paulo, Brazil

1 Introduction

Item response theory models (IRT; e.g., Lord and Novick, 1968) provide a link between a student's observed item responses on an assessment and a set of continuously-valued latent variables representing the student's ability on dimensions that underlie an assessment. Whereas unidimensional IRT models posit a single latent trait that is measured by all items of an assessment, multidimensional IRT models (Reckase, 2009) feature multiple latent traits, often where only a small set are measured by any one item. Although MIRT models have a long history, estimation of these models can be computationally intensive, depending on which estimation methods are used, and often have practical limits on the number of latent variables that can be estimated in an analysis. Despite this added complexity, studies have explored the advantages that multidimensional assessments hold over unidimensional assessments (Bradshaw et al., 2014). In recent years, artificial neural networks (ANN) have been employed in various areas of education research, including the tasks of knowledge tracing (Piech et al., 2015; Zhang et al., 2017), parameter estimation (Guo et al., 2017; Wu et al., 2020), and essay scoring (Taghipour and Ng, 2016). Each of these applications aim to measure student ability in some setting.

Most closely related to the present work, Guo et al. (2017) introduced the idea of using a neural network with autoencoding structure for parameter estimation. They make some constraints on network parameters in order to interpret a hidden layer of an autoencoder as estimates to student skills. It has previously been demonstrated by Converse et al. (2019) that using a VAE instead of a regular autoencoder yield large improvements. This is due to the assumption in IRT that the students' latent skills follow normal prior distribution, and how a VAE uses a regularization term to fit a hidden layer to this same prior distribution. The VAE approach (Curi et al., 2019) assumes that latent traits are independent—the work in this paper generalizes this approach to allow for more complicated prior distributions involving correlated latent traits.

Other recent publications suggest related approaches using neural networks for variational inference in IRT. Wu et al. (2020) uses an objective function similar to what we present here, but includes additional regularization terms. They use two separate neural networks—one for student ability parameters and the other for item parameters. Wang et al. (2020) presents a similar inference model which can be applied to other factor analysis models. Like other previous literature, both of these works assume that latent abilities are independent and regularizes student ability against a standard normal distribution.

In this study, we investigate the use of a specific ANN, a variational autoencoder, in the estimation of MIRT models with large numbers of correlated latent traits. Different from Wu et al. (2020), we obtain the item parameter estimates from trainable neural network weights. This work is an extension of preliminary results which have been previously published in a neural computing conference, where the method was first presented by Curi et al. (2019). There are three main contributions that this work provides.

First, we propose an alternative network architecture and regularization term in the loss function, which allows for correlated latent skills. As such, the VAE can be fit to a more general multivariate Gaussian prior distribution with a non-identity correlation matrix. This neural architecture can be applied in other fields outside of educational measurement. Secondly, this work performs experiments on four different datasets, both real and simulated, which are much larger than those used in Curi et al. (2019),

demonstrating the ability of the proposed method to handle high-dimensional data. Finally, we directly compare our results with those of traditional IRT parameter estimation methods, displaying the considerable advantages of our method as the number of latent skills increases.

Estimation of MIRT models (or models that are roughly equivalent to MIRT models, such as Item Factor Analysis models) dates to the 1970s. What now would be called limited information estimation methods—methods that rely upon a polychoric correlation matrix rather than raw data—were used in Christofferson (1975) and Fraser and McDonald (1988). Full information marginal maximum likelihood (MML) estimation of MIRT models is difficult stemming from the need to numerically integrate to marginalize across the latent variables. As the number of latent variables increases linearly, the number of computations in the marginalization process increases exponentially. Versions of MML algorithms have been developed to speed such estimation, such as the Expectation-Maximization (EM) algorithm (Bock and Aitkin, 1981), but still suffer from an exponential increase in computation with a linear increase in latent variables. More recently, hybrid MML algorithms have been developed to speed calculation of the multidimensional integral across latent variables. For instance, Cai (2009) implemented a Robbins Monro algorithm that sampled from the posterior distribution of the latent traits in order to provide an estimate in the Expectation step of an EM algorithm. Despite such advances, estimation of MIRT models with large numbers of latent traits remains elusive.

The importance of developing alternative approaches for parameter estimation that are more computationally efficient than traditional methods is seen in the analysis of large-scale data. Significantly increasing the number of items, individuals, and latent variables puts strain on the traditional estimation processes. In recent years, a number of educational technology companies have aimed to provide more personalized and skill-specific instruction to students—this requires quick and efficient methods to evaluate student learning at scale.

In the educational field, the evolution of psychometric models and computational resources are leading to the evaluation of a larger number of more specific abilities in online learning environments. Pardos et al. (2010), for instance, used a Bayesian network to model up to 106 Math skills over students that took the MCAS (Massachusetts Comprehensive Assessment System) assessment, a state-administered standardized test of English, math, science, and social studies for grades 3 to 10. In a similar application, ACT has created a holistic framework (Camara et al., 2015) to conceptualize academic performance as multidimensional. In another fields, statistical models that deal with high dimensional latent variables can be applied to quantitative genetic studies of gene expression, where the number of traits assayed per individual can reach the thousands (Runcie and Mukherjee, 2013).

Additionally, the running time of estimation algorithms is of interest to improve. When more individuals are present in the dataset, more parameters have to be estimated. In large-scale assessments, it is common to have millions of students taking an exam, making numerical calculations taxing. Running time is also an issue in Computerized Adaptive Testing (CAT) (Van der Linden and Glas, 2000). CAT selects items along the test tailoring their difficulty to the level of ability presented by the examinees according to their answers. Abilities are estimated on the fly, requiring the estimation process to not be sluggish.

In addition to our proposed model's effectiveness in the application area, it is of interest to the machine learning community because of the manner in which a variational autoencoder is used. Through modifications and constraints to the neural network architecture, we are able to interpret some of the trainable network parameters as estimates to

the item parameters and interpret the activations at a hidden network layer as estimates of student ability. This adds a great deal of interpretability to an otherwise black-box neural network. Besides adding explainability to a neural network, we also introduce a neural architecture that allows a VAE to learn a more general Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with correlated latent code, rather than the standard normal distribution.

The concern with correlations between traits is that measurement of abilities in IRT, or mental traits in general, long has assumed that traits measured from the same person will be correlated. In many social science fields, the development of multidimensional assessment inventories often hinges upon the question of whether traits are too highly correlated (i.e., they are essentially the same dimension) rather than if they are correlated at all.

1.1 Multidimensional item response theory

In MIRT, it is assumed that each student j possesses K latent abilities $\Theta_j = (\theta_{j1}, \dots, \theta_{jK})^\top$, which are not directly observable. In real-world applications, the available data is often student j 's response set to an exam with n items, represented as a binary vector $u_j \in \mathbb{R}^n$. An entry in u_j equal to 1 corresponds with a correct answer, and an entry equal to 0 corresponds with an incorrect answer to a particular item of the exam. Given student j 's assessment results u_j , the goal in IRT parameter estimation is to infer the latent traits Θ_j .

A naive approach to quantifying student knowledge is to look at the percentage of questions that the student answered correctly. This is what is done in most classrooms to assign letter grades. But this does not take into account the fact that each item on the assessment is different—both in difficulty and in content. For example, if one student answers only questions 1 and 4 incorrect, and another student answers only questions 3 and 7 incorrect, they have the same percentage score. But it is not likely that the two students share the same latent trait values. Questions 3 and 7 may have tested a different skill than items 1 and 4, and could vary greatly in difficulty level. For this reason, MIRT models are useful in educational measurement.

The Multidimensional Logistic 2-Parameter (ML2P) model gives the probability of students answering a particular question as a continuous function of student ability (McKinley and Reckase, 1980). There are two types of parameters associated with each item: a difficulty parameter b_i for item i , and a discrimination parameter $a_{ik} \geq 0$ for each latent trait k quantifying the level of ability k required to answer item i correctly. The ML2P model gives the probability of student j with latent abilities $\Theta_j = (\theta_{j1}, \dots, \theta_{jK})^\top$ answering item i correctly as

$$P(u_{ij} = 1 | \Theta_j; \mathbf{a}_i, b_i) = \frac{1}{1 + \exp \left[- \sum_{k=1}^K a_{ik} \theta_{jk} + b_i \right]} \quad (1)$$

Note that if $a_{ik} = 0$, then success on item i does not depend on ability k . A single item can assess multiple abilities, so that both $a_{ik} \neq 0$ and $a_{ij} \neq 0$ for $j \neq k$. While it is impossible to discern the quantities of item parameters without some estimation method, it is often feasible for field experts to identify the latent traits required for each item. A Q-matrix (Tatsuoka, 1983) $Q \in \mathbb{R}^{n \times K}$ defines the relationship between exam items and latent abilities. Each entry is binary, where $Q_{ik} = 1$ if ability k is useful in answering item i , and $Q_{ik} = 0$ otherwise.

1.2 Variational autoencoders

A variational autoencoder (VAE) is an unsupervised learning method with an architecture similar to a regular autoencoder. It encodes the input to a smaller dimension, and then decodes this simpler representation to reconstruct the original input. But a VAE makes the additional assumption that the low-dimensional representation of data follows some probability distribution (often $\mathcal{N}(0, I)$), and fits the encoded data to this distribution. The encoder maps each data point to a probability distribution, which can then be sampled from, and that sample is reconstructed through the decoder.

This reveals the primary use of a VAE as a generative model: after training a VAE on some data set X , samples from the encoded distribution can be fed forward through the decoder to obtain *new* data samples, which are similar to X . Note that only the decoder is used at test time.

Formally, we assume that data X are generated from a distribution $p_\beta(X = x|\Theta)$. In other words, for all $x \in X$, there exist some values of latent variables Θ such that the parameters β in the decoder will produce x . The encoder will map the input data x to the its corresponding Θ . By Bayes' rule, we can write the unknown posterior distribution of Θ as

$$f(\Theta|x) = \frac{P(X = x|\Theta)f(\Theta)}{P(X = x)} = \frac{P(X = x|\Theta)f(\Theta)}{\int P(X = x|\Theta)f(\Theta)d\Theta} \quad (2)$$

In order to evaluate $f(\Theta|x)$, we assume that the prior $f(\Theta)$ follows a probability distribution. In most VAE applications, $\mathcal{N}(0, I)$ is used, but this can be generalized to $\mathcal{N}(\mu, \Sigma)$. While $f(\Theta|x)$ is directly computable in theory, the integral in Eq. 2 becomes intractible as the dimension of Θ increases. A solution is to approximate $f(\Theta|x)$ with some other function $q_\alpha(\Theta|x)$ with parameters α , which takes an input x and maps it to a distribution of likely values of Θ .

In order to train the VAE to fit $q_\alpha(\Theta|x)$ as close to $f(\Theta|x)$ as possible, another term must be added to the neural network loss function. Such q_α is chosen by minimizing the Kullback-Liebler divergence (Kullback and Leibler, 1951) between all possible functions $q(\Theta|x) \in \Psi$ and $f(\Theta|x)$, where Ψ is a particular class of probability distributions. For example, Ψ may contain all d -dimensional multivariate Gaussian distributions. The KL-Divergence is given by

$$KL[q(\Theta|x)||f(\Theta|x)] = E_{\Theta \sim q(\Theta|x)}[\log q(\Theta|x) - \log f(\Theta|x)]. \quad (3)$$

This expression still contains the unknown $f(\Theta|x)$, but it can be shown as in Kingma and Welling (2014) that minimizing Eq. 3 while still reconstructing input data is equivalent to maximizing

$$E_{\Theta \sim q(\Theta|x)}[\log P(X = x|\Theta)] - KL[q(\Theta|x)||f(\Theta)] \quad (4)$$

which only requires the prior distribution $f(\Theta)$. The first term of Eq. 4 corresponds with the reconstruction of inputs, and the second term pushes the approximate latent distribution $q(\Theta|x)$ closer to the prior $f(\theta)$. In the case of a VAE, the class of functions Ψ refers to a particular neural network architecture of the encoder, and each distinct $q \in \Psi$ represents a different setting of the network weights and biases W . The VAE is trained by a gradient descent algorithm to minimize

$$\mathcal{L}(W) = \mathcal{L}_0(W) + \lambda KL[q(\Theta|x)||f(\Theta)], \quad (5)$$

where \mathcal{L}_0 is the cross-entropy loss function and λ is a regularization hyper-parameter.

2 Model description

We expand upon the work of Curi et. al. and use the ML2P-VAE model proposed by Curi et al. (2019). Though an unconventional use of a variational autoencoder, we show that the method is a flexible and versatile parameter estimation method. Certain modifications to the architecture of the neural network allow for interpretation of weights and biases in the decoder as item parameter estimates, and activation values in the encoded hidden layer as ability parameter estimates. This is a unique property, as neural networks are usually uninterpretable and function as a black box model.

The required modifications are as follows. The decoder of the autoencoder (or variational autoencoder) has no hidden layers. The non-zero weights in the decoder, connecting the encoded distribution to the output layer, are determined by a given Q -matrix (Tatsuoka, 1983). Thus these two layers are not densely connected. The output layer must use the sigmoidal activation function,

$$\sigma(z_i) = \frac{1}{1 + e^{-z_i}} \quad (6)$$

where $z_i = \sum_{k=1}^K w_{ki}\alpha_k + \beta_i$. Here, w_{ik} is the weight between the k -th and i -th nodes in the second-to-last and output layer, respectively. α_k is the activation of the k -th node in the second-to-last layer, and β_i is the bias value of the i -th node in the output layer. The form of Eq. 6 is identical to that of Eq. 1, and so these specifications allow interpretation of the decoder as an ML2P model.

Additional modifications are made to improve performance, but are not technically required in order to allow interpretation. These include forcing decoder weights to be non-negative (since discrimination parameters are non-negative) and fitting the VAE to a normal distribution. Previous work by Converse et al. (2019) has demonstrated the advantages that a VAE fit to a normal distribution holds over a regular autoencoder with similar structure as seen in Guo et al. (2017). In this work, we make additional adjustments to the network architecture in order to account for correlated latent traits. A visualization of the ML2P-VAE neural network modifications is shown in Fig. 1, where $q_\alpha(\Theta|x)$ and $p_\beta(x|\Theta)$ parameterize the encoder and decoder, respectively. The encoder outputs parameters μ_0 and L_0 , which describe the mean and covariance matrix of the latent distribution.

2.1 VAE implementation for correlated traits

There are many publicly available code examples of VAE implementations which fit the latent space to a standard normal distribution $\mathcal{N}(0, I)$. However, it is uncommon to train a VAE which assumes that the latent traits are *not* independent of one another. Because most VAE applications do not have any interpretable hidden layers, there is not much merit in fitting correlated latent variables. In fact, it is often beneficial to force them to be independent by fitting to $\mathcal{N}(0, I)$.

Building a VAE which learns correlated latent traits is a unique approach, and brings extra difficulties in the network architecture. This affects the two distinguishing features of a VAE: (1) sampling from the learned distribution and sending this sample through the decoder and (2) adding Kullback-Liebler Divergence to the network's loss function.

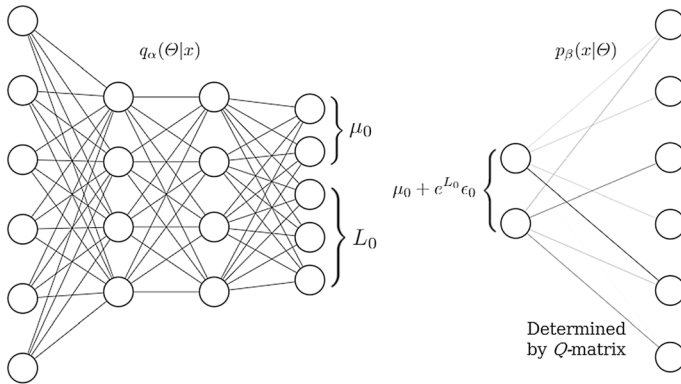


Fig. 1 Visualization of ML2P-VAE architecture for two correlated latent traits and six input items

After training a VAE, feeding a data point x_0 through the encoder results in a set of values that correspond to a probability distribution, namely a vector $\mu_0 \in \mathbb{R}^K$ and a matrix $\Sigma_0 \in \mathbb{R}^{K \times K}$. Sampling from $\mathcal{N}(\mu_0, \Sigma_0)$ requires a matrix G_0 such that $G_0 G_0^T = \Sigma_0$. Though G_0 is not unique, it is convenient to use the Cholesky decomposition of Σ_0 (Atkinson, 1989). Define $\epsilon_0 = (\epsilon_1, \dots, \epsilon_K)^T$ with each $\epsilon_i \sim \mathcal{N}(0, 1)$, for $1 \leq i \leq K$. Finally, generate the sample $z_0 = \mu_0 + G_0 \epsilon_0$, which will be sent through the decoder.

The KL-Divergence between two multivariate normal distributions of dimension K is given as

$$\mathcal{D}_{KL}[\mathcal{N}(\mu_0, \Sigma_0) || \mathcal{N}(\mu_1, \Sigma_1)] = \frac{1}{2} \left(\text{tr}(\Sigma_1^{-1} \Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1} (\mu_1 - \mu_0) - K + \ln \left(\frac{\det \Sigma_1}{\det \Sigma_0} \right) \right) \tag{7}$$

When using this in the loss function of a VAE, μ_1 and Σ_1 are constant, since we desire to fit the data to a known distribution $\mathcal{N}(\mu_1, \Sigma_1)$. Thus we only need to calculate Σ_1^{-1} one time, speeding computation time while training the neural network. But Eq. 7 computes $\ln \det \Sigma_0$, so we must have $\det \Sigma_0 > 0$. Ensuring $\det \Sigma_0 > 0$ is nontrivial; Σ_0 depends on the input x_0 , along with every weight and bias in the encoder. These weights and biases are updated throughout training, and often initialized randomly. If $\det \Sigma_0 \leq 0$ at any point during training, then it is not possible to compute the loss and gradient. We construct an encoder architecture which guarantees that $\det \Sigma_0 > 0$, regardless of the input x_0 or encoder parameters.

As before, the input to our neural network consists of n nodes, representing items on an assessment. After a sufficient number of hidden layers of sufficient size, the encoder outputs $K + K(K + 1)/2$ nodes. The first K nodes represent the mean vector μ_0 , and the remaining $K(K + 1)/2$ nodes are arranged into a lower triangular matrix $L_0 \in \mathbb{R}^{K \times K}$. We obtain the covariance matrix by using the matrix exponential: $G_0 = e^{L_0}$, so $\Sigma_0 = G_0 G_0^T = e^{L_0} \cdot (e^{L_0})^T$.

Theorem 1 Σ_0 under the previously described construction is symmetric, positive definite, and has positive determinant.

Proof Consider any lower triangular $L_0 \in \mathbb{R}^{K \times K}$. Define

$$G_0 := e^{L_0} = \sum_{n=0}^{\infty} \frac{L_0^n}{n!} = I + L_0 + \frac{1}{2}L_0^2 + \dots$$

G_0 is lower triangular, since addition and multiplication of matrices preserve this property. Further, G_0 is nonsingular, since $\det G_0 = \det e^{L_0} = e^{\text{tr}L_0} > 0$.

Set $\Sigma_0 := G_0 G_0^T$. Notice that Σ_0 is symmetric:

$$\Sigma_0^T = (G_0 G_0^T)^T = G_0 G_0^T = \Sigma_0.$$

Further, $\det \Sigma_0 = \det(G_0) \cdot \det(G_0^T) > 0$. Now for any nonzero $y \in \mathbb{R}^K$, we have

$$\langle \Sigma_0 y, y \rangle = y^T \Sigma_0 y = y^T G_0 G_0^T y = \langle G_0^T y, G_0^T y \rangle = \|G_0^T y\|_2^2 > 0$$

Therefore, Σ_0 is positive definite. □

Theorem 1 shows that we can interpret Σ_0 as a covariance matrix, and thus our encoder architecture maps a data point x_0 to a multivariate distribution $\mathcal{N}(\mu_0, \Sigma_0)$. Additionally, we are able to perform the sampling operation using G_0 , and it is always possible to compute KL-Divergence as in Eq. 7.

If latent traits are assumed to be correlated, then a full correlation matrix must be provided for the ML2P-VAE model. This corresponds to the fixed covariance matrix Σ_1 . ML2P-VAE is capable of estimating ability, discrimination, and difficulty parameters, but it does not estimate correlations between latent traits. Note that relaxing these restrictions on the entries of the covariance matrix Σ_1 would likely lead to an issue with model identification. Without any restrictions on the distribution of Θ , it is possible that for a given item and fixed student, multiple settings of θ_j , a_{ik} , and b_i can produce the same global minimum in learning Eq. 1 (Haberman, 2005). Our notion of fixing each entry in Σ_1 is similar to the more common constraint that $\Theta \sim \mathcal{N}(0, I)$.

3 Experimental testing

We compare variations of the ML2P-VAE method with three traditional parameter estimation techniques—Metropolis-Hastings Monte Carlo (MHRM), quasi-Monte Carlo Expectation-Maximization (QMCEM), and Monte Carlo Expectation-Maximization (MCEM)—on various data sets. We focus on these traditional estimation techniques instead of Guo et al. (2017) because previous work has demonstrated the advantages that a VAE holds over a regular autoencoder (Converse et al., 2019). We demonstrate that higher dimensions of latent traits do not significantly effect the runtime of the ML2P-VAE model. In this case, due to computational difficulties, traditional methods cannot be used, so we just present the estimates produced by ML2P-VAE.

We consider three scenarios for ML2P-VAE: (a) the best case scenario where we assume that the covariance matrix between all latent traits is known, (b) we don't know the exact covariance matrix, so it is estimated using other methods, and (c) we simply assume that all traits are independent. Throughout the rest of this section, these three methods are referred to as ML2P-VAE_{full}, ML2P-VAE_{est}, and ML2P-VAE_{ind}, respectively. Note that ML2P-VAE_{ind} is equivalent to the model proposed by Curi et al. (2019). In scenario (b), we multiply the response matrix (N students by n items) by the Q -matrix

(n items by K abilities). We then take the Pearson correlation of the columns of this $N \times K$ matrix to obtain an approximate correlation matrix in $\mathbb{R}^{K \times K}$ between abilities.

In all variations of ML2P-VAE, we train the neural network with the ADAM optimizer for 10 epochs and batch size 1 (pure stochastic gradient descent). We found that increasing batch size can improve runtime, but accuracy is adversely affected when the batch size is greater than 16. As such, we decided to use the smallest batch size possible, since runtime isn't an issue for the small networks we used. The regularization parameter in Eq. 5 was fixed at $\lambda = 1$. The specific encoder architecture of the neural network was dependent on the size of the data set. A simulated assessment with 6 latent abilities used two hidden layers of size 32 and 16. A real data set evaluating 3 latent abilities used two hidden layers of 16 and 8 nodes, and this same architecture was used for a simulated data set with 4 latent abilities. The largest network we used was for an assessment for 20 latent abilities, which utilized two hidden layers of size 64 and 32. In each network, we used a sigmoidal activation function in the encoder hidden layers and a linear activation function in the encoded distribution. As described earlier, the ML2P-VAE model requires the use of a sigmoidal activation function in the output layer of the decoder.

3.1 Data

We ran experiments on four data sets: (i) a simulated data set with 6 latent traits, 50 items, and 20,000 students; (ii) results from the Examination for the Certificate of Proficiency in English (ECPE) (Templin and Hoffman, 2013), a real data set with 3 latent traits, 28 items, and 2922 students; (iii) a simulated data set with a 20 latent traits 200 items, and 50,000 students; and (iv) a simulated data set with 4 latent traits, 27 items, and 3,000 students. Note that comparisons with traditional techniques are only possible for (i), (ii), and (iv) because of the large number of latent traits in (iii). It is also worth pointing out that true parameter values, for both students and items, are only available for simulated data.

When simulating data for (i) and (iii), we used Python's SciPy package to generate a symmetric positive definite matrix with 1s on the diagonal (correlation matrix) and all matrix entries non-negative. All latent traits had correlation values between 0 and 1. We assumed that each latent trait was mean-centered at 0. Then, we sampled ability vectors to create simulated students. We generated a random Q -matrix where each entry $q_{ij} \sim \text{Bern}(0.2)$. If a column $q_{i\cdot} = 0$ for each element after sampling from this Bernoulli distribution, then one random element was changed to a 1. This ensures that each item measured at least one trait. Discrimination parameters were sampled from a range so that $0.25 \leq MDISC_i \leq 1.75$ for each item i , and difficulty parameters were sampled uniformly from $[-3, 3]$. Finally, response sets for each student were sampled from the ML2P model using these parameters.

For data set (iv) we were more particular in selecting the Q -matrix and correlation matrix. Rather than generating these randomly, each entry in these matrices was chosen manually. Of the 4 skills in the correlation matrix, one of them is entirely independent of the other three. The other three latent abilities had correlations of 0.25, 0.1, and 0.15 between them. The correlation matrix was chosen in this way so that it is closer to the identity matrix, allowing the ML2P-VAE_{ind} variation to perform better. The Q -matrix was chosen so that it contained 16 "simple" items (items requiring only one skill), 6

items requiring 2 latent abilities, 4 items requiring 3 latent abilities, and one item requiring all 4 skills. In this way, each of the possible $\binom{4}{k}$ combinations is present in the Q -matrix, for $k \in \{1, 2, 3, 4\}$.

3.2 Results

A summary of each method's performance is given in Table 1. All experiments were conducted using Tensorflow for R and the ML2Pvae package (Converse, 2020) on a laptop computer with a 2.9 GHz Intel Core i7-7500U CPU. The results from traditional methods were obtained using default settings of the MIRT package (Chalmers, 2012). The ML2P-VAE methods used the previously described hyper-parameters, though both accuracy and runtime could likely be fine-tuned by changing the encoder architecture and implementing techniques such as early stopping.

Note that when comparing error measures in data set (i), the ML2P-VAE methods are competitive with traditional methods. In particular, assuming full knowledge of the latent trait covariances in ML2P-VAE yields discrimination, difficulty, and ability parameter estimates of similar accuracy to MHRM. When we relax our assumption of known latent trait correlation, the accuracy of parameter estimates understandably slip.

Although the ML2P-VAE methods are slightly less accurate than MHRM, they are much faster than traditional methods, especially as the number of latent traits increase. Much of this speedup is due to the fact that neural networks do not require numerical integration over the latent abilities. While quadrature or MCMC methods become infeasible on data sets much larger than (i), this is no cause for concern with ML2P-VAE. Note that for neural networks of this size (50-200 inputs and latent dimension 6-20), the longer runtime is more due to the number of data samples, rather than the size of the latent dimension. In fact, the largest neural network we used in these experiments, for data set (iii), only had 1,670 trainable parameters. In common applications, neural networks used for image classification often have hundreds of thousands or even millions of trainable parameters.

We can visualize some of the results in Figs. 2, 3, 4, and 5 for data sets (i), (ii), (iii), and (iv), respectively. Figure 2 shows the correlation between the true and estimated discrimination parameters for the ML2P-VAE_{full} and MHRM methods. We don't include such plots for the difficulty parameters, as all methods estimate each b_i with very high accuracy. From these figures, it appears that while MHRM obtains better results on smaller discrimination parameters, ML2P-VAE_{full} has less error on larger parameters, and the estimation error seems to be independent of the magnitude of the parameter. The other two ML2P-VAE methods do not obtain the same levels of accuracy as when assuming full knowledge of the latent ability correlations.

When examining the ECPE data, there are no "true" values of parameters, so we compare ML2P-VAE's results directly with MHRM's estimates. As seen in Table 1, the parameter estimates from QMCEM and MCEM are nearly identical to those of MHRM on the ECPE data. Of course, there is not a known covariance matrix between the three latent abilities, so we only analyze ML2P-VAE_{est} and ML2P-VAE_{ind}. While both methods perform similarly to MHRM in difficulty parameter estimates, we can see that the two yield different results when applied to discrimination and ability parameters. This is depicted in Fig. 3; we first note that while ML2P-VAE_{ind} gives accurate estimations for the green and black abilities (and the discrimination parameters associated with those abilities), the red ability estimates are all very near zero for every student. This tells us that the ML2P-VAE_{ind}

Table 1 Error measures for discrimination (a), difficulty (b), and ability (θ) parameters from various parameter estimation methods on three different data sets

Data Set	Method	a .RMSE	a .BIAS	a .COR	b .RMSE	b .BIAS	b .COR	θ .RMSE	θ .BIAS	θ .COR	Runtime	
(i) 6 abilities (Simulated)	MHRM	0.0693	0.0319	0.9986	0.0256	-0.0021	0.9999	0.7140	-0.0033	0.7006	1110s	
	QMCEM	0.1490	-0.0670	0.9939	0.0376	-0.0020	0.9998	0.7206	0.0023	0.6939	322s	
	MCEM	0.1497	-0.0633	0.9936	0.0383	0.0035	0.9997	0.7206	-0.0016	0.6938	1009s	
	ML2P-VAE _{full}	0.0705	0.0255	0.9985	0.0471	-0.0079	0.9996	0.6649	-0.0178	0.7476	343s	
	ML2P-VAE _{est}	0.1803	0.0871	0.9891	0.0640	-0.0131	0.9993	0.7109	0.0772	0.7082	364s	
	ML2P-VAE _{ind}	0.1218	-0.0004	0.9944	0.0597	-0.0145	0.9994	0.7222	0.0316	0.6928	252s	
	MHRM*	0.000*	0.000*	1.000*	0.000*	0.000*	1.000*	0.000*	0.000*	1.000*	1.000*	162s
	QMCEM	0.0159	0.0035	0.9999	0.0067	-0.0005	1.0000	0.0111	0.0007	0.9999	33s	
	MCEM	0.0228	0.0148	0.9998	0.0064	-0.0008	1.0000	0.0132	0.0026	0.9998	192s	
	ML2P-VAE _{full}	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
(ii) 3 abilities (ECPE)	ML2P-VAE _{est}	0.2794	0.2152	0.9713	0.1480	0.0951	0.9930	0.4430	-0.0628	0.8237	61s	
	ML2P-VAE _{ind}	0.3208	0.2184	0.9504	0.1540	0.0872	0.9932	0.3063	0.0100	0.9017	49s	
	MHRM	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	QMCEM	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	MCEM	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	ML2P-VAE _{full}	0.0780	0.0473	0.9983	0.0608	0.0054	0.9996	0.6145	0.0065	0.7893	1292s	
	ML2P-VAE _{est}	0.2992	-0.1304	0.9822	0.1655	0.1215	0.9987	0.7364	-0.0276	0.7257	961s	
	ML2P-VAE _{ind}	0.2043	0.0592	0.9792	0.0958	-0.0029	0.9992	0.7054	0.0747	0.7135	850s	
	MHRM	0.0953	-0.0158	0.9966	0.0614	-0.0101	0.9988	0.6325	0.0118	0.7697	94s	
	QMCEM	0.0938	-0.0160	0.9967	0.0614	-0.0179	0.9989	0.6326	0.0154	0.7696	29s	
(iv) 4 abilities (Simulated)	MCEM	0.0951	-0.0138	0.9966	0.0644	-0.0199	0.9987	0.6326	0.0150	0.7696	196s	
	ML2P-VAE _{full}	0.1326	0.0780	0.9960	0.0872	-0.0311	0.9978	0.6384	0.0210	0.7648	37s	
	ML2P-VAE _{est}	0.2526	0.2106	0.9883	0.1035	-0.0337	0.9980	0.6897	-0.0256	0.7182	38s	
	ML2P-VAE _{ind}	0.1658	0.1099	0.9939	0.0944	-0.0254	0.9976	0.6474	-0.0397	0.7579	30s	

Note that in data set (ii), there are no true values, so MHRM estimates are accepted as true. In data set (iii), only ML2P-VAE methods are capable of estimating such high-dimensional latent traits

Discrimination Parameter Estimates

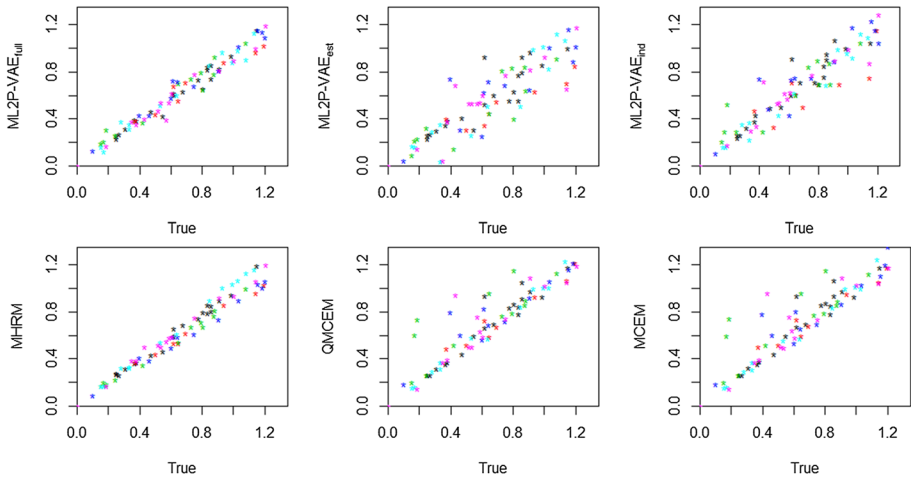


Fig. 2 Correlation plots of discrimination parameter estimates for data set (i) with 50 items and 6 latent traits. ML2P-VAE estimates are on the top row, and traditional method estimates are on the bottom row. Each color represents discrimination parameters relating one of the 6 latent skills

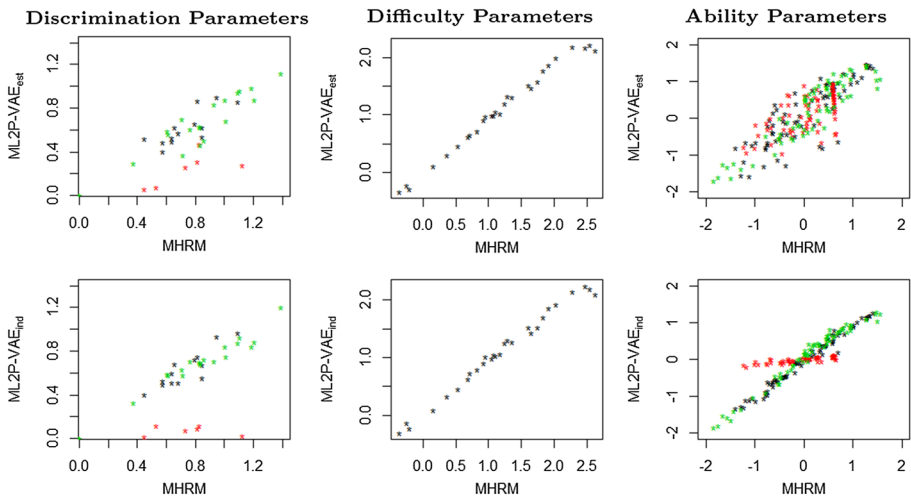


Fig. 3 Estimates from ML2P-VAE methods plotted against “accepted” MHRM estimates from ECPE data. The three colors in the left and right plots represent discrimination and ability parameters associated with the three latent traits

method found that the red ability has no effect on exam performance. On the other hand, ML2P-VAE_{est} captures the general trend of the MHRM ability parameters, but the estimates have much more variance. The discrimination parameter estimates also show some correlation, but each of the three abilities are on a different scale.

While estimating parameters for data set (iii), the dimension of the latent traits (\mathbb{R}^{20}) is too large for traditional methods, so we only study the three ML2P-VAE techniques.

Discrimination and Ability Parameter Estimates

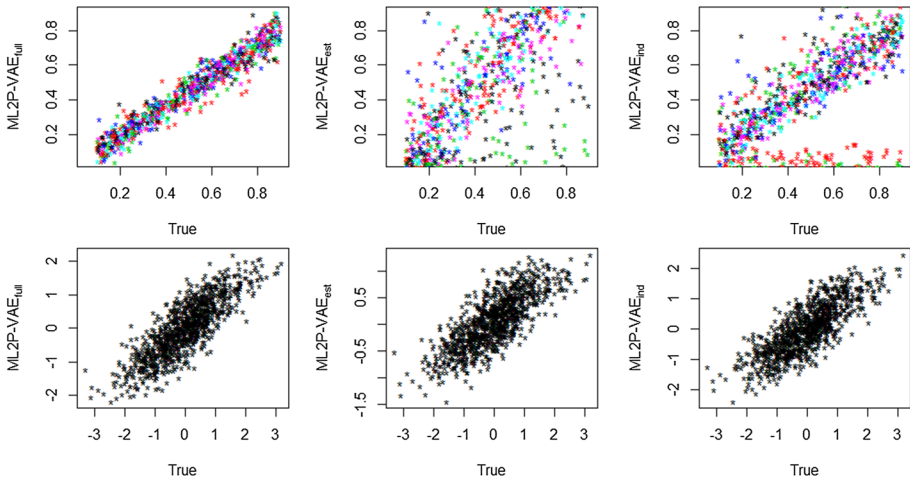


Fig. 4 ML2P-VAE parameter estimates for data set (iii) with 200 items and 20 latent traits. The top row shows discrimination parameter correlation, and the bottom row shows the correlation of a single ability parameter. Each color corresponds to discrimination parameters related to one of the 20 latent traits

Discrimination Parameter Estimates

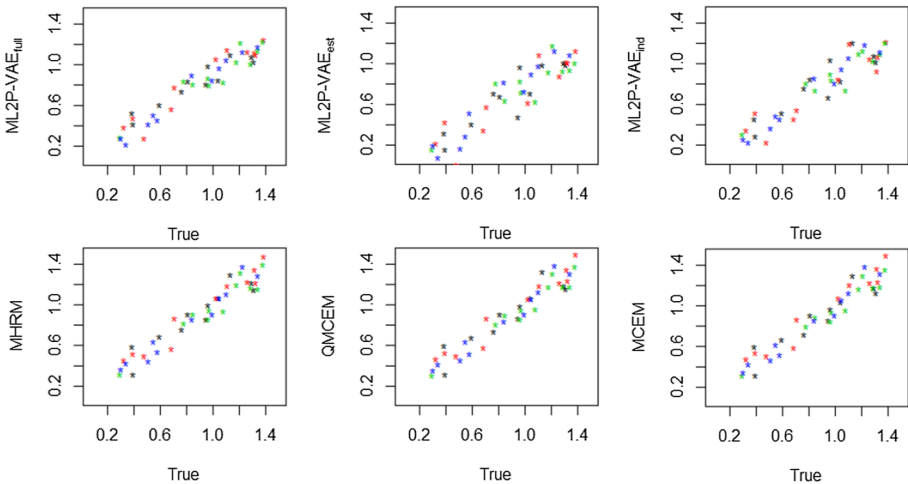


Fig. 5 Discrimination parameter estimates for data set (iv) with 27 items and 4 latent skills. The top row shows estimates from ML2P-VAE methods, and the bottom row gives estimates yielded by traditional methods. The four colors correspond to discrimination parameters associated with one of the four skills

All three of these methods estimate the difficulty parameters with high accuracy. Similar to data set (i), we again observe that the ML2P-VAE_{full} error seems to be independent of the size of the discrimination parameter, a promising trend. However, ML2P-VAE does not perform as well when we do not have full knowledge of the latent ability correlation matrix. The discrimination parameter estimates for ML2P-VAE_{est} seem to have no pattern.

Upon closer inspection, it can be seen that the discrimination parameter estimates associated with a particular ability are correlated, but each ability is on a different scale.

The discrepancy between ML2P-VAE_{full} and ML2P-VAE_{est} can be attributed to a poorly estimated covariance matrix. For this data set, the covariance matrix obtained by the method described previously *greatly* overestimates every correlation between latent traits; the average signed bias in the correlation matrix estimation is -0.61 , and even the closest correlation estimation has signed bias -0.26 . Finding a better method to compute an approximate correlation matrix could greatly improve this method.

The estimates for data set (iii) produced by ML2P-VAE_{ind} display the same behavior observed in data set (ii). Two of the abilities have discrimination parameters estimated near zero, meaning ML2P-VAE_{ind} deemed these abilities to have no relation with performance on the assessment. But in contrast to the ECPE data, data set (iii) was simulated and we know that this is not true. Outside of this issue, the other discrimination parameters were reasonably estimated, showing clear correlation with the true values on near a 1:1 scale.

Though ML2P-VAE_{est} and ML2P-VAE_{ind} have trouble converging to the true discrimination parameters, they are still able to obtain quality estimates to the ability parameters. The values in Table 1 for θ in data set (iii) are comparable to those of data set (i). The plots in Fig. 4 show this high correlation in all three ML2P-VAE variants.

In data set (iv), the advantages of ML2P-VAE methods are less apparent. The runtime difference is much smaller, since traditional methods do not struggle so much when integrating over a smaller latent dimension of size 4. This also affects the accuracy of parameter estimates. The latent skill estimates are better in data set (iv) than those of data set (i) for all methods, but particularly the traditional methods. For latent ability θ and item difficulty b , all six methods produced similar estimates, and so these correlation plots are omitted. As seen in Table 1, the corresponding error measures are very close, though traditional methods are slightly more accurate.

A comparison between the data set (iv) discrimination parameter estimates is shown in Fig. 5, which clearly visualizes the values in Table 1. Though all ML2P-VAE methods produce highly correlated estimates, they also tend to underestimate the true values. This is most apparent in the plot for ML2P-VAE_{est} and in the relative bias values in Table 1. While traditional parameter estimation results may be more desirable in data set (iv), this demonstrates that the ML2P-VAE methods are most useful when the number of latent abilities is large.

3.3 Effect of training data size

A common criticism of neural networks is that they are computationally intensive and training them with a gradient descent based algorithm (a first order method) can take a long time. They also require large amounts of data. As mentioned before, the architecture used in this application results in a relatively small neural network.

The longer runtimes in Table 1 for data set (iii) can be attributed more to the fact that there were 50,000 data samples, rather than the large latent dimension. The left plot of Fig. 6 displays the relation between the size of the training data and estimation accuracy. We see that the error does not decrease very much after the number of training samples becomes greater than 20,000—less than half of the available simulated data. The right plot of Fig. 6 shows that training time grows linearly with the size of training data.

Both plots in Fig. 6 demonstrate the trade-off between accuracy and speed, as well as highlighting that ML2P-VAE methods can still be viable, even if the data size is not

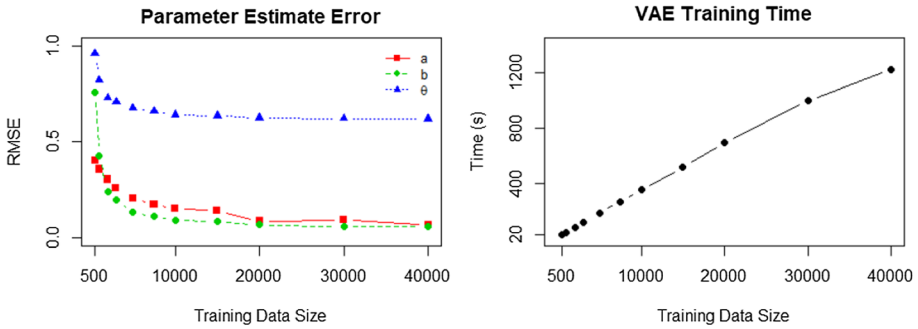


Fig. 6 Performance of ML2P-VAE_{full} on data set (iii) when trained on data sets of increasing size. The left plot gives the test RMSE after using different sizes of training data, and the right plot shows the time required to train the neural network

exceptionally large. This is particularly true in estimating the ability parameter Θ . Estimating the difficulty parameters b is manageable with a smaller data set, while discrimination parameters require a large amount of training data to obtain quality estimates.

4 Discussion

4.1 Future work

This work introduces additional paths for continued research. One important topic involves analyzing the convergence of ML2P-VAE methods. It is important to find conditions which guarantee that the estimates for the discrimination and difficulty parameters will converge to their respective true values. Based on the results shown in this work, it seems likely that convergence will require full knowledge of the covariances among latent traits. In each data set, we saw that either using an inaccurate estimated covariance matrix or simply assuming that latent traits are independent results in inaccurate parameter estimates. Another possible factor in ML2P-VAE's convergence is the sparsity of the Q -matrix. If $Q_{ik} = 1$ for all i, k , then interpretation of the encoded hidden layer as estimates to ability parameters and weights/biases in the decoder as discrimination/difficulty parameter estimates may not be possible.

In real applications, it is unlikely that the exact correlations between latent abilities are available, so an approximate covariance matrix would need to be used instead. The experiments in this work imply that convergence likely relies on knowledge of an accurate covariance matrix among latent traits, thus it is important to develop better methods of estimating this covariance matrix.

The methods used in our study differ from other sampling methods that can be used in the estimation of MIRT models. Specifically, our methods seek to find optimal values for item and person parameters of MIRT models via optimization of a loss function. In this respect, they more closely resemble marginal maximum likelihood methods that seek to find a global maxima across a likelihood function that has been marginalized across the person parameters. Alternatively, other methods rely on sampling from a Bayesian posterior distribution. Such estimation methods include Gibbs sampling, Hamiltonian Monte Carlo, and Metropolis-Hastings algorithms. These algorithms differ in their approach in

that rather than optimize a loss function (or maximize a likelihood function), they sample values from the posterior distributions of the item and person parameters then summarize these distributions using their mean (expected a posteriori) or mode (maximum a posteriori). Similarly, the Robbins-Monroe algorithm presents a hybrid approach, iterating between estimation person parameters using a sampling-based Bayesian algorithm and then, based on summaries of those parameters, maximizing likelihoods for each of the item parameters. It may be worthwhile to compare ML2P-VAE methods with such sampling-based parameter estimation approaches in the future.

It is also possible that the ML2P-VAE method can be extended to estimating the parameters in the Multidimensional Logistic 3-Parameter model (Birnbaum, 1968), which introduces a guessing parameter for each item. Implementing a guessing parameter into the VAE framework is trivial. However, since many other parameter estimation methods struggle in estimating a 3-parameter model (Baker and Kim, 2004), “ML3P-VAE” may face the same issue.

4.2 Conclusion

ML2P-VAE is a novel technique which allows IRT parameter estimation of correlated high-dimensional latent traits. This requires a VAE architecture capable of fitting a more general multivariate Gaussian distribution, rather than a standard normal distribution. Where other estimation methods rely on numerical integration or MCMC methods, which become infeasible for large numbers of latent abilities, ML2P-VAE trains a neural network using a gradient descent based optimization method. While this technique introduces hundreds or thousands of trainable parameters, the parameters in the decoder can be interpreted as estimates to discrimination and difficulty parameters. The individual parameters in the encoder do not represent anything concrete, but together, they learn a function which maps a student’s response set to a distribution representing the student’s latent ability.

All of these parameters are trained simultaneously by optimizing a single loss function. After training the neural network, the discrimination and difficulty parameter estimates are immediately available, and the ability parameter estimates are easily obtained at test time by feeding forward response sets through the encoder. Note that the estimates for θ_j are not directly trainable parameters of the neural network.

Of course, the most accurate ML2P-VAE method makes the strongest and most restrictive assumption; that the exact correlation quantities between latent abilities is known. This may be impractical in applications, and for this reason the other ML2P-VAE methods must also be closely examined. In theory, using a covariance matrix that is estimated from the data should yield better results than assuming all traits are independent. But if this estimated matrix is inadequate, the accuracy of parameter estimates suffers heavily. A possible way to remedy this is to adjust the weight λ of the KL-Divergence in the VAE loss function. Decreasing this hyper-parameter gives more emphasis on reconstructing inputs, rather than fitting data to an estimated distribution which may be flawed.

ML2P-VAE methods are most useful on high-dimensional data. But even when applied to smaller data sets where traditional techniques are feasible, the results from ML2P-VAE are competitive. They are significantly faster in runtime, and yield similar error measures. When estimating difficulty parameters, the improvement gained from using traditional methods is incredibly small. Estimates for students’ latent abilities are often more accurate when using ML2P-VAE methods. This is especially interesting, as the estimates θ_j are not updated in the iterations of a gradient descent algorithm, while the estimates to a_{ik} and b_i

are. In all, these results show the versatility of ML2P-VAE methods in estimating item and ability parameters from a variety of data sets.

References

- Atkinson, K. (1989). *An introduction to numerical analysis*. Wiley.
- Baker, F., & Kim, S. (2004). *Item response theory parameter estimation techniques* (2nd ed.). Taylor & Francis Group.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In: Lord, F., & Novick, M. (Eds.), *Statistical theories of mental test scores* (pp. 395–479).
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4), 443–459.
- Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, 33(1), 2–14.
- Cai, L. (2009). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins–Monro algorithm. *Psychometrika*, 75(1), 33–57. <https://doi.org/10.1007/s11336-009-9136-x>.
- Camara, W., O'Connor, R., Mattern, K., & Hanson, M. A. (2015). Beyond academics: A holistic framework for enhancing education and workplace success. *act research report series*. 2015 (4). ACT, Inc.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40(1), 5–32.
- Converse, G. (2020). ML2Pvae: Variational autoencoder models for IRT parameter estimation. <https://CRAN.R-project.org/package=ML2Pvae>, r package version 1.0.0.
- Converse, G., Curi, M., & Oliveira, S. (2019). Autoencoders for educational assessment. In: *International conference on artificial intelligence in education (AIED)*.
- Curi, M., Converse, G., Hajewski, J., & Oliveira, S. (2019). Interpretable variational autoencoders for cognitive models. In: *International joint conference on neural networks (IJCNN)*.
- Fraser, C., & McDonald, R. (1988). Noharm: Least squares item factor analysis. *Multivariate Behavioral Research*, 23(2), 267–269.
- Guo, Q., Cutumisu, M., & Cui, Y. (2017). A neural network approach to estimate student skill mastery in cognitive diagnostic assessments. In: *10th International conference on educational data mining*.
- Haberman, S. J. (2005). Identifiability of parameters in item response models with unconstrained ability distributions. Tech. Rep. RR-05-24, Research and Development, ETS.
- Kingma, D., & Welling, M. (2014). Auto-encoding variational Bayes. In: *International conference on learning representations*.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.
- Lord, F., & Novick, M. R. (1968). *Statistical theories of mental test scores*. IAP.
- McKinley, R., & Reckase, M. (1980). The use of the general Rasch model with multidimensional item response data. American College Testing.
- Pardos, Z. A., Heffernan, N. T., Anderson, B., Heffernan, C. L., & Schools, W. P. (2010). *Using fine-grained skill models to fit student performance with bayesian networks*. *Handbook of educational data mining* (p. 417).
- Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In: *Advances in neural information processing systems* (pp. 505–513).
- Reckase, M. D. (2009). Multidimensional item response theory models. In: *Multidimensional item response theory* (pp. 79–112). Springer.
- Runcie, D. E., & Mukherjee, S. (2013). Dissecting high-dimensional phenotypes with bayesian sparse factor analysis of genetic covariance matrices. *Genetics*, 194(3), 753–767.
- Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In: *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882–1891).
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4), 345–354. <https://doi.org/10.1111/j.1745-3984.1983.tb00212.x>.
- Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using mplus. *Educational Measurement: Issues and Practice*, 32, 37–50.
- Van der Linden, W., & Glas, C. (2000). *Computerized adaptive testing: Theory and practice*. Springer.

- Wang, Z., Gu, Y., Lan, A., & Baraniuk, R. (2020). Varfa: A variational factor analysis framework for efficient bayesian learning analytics. [arXiv:2005.13107](https://arxiv.org/abs/2005.13107).
- Wu, M., Davis, R., Domingue, B., Piech, C., & Goodman, N. (2020). Variational item response theory: Fast, accurate, and expressive. In: *Proceedings of the 13th international conference on educational data mining (EDM2020)* (pp. 257–268)
- Zhang, J., Shi, X., King, I., & Yeung, D. Y. (2017). Dynamic key-value memory networks for knowledge tracing. In: *26th International world wide web conference (WWW 2017)* (pp. 765–774).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.