UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Using VAE for Incomplete Educational Data

Claudia Evelyn Escobar Montecino

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

Claudia Evelyn Escobar Montecino

Using VAE for Incomplete Educational Data

Thesis submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Doctor in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Mariana Curi

USP – São Carlos
May 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

E74u

Escobar Montecino, Claudia Evelyn
Using VAE for Incomplete Educational Data / Claudia Evelyn Escobar Montecino; orientadora Mariana Cúri. -- São Carlos, 2022.
59 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Variational Autoencoder. 2. Autoencoder. 3. Item response theory. 4. Incomplete educational data. 5. Neural networks. I. Cúri, Mariana, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Claudia Evelyn Escobar Montecino

Usando VAE para Dados Educacionais Incompletos

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutora em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Mariana Curi

USP – São Carlos
Maio de 2023

To Caio, Javier, Guilherme and Luna.

ACKNOWLEDGEMENTS

Thinking about the acknowledgments when completing my doctorate makes me remember all the way I've come here. Starting with the professional and cultural change I made when I came to Brazil. In this sense, I thank God first, because it has been 12 years of innumerable experiences and learning.

I would like to thank my companion in all battles, my husband, colleague and friend Caio Pena, for his patience, support and dedication over the years to deal with me, our children, work and his doctorate.

To my co-workers and friends at the PIPGEs, many of them current doctors, Ana Paula, Jardel, Oilson and Vitor for the productive study afternoons in room 57.

I thank my advisor Dr. Mariana Cúri, for her patience, support and sharing of knowledge since the first day we started working together. To Guilherme Freire for collaboration with the codes and computational help.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

“Queda prohibido no buscar tu felicidad,
no vivir tu vida con una actitud positiva,
no pensar en que podemos ser mejores,
no sentir que sin ti este mundo no sería igual.”
(Pablo Neruda)

RESUMO

ESCOBAR, C. **Usando VAE para Dados Educacionais Incompletos**. 2023. 70 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Em Psicometria, e em particular em avaliações educacionais, é comum encontrar bases de dados incompletas. A falta de tempo, esquecimento do conteúdo envolvido, nervosismo ou mesmo o formato da prova são alguns dos motivos pelos quais um indivíduo pode deixar itens sem responder em uma avaliação. Neste contexto, é importante a existência de métodos de estimação para modelos psicométricos que lidem com dados faltantes e sejam afetados o menos possível pela ausência de informação naqueles itens não respondidos. Num cenário de pequena dimensão, métodos tradicionais de estimação para modelos de Teoria de Resposta ao Item (TRI), por exemplo, são adequados para situações com dados completos e incompletos. No entanto, para situações de alta dimensionalidade, como em avaliações que envolvem muitas competências e habilidades latentes, os métodos tradicionais não são computacionalmente eficientes ou mesmo incapazes de obter estimativas para tantos parâmetros. Aprendizagem profunda vem sendo adaptada de forma a incorporar modelos de TRI e fazer previsões e estimações a partir de grandes bancos de dados, de alta dimensionalidade. Neste trabalho, aprofundamos a investigação de Curi (2019), que definiu um Modelo Logístico de Dois Parametros (ML2P) na arquitetura de um autoencoder variacional (VAE) como uma proposta para solucionar o problema de estimação dos muitos parâmetros do modelo. Realizamos um estudo de simulação para comparar duas variações de redes neurais profundas, autoencoders (AE) e VAE, definidas com um modelo ML2P no decodificador, para situações com um número grande de traços latentes e dados completos. Após resultados favoráveis do VAE, propomos uma extensão do mesmo (IVAE) para poder fazer previsões em casos de dados faltantes e, assim, tornar o modelo mais geral e útil na prática. Simulações do modelo proposto foram realizadas sob diferentes cenários para investigar a eficiência do novo método na recuperação dos parâmetros. Comparações dos resultados com uma das metodologias atualmente mais indicadas em TRI para lidar numa situação de maior dimensionalidade, a máxima verossimilhança conjunta, também foram feitas, além da aplicação a um caso real de alta dimensão e com dados faltantes.

Palavras-chave: Autoencoder variacional. Autoencoder. Teoria da resposta ao item. Dados ausentes. Dados Educacionais Incompletos. Redes neurais.

ABSTRACT

ESCOBAR, C. **Using VAE for Incomplete Educational Data**. 2023. 70 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

In Psychometrics, especially in educational assessments, incomplete databases are common. An individual may leave items unanswered in an assessment due to lack of time, forgetting the content involved, nervousness, or even the test design. In this context, it is crucial to have estimation methods for psychometric models that deal with missing data and are as little affected as possible by the lack of information on those unanswered items. In a small-scale scenario, traditional estimation methods for Item Response Theory (IRT) models, for example, are suitable for situations with complete and incomplete data. However, traditional methods are not computationally efficient or cannot obtain estimates for many parameters, such as assessments involving many latent skills and abilities. Deep learning has been adapted to incorporate IRT models and make predictions and estimates from large, high-dimensional databases. In this work, we deepen the investigation of Curi (2019), who defined a Two Parameter Logistic Model (ML2P) in the architecture of a variational autoencoder (VAE) as a proposal to solve the problem of estimating the many parameters of the model. We performed a simulation study to compare two variations of deep neural networks, autoencoders (AE) and VAE, defined with an ML2P model in the decoder, for situations with a large number of latent traces and complete data. After favorable results of the VAE, we propose an extension of the same (IVAE) to make predictions in cases of missing data and, thus, make the model more general and useful in practice. Simulations of the proposed model were performed under different scenarios to investigate the efficiency of the new method in recovering the parameters. Comparisons of the results with one of the methodologies currently most indicated in IRT to deal with a situation of greater dimensionality, the joint maximum likelihood, were also made, in addition to the application to a real case of high dimension with missing data.

Keywords: Variational autoencoder. Autoencoder. Item response theory. Missing data. Incomplete Educational Data. Neural networks.

LIST OF FIGURES

Figure 1 – Autoencoder with a hidden layer. 28
Figure 2 – Variational autoencoder structure example. 29
Figure 3 – VAE with the decoder as a ML2P model 33
Figure 4 – Representation of the VAE algorithm 38
Figure 5 – Representation of the IVAE algorithm with one iteration. 39
Figure 6 – Comparison of the correlations regarding θ, a, and b, respectively, for three methods in the six different scenarios. 43
Figure 7 – Comparison of the mean square error regarding θ, a and b, respectively, for three methods in the six different scenarios. 44
Figure 8 – Comparison of bias regarding θ, a, and b, respectively, for three methods in the six different scenarios. 44
Figure 9 – Actual versus estimated discrimination parameter by VAE, AE, and JML methods, respectively, for the simulation with 160 items and 10000 individuals. 45
Figure 10 – Real versus estimated difficulty parameter by VAE, AE, and JML methods respectively, for the simulation with 160 items and 10000 individuals. . . . 45
Figure 11 – Real versus estimated latent trait for the scenario Sim1 46
Figure 12 – Actual versus estimated latent trait for the scenario Sim2 46
Figure 13 – Real versus estimated latent trait for the scenario Sim3 47
Figure 14 – Real versus estimated latent trait for the scenario Sim4 47
Figure 15 – Real versus estimated latent trait for the scenario Sim5 48
Figure 16 – Real versus estimated latent trait for the scenarios Sim6 respectively 48
Figure 17 – Actual θ_7 versus θ_7 estimated for Sim6 in replica 1 (top) and replica 3 (bottom) for the three methods. 49
Figure 18 – From left to right: Probability of correctly answering item 1 as a function of the latent trait θ_7 and the probability of correctly answering item 2 as a function of the latent trait θ_8 in the scenario Sim6 49
Figure 19 – From left to right: Probability of correctly answering item 3 as a function of the latent trait θ_7 and the probability of correctly answering item 4 as a function of the latent trait θ_{10} in the scenario Sim6 50
Figure 20 – From left to right: Probability of correctly answering item 5 as a function of the latent trait θ_{12} and the probability of correctly answering item 6 as a function of the latent trait θ_{16} in the scenario Sim6 50

Figure 21 – Correlation and Mean Quadratic Error for the MIRT, VAE, and IVAE methods for the three dimension 3 scenarios of this section. 51
Figure 22 – MIRT vs IVAE for complete data 52
Figure 23 – MIRT vs IVAE for incomplete data. 52
Figure 24 – MIRT vs IVAE for semi-incomplete data. 52
Figure 25 – Correlations for 20 iterations of the Semi_Incomplete_Dim3 and the Incomplete_Dim3 respectively. 53
Figure 26 – Mean square error for 20 iterations of the Semi_Incomplete_Dim3 and the Incomplete_Dim3 respectively. 53
Figure 27 – Comparison between the averages of the correlations for the discrimination and difficulty parameters and for the latent traits. For the IVAE and JML methods in the different scenarios with and without missing data 55
Figure 28 – Comparison between the averages of the RMSE for the discrimination and difficulty parameters and for the latent traits. For the IVAE and JML methods in the different scenarios with and without missing data 56
Figure 29 – Comparison between the averages of the AVB for the discrimination and difficulty parameters and for the latent traits. For the IVAE and JML methods in the different scenarios with and without missing data 57
Figure 30 – Actual versus estimated discrimination parameter by IVAE and JML methods respectively. 58
Figure 31 – Actual versus estimated discrimination parameter by IVAE and JML methods respectively. 58
Figure 32 – Actual versus estimated difficulty parameter by IVAE and JML methods respectively. 59
Figure 33 – Actual versus estimated difficulty parameter by IVAE and JML methods respectively. 59
Figure 34 – Real versus estimated latent trait by IVAE and JML methods respectively. . 60
Figure 35 – Real versus estimated latent trait by IVAE and JML methods respectively. . 60
Figure 36 – Comparison of discrimination parameter estimates, for complete data, with 10% and 20% of missing data. Methods used IVAE, JML, and MIRT. 62
Figure 37 – Comparison of difficulty parameter estimates, for complete data, with 10% and 20% of missing data. Methods used IVAE, JML, and MIRT. 62
Figure 38 – Comparison between latent traits estimated by JML and IVAE 63
Figure 39 – Comparison between latent traits estimated by MML and JML 63
Figure 40 – Comparison between latent traits estimated by MML and IVAE 64
Figure 41 – Comparison between latent traits estimated by JML and IVAE to complete data 64
Figure 42 – Comparison between latent traits estimated by JML and IVAE with 10% missing data 65

Figure 43 – Comparison between latent traits estimated by JML and IVAE with 20% missing data 65
Figure 44 – Estimated probability of getting the item 6 right as a function of θ_{10} 66

LIST OF TABLES

Table 1 – Scenarios simulated to compare the AE, VAE, and JML. 42
Table 2 – Mean of the correlations of the a, b and θ. 43
Table 3 – Mean of the root mean square error of the a, b and θ. 43
Table 4 – Mean of the AVB of the a, b and $theta$. 44
Table 5 – Scenarios simulated to compare IVAE with MIRT. 51
Table 6 – Comparation of correlations and mean square error of estimated parâmeters by different methods 51
Table 7 – Scenarios simulated to compare IVAE with MIRT. 54
Table 8 – Mean of the correlations of the a, b and θ. 55
Table 9 – Mean of the root mean square error of the a, b and θ. 56
Table 10 – Mean of bias of the a, b and θ. 57
Table 11 – Number of items that evaluates each latent trait of the Mathematics ACT (dimension 4) 61
Table 12 – Number of items that evaluates each latent trait of the Mathematics ACT (dimension 22) 64

CONTENTS

1 INTRODUCTION 23

2 BACKGROUND 27
2.1 Autoencoder 27
2.2 Variational Autoencoder 29
2.3 Multidimensional Logistic 2-parameter Model 30
2.4 Marginal Maximum Likelihood and Expected a Posteriori 31
2.5 Joint Maximum Likelihood 32
2.6 AE and VAE combined with the ML2P 32
2.7 VAE versus EM algorithm 34
2.8 Stochastic Gradient Descent (SGD) 34

3 PROPOSAL: IVAE, AN ADAPTATION OF VAE FOR MISSING DATA 37
3.1 VAE for missing data 37

4 SIMULATIONS 41
4.1 AE vs VAE vs JML to complete data 41
4.2 IVAE vs MIRT 50
4.3 IVAE with iterations 52
4.4 IVAE for high latent trait dimension 53

5 REAL APLICATION 61
5.1 Four dimensions 61
5.2 Twenty-two dimensions 64

Bibliography 69

CHAPTER

1

INTRODUCTION

As an alternative to the parameter estimation currently proposed in the item response theory (IRT) model literature, we present two machine learning algorithms capable of solving estimation problems of high-dimensional latent traits from observed data, which may be complete or with missing data. This problem, named the curse of dimensionality, has been a significant challenge for the conventional estimation methods used in Psychometrics, based on the expectation–maximization (EM) algorithm and Markov chain Monte Carlo (MCMC) methods, which do not deal well with high latent trait dimensions. Therefore, several authors are looking for improvements in this regard.

IRT [Reckase 2009] is a test evaluation theory that proposes models to relate the probability of the responses that an individual gives to a set of questions/items with the proficiencies (latent traits) related to the underlying construct. Until the 1950s, the way to rank people was through the raw scores resulting from a test. This methodology, named Classical Test Theory (CTT), left much room to compare people who responded to different tests or assess learning over time. On the other hand, IRT models the response probability of each item as a function of the latent traits at stake and considers, for example, the difficulty and discriminating power of the item. This characteristic complements the classic assessment methods and allows comparisons among different evaluation results and over-time contrasts.

Intuitively, the model establishes a common metric for the latent traits of individuals. Proficiency in a particular competence can be calculated considering the difficulties of the items that the individual answers correctly, enabling the comparison of the same competence between many individuals, even though they do not respond precisely to the same test.

With the advancement of computing and the ability to manipulate large volumes of data, traditional psychometric methods have gained some extensions. Cai 2017 proposes modifying the Metropolis-Hastings algorithm (MH-RM) for high-dimensional maximum marginal likelihood in exploratory item factor analysis and exemplifies it with five dimensions. Despite mentioning

having applied MH-RM to problems with dimensions greater than five and recognizing that more research is needed to cover other situations. Some years later, Chen et al. 2019 proposed a constrained joint maximum likelihood estimator (CJMLE) for high dimension, capable of dealing with more than ten latent traits and more than ten thousand respondents, in addition to allowing the presence of missing data. Their results are valid under an asymptotic setting in both the numbers of items and individuals growing to infinity.

In parallel with the evolution of deep learning, automatic models that learn patterns have gained notoriety due to improved productivity and cost reduction when working with large databases. Although most of these methods are mainly explored in image compression or analysis, some variations have been recently established and allow estimating the parameters of an IRT model to escape the "curse of dimensionality".

Curi *et al.* 2019 proposed a new Variational Autoencoder (VAE) incorporating the matrix Q (which connects each item, represented by a node in the output layer, only to the latent variables measured by it) and a Multidimensional 2-Parameter Logistic model (M2PL) on the decoder. However, this method does not deal with missing data. Converse 2019 proposed an analogous model using an Autoencoder (AE) architecture and compared the two methods (VAE and AE) for simulated data using three-dimensional continuous latent variables. Following this line of thought, in the present work, we compared the AE and VAE methods (with Q-matrix and ML2P in the decoder) and the CJMLE in several scenarios with higher latent trace dimensions.

Other authors have recently worked with Deep Learning in educational assessment. Urban and Bauer 2021 extends Curi's work by introducing a model that can be applied to exploratory analysis of polytomous item response data in the frequentist configuration. Wu *et al.* 2020 also presents an extension of Curi´s method in the Bayesian setting. Additionally, Maris and Bechger 2000 represented the Boltzmann machines, a type of stochastic recurrent neuronal network, as multidimensional item response theory (MIRT) models, showing them as a rich class of generative models.

The results of the present work corroborated the findings of Converse 2019, which indicated better estimation results with VAE. The theme of this dissertation will focus on VAE model (the best in the previous comparisons) to extend the estimation method to incomplete data sets, a more realistic scenario.

It is common to find unanswered items in both educational and psychological tests. Omission of responses due to lack of knowledge, lack of time, lack of interest in responding, or simply because the item was not in the test by design when the individuals being evaluated respond to tests with different items. Whatever the reason for missing data, they make the data analysis difficult and compromise the quality of decision-making. In this regard, it is imperative to extend deep learning methods proposed as IRT model parameter estimation methods to more realistic scenarios dealing with incomplete data.

Some authors have proposed deep learning methods that deal with missing data. Nelwamondo 2007, for example, makes a comparison between the expectation maximization (EM) algorithm and the auto-associative neural network and genetic algorithm (GA) combination. Yoon 2018 proposes a method to impute data by adapting the Generative Adversarial Nets (GAN). In the case of VAE, Cardoso *et al.* 2020 and Boquet *et al.* 2020 propose VAE to solve two problems with missing data, one related to images and the other to road traffic.

While MCMC and EM [Takahashi 2017] deal well with missing data in the literature of IRT models, they require that alternative methods be utilized when high dimensions are present. CJMLE and MH-RM are the most frequently referred alternative methods in Psychometrics, conditional to an asymptotic setting for items and individuals or not-too-high dimensional situations, respectively. To enhance these possibilities, we extend the study of Curi *et al.* 2019 and propose a VAE as a machine learning alternative to deal with missing data in educational assessment. First, we modify the objective function optimized to obtain the item parameters and latent trait estimates to make predictions in the presence of missing data. Subsequently, the imputation of the missing data is proposed through its own VAE.

This work has three contributions: (i) to explain how the two DL methods (AE and VAE) can be linked to the ML2P model of IRT in terms of their parameters and traditional estimation methods, (ii) to compare the performance of AE and VAE in estimating ML2P model parameters for high latent trait dimension, and (iii) to propose a modification in the VAE structure to address missing data in the estimation of ML2P model parameters for high latent trait dimensions. The third contribution is the most significant, while the other two fill a gap in the existing literature about VAE and IRT.

In chapter 2, we begin with the presentation and comparison of two Deep Learning (DL) methods, AE and VAE, and the relationship that we can establish between them and a high latent trait dimensional ML2P model. This relationship is explored for different scenarios with complete data inspired by a real high-dimensional assessment in Brazil. On the other hand, also in chapter 2, the traditional methods of estimating the parameters of the items in IRT are presented, both for low and high latent trait dimensions (Marginal Maximum Likelihood and Joint Maximum Likelihood, respectively). Then, we'll look at the math behind the Stochastic Gradient Descent (SGD), the algorithm used by AE and VAE to optimize the objective function and estimate model parameters. The relationship between the VAE and the Expectations Maximization Algorithm (EM) is highlighted, which will help construct a technique that adapts the VAE to deal with missing data. In chapter 3, we propose a new imputation strategy in a VAE to address missing data in estimating latent traits and item parameters for ML2P model. In chapter 4, a simulation study is presented, comparing the methods in practice for several scenarios for complete and incomplete data. We extend this study in chapter 5 to a real problem. Finally, the conclusions and future intentions for this research are presented.

CHAPTER

2

BACKGROUND

We can understand machine learning (ML) as a discipline that automatically learns patterns from data through neural networks. When said analysis is deeper and involves a large volume of data, we speak of deep learning (DL). In this case, we can fit the theory involving chatbots, which is much discussed today.

IRT models are used to analyze data from educational and psychological tests. They assume that each item on a test measures one or more latent trait(s) (e.g., reading comprehension, mathematical ability, personality traits) and that the probability of the response to an item depends on the item parameters and the abilities of the test-takers.

One of the contributions of this work is the comparison between two deep learning methods - the AE and the VAE - for estimating parameters in an IRT model within a high dimensional latent trait space. The next chapter will focus on this comparison. To aid comprehension, we will introduce the main characteristics of each method and explain their relationship to an IRT model, the ML2P, in this chapter.

We will also present the two commonly used methods for estimating item parameters and latent traits in low- and high-dimensional IRT models: marginal maximum likelihood (MMLE) and joint maximum likelihood (JMLE) estimation, respectively.

2.1 Autoencoder

Autoencoder is a type of neural network used to explain the behavior of unlabeled data, making it an unsupervised method. We can represent an AE [Goodfellow 2017] through a system of interconnected nodes with an input layer, one or more hidden layers, and an output layer. Through computational training, we reduce the dimensionality of the network's input to recover it in the best way possible. Unlike a neural network in general, an AE has the same number of nodes in the input and output layers. In Figure 1, we have a representation of an Autoencoder

with a hidden layer.

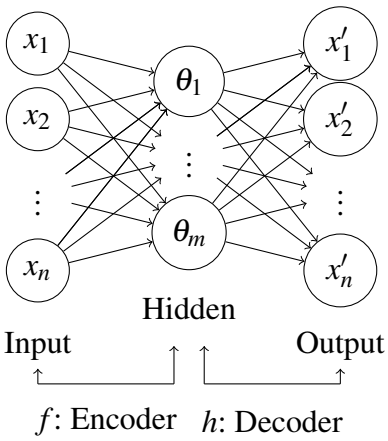

Figure 1 – Autoencoder with a hidden layer.

To understand how the layers of an AE relate, we will consider the case where $x = \{x_i\}_{i=1}^{n}$ is the input we want to encode by means of a $\theta = \{\theta_d\}_{d=1}^{m}$ node. Note that x and θ in the encoding are related through a linear or nonlinear function s, called activation function, applied to the weighted sum, by a weight matrix $W = \{w_{di} : d \in 1\ldots,m; i \in 1\ldots,n\}$, of the components of x plus a bias vector b (see equation (2.1)). The decoding from θ to get x' (that should be as close as possible to x) happens in the same way as described in the encoding (see equation (2.2)).

$$\theta_d = f(x_i) := s^{(1)}\left(\sum_{i=1}^{n} w_{id}^{(1)} x_i + b_d^{(1)}\right) \quad d \in 1,\ldots,m \tag{2.1}$$

$$x'_i = h(\theta_d) := s^{(2)}\left(\sum_{d=1}^{m} w_{id}^{(2)} \theta_d + b_i^{(2)}\right) \quad i \in 1,\ldots,n \tag{2.2}$$

Using Autoencoder, the goal is to minimize some loss function denoted by $\mathscr{L}$, which compares the difference between x and x', given the activation functions, the number of layers, and the number of nodes in each layer. If S is a set of sample unities, the expression is given by:

$$\mathscr{J}(W,b;S) = \sum_{x\in S} \mathscr{L}(x,(h\circ f)(x)) = \sum_{x\in S} \mathscr{L}(x,x') \tag{2.3}$$

It is worth mentioning that when the autoencoder has only one fully connected hidden layer, with a linear activation function and the loss function defined as the quadratic error, the weights obtained by training the network are equivalent to those obtained by Principal Component Analysis (PCA).

In the present work, x will represent the right or wrong answers of a group of individuals to items in a test, so x is binary, and the loss function used is the binary cross-entropy:

$$\mathscr{J}(W,b;S) = -\frac{1}{n}\sum_{i=1}^{n} x_i \log(p(x_i)) + (1-x_i)\log(1-p(x_i)) \tag{2.4}$$

$$= -\frac{1}{n}\sum_{i=1}^{n} x_i \log(x_i') + (1-x_i)\log(1-x_i') \tag{2.5}$$

where $p(x_i)$ is the probability of success of the Bernoulli variable.

2.2 Variational Autoencoder

A Variational autoencoder is an autoencoder in which the intermediate layer nodes represent the parameters of a pre-assumed probability distribution for the latent variables. The decoder propagates the values generated from this probability density with parameters equal to the nodes of the intermediate hidden layer.

Suppose that in Figure 1, the θ is a set of latent variables that through some random process produces the observations x. The purpose of the VAE is to approximate the posterior distribution of θ given the observation x, $q(\theta|x)$ by another probability distribution $g(\theta|x)$. This is because when we have a large latent space, the estimate of the real posterior distribution of $\theta|x$ is unfeasible [Kingma D. 2014]. An illustration is depicted in Figure2

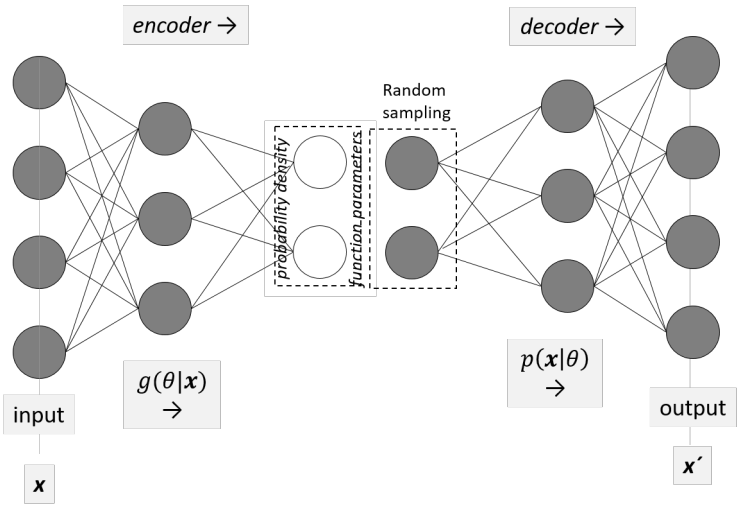

Figure 2 – Variational autoencoder structure example.

To obtain the objective approximation of the VAE, we will minimize the Kullback-Leibler divergence between said functions $g(\theta|x)$ and $q(\theta|x)$, expressed by the equation (2.6).

$$KL(g(\theta|x)\|q(\theta|x)) = E_{\theta\sim g(\theta|x)}[\log g(\theta|x) - \log q(\theta|x)] \tag{2.6}$$

To be able to reach a feasible solution, we need this divergence not to depend on the unknown distribution $q(\theta|x)$. To this end, we will work with equation (2.7).

$$\begin{aligned}
KL(g(\theta|x)\|q(\theta|x)) &= E_{\theta\sim g(\theta|x)}[\log g(\theta|x)] - E_{z\sim g(\theta|x)}[\log q(\theta|x)] \\
&= E_{\theta\sim g(\theta|x)}[\log g(\theta|x)] - E_{\theta\sim g(\theta|x)}\left[\log \frac{q(\theta,x)}{p(x)}\right] \\
&= E_{\theta\sim g(\theta|x)}[\log g(\theta|x)] - E_{\theta\sim g(\theta|x)}[\log q(\theta,x)] + E_{\theta\sim g(\theta|x)}[\log p(x)] \\
&= E_{\theta\sim g(\theta|x)}[\log g(\theta|x)] - E_{\theta\sim g(\theta|x)}[\log p(x|\theta)p(\theta)] + E_{\theta\sim g(\theta|x)}[\log p(x)] \\
&= E_{\theta\sim g(\theta|x)}[\log g(\theta|x)] - E_{\theta\sim g(\theta|x)}[\log p(x|\theta)] - E_{\theta\sim g(\theta|x)}[p(\theta)] \\
&\quad + E_{\theta\sim g(\theta|x)}[\log p(x)] \\
&= -E_{\theta\sim g(\theta|x)}[\log p(x|\theta)] + KL(g(\theta|x)\|p(\theta)) + E_{\theta\sim g(\theta|x)}[\log p(x)]
\end{aligned}$$

As $p(x)$ does not depend on θ, minimizing (2.6) is equivalent to maximizing (2.7).

$$E_{\theta\sim g(\theta|x)}[\log p(x|\theta)] - KL(g(\theta|x)\|p(\theta)) \tag{2.7}$$

Having made these assumptions we trained the Variational Autoencoder by means of the stochastic gradient descending method, explained later in this chapter.

2.3 Multidimensional Logistic 2-parameter Model

The ML2P is a model of IRT widely used in Psychometrics: in the Graduate Record Examination (GRE) 11 and in the Programme for International Student Assessment (PISA) 16, for example. This model defines the probability P that an individual $j \in J$ responds correctly to an item $i \in I$, i.e. $x_{ji} = 1$, given that the individual possesses the latent trait vector $\theta_j = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{jm})$ of dimension m. This probability is given by equation (2.8).

$$P(x_{ji} = 1|\theta_j) = \frac{1}{1 + \exp\left(-\sum_{d=1}^{m} a_{di}\theta_{jd} + b_i\right)}, \tag{2.8}$$

where the discrimination parameter a_{di} relates the latent trait d and the item i, and b_i is the difficulty parameter[1] related to the item i. The relationship between the latent traits and the items can be defined by a Q-matrix, of dimension $m \times I$, with elements $Q_{di} = 1$ if the item i needs the ability d to be answered correctly, and $Q_{di} = 0$, otherwise. It is not uncommon to know

[1] Note that b in the equation 2.8 represents the difficulty parameter in the IRT model, while b in the 2.1 section represents the Autoencoder bias. We use the same notation for both, as it is usual in each area, and also as it will be useful to incorporate ML2P in VAE.

which items are related to which latent variables, especially in the educational area. In this way, equation (2.8) can be rewritten by:

$$P(x_{ji} = 1|\theta_j) = \frac{1}{1 + \exp\left(-\sum_{d=1}^{m} a_{di} Q_{di} \theta_{jd} + b_i\right)}, \tag{2.9}$$

In IRT models, there are various ways to estimate parameters, such as using the likelihood function, Bayesian methods, and Markov Chain Monte Carlo (MCMC) methods. This work will compare the proposed approach to the two most commonly referenced methods for low- and high-dimensional latent variable vectors, presented below.

2.4 Marginal Maximum Likelihood and Expected a Posteriori

The Marginal Maximum Likelihood (MML) is the most commonly used method for estimating item parameters in IRT models. Proposed by Bock and Aitkin in 1981, it is based on the expectation–maximization algorithm. Integrating out the latent trait parameters of the likelihood function, the item parameters are estimated based only on the marginal probabilities of the observed responses.

There are two initial assumptions to the estimation: (i) an individual's responses to items are independent of one another, and (ii) that the item responses of a given individual are independent given his/her latent trait value. Under these assumptions, the marginal likelihood of the observed data can be written as:

$$\prod_j \int \prod_i P_{ji}^{x_{ji}} (1 - {}_{ji})^{1 - x_{ji}} dF(\theta),$$

where P_{ji} is the IRT model, given in (2.8), and the latent variables θ are considered random effects sampled from some larger distribution, $F(\theta)$.

The marginal likelihood is maximized with respect to the item parameters to derive their MML estimates.

It is important to highlight that location and scale constraints are required to identify the model. They can either be placed on the mean and standard deviation of the latent trait distribution, F, or on the item parameters. It is very usual to assume that F is the standard normal distribution.

MML requires numerical approximation integration techniques. And, when the latent trait dimension is high, resulting in multiple integrals to approximate, it becomes unfeasible in practice.

In the second step, given obtained estimates of the item parameters, the latent traits can be estimated either via maximum likelihood estimation (MLE) or using Bayesian methods such as maximum a posteriori (MAP) estimation or expected a posteriori (EAP) estimation.

All these methods are implemented in MIRT package in software R. The default options are MML for item parameter estimation and EAP for latent trait estimation, considering N(0,1) as the prior distribution.

2.5 Joint Maximum Likelihood

In contrast to MML method, in the JML, only one- and two-dimensional numerical integrals need to be evaluated even for high-dimensional cases, making it more computationally efficient.

In the JML method, proposed originally in the 60s, item and latent trait parameters are treated as fixed effects. The likelihood function is then maximized with respect to all of them. However, when the number of individuals tries to infinity, and the number of items is fixed, the number of parameters in the joint likelihood function also grows to infinity, which makes the JML inconsistent.

Chen 2019 proposes a Constrained Joint Maximum Likelihood Estimator (CJMLE), which we refer to as JML for simplicity in this text. In summary, a constraint on the Euclidian norms of both the item and individual parameters is proposed to guarantee a solution for the optimization problem and, consequently, allow estimation for items or persons with perfect scores. It is shown that the estimates are asymptotically consistent when both the numbers of individuals and items grow to infinity. Additionally, the proposed algorithm is suitable for high dimensional scale due to the possibility of updating the parameters in parallel: updating individual parameters given item parameters and vice versa. For more details on this method, see Chen 2019.

2.6 AE and VAE combined with the ML2P

Inspired by the work of Guo 2017, where the authors proposed the use of an AE to obtain probability estimates of master latent classes for cognitive diagnostic assessments, Curi *et al.* 2019 proposed a new method for parameter estimation in IRT. They defined a VAE with the decoder as a ML2P model and illustrated the utility of the method to obtain the parameter estimators of the IRT model, useful especially for high latent trait dimensions. In Converse 2019, this last model was compared with an analogous adaptation established between AE and an IRT model, considering three latent traits. In the present work, we extend these comparisons to a greater dimension of the latent trait space.

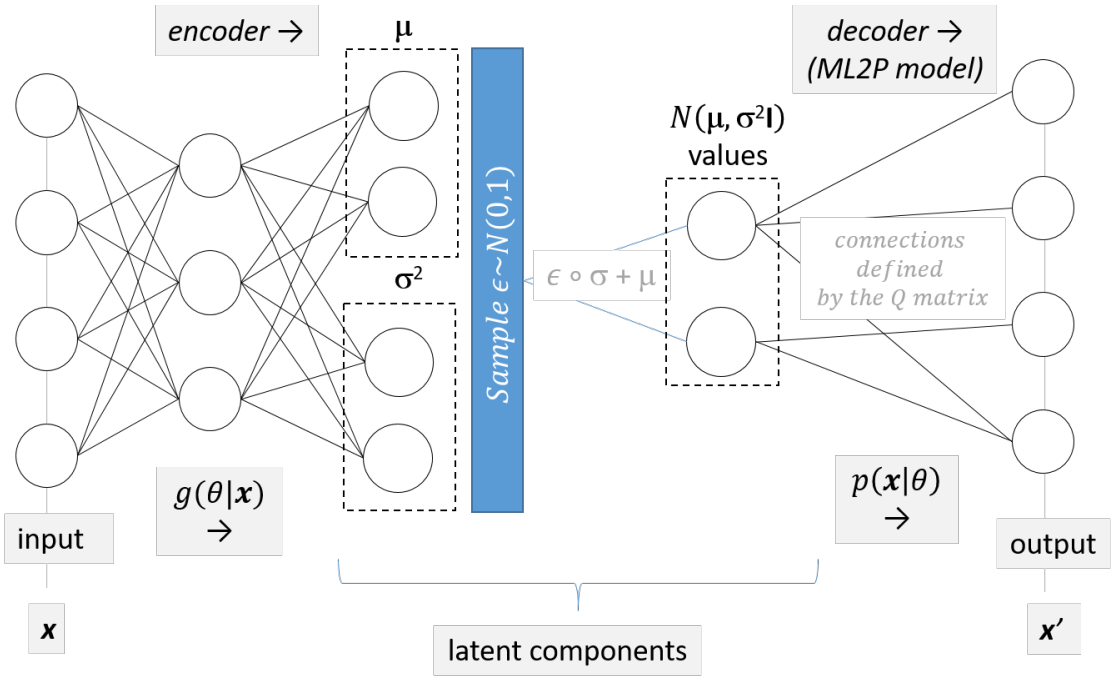

Figure 3 – VAE with the decoder as a ML2P model

The VAE and the AE in this work, as well as in the model of Curi *et al.* 2019, have no hidden layers in the decoder and have a sigmoidal activation function that joins the latent trait layer nodes θ with the output layer x', and with some weights, w, forced to be zeros by the Q-matrix (see figure 3). In this way, the equation (2.2) can be re-written as:

$$x' = \frac{1}{1+\exp\left(-\sum_{d=1}^{D} w_{di}\theta_{jd} + b_i\right)} \tag{2.10}$$

From the relation between the equation (2.8) and the equation (2.10), we can interpret the weights w_{di} of the decoder as estimates for the discrimination parameters a_{dj} and the bias b_i as estimates for the parameters of difficulty b_i of a model ML2P.

The relationship between the latent traits and the items can be defined by a Q-matrix, of dimension $m \times I$, where $Q_{di} = 1$ if the item i needs the ability d to be answered correctly, and $Q_{di} = 0$, otherwise. This is important in order to identify the decoder part of the network and avoid (or preclude) multiple solutions.

This parallel drawn between IRT models and VAE is interesting in two ways: (I) to include some interpretation for the intermediate hidden layer, and (ii) to make computation feasible for high dimensional cases. For the educational area, the observed responses x correspond to the dichotomous responses of an individual to the group of items, which will be called input due to the role it will play within the structure of the network. If there is no hidden layer in the decoder, $p(x|\theta)$ may correspond to the product of Bernoulli distributions depending continuously

on latent traits, θ. And for θ, which corresponds to the latent abilities of each individual located in the hidden layer of the network, we assume $p(\theta)$ with distribution $N(0,\mathbb{I})$, where $\mathbb{I}$ is the identity matrix, as is usual in VAE and MML for IRT estimation.

2.7 VAE versus EM algorithm

The EM algorithm Dempster 1977 is an iterative method to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, with unobserved latent variables.

In this iteration, this method is considered an arbitrary distribution for the underserved data, and once this assumption is made, the model parameters are estimated, so the unobserved data distribution is updated (with the last estimated parameters) and the parameters are estimated again. This process is repeated until some criterion is met.

If we approach $q(\theta|x)$ by $p(\theta)$ (instead of $g(\theta|x)$) in the equation 2.7, aggregate again the term $E_{\theta\sim g(\theta|x)}\log p(x)$ that was taken for not relying on θ and reorganized the terms we have:

$$\begin{aligned} KL(g(\theta|x)\|q(\theta|x)) &= -E_{\theta\sim g(\theta|x)}[\log p(x|\theta)] + KL(g(\theta|x)\|p(\theta)) + E_{\theta\sim g(\theta|x)}[\log p(x)] \\ &\Updownarrow \\ KL(p(\theta)\|q(\theta|x)) &= -E_{\theta\sim p(\theta)}[\log p(x|\theta)] + KL(p(\theta)\|p(\theta)) + E_{\theta\sim p(\theta)}[\log p(x)] \end{aligned}$$

Which is equivalent to the equation (2.11) that is optimized in EM algorithm.

$$E_{\theta\sim p(\theta)}[\log p(x|\theta)] = E_{\theta\sim p(\theta)}[\log p(x)] - KL(p(\theta)\|q(\theta|x)) \tag{2.11}$$

So we can consider the Variational Autoencoder as a generalization of the EM algorithm.

2.8 Stochastic Gradient Descent (SGD)

The SGD is the optimization method adopted for the deep learning techniques.

Remember that the Gradient Descent (GD) is an iterative method that searches for the minimum of a function. To use this method, we need a point, the gradient of the function at that point, and to stipulate the step size we will take to get a new point.

Suppose that we have N observed data of the form $(x_i, x_i')_{i=1}^N$ and we want to find $f_{\omega^*} \in \mathscr{F}$, with $\mathscr{F}$ a family of functions parameterized by the vector ω, where f_{ω^*} minimizes the loss J when we approach x_i' by $f_\omega(x_i)$.

Each iteration t of the gradient descent will be defined as in the equation (2.12), where α is the step size that will be given toward a local minimum of the function:

$$\omega_{t+1} = \omega_t - \alpha \nabla_\omega J(\omega_t), \tag{2.12}$$

The objective function gradient is obtained from the mean evaluated in each of the N sample data. As we can imagine, if N is too large this method can be very expensive:

$$\nabla_\omega J(\omega_t) = \sum_{i=1}^{N} \frac{\nabla_\omega J(f_{\omega_t}(x_i), x_i')}{N} \tag{2.13}$$

In our case, we need to minimize the neural network loss function that we are proposing to estimate parameters of a ML2P model, incorporated in the decoder, as well as the weights of the encoder neural network of the VAE. Since we can usually have large datasets, we will use a modified version of the GD that can handle such situations.

The SGD (or sometimes called Mini Batch Gradient Descent) is a variation of the GD method, which instead of calculating the objective function gradient across the entire dataset, calculates the gradient based on a random subset of the original database at each iteration. This is why SGD is the algorithm commonly used to optimize objective functions in VAEs, because each iteration of the SGD is more computationally economical than the iterations of the GD. Therefore, in this case $\nabla_\omega J(\omega_t)$ is defined as in the equation 2.14. Where $\mathcal{N}_t$ is some subset of $\{x_1, \ldots, x_N\}$.

$$\nabla_\omega J(\omega_t) = \sum_{i \in \mathcal{N}_t} \frac{\nabla_\omega J(f_{\omega_t}(x_i), x_i')}{N_t} \tag{2.14}$$

The random subset $\mathcal{N}_t$ is called batch. In each iteration t of the SGD, the gradient $\nabla_\omega J$ is calculated in a different batch. After several batches, when all the data set has been used, we say that we have completed an epoch. In the case of the GD, the batch is the complete dataset. Therefore, each iteration of the GD corresponds to an epoch, while in the SGD we need several iterations to use the entire dataset and complete a single epoch.

To understand more about this method, such as convergence conditions, see Bottou 2010.

CHAPTER

3

PROPOSAL: IVAE, AN ADAPTATION OF VAE FOR MISSING DATA

In the previous chapter, we saw, among other methods, how the VAE can be adapted to estimate the parameters and latent traits of an M2PL. In this chapter, we will propose its generalization to use in cases with missing data. This is crucial because it is a computationally economical method to deal with high-dimension latent trait vector, many items and respondents in the presence of incomplete data, which is very common in practice.

3.1 VAE for missing data

We modified the VAE to be able to estimate parameters even in the presence of missing data, based on the motivation provided by the EM algorithm.

As a first attempt, we propose modifications to the objective function (2.7) to estimate the parameters, disregarding the loss given by missing data.

Let us consider a database associated with the responses of a group of individuals to a test. Represented by a matrix of entries 0 when the item was answered wrongly, 1 if the item was answered correctly and - 1 if the subject did not answer a given item.

We can assume, as seen in section 2.2, that the decoder $p(x|\theta)$ is a product of Bernoulli distributions that depend continuously on the latent traits θ and a priori $p(\theta)$, assumed as the $N(0,\mathbb{I})$ distribution.

Thus, the objective function (2.7) that is optimized by VAE is in the form of (3.1) and (3.2).

$$KL(g(\theta|x)\|q(\theta|x)) = E_{\theta\sim g(\theta|x)}[\log p(x|\theta)] - KL(g(\theta|x)\|p(\theta)) \tag{3.1}$$

$$E_{\theta\sim g(\theta|x)}[\log p(x|\theta)] = -\frac{1}{N}\sum_{i=1}^{N} x_i \log(x'_i) + (1-x_i)\log(1-x'_i), \tag{3.2}$$

where N is the size of the sample being considered.

The equation (3.2) works perfectly if we have complete data, i.e., all individuals answer all items. But in the case of missing data, we would like to codify the missing value as "-1" (some different value from the real observations, coded as 0 and 1), for instance, and eliminate this observation from the calculation of the loss function. For this reason, we adapted equation (3.2) so that the summation also considers missing answers inputted as "-1" in the following way:

$$E_{\theta\sim g(\theta|x)}[\log p(x|\theta)] = -\frac{1}{N}\sum_{i=1}^{N} (0.5x_i^2 + 0.5x_i)\log(x'_i) + (1-x_i^2)\log(1-x'_i) \tag{3.3}$$

Note that in the presence of complete data (3.2) e (3.3) are equivalent. In contrast, when $x_i = -1$, the element in the sum operator equals 0, i.e., is not considered in the loss function. For this reason, we will continue calling this method VAE (see Figure 4).

Figure 4 – Representation of the VAE algorithm

To run and train a VAE in Figure 4, we first define the estimation problem, giving the network the right, wrong or missing answers of all individuals to all test items through a matrix of 1, 0, and -1, respectively. In the decoder, the Q matrix defines the existent connections among the nodes of the last two layers of the network, representing which latent traits evaluate each item. On the other hand, we must define the number of lots selected from the sample set to carry out the stochastic optimization method that will minimize the loss function. Then we define the network architecture in the encoder and the decoder, the number of layers and nodes, and the activation functions that relate these layers. In this work, we use the activation functions *tanh* in the encoder, which has one hidden layer, and *sigmoid* in the decoder, mandatory to represent the IRT model. We define a function that restricts the network weights from the Q matrix. And finally, we train the network through the optimization function SGD. We can retrieve the weights of the

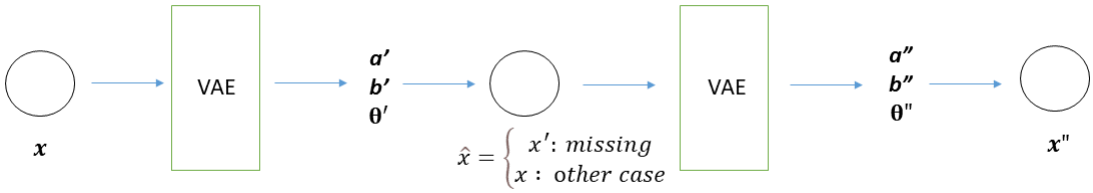

Figure 5 – Representation of the IVAE algorithm with one iteration.

network decoder and the latent trait layer, which will be the parameter estimates of difficulty, discrimination, and latent traits.

We have proposed a second solution for handling missing data, viewing the VAE as a generalization of the EM algorithm, as shown in equation (2.7). Analogously, we can extend the way the EM algorithm works. For this purpose, we will use the VAE with the modified loss function (3.3) to estimate the latent traits of an incomplete database. Initially, the current estimated latent traits are used to impute individuals' missing responses in the second step of the proposed algorithm. If the output of the VAE id is greater than 0.5, the imputed value is equal to 1. Otherwise, it is equal to 0. Then the network is retrained with the complete imputed database to obtain the final estimates for the latent abilities. We will name this algorithm Incomplete Variational Autoencoder (IVAE); see Figure 5. This process can be performed iteratively until the estimates stabilize or by defining another stopping criterion.

Note that the first solution, proposed in Figure 4, is a particular case of the second proposal, in Figure 5, with no network retraining with the imputed database.

Both the VAE represented in Figure 4 and the IVAE represented in Figure 5 will be executed through codes implemented in Python in a single machine with an Intel(R) Core(TM) i7-8565U CPU @1.80GHz (8CPUs). The codes, graphics, and databases implemented and used in this work can be found in the repository 9.

CHAPTER

4

SIMULATIONS

In this chapter, we will present some simulations to show DL methods are comparable to the ones currently used to estimate parameters and latent traits of an M2PL model.

Initially, we present analysis for complete data case with latent trait dimension 18 and compare the AE with the VAE, as they were the first two methods that we started to explore to present the proposal for this work. As we are in a high-dimensional case, we will also compare them with JML, the method currently recommended in the literature for IRT estimation in these scenarios.

Next, we will demonstrate the efficiency of our innovative technique, IVAE, in managing missing data scenarios. We will compare its performance with the MML method in a low-dimensional setting and examine parameter recovery quality in a high-dimensional case.

4.1 AE vs VAE vs JML to complete data

Our first simulation study was inspired by the Brazilian test ANA Microdados 2014, National Literacy Assessment developed by INEP in 2013, to inform the levels of mathematical literacy of 3rd-year public school students. We simulated 50 replicates of the responses of 5000, 10000, and 20000 individuals, for two tests: one with 80 and the other with 160 items, both involving 18 latent traits. This organization formed the six scenarios described in Table 1. Discrimination and difficulty parameters were also independently generated from N(0,0.3) and N (0,0.6), respectively.

ANA has 80 simple items that evaluate 18 skills. An item is called simple when it evaluates only one skill. The relationship between the items and the skills is given by a Q-matrix formulated by INEP researchers.

We trained the AE and VAE to calculate the estimates of a_{di}, b_i, and θ_{jd} for the six proposed scenarios and each of the 50 replicates. The structure of the networks followed the

Scenario	Itens	Individuals
Sim1	80	5000
Sim2	80	10000
Sim3	80	20000
Sim4	160	5000
Sim5	160	10000
Sim6	160	20000

Table 1 – Scenarios simulated to compare the AE, VAE, and JML.

description in the previous Chapter: one hidden layer in the encoder with the tangent hyperbolic activation functions in the encoder, an equal number of nodes in the input and output layers (that corresponds to the number of items in the test), one latent variable layer to represent 18 latent traits (skills), no hidden layer in the decoder, and connections to the output layer defined by the Q matrix, using a sigmoid activation function. In addition, the number of nodes in the hidden layer of the encoder was established as half the size of the input. Alternative structures in the encoder were explored (more or no hidden layers, varying the number of nodes), but with no significant improvement in the final accuracy. In this sense, we decided to fix this structure for all the simulations.

Some indices were computed to study and compare the recovery of parameters through the proposed methods: AE, VAE, and JML, the currently used method proposed by Chen to estimate parameters and latent traits of high-dimensional IRT models. These indices are the Pearson coefficient correlation (Corr), the root mean square error (RMSE), and the absolute value of the bias (AVB). Equations (4.1) and (4.2) state the notation, where π_l is the real value of a parameter, $\hat{\pi}_{lr}$ is its respective estimate obtained in the replica r, and $\hat{\bar{\pi}}_l = \frac{1}{50}\sum_{r=1}^{5}\hat{\pi}_{lr}$.

$$RMSE = \sqrt{\frac{1}{50}\sum_{r=1}^{50}(\hat{\pi}_{lr} - \pi_l)^2} \tag{4.1}$$

$$AVB = |\hat{\bar{\pi}}_l - \pi_l| \tag{4.2}$$

In Figure 6 and Table 2, we can see the means of the correlations between the actual parameters and the estimates obtained in each scenario for each method. The prominence of the VAE, about the other two methods, is noted once its correlation means are closer to 1, for the model parameters: the latent traits, and the item parameters of discrimination and difficulty.

It is also observed that increasing the number of items from 80 to 160 improves the quality of the estimates of the three methods.

On the other hand, in Figure 7 and Table 3, we can see that each of the three methods stands out negatively in the mean squared error of each of the three parameters. AE is the worst

Scenario	$Corr_a$ AE	$Corr_a$ VAE	$Corr_a$ JML	$Corr_b$ AE	$Corr_b$ VAE	$Corr_b$ JML	$Corr_\theta$ AE	$Corr_\theta$ VAE	$Corr_\theta$ JML
Sim1	0.535	0.643	0.362	0.898	0.979	0.935	0.653	0.663	0.603
Sim2	0.439	0.555	0.348	0.860	0.975	0.924	0.633	0.654	0.587
Sim3	0.522	0.644	0.208	0.894	0.969	0.923	0.601	0.655	0.585
Sim4	0.872	0.869	0.636	0.937	0.985	0.933	0.787	0.784	0.695
Sim5	0.877	0.923	0.684	0.901	0.978	0.926	0.781	0.782	0.687
Sim6	0.925	0.961	0.633	0.910	0.983	0.918	0.787	0.790	0.592

Table 2 – Mean of the correlations of the a, b and θ.

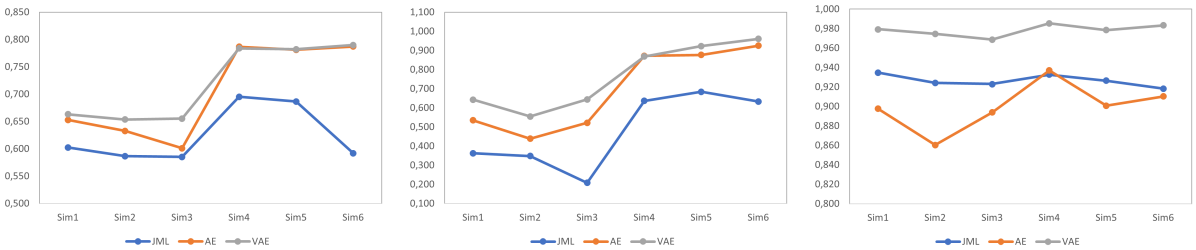

Figure 6 – Comparison of the correlations regarding θ, a, and b, respectively, for three methods in the six different scenarios.

for the latent trait estimation, JML is the worst for the discrimination parameters, and VAE is the one that shows the worst result in the difficulty parameter. Considering that the correlation means between the estimates and the actual difficulty parameters are high for the 3 methods, we conclude that the one with the best result in general in the mean squared error metric would be the VAE, as it has better errors both in the latent trait and in the discrimination parameter.

Scenario	$RMSE_a$ VAE	$RMSE_a$ AE	$RMSE_a$ JML	$RMSE_b$ VAE	$RMSE_b$ AE	$RMSE_b$ JML	$RMSE_\theta$ VAE	$RMSE_\theta$ AE	$RMSE_\theta$ JML
Sim1	0.331	0.346	16.476	1.098	1.063	1.142	1.015	1.365	0.866
Sim2	0.471	0.430	20.193	1.111	1.077	1.067	0.973	1.392	0.885
Sim3	0.766	0.710	26.766	1.199	1.174	1.199	0.949	1.734	0.886
Sim4	0.400	0.486	7.132	0.905	0.919	1.004	1.098	1.506	0.726
Sim5	0.199	0.369	7.945	0.933	0.937	0.960	0.932	1.392	0.739
Sim6	0.106	0.364	10.457	0.932	0.929	0.949	0.805	1.379	0.797

Table 3 – Mean of the root mean square error of the a, b and θ.

In Figure 8 and in Table 4, we have the mean of the bias measure for the three methods for the item and individual parameters. In general, we see that the VAE delivers better estimates, as it is the only one that keeps the bias value closer to zero for most of the cases. It is important to highly that Figures 6, 7 and 8 and Tables 2, 3 and 4 represent the mean of the indexes among each type of parameters.

Complementing the comparative study of the three methods, we have Figures 9, 10, and from figure 11 to figure and 16 that show the relationship between the actual parameters and their estimates obtained by the three methods for one of the simulations. In this case, we hope

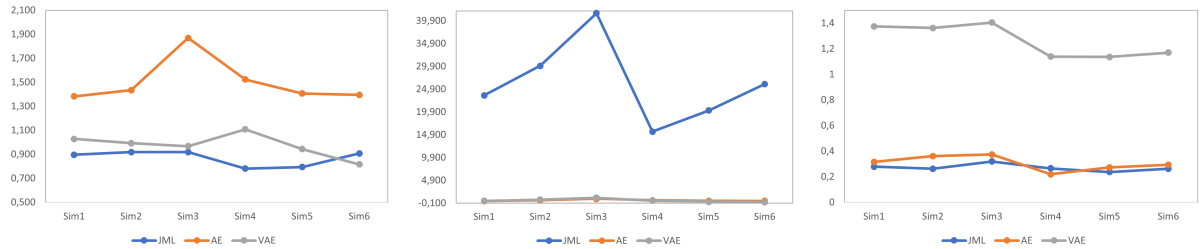

Figure 7 – Comparison of the mean square error regarding θ, a and b, respectively, for three methods in the six different scenarios.

Scenario	AVB_a VAE	AVB_a AE	AVB_a JML	AVB_b VAE	AVB_b AE	AVB_b JML	AVB_θ VAE	AVB_θ AE	AVB_θ JML
Sim1	0.123	0.159	12.627	1.092	1.028	1.122	0.192	0.369	0.339
Sim2	0.302	0.189	15.988	1.107	1.037	1.050	0.203	0.360	0.364
Sim3	0.533	0.356	21.852	1.196	1.146	1.186	0.191	0.479	0.362
Sim4	0.386	0.481	5.921	0.901	0.897	0.993	0.390	0.732	0.275
Sim5	0.189	0.365	7.629	0.931	0.923	0.958	0.256	0.638	0.288
Sim6	0.099	0.363	9.638	0.930	0.914	0.943	0.188	0.652	0.365

Table 4 – Mean of the AVB of the a, b and $theta$.

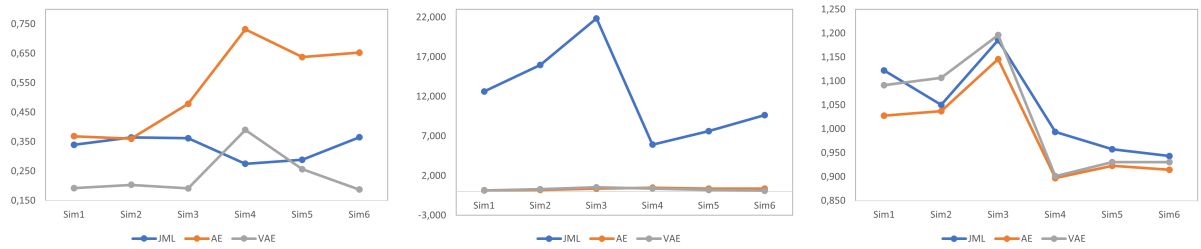

Figure 8 – Comparison of bias regarding θ, a, and b, respectively, for three methods in the six different scenarios.

that the points in the graphs arrive as close as possible to an identity line. We have such a graph for a, b and θ in each of the six scenarios, in each of the 50 replicates, and for each method, we chose only a few graphs to illustrate the behavior of the methods. Note that in all figures mentioned above, the VAE estimates appear first, then the AE, and finally the JML.

Although Figure 9 corresponds to Sim5, replica 1, it well represents the behavior of the three methods for estimating the discrimination parameters in each scenario and each replica. We can see that the VAE presents points closer to a line of identity, while the JML tends to provide estimates of the discrimination parameter that are larger than the actual value of the parameter. This shows the advantage of VAE in estimating these parameters.

On the other hand, and as previously mentioned, the three methods, in all scenarios and all replicates, are capable of estimating very well the b-difficulty parameters, or intercept of the model, as we can see in Figure 10.

In Figures 11, 12 and 13 we have from left to right the graphs related to VAE, AE, and

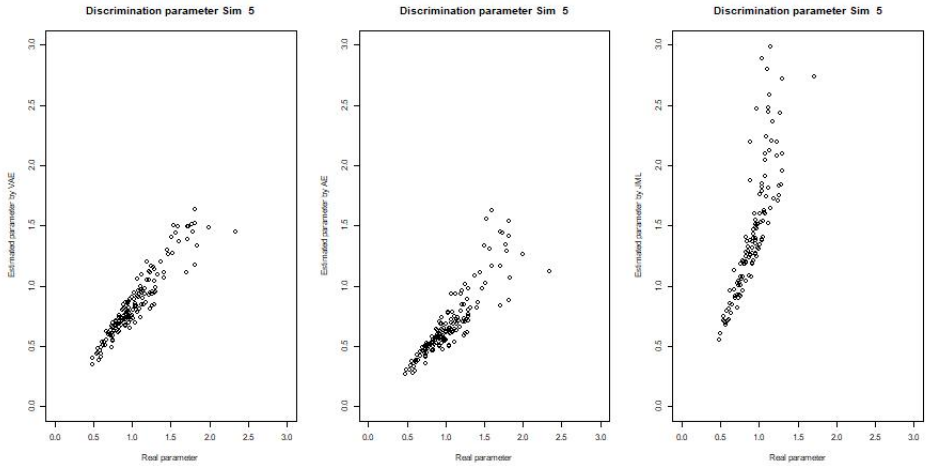

Figure 9 – Actual versus estimated discrimination parameter by VAE, AE, and JML methods, respectively, for the simulation with 160 items and 10000 individuals.

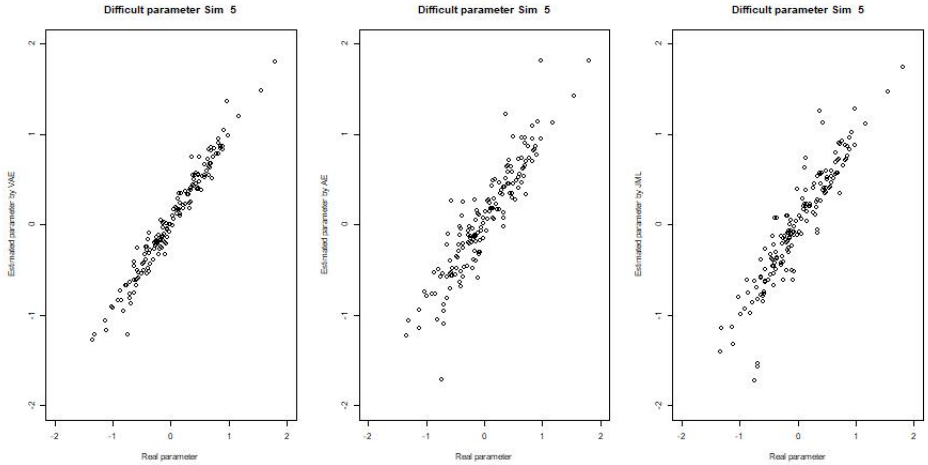

Figure 10 – Real versus estimated difficulty parameter by VAE, AE, and JML methods respectively, for the simulation with 160 items and 10000 individuals.

JML with the dispersion plot of estimate and actual latent trait values for Sim1, Sim2, and Sim3, respectively.

As we can see, the AE and JML are the ones that present more instability in the estimates because, in the different scenarios, Sim1, Sim2, and Sim3 the representations of the estimates seem to worsen when we increase the number of individuals, which does not happen with the VAE, which in the three scenarios has a similar behavior.

The instability in the estimates observed in scenarios Sim1, Sim2, and Sim3 remains in scenarios Sim4, Sim5, and Sim6. We can see that the VAE provides better estimates, as the points are more like an identity line, and are always in the range [-4.4], which is where the actual skill is found, as opposed to AE and JML, which apparently get worse as the amount of data

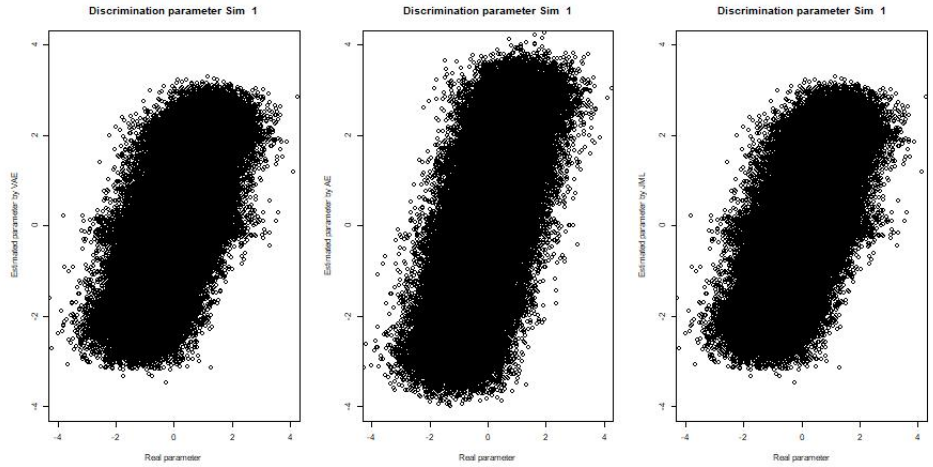

Figure 11 – Real versus estimated latent trait for the scenario Sim1

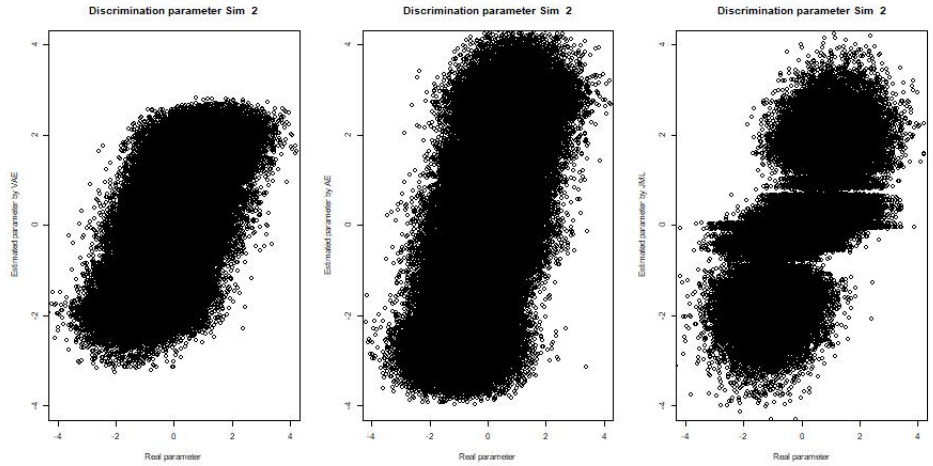

Figure 12 – Actual versus estimated latent trait for the scenario Sim2

involved increases.

One aspect draws attention in Figure 16 for the JML method in Sim6: some latent trait estimates are inversely correlated with their real value.

According to an in-depth analysis, we find that the inversion is only happening for the seventh dimension, θ_7, only for some replicates, only in Sim6, and only for the JML method, as we can see in the figure 17. It represents the dispersion plot between estimates and actual values for the latent traits of dimension 7, by the three methods, for replicates 1 and 3, respectively. This behavior indicates a JML identifiability issue.

According to our analysis of latent traits estimation by dimension, we found that dimensions that had more associated items obtained better estimates in all methods. As was the case of θ_4, θ_7, and θ_9, with 6, 7, and 7 items associated with the 80-question tests and 12, 14,

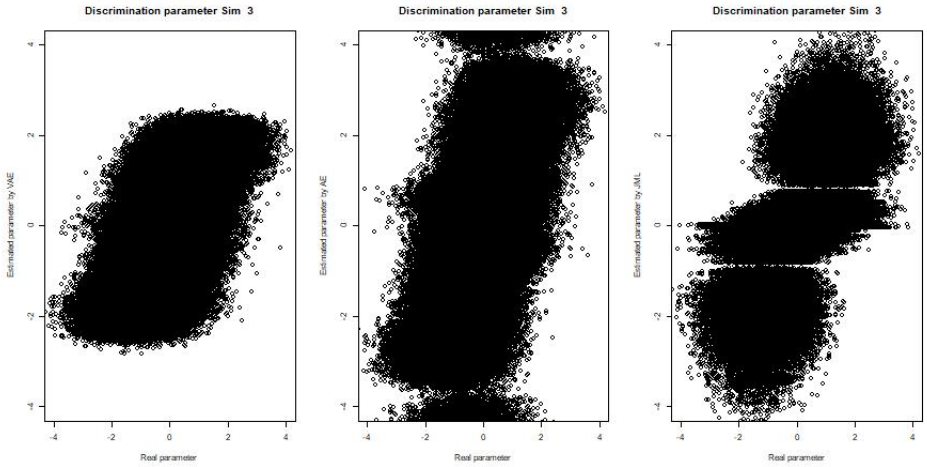

Figure 13 – Real versus estimated latent trait for the scenario Sim3

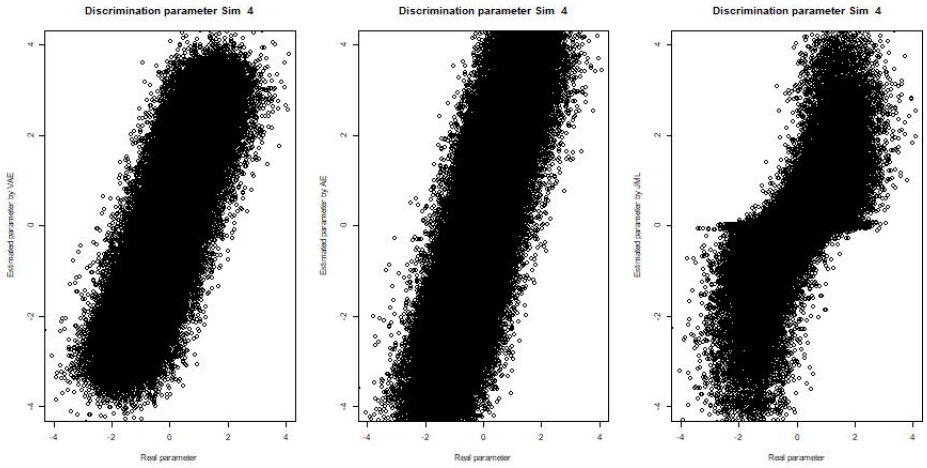

Figure 14 – Real versus estimated latent trait for the scenario Sim4

and 14 items in the 160-question tests, which reached correlations of 0.873. They had the best correlations. In contrast, the latent traits θ_2 that have 3 and 6 associated items in the tests of 80 and 160 questions respectively, had the worst correlations, around 0.474.

Another fact that caught our attention was that while the AE and VAE took between 0 and 2 minutes to estimate all parameters and latent traits in each scenario and each replica, the JML took from 10 minutes, for scenarios with 5 thousand individuals, to nearly 4 hours for scenarios with 20,000 subjects and 160 items. This shows an additional advantage for the neural network methods, as the three methods were run on conventional and equivalent notebooks.

In the next illustrations, we have a graphical representation of the estimated probability of correctly answering an item as a function of the estimated latent trait evaluated in that item, for each method, and we compare it with the same representation made from real data. We can

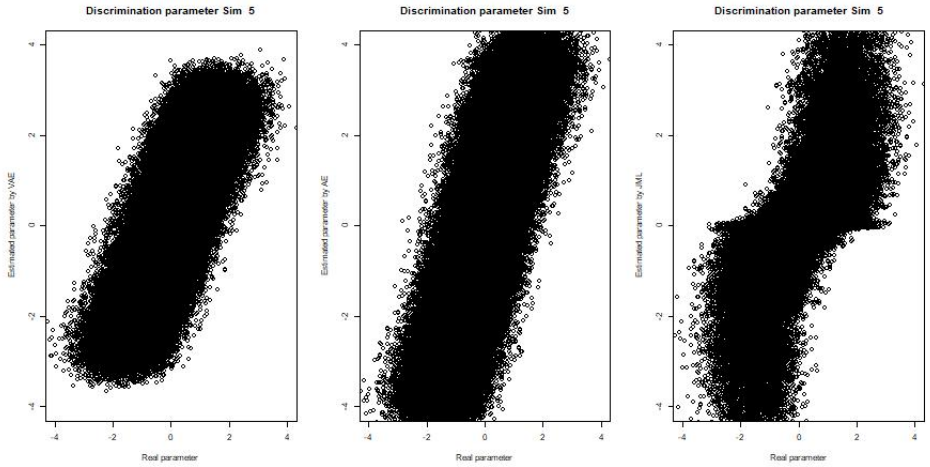

Figure 15 – Real versus estimated latent trait for the scenario Sim5

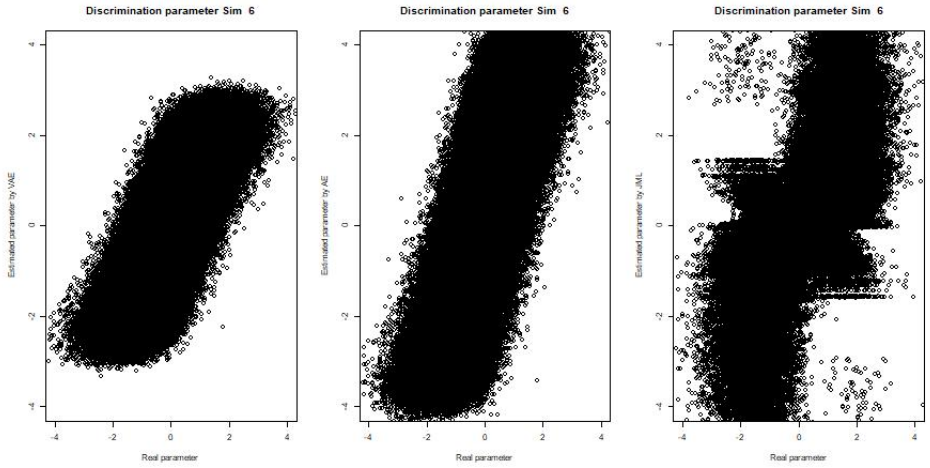

Figure 16 – Real versus estimated latent trait for the scenarios Sim6 respectively

have a clear view of which method is delivering better estimates when we find the curve of the method that comes closest to the real curve, which is represented in red color. We chose only a few items estimated in the first replica, and in the Sim6 scenario to illustrate this comparison.

In Figures 18, 19, and 20, the probability of correcting the item as a function of the real latent trait is represented in red. And as we can see, the VAE represented in blue color is the closest to the real graph, so it is the one that delivers the best estimates in the Sim6 scenario that involves 160 items, 18 latent traits, and 20000 individuals. The graphics for the other scenarios deliver similar results, which is why they were not included in this text.

Although we are quite satisfied with the VAE method and consider it a promising method to be used in Psychometrics to estimate high-dimensional latent traits, due to the quality of the estimates and the speed in achieving them, there is still much to improve and explore. We hope,

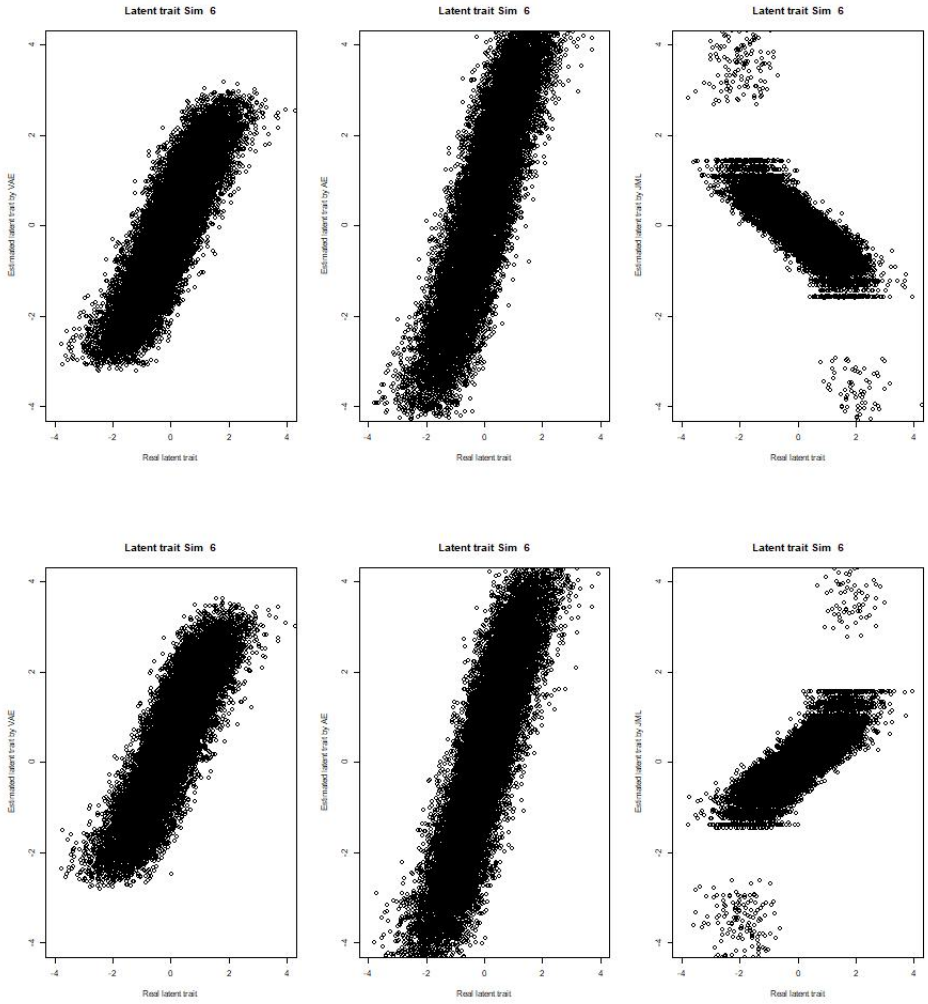

Figure 17 – Actual θ_7 versus θ_7 estimated for Sim6 in replica 1 (top) and replica 3 (bottom) for the three methods.

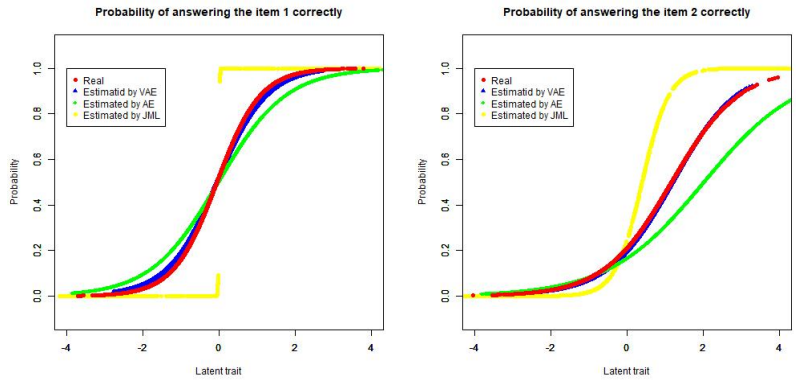

Figure 18 – From left to right: Probability of correctly answering item 1 as a function of the latent trait θ_7 and the probability of correctly answering item 2 as a function of the latent trait θ_8 in the scenario Sim6

for example, that some customizations in its architecture can solve the difficulties it presents in estimating the parameters at the extremes.

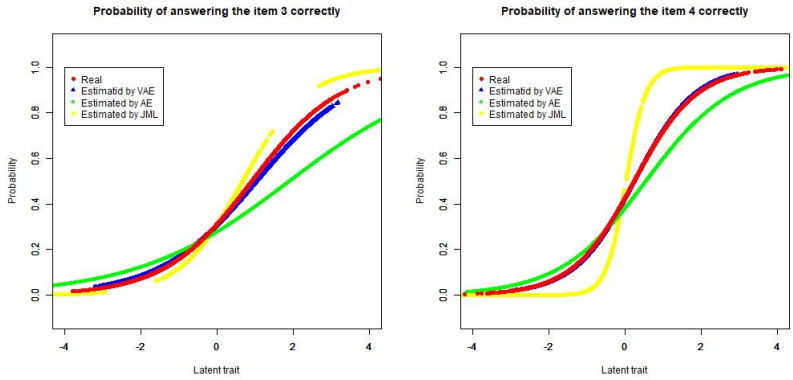

Figure 19 – From left to right: Probability of correctly answering item 3 as a function of the latent trait θ_7 and the probability of correctly answering item 4 as a function of the latent trait θ_{10} in the scenario Sim6

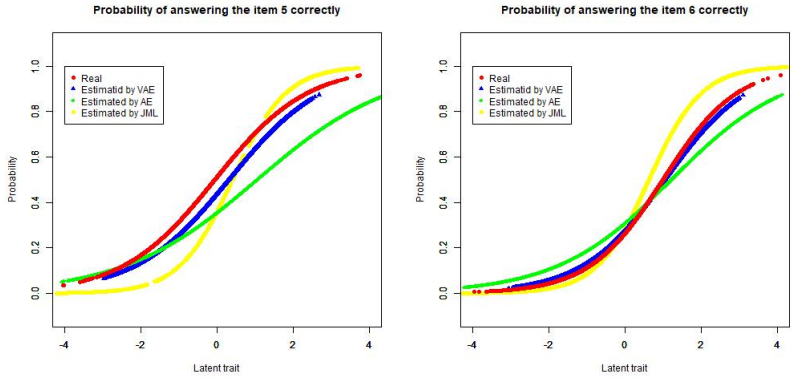

Figure 20 – From left to right: Probability of correctly answering item 5 as a function of the latent trait θ_{12} and the probability of correctly answering item 6 as a function of the latent trait θ_{16} in the scenario Sim6

4.2 IVAE vs MIRT

MIRT is a package in the R programming language to estimate multidimensional item response theory parameters for exploratory and confirmatory models using maximum-likelihood methods. It adopts MML based on the EM as the default method for estimating item parameters, and EAP for estimating a low-dimensional latent trait vector, up to 3 or 4 dimensions approximately.

Therefore, to enhance the study of the quality of the proposed model, we will use a database of a test that assesses three latent abilities. In Table 5, we define the simulations considered in this section to perform the comparisons among IVAE and MIRT estimation results.

To assemble the scenarios in Table 5, we used the database of a test with 28 items, which assess 3 skills, answered by 10,000 individuals, and we called this first scenario with the name Complete_Dim3. From it, we randomly delete four responses from each individual, which we call Incomplete_Dim3. Additionally, we randomly delete a maximum of four responses from each individual, which we call Semi complete_Dim3. Then we estimate the parameters for the

Scenario	Individuals	Skills	Answered items	Total items
Complete_Dim3	10000	3	28	28
Incomplete_Dim3	10000	3	24	28
Semi_complete_Dim3	10000	3	≥ 24	28

Table 5 – Scenarios simulated to compare IVAE with MIRT.

three scenarios by the traditional MIRT method, the proposed VAE, and IVAE method. Note that in the case where the database is complete, the VAE and IVAE coincide.

Scenario	$Corr_\theta$ MIRT	$Corr_\theta$ VAE	$Corr_\theta$ IVAE	$RMSE_\theta$ MIRT	$RMSE_\theta$ VAE	$RMSE_\theta$ IVAE
Incomplete_Dim3	0.8015	0.6238	0.7548	0.5993	0.8182	0.7017
Semi_Incomplete_Dim3	0.8145	0.6335	0.7873	0.5815	0.8080	0.6556
Complete_Dim3	0.8248	0.8179	0.8179	0.5667	0.6178	0.6178

Table 6 – Comparation of correlations and mean square error of estimated parâmeters by different methods

Table 6 and figure 21 present the correlations between the latent traits estimated by the three previously mentioned methods and the real latent traits. The proportion of missing data affects IVAE much more than MIRT. In the Incomplete_Dim3 scenario where we have approximately 14.3% missing data, the IVAE correlations for the latent traits dropped from 0.8179 to 0.7548, while the MIRT ones dropped from 0.8248 to 0.8015. But the correlations associated with the IVAE show how the adaptation made to the VAE improved the estimates since in the case of the VAE the drop was much more abrupt, changing from 0.8179 to 0.6238.

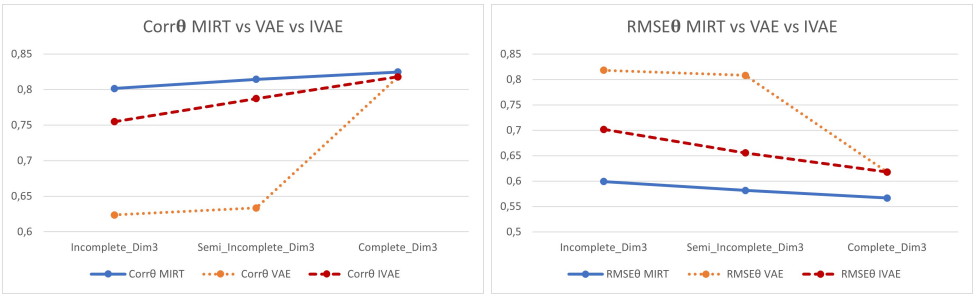

Figure 21 – Correlation and Mean Quadratic Error for the MIRT, VAE, and IVAE methods for the three dimension 3 scenarios of this section.

For better visualization of the results, we can see Figures 22, 23 and 24 which represent a comparison between the real latent traits and the actual parameter, and those estimated by MIRT and IVAE in the three scenarios presented in Table 5. Figure 22 shows how the two methods achieve better estimates when the database is complete, while the figures 23 and 24 show the same comparison for incomplete and semi-complete cases respectively. The orange dots correspond to the traditional method and the blue triangles correspond to the IVAE.

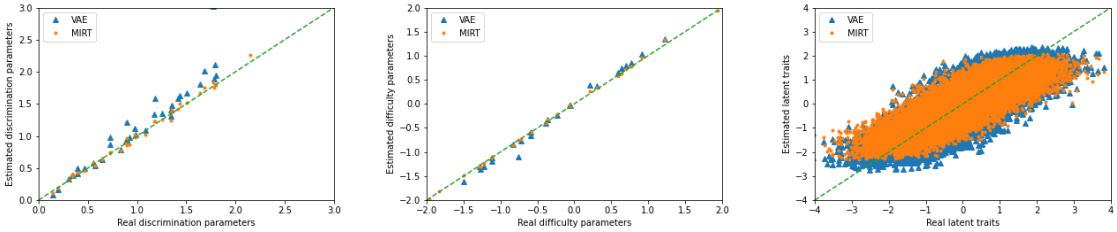

Figure 22 – MIRT vs IVAE for complete data

In Figure 23 where we have 14.3% of missing data, we can see how the orange dots representing MIRT are closer to the identity line than the blue triangles representing IVAE. Difference that is smaller in Figure 24 when missing data is at most 14.3%.

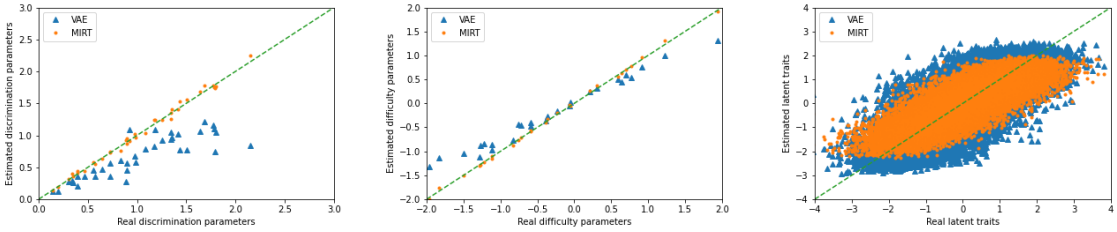

Figure 23 – MIRT vs IVAE for incomplete data.

As we can see, the IVAE method that uses data imputation and parameter estimation both through the VAE delivers better estimates than the VAE with modified loss without data imputation. It is not as good as the traditional MIRT method for low-dimensional latent traces, as was expected. But the advantage of this method is that it manages to work with large databases and with high-dimensional latent traces, which the traditional MIRT method cannot

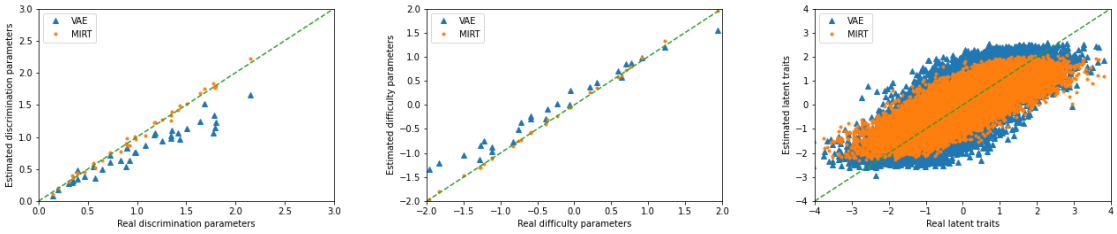

Figure 24 – MIRT vs IVAE for semi-incomplete data.

4.3 IVAE with iterations

Note that it is possible to replace the missing values from the input database each time we obtain new estimates for the latent traits. To compare whether this practice leads to better results for the estimates after imputing the data several times, we compared the correlations

and errors for different numbers of iterations of the IVAE. This test will be performed using the Semi_Incomplete_Dim3 and Incomplete_Dim3 scenarios.

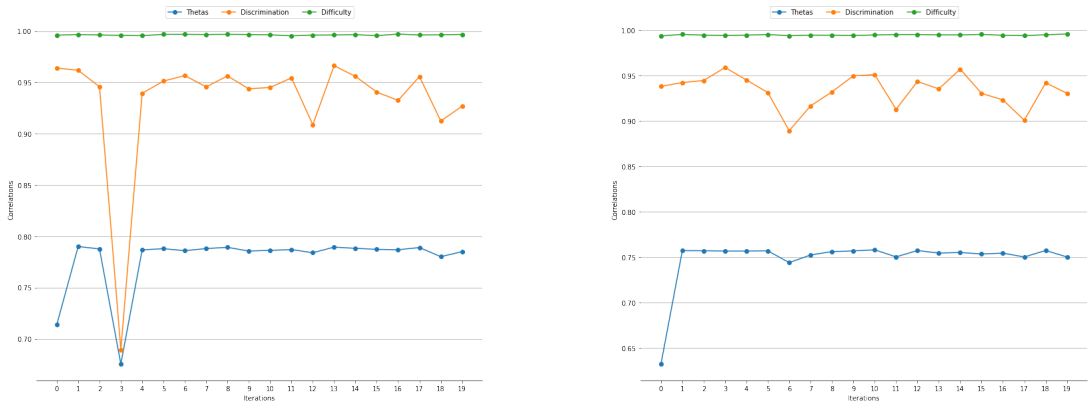

Figure 25 – Correlations for 20 iterations of the Semi_Incomplete_Dim3 and the Incomplete_Dim3 respectively.

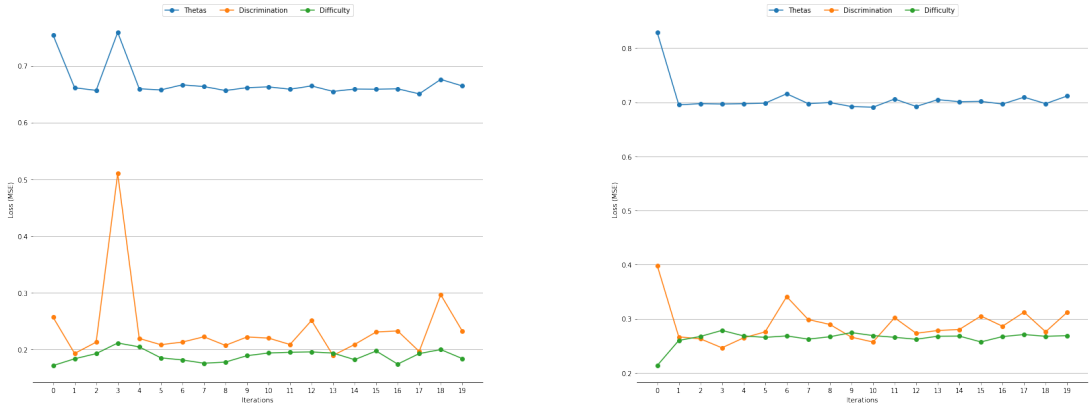

Figure 26 – Mean square error for 20 iterations of the Semi_Incomplete_Dim3 and the Incomplete_Dim3 respectively.

Please observe Figures 25 and 26. It is important to note that only the VAE's correlation and mean squared error are shown in iteration 0. The metrics related to the IVAE are displayed from the first iteration onward. Despite seeing a valley in the correlation graph in iteration 3 and a mountain in the mean squared error graph, which shows the IVAE worse than the VAE in this iteration, in general, we can see that the IVAE stabilizes and obtains better estimates than the VAE. As we see stability in the IVAE estimates with several iterations. These results ensure our choice of considering the estimates after a single iteration for the IVAE method.

4.4 IVAE for high latent trait dimension

In this section, we will show the behavior of IVAE involving higher dimensional latent features, where traditional methods are not able to produce any result.

To assemble the scenarios studied in this section, we considered the Sim1, Sim2, Sim4, and Sim5 from section 4.1 and randomly excluded 12.5% and 25% of the responses of each

individual. The new scenarios are named Sim1_Rde80, Sim2_Rde80 for R= 70,60 and Sim4_-Pde160 and Sim5_Pde160 for P=140, 120, as depicted in table 7.

Scenario	Individuals	Skills	Answered items	Total items
Sim1	5000	18	80	80
Sim1_70de80	5000	18	70	80
Sim1_60de80	5000	18	60	80
Sim2	10000	18	80	80
Sim2_70de80	10000	18	70	80
Sim2_60de80	10000	18	60	80
Sim4	5000	18	160	160
Sim4_140de160	5000	18	140	160
Sim4_120de160	5000	18	120	160
Sim5	10000	18	160	160
Sim5_140de160	10000	18	140	160
Sim5_120de160	10000	18	120	160

Table 7 – Scenarios simulated to compare IVAE with MIRT.

We simulated 10 replicates of responses for a test under the conditions of scenarios Sim1, Sim2, Sim4, and Sim5 and arbitrarily excluded 12.5% (25%) responses from each individual in the first replica, as it is natural that different individuals can have different items without responding. Then, to delete the responses in the other 9 replicates, we follow what was deleted in the first one, as we want to create 10 possible responses from the same individual for the same test, and thus we obtain 10 replicates for Sim1_70de80, Sim2_70de80, Sim4_140de160, and Sim5_140de160 respectively (Sim1_60de80, Sim2_60of80, Sim4_120of160, and Sim5_-120of160).

We estimated the parameters and latent traits using the IVAE and compared them with the JML, as a competitive method proposed in the literature for high dimension. In Tables 8, 9, and 10 and Figures 27, 28, and 29, we present the metrics related to said comparison. Remember that in the complete case, the IVAE coincides with the VAE.

In general, we can see that from the complete case to the incomplete one, in both methods, we have a small decrease in the correlation and an increase in the RMSE and AVB, which is natural since we have less information to estimate the same amount of parameters and latent traits. We also observed, once again, an advantage in IVAE with respect to JML, as expected since JML works really well in scenarios with the number of individuals tending to infinity.

In Tables 9 and 10, we observed that the RMSE and AVB of the a discrimination parameter are very high in the case of JML, which we can see as a huge disadvantage of this method compared to what we are proposing. In contrast, IVAE presents higher RMSE for θ, but the differences are much less expressive. These results stimulate further studies to continue enhancing VAE method proposals for psychometric assessment.

Scenario	$Corr_a$ IVAE	$Corr_a$ JML	$Corr_b$ IVAE	$Corr_b$ JML	$Corr_\theta$ IVAE	$Corr_\theta$ JML
Sim1	0.624	0.341	0.978	0.934	0.660	0.603
Sim1_70de80	0.325	0.399	0.984	0.917	0.556	0.517
Sim1_60de80	0.204	0.508	0.984	0.905	0.464	0.554
Sim2	0.555	0.348	0.975	0.924	0.654	0.587
Sim2_70de80	0.426	0.290	0.977	0.885	0.581	0.563
Sim2_60de80	0.292	0.427	0.975	0.893	0.450	0.540
Sim4	0.869	0.636	0.985	0.933	0.784	0.695
Sim4_140de160	0.840	0.654	0.987	0.938	0.741	0.688
Sim4_120de160	0.747	0.659	0.988	0.944	0.674	0.656
Sim5	0.923	0.684	0.978	0.926	0.782	0.687
Sim5_140de160	0.907	0.680	0.987	0.917	0.737	0.661
Sim5_120de160	0.858	0.690	0.989	0.927	0.667	0.631

Table 8 – Mean of the correlations of the a, b and θ.

Figure 27 – Comparison between the averages of the correlations for the discrimination and difficulty parameters and for the latent traits. For the IVAE and JML methods in the different scenarios with and without missing data

To better illustrate the comparison between IVAE and JML, we present some graphs of the results. As we did before, because there is not a big difference in the graphics between replicas and scenarios, only the graphics referring to replica 1 of scenarios Sim1_70de80 and Sim5_120de160 follow.

Scenario	$RMSE_a$ IVAE	$RMSE_a$ JML	$RMSE_b$ IVAE	$RMSE_b$ JML	$RMSE_\theta$ IVAE	$RMSE_\theta$ JML
Sim1	0.331	16.476	0.135	0.232	1.015	0.866
Sim1_70de80	0.488	17.602	0.146	0.263	1.115	1.223
Sim1_60de80	0.545	15.538	0.201	0.274	1.162	0.862
Sim2	0.430	20.193	1.077	1.067	1.392	0.885
Sim2_70de80	0.414	21.493	0.128	0.281	0.956	0.752
Sim2_60de80	0.480	18.836	0.182	0.259	1.059	0.770
Sim4	0.486	7.132	0.919	1.004	1.506	0.726
Sim4_140de160	0.390	7.218	0.110	0.161	0.883	0.605
Sim4_120de160	0.439	8.852	0.166	0.160	0.940	0.640
Sim5	0.369	7.945	0.937	0.960	1.392	0.739
Sim5_140de160	0.192	8.854	0.098	0.171	0.748	0.628
Sim5_120de160	0.254	10.649	0.155	0.163	0.819	0.663

Table 9 – Mean of the root mean square error of the a, b and θ.

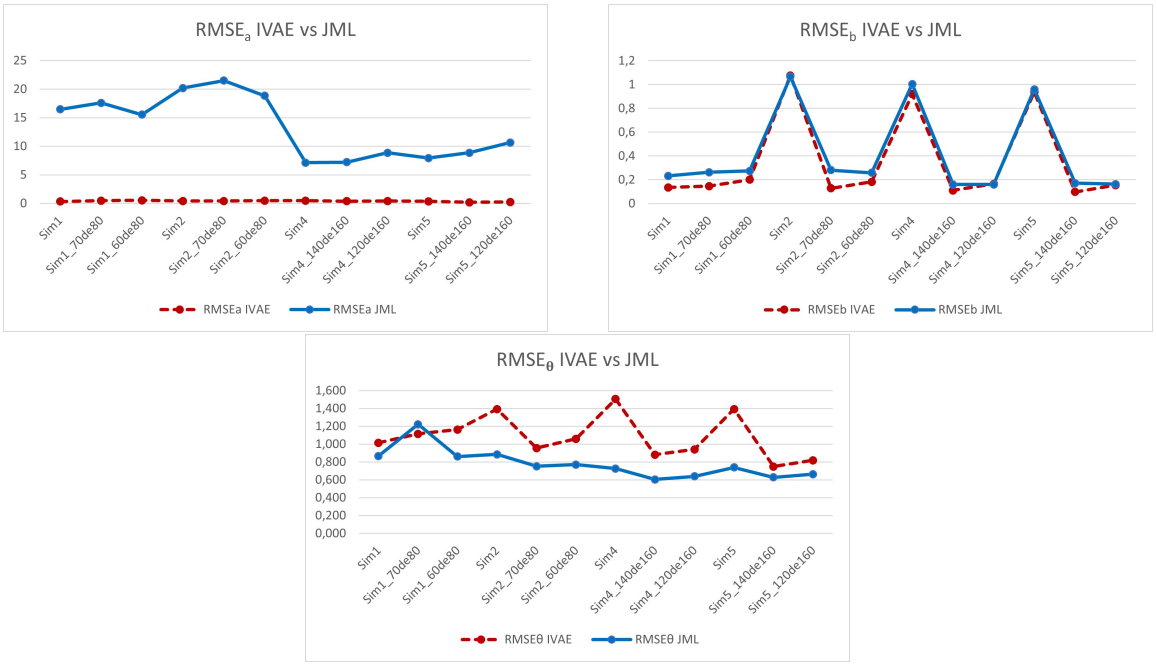

Figure 28 – Comparison between the averages of the RMSE for the discrimination and difficulty parameters and for the latent traits. For the IVAE and JML methods in the different scenarios with and without missing data

Figure 30 and 31, we can see the relationship between the discrimination parameters estimated by each method and the real values for scenarios Sim1_70de80 and Sim5_120de160 respectively. Here, as in the previous tables, we can see how the two methods improve the estimates when we increase the amount of data involved.

Scenario	AVB_a IVAE	AVB_a JML	AVB_b IVAE	AVB_b JML	AVB_θ IVAE	AVB_θ JML
Sim1	0.123	12.627	1.092	1.122	0.192	0.339
Sim1_70de80	0.291	13.209	0.124	0.197	0.520	0.960
Sim1_60de80	0.361	13.275	0.184	0.230	0.608	0.509
Sim2	0.189	15.988	1.037	1.050	0.360	0.364
Sim2_70de80	0.169	17.519	0.102	0.249	0.790	0.734
Sim2_60de80	0.189	17.101	0.160	0.243	0.845	0.757
Sim4	0.481	5.921	0.897	0.993	0.732	0.275
Sim4_140de160	0.375	6.996	0.102	0.157	0.786	0.605
Sim4_120de160	0.421	8.271	0.160	0.151	0.807	0.639
Sim5	0.365	7.629	0.923	0.958	0.256	0.288
Sim5_140de160	0.183	8.760	0.093	0.169	0.696	0.628
Sim5_120de160	0.234	10.421	0.153	0.159	0.745	0.662

Table 10 – Mean of bias of the a, b and θ.

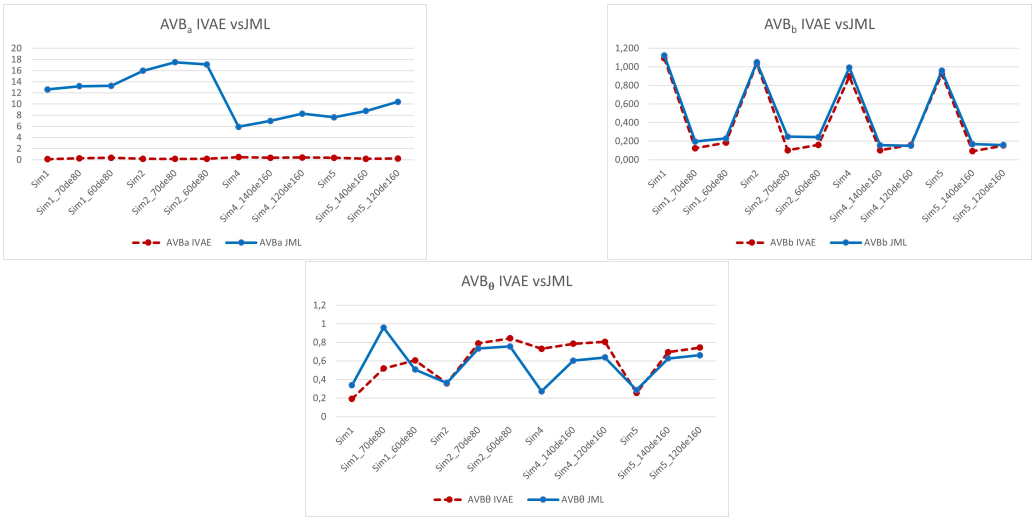

Figure 29 – Comparison between the averages of the AVB for the discrimination and difficulty parameters and for the latent traits. For the IVAE and JML methods in the different scenarios with and without missing data

Figure 32 and 33, represent the relationship between the difficulty parameters estimated by each method and the real values for the same scenarios. We can see how both methods deliver good estimates, as expected, due to the fact that b is the intercept of the model.

In the latent traits, in figures 34 and 35, we can see how the JML seems to be lost in some dimensions, which may be related to the problem of lack of identifiability that the method has, as mentioned in the previous sections. IVAE, on the other hand, manages to recover latent traits considerably better, turning it into a promising method for IRT models.

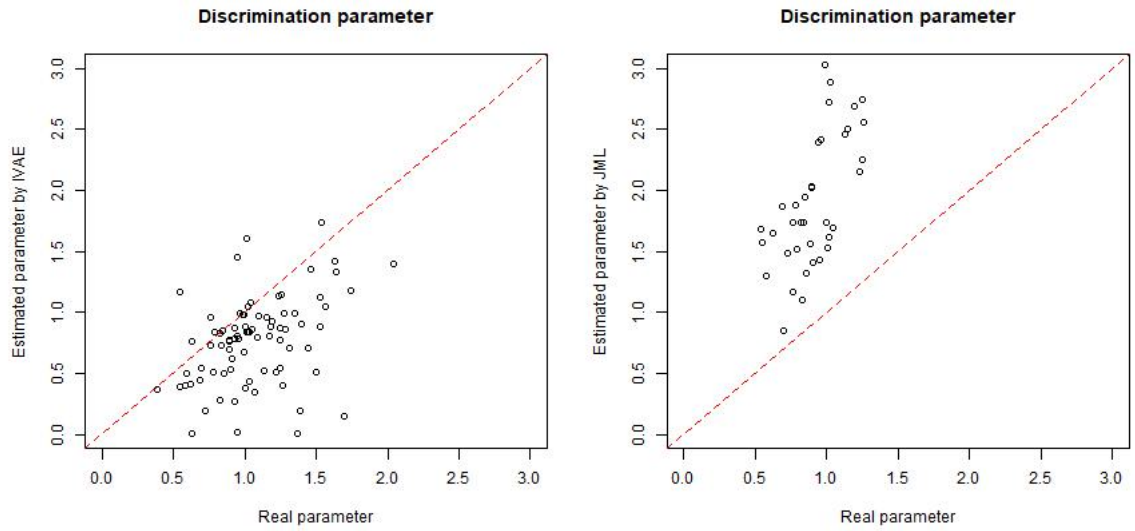

Figure 30 – Actual versus estimated discrimination parameter by IVAE and JML methods respectively.

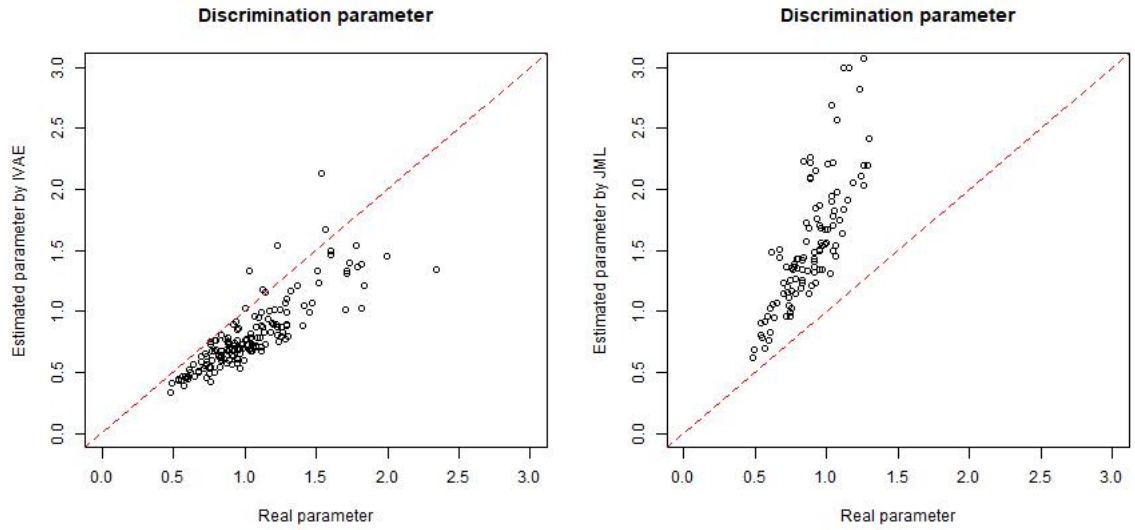

Figure 31 – Actual versus estimated discrimination parameter by IVAE and JML methods respectively.

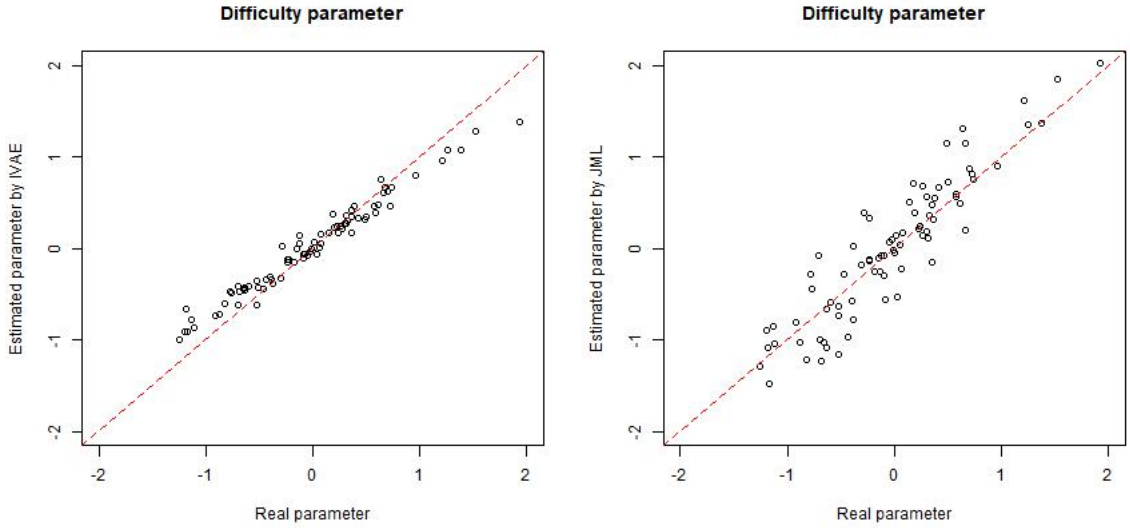

Figure 32 – Actual versus estimated difficulty parameter by IVAE and JML methods respectively.

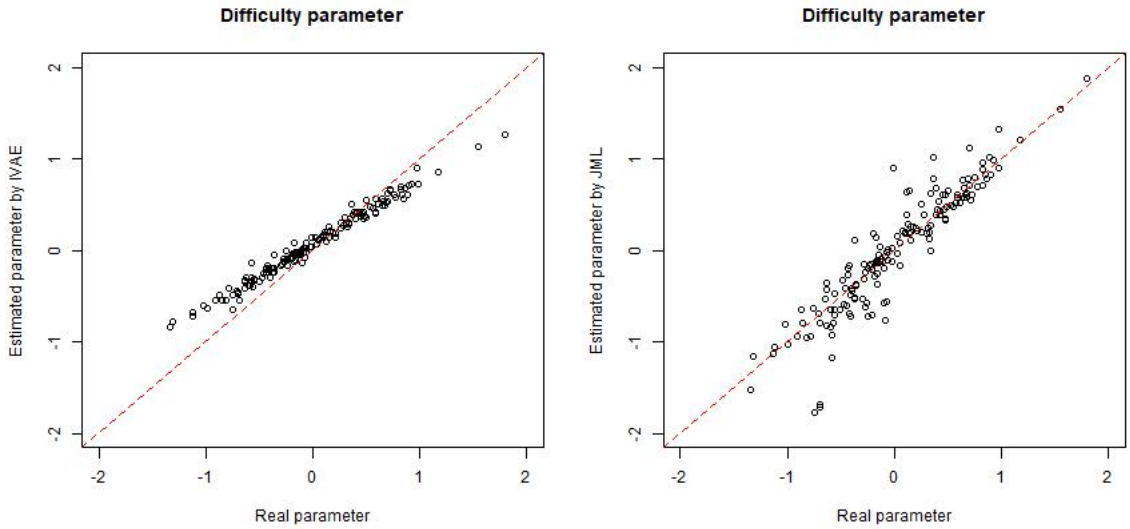

Figure 33 – Actual versus estimated difficulty parameter by IVAE and JML methods respectively.

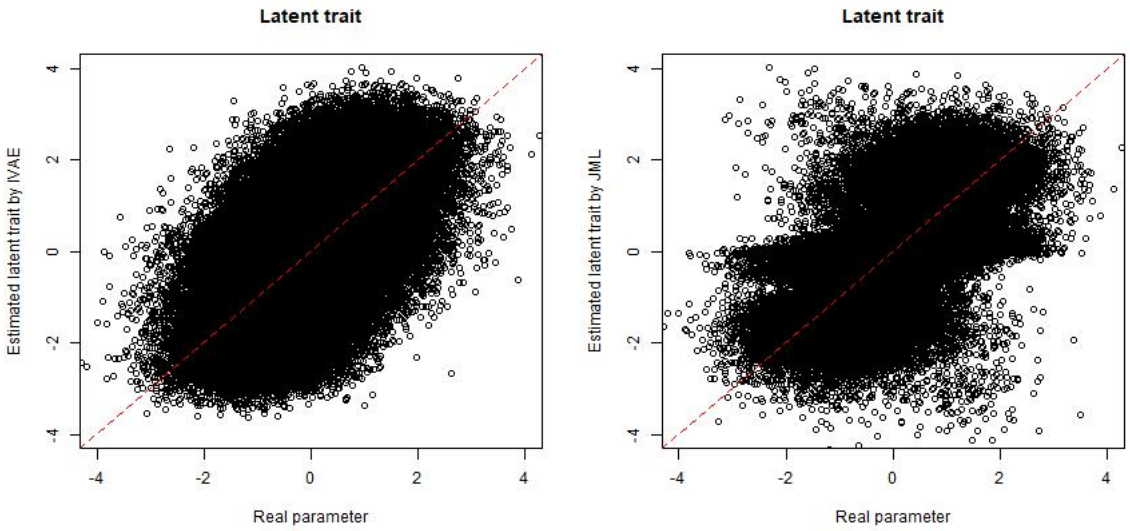

Figure 34 – Real versus estimated latent trait by IVAE and JML methods respectively.

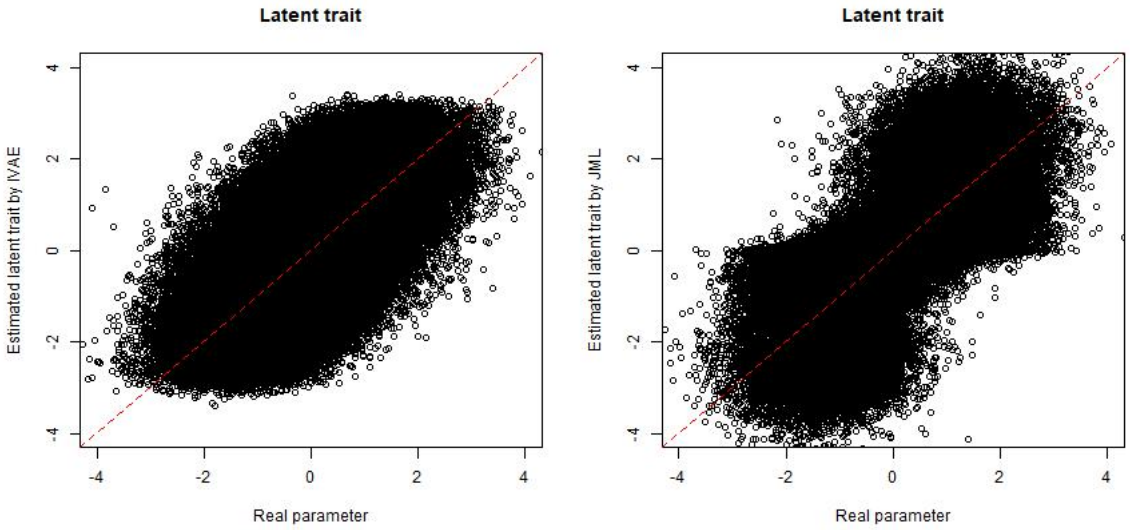

Figure 35 – Real versus estimated latent trait by IVAE and JML methods respectively.

CHAPTER

5

REAL APLICATION

In this chapter, we apply the proposed method to a real data Math assessment obtained from practice simulations for the ACT test.

ACT corresponds to a group of tests used as part of the admission process to American universities, similar to the Brazilian ENEM. We consider the set of the Math items, which have 60 items related to 22 latent traits, which can be regrouped into 4 more general constructs.

The methods used to estimate the parameters of this real application will be IVAE and JML to 22 dimensions of the latent feature, and IVAE, JML, and MIRT in the case of 4 dimensions.

5.1 Four dimensions

In this section, we will consider the responses of 4898 individuals to the ACT practice test in Mathematics. The test assesses the skills: Geometry and Measurement (GM), Number and Quantity (NQ), Operations, Algebra and Functions (OAF), and Statistics and Probability (SP). A group of items assesses each skill according to Table 11.

Latent trait	Number of Items
θ_1 (GM)	19
θ_2 (NQ)	4
θ_3 (OAF)	28
θ_4 (SP)	9

Table 11 – Number of items that evaluates each latent trait of the Mathematics ACT (dimension 4)

We estimate the skills and parameters of discrimination and the difficulty of the M2LP using three different methods. IVAE for being the proposal of this work, JML for being the method currently used in cases of estimating a high-dimensional M2LP, and MML and EAP via MIRT for being the most used method for estimating IRT in low-dimensional latent trait,

since we are working with dimension 4. This last method will be called MIRT (or M2PL) in the graphics because it is the name of the package in R where this method is implemented as a particular case (or because is the most used method in this case).

In parallel to the aforementioned study, we will randomly delete 10% and 20% from the subjects' responses to compare the methods in the presence of missing data.

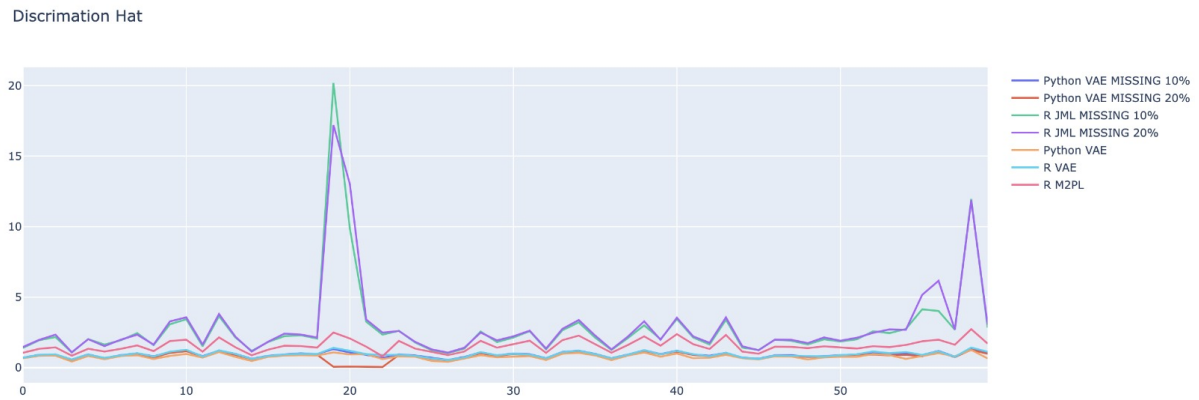

Figure 36 – Comparison of discrimination parameter estimates, for complete data, with 10% and 20% of missing data. Methods used IVAE, JML, and MIRT.

In Figures 36 and 37, we have the position of the estimated parameter on the x-axis. In this case, we have 60 discrimination and difficulty parameters to estimate, so we organize them in positions 1 to 60 and leave the respective estimated value for each of them on the y-axis, with different colors for each method.

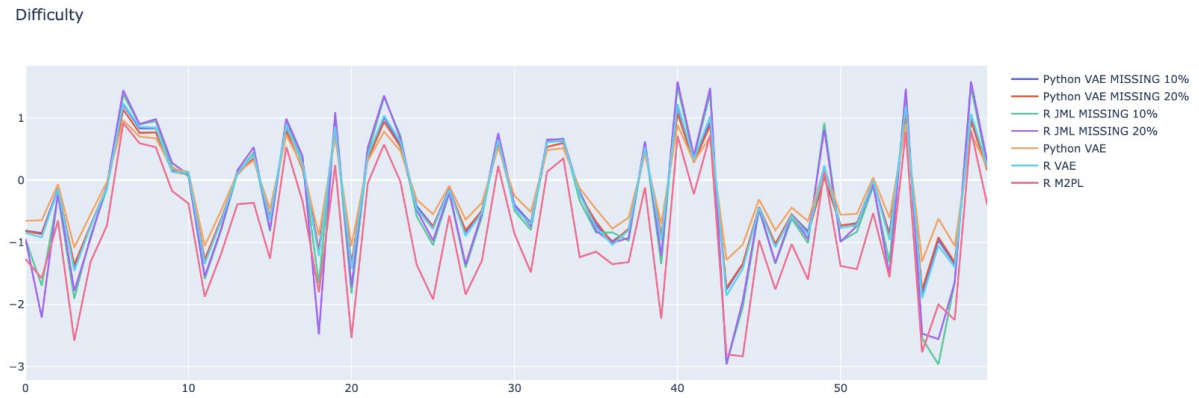

Figure 37 – Comparison of difficulty parameter estimates, for complete data, with 10% and 20% of missing data. Methods used IVAE, JML, and MIRT.

In the legend of the graphs of the figures 36 and 37, we can read VAE (understood IVAE in the presence of missing data). On the other hand, note that the R VAE and Python VAE are the same methods, implemented in two different programming languages.

The dark pink line in the middle in Figure 36 and at the bottom in Figure 37 corresponds to the estimates for the discrimination and difficulty parameters obtained through the MML (the

most used model for estimating M2LP), so in this section, it will be our reference to be reached by the other methods.

We can see that the estimates of both discrimination and difficulty of the two methods (IVAE and JML) are proportional to the estimates given by the MML. We can see a relationship between them despite being on different scales.

In the following figures 38, 39, and 40 we see the comparison between the average estimates of two methods for the four latent traits (θ_1, θ_2, θ_3, and θ_4). In these, we can see that the quality of the IVAE estimate in relation to the MML is comparable with that of the JML in relation to the same MML.

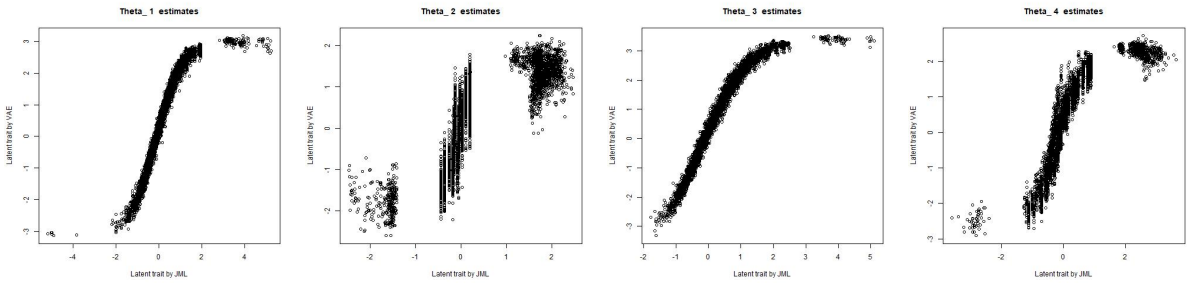

Figure 38 – Comparison between latent traits estimated by JML and IVAE

We can see, for example, that both IVAE and JML methods presented worse estimates, in relation to MML, for θ_2, which according to Table 11 corresponds to the latent trait that evaluates fewer items, namely 4, which, as we have seen through the simulation study and as has been commented throughout this work, is not a scenario that favors the quality of the estimates of either method.

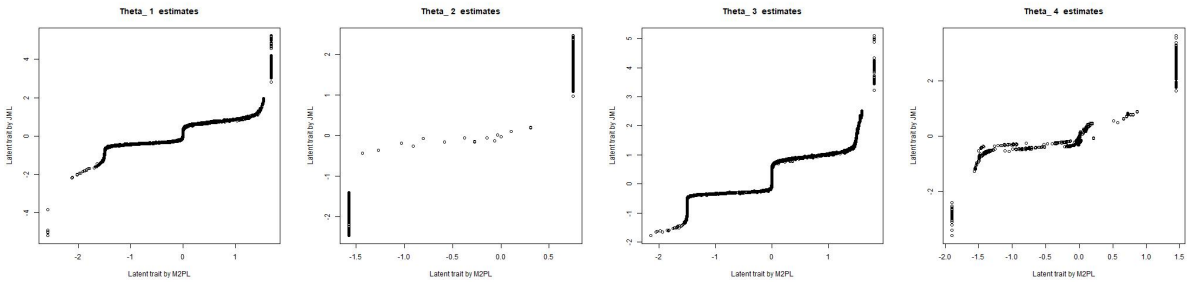

Figure 39 – Comparison between latent traits estimated by MML and JML

On the other hand, θ_1 and θ_3 seem to present better estimates, if we consider the relationship of the methods with the MML, these latent traits are precisely those that involve more items, namely 19 and 28 respectively.

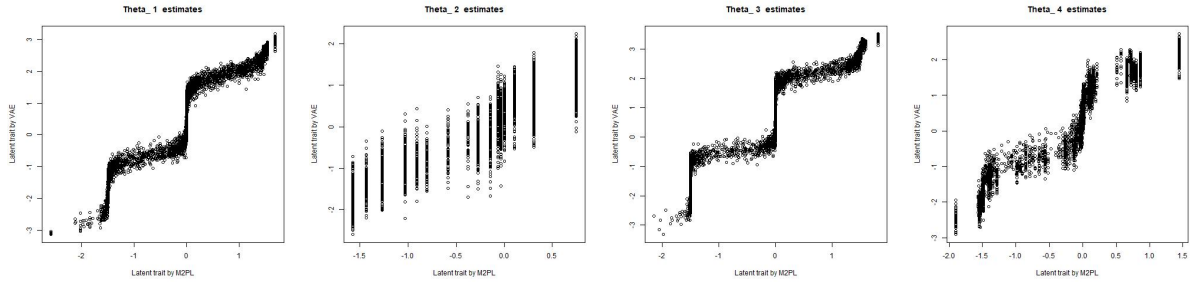

Figure 40 – Comparison between latent traits estimated by MML and IVAE

5.2 Twenty-two dimensions

Analogously to what was done in the previous section, in this section, we will consider the same responses of the 4898 individuals to the ACT in mathematics, but now looking at this test from a more specific point of view, which allows us to organize it as assessing 22 skills, according to Table 12.

Latent trait	Items Number	Latent trait	Items Number	Latent trait	Items Number
θ_1	3	θ_9	1	θ_{17}	1
θ_2	6	θ_{10}	10	θ_{18}	1
θ_3	7	θ_{11}	1	θ_{19}	1
θ_4	1	θ_{12}	4	θ_{20}	3
θ_5	2	θ_{13}	1	θ_{21}	2
θ_6	1	θ_{14}	1	θ_{22}	4
θ_7	2	θ_{15}	1		
θ_8	1	θ_{16}	6		

Table 12 – Number of items that evaluates each latent trait of the Mathematics ACT (dimension 22)

We estimated M2LP skills and discrimination and difficulty parameters using two different methods. IVAE for being the proposal of this work and JML for being the method currently used in cases of estimating a high-dimensional M2LP In parallel with the aforementioned study, we will randomly exclude 10% and 20% from the subjects' responses to compare the methods in the presence of missing data.

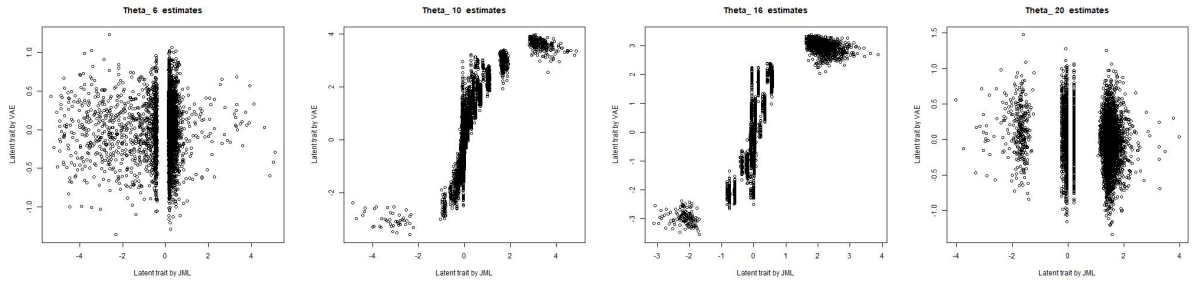

Figure 41 – Comparison between latent traits estimated by JML and IVAE to complete data

We chose 4 out of 22 latent traits to show the comparison between JML and IVAE methods. These are the θ_6, θ_{10}, θ_{16} and θ_{20} arranged in the same order in the figures 41, 42 and 43 . They are representative of the number of related items.

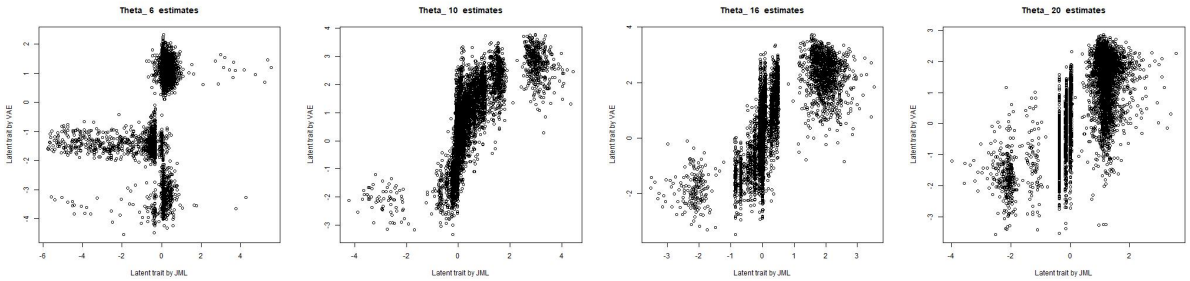

Figure 42 – Comparison between latent traits estimated by JML and IVAE with 10% missing data

As we can see in the figures 41, 42 and 43, the IVAE distributes the latent trait estimates more than the JML. We can also observe that the relationship between the two methods for the latent features θ_{10} and θ_{16} is more defined than in the cases of θ_6 and θ_{20}, this is due to the number of items that evaluate each chosen latent trait, as we can see in table 12, the latent traits θ_{10} and θ_{16} have more items evaluating them, 10 and 6 respectively, while θ_6 and θ_{20} are evaluating only 1 and 3 items respectively.

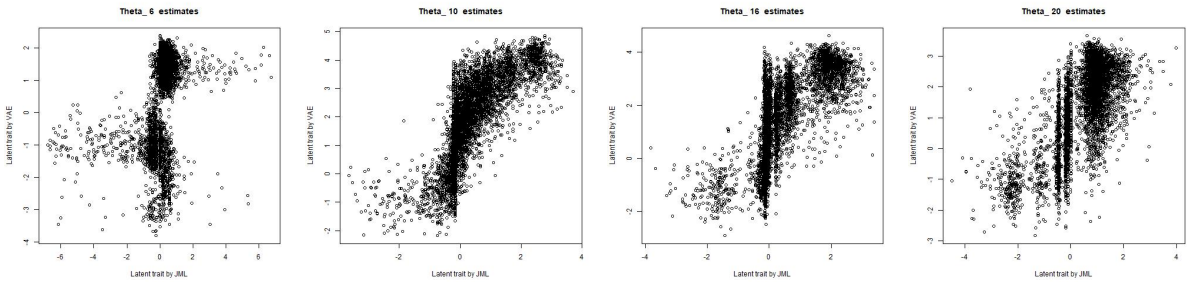

Figure 43 – Comparison between latent traits estimated by JML and IVAE with 20% missing data

Another way to understand the estimates given by both methods, JML and IVAE, is with the estimated probability of correcting an item, we can calculate this probability through the estimated parameters of the items and the estimated probability of correcting each individual.

We can observe the characteristic format of the probability of correcting an item of an IRT model for both methods. We see a better distribution achieved by the IVAE estimate, whereas, in the case of the JML, we observe some gaps, theta values that would not have been obtained by any individual, which does not mean that there is an error in the method, considering that this probability corresponds to the estimated one from a sample of just over 4500 individuals. But yes, we can conclude that the LAVI can be compared to methods currently accepted by the literature for estimating parameters and latent traits jointly for high-dimensional M2LP.

To see all graphs related to this simulation study, see 9.

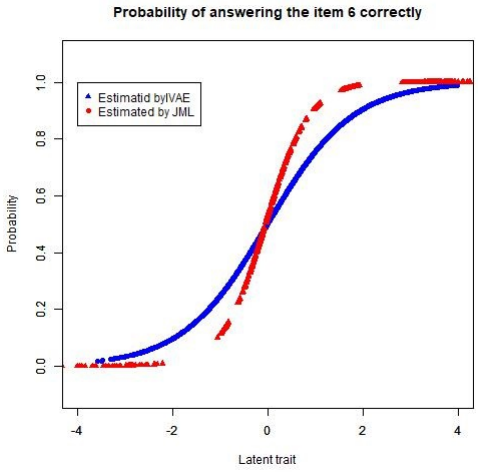

Figure 44 – Estimated probability of getting the item 6 right as a function of θ_{10}

CONCLUSION

After comparing two DL methods capable of dealing with large databases, we made a more detailed study of the VAE method that presented the best simulation results. We propose a VAE method of estimation for psychometric models capable of handling high-dimensional estimation both for complete data and in the presence of missing data. Also, we clarify a connection between the VAE and the EM algorithm, the most used classical method for estimating IRT parameters. We modify the loss function that the SGD minimizes to get the VAE estimation, to deal with missing data as a first approach. Enhancing it, we propose an extension, the IVAE, as an imputation method from the VAE to improve the quality of the estimates in the presence of missing data.

BIBLIOGRAPHY

BOQUET, G.; MORELL, A.; SERRANO, J.; VICARIO, J. L. A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection. **Transportation Research Part C: Emerging Technologies.**, 2020. Citation on page 25.

BOTTOU, L. Large-scale machine learning with stochastic gradient descent. **Proc. 19th Int. Conf. Comput. Statist.**, 2010. Citation on page 35.

CAI, L. High-dimensional exploratory item factor analysis by a metropolis–hastings robbins–monro algorithm. **PSYCHOMETRIKA**, 2017. Citation on page 23.

CARDOSO, R.; CRISTO, J.; PEREIRA, J.; PEREIRA, P.; HENRIQUES, P. Missing image data imputation using variational autoencoders with weighted loss. **European Symposium on Artificial Neural Networks**, 2020. Citation on page 25.

CHEN, Y.; LI, X.; ZHANG, S. Joint maximum likelihood estimation for high-dimensional exploratory item response analysis. **PSYCHOMETRIKA**, 2019. Citations on pages 24, 32, and 42.

CONVERSE, G.; CURI, M.; OLIVEIRA, S. Autoencoders for educational assessment. **20th Annual Conference on Artificial Intelligence in Education (AIED)**, 2019. Citations on pages 24 and 32.

CURI, M.; CONVERSE, G.; HAJEWSKI, J.; OLIVEIRA, S. Interpretable variational autoencoders for cognitive models. **International Joint Conference on Neural Networks (IJCNN)**, 2019. Citations on pages 24, 25, 32, and 33.

DEMPSTER, A.; N., L.; RUBIN, D. Maximum likelihood from incomplete data via the em algorithm. **Journal of the Royal Statistical Society: Series B.**, 1977. Citation on page 34.

FREIRE, G. and Escobar, C. Github, 2023. Available: <https://github.com/ClaudiaEscobarM0210/VAETeseDoutorado>. Accessed: 2023, May. Citations on pages 39 and 65.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep learning. **MIT Press**, 2017. Citation on page 27.

GRADUATE Record Examination (GRE). Educational Testing Service (ETS), 2022. Available: <https://www.ets.org/gre>. Accessed: 2022, March. Citation on page 30.

GUO, Q.; CUTUMISU, M.; CUI, Y. A neural network approach to estimate studentskill mastery in cognitive diagnostic assessments. **10th International Conference on Educational Data Mining**, 2017. Citation on page 32.

KINGMA D., W. M. Auto-encoding variational bayes. international conferenceon learning representations. **ICLR**, 2014. Citation on page 29.

MARIS, G.; BECHGER, T. Boltzmann machines as multidimensional item response theory models. **Journal of Machine Learning Research 1**, 2000. Citation on page 24.

NELWAMONDO, F.; MOHAMED, S.; MARWALA, T. Missing data: A comparison of neural network and expectation maximization techniques. **Current Science**, 2007. Citation on page 25.

PROGRAMME for International Student Assessment (PISA). Organisation for Economic Co-operation and Development (OECD), 2022. Available: <https://www.oecd.org/pisa/>. Accessed: 2022, March. Citation on page 30.

RECKASE, M. Multidimensional item response theory. **Springer**, 2009. Citation on page 23.

TAKAHASHI, M. Statistical inference in missing data by mcmc and non-mcmc multiple imputation algorithms: Assessing the effects of between-imputation iterations. **Data Science Journal**, 2017. Citation on page 25.

URBAN, C.; BAUER, D. A deep learning algorithm for high-dimensional exploratory item factor analysis. **Psychometrika**, 2021. Citation on page 24.

WU, M.; DAVIS, R.; DOMINGUE, B.; PIECH, C.; GOODMAN, N. Variational item response theory: Fast, accurate, and expressive. **Proceedings of the 13th International Conference on Educational Data Mining**, 2020. Citation on page 24.

YOON, J.; JORDON, J.; SCHAAR, M. Gain: Missing data imputation using generative adversarial nets. **Proceedings of the 35 th International Conference on Machine Learning**, 2018. Citation on page 25.


ICMC
USP
SÃO CARLOS