ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202009.00058

Variational Inference Applied to Educational Measurement Models

Authors: Dai Liyang, Dai Liyang

Date: 2020-12-26T00:00:00+00:00

Abstract

Variational inference is commonly employed in machine learning, constituting a parameter estimation algorithm that can be implemented with minimal code while achieving relatively fast computational speeds. This study demonstrates the application of black-box variational inference and amortized variational inference to educational measurement models, developing latent variable networks capable of generating arbitrary variance-covariance matrices, neural correlation matrices that can serve as prior distributions, attribute mastery pattern networks based on sigmoid or softmax functions, among other innovations. Experimental results indicate that the parameter estimation performance of variational inference on item response theory and certain DINA models has reached a state-of-the-art level. The research demonstrates that variational inference can substantially facilitate researchers in developing novel educational measurement models and is also relatively suitable for general users in practical application scenarios.

Full Text

Variational Inference for Educational Measurement Models

Survey and Data Center, East China Normal University

Abstract

Variational inference, commonly used in machine learning, is a parameter estimation algorithm that achieves fast computation with minimal code implementation. This study demonstrates the application of black-box variational inference and amortized variational inference to educational measurement models, developing latent variable networks capable of generating arbitrary variance-covariance matrices, neural correlation matrices that can serve as prior distributions, and attribute mastery pattern networks based on sigmoid or softmax

functions. Experimental results show that variational inference achieves state-of-the-art performance in parameter estimation for item response theory and certain DINA models. The study indicates that variational inference can greatly assist researchers in developing new educational measurement models and is also suitable for practical applications by general users.

Keywords: variational inference; item response theory; cognitive diagnosis model; neural network

1 Introduction

The research and application of educational measurement models depend heavily on the development of parameter estimation algorithms. Item Response Theory (IRT) and Cognitive Diagnosis Models (CDM) are the mainstream educational measurement models today, with commonly used parameter estimation methods including MCMC algorithms and EM algorithms. MCMC is a widely applied parameter estimation algorithm for IRT and CDM [**?**, **?**, **?**, **?**]. The advantage of MCMC is that software such as STAN [**?**, **?**] and JAGS [**?**, **?**] are available for researchers, who only need to write minimal code to apply MCMC for parameter estimation. Consequently, researchers frequently use MCMC to develop models with numerous parameters and complex structures; for instance, the HO-DINA model was initially estimated using MCMC [**?**, **?**]. The disadvantage of MCMC is its high computational time and memory consumption, requiring researchers to have considerable patience and computing power when developing new models. From a user's perspective, these drawbacks make MCMC difficult to apply in educational measurement production environments.

The EM algorithm is another commonly used parameter estimation algorithm for IRT and CDM [**?**, **?**, **?**, **?**, **?**], which has achieved excellent results in application software development [**?**, **?**, **?**, **?**]. A well-known EM algorithm is BAEM developed by R. D. Bock et al. (1982). The EM family also includes adaptive quadrature algorithms [**?**, **?**], Laplace approximation algorithms [**?**, **?**], and Monte Carlo EM (MCEM) algorithms [**?**, **?**], which were proposed to handle high-dimensional latent variable IRT models. However, Cai (2010) pointed out that these algorithms remain unsuitable for high-dimensional IRT models, leading to the development of MHRM, which can be viewed as a stochastic EM algorithm [**?**, **?**]. Although stochastic optimization-based EM algorithms have been applied to educational measurement models [**?**, **?**], these studies did not employ mini-batch optimization and variance reduction techniques, making parameter estimation with EM algorithms difficult for large samples. Finally, EM algorithms require certain mathematical and coding proficiency from researchers, making it challenging to apply EM algorithms to develop models with many parameters and complex structures. A typical case is the HO-DINA model mentioned earlier, for which EM algorithm implementation only appeared years later [**?**, **?**].

With the development of artificial intelligence, computer researchers have devel-

oped variational inference algorithms to solve parameter estimation for Bayesian models with large samples and many parameters, introducing these algorithms to statisticians [?, ?], who have begun using variational inference [?, ?]. Researchers in educational measurement have also suggested using variational inference for parameter estimation [?, ?]. With the advancement of probabilistic programming software, the barrier to entry for variational inference has continuously lowered, to the point where researchers can complete parameter estimation programs with minimal code, similar to MCMC. This surpasses EM algorithms in terms of ease of use, while variational inference's computation time is less than MCMC algorithms [?, ?], making variational inference potentially very suitable for parameter estimation in educational measurement models.

Currently, research on applying variational inference to educational measurement models is limited [?, ?]. Most studies involving variational inference in educational measurement models focus on coordinate ascent variational inference [?, ?, ?, ?, ?, ?], with relatively few studies on black-box variational inference and amortized variational inference, which are based on machine learning concepts and can be implemented using probabilistic programming software. Existing research is limited to IRT parameter estimation [?, ?, ?, ?, ?], while CDM remains completely unexplored (Minka (2009) once demonstrated DINA model parameter estimation under the name of variational inference, but actually used expectation propagation). Coordinate ascent variational inference [?, ?] requires researchers to manually derive analytical expressions for parameter distribution expectations, which presents a certain technical barrier and is only suitable for models where such analytical expressions can be derived. This algorithm has poor generality and usability. Black-box variational inference and amortized variational inference can be coded based on probabilistic programming software (Pyro [?, ?], Edward [?, ?], and PyMC3 [?, ?], etc.), offering higher usability. The main issues with black-box variational inference and amortized variational inference in educational measurement research are: first, the selection of loss function gradients. The gradient calculations in [?, ?] and [?, ?] are based on the reparameterization trick [?, ?], but whether CDM can apply the reparameterization trick for parameter estimation remains to be studied. Additionally, there is currently a lack of research applying REINFORCE gradients. Second, the studies by [?, ?] and [?, ?] assume that the posterior variance-covariance matrix of IRT latent variables is diagonal, which contains significantly less information than an arbitrary variance-covariance matrix. Third, [?, ?] and [?, ?] did not conduct parameter recovery experiments, leaving the issue of IRT parameter recovery to be investigated.

In summary, variational inference is a highly promising parameter estimation method for educational measurement models. This study further demonstrates the application of black-box variational inference and amortized variational inference to IRT and fills the application gap for these methods in CDM.

2.1 Loss Function

The principle of variational inference is to approximate complex true distributions with simple distributions. The loss function of variational inference is called ELBO (Evidence Lower Bound), with the specific formula:

$$\text{ELBO} \equiv \mathbb{E}_{q_\phi(\mathbf{z})}[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z})]$$

This formula derives from the KL divergence between the simple computable distribution and the complex true distribution:

$$\text{KL}(q_\phi(\mathbf{z}) \| p_\theta(\mathbf{z}|\mathbf{x})) = \log p_\theta(\mathbf{x}) - \text{ELBO}$$

In the above formulas, ϕ represents variational parameters. For IRT, $\phi = \{\mu, \sigma\}$, where μ is the location parameter of the normal distribution and σ is the scale parameter. For CDM, $\phi = \{p\}$, where p can be the parameter of a Bernoulli distribution or a categorical distribution.

2.2 Mean Field

Variational inference uses simple approximate distributions to fit true distributions, typically expressed in mean field form. Assuming parameters are ϕ_i with distributions $q_i(\phi_i)$, the mean field is:

$$q(\phi) = \prod_i q_i(\phi_i)$$

The advantages of mean field are: first, it simplifies the form of the approximate distribution, and second, it enables loss backpropagation using dynamic computation graphs [**?**, **?**].

2.3 Coordinate Ascent Variational Inference

[**?**, **?**] and [**?**, **?**] applied coordinate ascent variational inference to estimate parameters for 2-parameter probit models, while [**?**, **?**, **?**] used coordinate ascent variational inference for DINA and MC-DINA model parameter estimation. Coordinate ascent variational inference is similar to Gibbs sampling, continuously computing parameter expectations with the simplified form:

$$\log(\phi_i(\theta_i)) \propto \mathbb{E}_{\phi_i \notin \phi}(\log(f(y|\phi)) + \log(p(\phi)))$$

The disadvantage of coordinate ascent variational inference is its difficulty in generalizing to multi-parameter IRT and IRT based on logit functions (strictly speaking, it can be handled approximately, but this is quite complex [**?**, **?**]).

2.4 Black Box Variational Inference

Black box variational inference is an algorithm proposed by [**?**, **?**]. English Liulishuo [**?**, **?**] and [**?**, **?**] used probabilistic programming software Edward and Pyro to implement black box variational inference parameter estimation for unidimensional IRT. The parameter estimation process for black box variational inference involves: first, computing the gradient with respect to variational parameters ϕ:

$$\nabla_{\phi}\text{ELBO} = \mathbb{E}_{q_{\phi}(\mathbf{z})}\{[\nabla_{\phi}\log q_{\phi}(\mathbf{z})][\log \mathbf{p}(\mathbf{x},\mathbf{z}) - \log \mathbf{q}_{\phi}(\mathbf{z})]\}$$

Sampling z_s from distribution $q_{\phi}(z)$, where $s = \{1, 2, 3, \ldots, S\}$, yields the approximate gradient of variational parameters:

$$\nabla_{\phi}\text{ELBO} = \sum\{[\nabla_{\phi}\log q_{\phi}(z_s)][\log p(\mathbf{x},\mathbf{z}_s) - \log \mathbf{q}_{\phi}(\mathbf{z}_s)]\}$$

Then update the variational parameters:

$$\phi = \phi + \rho\sum\{[\nabla_{\phi}\log q_{\phi}(z_s)][\log p(\mathbf{x},\mathbf{z}_s) - \log \mathbf{q}_{\phi}(\mathbf{z}_s)]\}$$

where ρ is the Robbins-Monro coefficient, primarily used in stochastic optimization. Repeat the above steps until variational parameters ϕ converge.

2.5 Amortized Variational Inference

The disadvantage of black box variational inference is that the number of latent variable parameters explodes with large samples, and when encountering new samples, black box variational inference must relearn from scratch. Therefore, researchers proposed amortized variational inference [**?**, **?**]. A typical example of amortized variational inference is the variational autoencoder [**?**, **?**], which uses neural networks as generating functions for distribution parameters, leveraging the universal approximation property of neural networks [**?**, **?**]. Using amortized variational inference, the approximate distribution $q_{\phi_i}(z_i)$ is rewritten as $q_{f(x_i)}(z_i)$. Amortized variational inference is commonly applied to artificial intelligence tasks such as image generation.

2.6 Reparameterization

Educational measurement models contain both item parameters and latent variable parameters, so computing gradients encounters the following expression:

$$\nabla_{\theta,\phi}\text{ELBO} = \nabla_{\theta,\phi}\mathbb{E}_{q_{\phi}(\mathbf{z})}[\log p_{\theta}(\mathbf{x},\mathbf{z}) - \log q_{\phi}(\mathbf{z})]$$

Directly computing this gradient is difficult, requiring tricks such as reparameterization. For normal distributions, the reparameterization method assumes

$\mathbf{z} \sim \mathcal{N}(\mu, \sigma)$. Let $\epsilon \sim \mathcal{N}(0, 1)$, then $\mathbf{z} = \epsilon * \sigma + \mu$, and the gradient of ELBO becomes:

$$\nabla_{\theta,\mu,\sigma}\text{ELBO} = \mathbb{E}_{q(\epsilon)}[\nabla_\theta \log p_\theta(\mathbf{x}, \mathbf{z}) - \nabla_{\mu,\sigma} \log q_{\mu,\sigma}(\epsilon * \mu + \sigma)]$$

This reparameterization method can be applied to IRT, and the VIBO algorithm [**?**, **?**] is based on this method. For discrete latent variables, there exists an algorithm called Gumbel-Softmax [**?**, **?**]. The Gumbel-Softmax process assumes a two-dimensional vector $\mathbf{v}$, samples G_1, G_2 from the standard Gumbel distribution, adds them to obtain a new vector $\mathbf{v}' = [v_1 + G_1, v_2 + G_2]$, and computes the final category probabilities through the softmax function, i.e., $\sigma_\tau(v_i)$, where τ is the temperature parameter. CDM can theoretically apply the HARD mode of this method (Straight-Through Gumbel-Softmax).

2.7 REINFORCE

REINFORCE is another method for computing gradients. The mathematical form of the REINFORCE gradient is:

$$\nabla_{\theta,\phi}\text{ELBO} = \mathbb{E}_{q_\phi(\mathbf{z})}\{\nabla_\phi \log q_\phi(\mathbf{z})[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z})] + \nabla_\theta \log p_\theta(\mathbf{x}, \mathbf{z}) - \nabla_\phi \log q_\phi(\mathbf{z})\}$$

This method can be used for both continuous and discrete latent variable gradient computation. It is commonly used for reinforcement learning in artificial intelligence [**?**, **?**]. The disadvantage is the large variance of stochastic gradients, though variance reduction techniques can be applied to mitigate this issue.

2.8 Stochastic Optimization and Variance Reduction

Variational inference is often applied to large-scale datasets, leading researchers to develop stochastic optimization and variance reduction techniques [**?**, **?**, **?**]. Stochastic optimization involves sampling mini-batches from the data to compute stochastic gradients each iteration. To reduce the variance of stochastic gradients, researchers have developed variance reduction methods including Rao-Blackwellization and Control Variates [**?**, **?**].

3.1 Models

In the model formulas, subscript i represents the sample index, subscript j represents the item (question) index, and y_{ij} is the input data (response data).

The selected models for experiments are 2-4 parameter IRT models. The mathematical form of IRT models is:

$$P(y_{ij}|X_i) = c_j + \frac{d_j - c_j}{1 + \exp(X_i \mathbf{a}_j + b_j)}, \quad 0 < c_j < d_j < 1$$

where X_i is the latent variable.

The selected models for experiments are the DINA model and HO-DINA model. The mathematical form of the DINA model is:

$$P(y_{ij}|\alpha_i) = g_j^{1-\eta_{ij}}(1-s_j)^{\eta_{ij}}, \quad 0 < g_j < 1, 0 < s_j < 1, 0 < g_j + s_j < 1$$

or Torre's (2011) reparameterized version:

$$P(y_{ij}|\alpha_i) = f_j + d_j\eta_{ij}, \quad 0 < f_j < 1, 0 < f_j + d_j < 1$$

where $\eta_{ij} = \prod_k \alpha_{ik}^{q_{jk}}$, $\alpha_i = [\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK}]$.

The mathematical form of the HO-DINA model is:

$$P(\alpha_{ik}|\theta_i) = \frac{\exp(\theta_i\lambda_{1k} + \lambda_{0k})}{1 + \exp(\theta_i\lambda_{1k} + \lambda_{0k})}, \quad \alpha_{ik} \sim \text{Bernoulli}(P(a_{ik}|\theta_i))$$

where α_i is the discrete latent variable (attribute mastery pattern) and q_{jk} are elements of the Q-matrix.

The experimental code is written based on Pyro and PyTorch, with loss function gradients based on REINFORCE. Reparameterization gradients (based on the Gumbel-Softmax method) are only experimented with in CDM, and variance reduction is based on the Rao-Blackwellization method. By default, latent variables are treated as random parameters and item parameters as deterministic parameters, with only one sample drawn when performing Monte Carlo sampling of latent variables.

3.2.1 Black Box Variational Inference

For unidimensional IRT models, the prior distribution of latent variables is $x_i \sim \mathcal{N}(0, 1)$, and the posterior distribution is $x_i \sim \mathcal{N}(\mu_i, \sigma_i)$. This method is consistent with [**?**, **?**].

For multidimensional IRT models, the prior distribution of latent variables is $\mathbf{X}_i \sim \mathcal{N}(0, \mathbf{I})$. The posterior distribution is $\mathbf{X}_i \sim \mathcal{N}(\mu_i, \Sigma_i)$, where Σ_i is an arbitrary variance-covariance matrix.

For DINA models, we first reference Culpepper's (2015) Gibbs algorithm implementation, where $\alpha_i = [\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{im}]$, with $\alpha_{i0} = [0, 0, \ldots, 0]$, $\alpha_{i1} = [1, 0, \ldots, 0]$, etc. If the dimension of the attribute mastery pattern is K, the α_i list contains 2^K elements. We randomly sample α_{ih} from α_i, where $\alpha_{ih} \sim$ categorical($[p_{i0}, p_{i1}, \ldots, p_{im}]$). The prior distribution of α_{ih} is $\alpha_{ih} \sim$ categorical($[1, 1, \ldots, 1]$). This setting is only used in low-dimensional attribute mastery pattern experiments. Second, the prior distribution of attribute mastery patterns is $\alpha_{ik} \sim$ Bernoulli(0.5), and the posterior distribution

is $\alpha_{ik} \sim \text{Bernoulli}(p_{ki})$. This setting is only used in high-dimensional attribute mastery pattern experiments.

For HO-DINA models, we similarly reference Culpepper's (2015) Gibbs algorithm implementation. The prior distribution of higher-order traits is $\theta_i \sim \mathcal{N}(0,1)$, and the posterior distribution is $\theta_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Let $\alpha_{ih} = [\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik}, \ldots]$, then $p_{ih} = \prod_k P(\alpha_{ik}|\theta_i)$. We then randomly sample attribute mastery patterns α_{ih} from the α_i list, where $\alpha_{ih} \sim \text{categorical}([p_{i0}, p_{i1}, \ldots, p_{im}])$.

3.2.2 Amortized Variational Inference

For unidimensional IRT models, consistent with the approach in [**?**, **?**], the prior distribution is standard normal, and the posterior distribution parameter generator model is:

$$\begin{aligned} \mathbf{h}_i &= g(\mathbf{y}_i) \\ \mu_i &= \mathbf{w}_\mu \mathbf{h}_i + b_\mu \\ \log \sigma_i^2 &= \mathbf{W}_\sigma \mathbf{h}_i + \mathbf{b}_\sigma \end{aligned}$$

where g is the activation function.

[Figure 1: see original paper]

For multidimensional IRT models, the posterior distribution location parameter generator model is:

$$\begin{aligned} \mathbf{h}_i &= g(\mathbf{y}_i) \\ \mu_i &= \mathbf{w}_\mu \mathbf{h}_i + \mathbf{b}_\mu \end{aligned}$$

The posterior distribution variance-covariance matrix parameter generator is:

$$\begin{aligned} \text{tril}(\mathbf{L}_i^*) &= \mathbf{W}_L \mathbf{h}_i + \mathbf{b}_L \\ \mathbf{L}_i &= \text{tril\_}(\mathbf{L}_i^*) + \exp(\text{diag}(\mathbf{L}_i^*)) \\ \Sigma_i &= \mathbf{L}_i \mathbf{L}_i^T \end{aligned}$$

where tril represents taking the lower triangular elements of a matrix and tril_ represents taking the lower triangular elements excluding the diagonal.

Another approach shares the variance-covariance matrix across latent variables, with the generator form:

$$\begin{aligned} \text{tril}(\mathbf{L}^*) &= \sum_i \text{tril}(\mathbf{L}_i) \\ \mathbf{L} &= \text{tril\_}(\mathbf{L}^*) + \exp(\text{diag}(\mathbf{L}^*)) \\ \Sigma &= \mathbf{L}\mathbf{L}^T \end{aligned}$$

where N is the mini-batch size.

For prior distribution settings, in addition to $\mathbf{X}_i \sim \mathcal{N}(0, \mathbf{I})$, amortized variational inference also introduces $\mathbf{X}_i \sim \mathcal{N}(0, \Omega)$ or $\mathbf{X}_i \sim \mathcal{N}(0, \Omega_i)$, where Ω is a correlation matrix. The computation and constraints of Ω follow the scheme in Stan software [**?**, **?**]. When Ω is computed using neural networks, we refer to it as a neural correlation matrix.

[Figure 2: see original paper]

This form is inspired by the fact that the prior distribution of discrete variational autoencoders can be restricted Boltzmann machines. $\mathbf{X}_i \sim \mathcal{N}(0, \mathbf{I})$ is equivalent to L_2 regularization, so the prior distribution constrains dimensionless latent variable parameters [**?**, **?**]. The diagonal of the correlation matrix ensures the scale constraint of latent variable parameters, while the free estimation of correlation coefficients allows for the incorporation of additional information.

[Figure 3: see original paper]

[Figure 4: see original paper]

For DINA models, the posterior distribution generator referencing Culpepper (2015) is:

$$\begin{aligned} \mathbf{h}_i &= g(\mathbf{y}_i) \\ \mathbf{p}_{im} &= \text{softmax}(\mathbf{h}_i) \end{aligned}$$

The posterior distribution generator consistent with the second form of black box variational inference is:

$$\begin{aligned} \mathbf{h}_i &= g(\mathbf{y}_i) \\ p_{ik} &= \text{sigmoid}(\mathbf{h}_i) \end{aligned}$$

The prior distribution is consistent with black box variational inference.

For HO-DINA models, the posterior distribution generator and prior distribution are consistent with the unidimensional item response model.

All missing data in the experiments are treated as ignorable missing data. Samples with missing data are not removed unless all data for that sample are missing. When missing data are fed into neural network calculations, they are assigned a value of -1.

The experiments adopt the multidimensional IRT model identification method consistent with [**?**, **?**], which is also used by flexmirt and the R package mirt.

3.2.5 Differences from VIBO

VIBO [**?**, **?**] is another variational inference algorithm, and existing educational measurement model parameter estimation algorithms based on black box variational inference and amortized variational inference can be viewed as VIBO. Regarding loss gradients, VIBO is based on reparameterization, while our experiments are based on REINFORCE (except for CDM where Gumbel-Softmax reparameterization is experimented with). VIBO cannot be applied to CDM, but our experiments can. For multidimensional IRT latent variable posterior distributions, VIBO's variance-covariance matrix is diagonal, while our experiments use arbitrary matrices. For multidimensional IRT latent variable prior distributions, VIBO's variance-covariance matrix is an identity matrix, while our experiments use both identity matrices and neural correlation matrices. Regarding missing data, VIBO is relatively complex, while our implementation is simpler.

3.3.1 LSAT

LSAT is response data published by [**?**, **?**] for testing IRT models, sourced from the Law School Admission Test of the Law School Admission Council. LSAT contains 1,000 samples and 5 items.

3.3.2 PISA

PISA is a study conducted by OECD to evaluate reading, mathematics, and science abilities of 15-year-old students. The experiments selected the PISA science test data cleaned by [**?**, **?**], with dichotomous scoring, 519,334 samples, and 183 items. Among these, 73,283 samples had completely empty data and were removed, leaving 446,051 samples containing 69,014,909 missing data points, totaling approximately 85% missing data with only 15% valid data.

3.3.3 ECPE

ECPE stands for Examination for the Certificate of Proficiency in English, containing 2,922 samples, 28 items, and 3 attributes, which has been used in studies such as [**?**, **?**].

3.4 Simulated Data

Unless otherwise specified, the number of items for IRT models is set to 50, and for CDM models to 100. Each experiment simulates 30 replications.

3.4.1 Item Response Theory

$a_{kj} \sim \text{uniform}(0.5, 3)$, $b_j \sim \mathcal{N}(0, 1)$, $c_j \sim \text{uniform}(0.05, 0.2)$, $d_j \sim \text{uniform}(0.8, 0.95)$, where a_{kj} is the slope on dimension k, and $\mathbf{X}_i \sim \mathcal{N}(0, \mathbf{I})$ or

$\mathbf{X}_i \sim \mathcal{N}(0, \begin{bmatrix} 1 & \text{corr} \\ \text{corr} & 1 \end{bmatrix})$, with corr taking values 0.3, 0.5, and 0.7.

3.4.2 Cognitive Diagnosis Models

First scheme: $g_j \sim \text{uniform}(0, 0.3)$, $s_j \sim \text{uniform}(0, 0.3)$, $q_{kj} \sim \text{Bernoulli}(0.5)$, $\alpha_{ik} \sim \text{Bernoulli}(0.5)$, $\lambda_{i0} \sim \mathcal{N}(0, 1)$, $\lambda_{i1} \sim \text{uniform}(0.5, 3)$, with attribute mastery pattern dimension set to 5.

Second scheme: Referencing the simulation design of [**?**, **?**], $g_j \sim \text{uniform}(0, 0.3)$, $s_j \sim \text{uniform}(0, 0.3)$, with the Q-matrix set as $Q = [\mathbf{I}_K \quad \mathbf{Q}_1 \quad \mathbf{Q}_2]$. $\mathbf{I}_K$ is a K-dimensional identity matrix. $\mathbf{Q}_1 \in \{0, 1\}^{K \times K}$ has elements equal to 1 at positions (i, i) for $i = 1, 2, \ldots, K$ and at positions $(i, i+1)$ for $i = 1, 2, \ldots, K-1$, with all other elements being 0. $\mathbf{Q}_2 \in \{0, 1\}^{K \times K}$ has elements equal to 1 at positions (i, i) for $i = 1, 2, \ldots, K$, at positions $(i, i+1)$ for $i = 1, 2, \ldots, K-1$, and at positions $(i, i-1)$ for $i = 2, 3, \ldots, K$, with all other elements being 0. The attribute mastery pattern is set as $\theta_i = [\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}]$, with $\theta_i \sim \mathcal{N}(0, \Omega)$, where Ω is a matrix with diagonal elements equal to 1 and all other elements equal to ρ, with experiments setting $\rho = 0.3$. This setting is only used for high-dimensional attribute mastery pattern experiments.

3.4.3 Missing Data Simulation

The missing data simulation sets 90% of data as missing.

3.5 Comparison Software and Algorithms

The main comparison software for IRT experiments is flexmirt 3.6.2 [**?**, **?**] and VIBO developed by [**?**, **?**]. The comparison algorithms are BAEM, MHRM, and VIBO. For the four-parameter model, the R package MIRT 1.32.8 [**?**, **?**] is used instead of flexmirt. The comparison software for CDM experiments is the R package GDINA 2.8.0, with the EM algorithm as the comparison algorithm.

4.1 Real Data

For real data, only amortized variational inference is used for parameter estimation. The evaluation metric is AUC [**?**, **?**]. All real data are split into test and validation sets with a split ratio of 8:2.

Two-parameter IRT is applied to analyze the data.

Table 1 shows the LSAT parameter estimation fit statistics with neural correlation matrix priors.

Table 2 shows the PISA parameter estimation fit statistics with neural correlation matrix priors.

Tables 1 and 2 show: (1) prior distributions with neural correlation matrices achieve higher fit performance than identity matrices; (2) high-dimensional mod-

els achieve higher fit performance than low-dimensional models. Overall, variational inference demonstrates excellent fit performance on LSAT and PISA data.

[Figure 5: see original paper]

4.1.2 Cognitive Diagnosis Models

Table 3 shows the ECPE parameter estimation fit statistics for higher-order trait dimensions in HO-DINA. Table 3 shows that the DINA model performs better than HO-DINA. Overall, variational inference shows good fit performance on ECPE data.

4.2.1 Item Response Theory

In the experimental results, BBVI represents black box variational inference, AI represents amortized variational inference, and a, b, c, d, x represent the corresponding parameters in IRT, while g, s, λ_0, λ_1 represent the corresponding parameters in CDM. Numbers in tables represent the mean of RMSE, with numbers in parentheses representing the standard deviation of RMSE.

Table 4 shows the recovery RMSE for unidimensional IRT parameter estimation. MHRM is not shown in Table 4 because it performed unstably in unidimensional multi-parameter models (parameter estimation failed twice). Table 5 shows the recovery RMSE for multidimensional IRT parameter estimation. BAEM is not shown in Table 5 because it is too time-consuming for multidimensional models. In Table 6 , flexmirt software sets latent variable covariances as freely estimated, while variational inference uses neural correlation matrices as prior distributions.

Tables 4, 5, 6, and 7 show: (1) for dimension-correlated simulated data, variational inference has significant advantages and lower computational time than MHRM; (2) for medium-sized datasets, variational inference has certain advantages and much lower computational time than BAEM and MHRM; (3) for unidimensional multi-parameter models, variational inference has certain advantages but higher computational time than BAEM; (4) for multidimensional models, variational inference performs essentially the same as MHRM but with lower computational time; (5) the VIBO algorithm performs extremely poorly in two-dimensional models, confirming previous speculation that VIBO is not suitable for multidimensional IRT models.

4.2.2 Reparameterization vs REINFORCE for DINA Models

Table 8 compares reparameterization versus REINFORCE gradients for DINA models. Table 8 shows that the Gumbel-Softmax reparameterization method is difficult to apply to DINA models, and Figure 7 [Figure 7: see original paper] also verifies this result. The right panel of Figure 6 [Figure 6: see original paper]

shows the AUC values between the latent variable network output values and the true attribute mastery pattern values.

[Figure 6: see original paper]

4.2.3 DINA and HO-DINA Models

Table 9 shows the recovery RMSE for DINA parameter estimation. Table 10 shows the recovery RMSE for HO-DINA parameter estimation. Tables 9 and 10 show that variational inference achieves parameter recovery accuracy essentially equivalent to EM for CDM models, but with much higher computational time than EM.

4.2.4 High-Dimensional DINA Models

Table 11 shows the recovery RMSE for high-dimensional DINA model parameter estimation. The space complexity of EM for high-dimensional attribute mastery pattern DINA models is $O(2^N)$, making EM difficult to apply to high-dimensional DINA. Therefore, the experiments also tried MCMC algorithms based on random walks, finding that MCMC algorithms take 3-10 times longer than variational inference (approximately 3-10 hours). Thus, using variational inference for high-dimensional attribute pattern parameter estimation is a more economical choice. Figure 7 [Figure 7: see original paper] shows the recovery degree for high-dimensional attribute mastery patterns.

[Figure 7: see original paper]

4.2.5 Missing Data

The IRT missing data experiment sets the number of items to 500 with a unidimensional two-parameter model and randomly missing 90% of data. The HO-DINA missing data experiment sets the number of items to 500 with randomly missing 90% of data.

Table 12 shows the recovery RMSE for IRT missing data parameter estimation. Table 13 shows the recovery RMSE for HO-DINA missing data parameter estimation. Tables 12 and 13 show that variational inference can maintain good parameter recovery even when handling 90% missing data.

5 Conclusion and Outlook

Both real data experiments and simulated data experiments demonstrate that variational inference achieves high prediction performance and parameter recovery performance in educational measurement models. The latent variable network models and neural correlation matrices developed in this study demonstrate the flexibility and extensibility of variational inference. Researchers can arbitrarily develop desired models using the universal approximation property of neural networks or the simplicity of black box variational inference. The study

does not present code due to space limitations, but the code is open-source. Researchers will find through the open-source code that writing variational inference parameter estimation programs is not much different from writing MCMC parameter estimation programs. Variational inference has great potential for application in educational measurement research or experimental environments and can help researchers develop new educational measurement models.

The algorithms designed in this experiments achieve state-of-the-art performance in the IRT domain, leading or matching flexmirt in both running time and parameter recovery. Therefore, we strongly recommend that researchers use variational inference to develop new IRT models, and also recommend that general users apply variational inference algorithms in practical scenarios. The algorithms designed in this experiments perform adequately in CDM, only surpassing EM in high-dimensional attribute mastery patterns. Nevertheless, we still recommend that researchers use variational inference to develop new CDM models, as variational inference can save researchers development time. As for general users, we recommend using the EM algorithm implemented in GDINA.

Although this study demonstrates that variational inference algorithms have great potential in the field of educational measurement, several issues remain. First, although this study presents neural correlation matrices and obtains good results through simulated and real data experiments, theoretical proof is lacking. Second, the algorithms designed in this experiments have long running times in CDM, which may be related to the use of REINFORCE gradients (which have high variance), while the use of reparameterization methods in CDM is not ideal. The application of variational inference in CDM requires further exploration. Third, regarding the application of normalizing flows, the distribution of latent variables may not be simple normal distributions, so normalizing flows may be needed to optimize parameter distributions. The experiments actually tested neural autoregressive flows but did not find their superiority, so normalizing flows are not elaborated in the main text and require future research. Fourth, this study demonstrates stochastic optimization for variational inference, but MCMC and EM can also apply stochastic optimization, namely stochastic gradient MCMC and stochastic EM algorithms, though no relevant studies have applied these two parameter estimation techniques to educational measurement models. Finally, we hope more AI technologies like variational inference can be applied to the field of education.

References

Baker, F. B., & Kim, S.-H. (2004). Item Response Theory: Parameter Estimation Techniques.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., . . . Goodman, N. D. (2019). Pyro: deep universal probabilistic programming. Journal of Machine Learning Research, 20(28), 973-978.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Infer-

ence: A Review for Statisticians. journal of the american statistical association, 112(518), 859-877. doi:10.1080/01621459.2017.1285773

Bock, R. D., & Aitkin, M. (1982). Marginal maximum likelihood estimation of item parameters. Psychometrika, 47(3), 369-369. doi:10.1007/BF02294168

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-Information Item Factor Analysis. applied psychological measurement, 12(3), 261-280. doi:10.1177/014662168801200305

Cai, L. (2010). High-Dimensional Exploratory Item Factor Analysis by a Metropolis-Hastings Robbins-Monro Algorithm. Psychometrika, 75(1), 33-57. doi:DOI 10.1007/s11336-009-9136-x

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A Probabilistic Programming Language. journal of statistical software, 76(1), 1-32. doi:10.18637/JSS.V076.I01

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. journal of statistical software, 48(1), 1-29. doi:10.18637/JSS.V048.I06

Chen, L. (2017). Fast Item Response Theory (IRT) Analysis by using GPUs. Retrieved from https://on-demand.gputechconf.com/gtcdc/2017/presentation/dc7176-lei-chen-fast-item-response-theory-irt-model-estimation-by-using-gpus.pdf

Cho, A. E., Wang, C., Zhang, X., & Xu, G. J. (2020). Gaussian variational estimation for multidimensional item response theory. British Journal of Mathematical & Statistical Psychology. Retrieved from ://WOS:000577623600001

Chung, S., & Houts, C. (2020). flexMIRT: A Flexible Modeling Package for Multidimensional Item Response Models. measurement interdisciplinary research & perspective, 18(1), doi:10.1080/15366367.2019.1693825

Culpepper, S. A. (2015). Bayesian Estimation of the DINA Model With Gibbs Sampling. Journal of Educational and Behavioral Statistics, 40(5), 454-476. Retrieved from ://WOS:000363883900002

Curi, M., Converse, G. A., Hajewski, J., & Oliveira, S. (2019, 14-19 July 2019). Interpretable Variational Autoencoders for Cognitive Models. Paper presented at the 2019 International Joint Conference on Neural Networks (IJCNN).

Feng, Y., Habing, B. T., & Huebner, A. (2014). Parameter Estimation of the Reduced RUM Using the EM Algorithm. applied psychological measurement, 38(2), 137-150. doi:10.1177/0146621613502704

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least Squares Item Factor Analysis. multivariate behavioral research, 23(2), 267-269. doi:10.1207/S15327906MBR2302_9

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. J. Mach. Learn. Res., 14(1), 1303–1347.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. neural networks, 2(5), 359-366. doi:10.1016/0893-6080(89)90020-8

Huber, P., Ronchetti, E., & Victoria-Feser, M.-P. (2004). Estimation of Generalized Linear Latent Variable Models. journal of the royal statistical society series b statistical methodology, 66(4), 893-908. doi:10.1111/J.1467-9868.2004.05627.X

Hui, F. K. C., Warton, D. I., Ormerod, J. T., Haapaniemi, V., & Taskinen, S. (2017). Variational Approximations for Generalized Linear Latent Variable Models. Journal of Computational and Graphical Statistics, 26(1), 35-43. doi:10.1080/10618600.2016.1164708

Imai, K., Lo, J., & Olmsted, J. (2016). Fast Estimation of Ideal Points with Massive Data. american political science review, 110(4), 631-656. doi:10.1017/S000305541600037X

Jang, E., Gu, S., & Poole, B. (2017). Categorical Reparameterization with Gumbel-Softmax. Paper presented at the International Conference on Learning Representations.

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. Paper presented at the International Conference on Learning Representations.

Li, C., Ma, C., & Xu, G. (2020). Learning Large Q-matrix by Restricted Boltzmann Machines. In.

Linden, W. J. v. d. (2016). Handbook of Item Response Theory, Volume Two: Statistical Tools. In.

Lu, J., Zhang, J., & Tao, J. (2018). Slice–Gibbs sampling algorithm for estimating the parameters of a multilevel response model. Journal of Mathematical Psychology, 82, doi:https://doi.org/10.1016/j.jmp.2017.10.005

Luo, Y., & Jiao, H. (2017). Using the Stan Program for Bayesian Item Response Theory. Educational and Psychological Measurement, 78(3), 384-408. doi:10.1177/0013164417693666

Ma, W., & Torre, J. d. l. (2020). GDINA: An R Package for Cognitive Diagnosis Modeling. journal of statistical software, 93(1), 1-26. doi:10.18637/JSS.V093.I14

McClish, D. K. (1989). Analyzing a portion of the ROC curve. medical decision making, 9(3), 190-195. doi:10.1177/0272989X8900900307

Meng, X.-L., & Schilling, S. (1996). Fitting Full-Information Item Factor Models and an Empirical Investigation of Bridge Sampling. journal of the american statistical association, 91(435), 1254-1267. doi:10.1080/01621459.1996.10476995

Minka, T. (2009). Automating Variational Inference for Statistics and Data Minin. Invited talk at the 74th Annual Meeting of the Psychometric

Society (IMPS 2009). Retrieved from https://www.microsoft.com/en-us/research/publication/automating-variational-inference-statistics-data-mining/

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. D. (2016). Bayesian Prior Choice in IRT Estimation Using MCMC and Variational Bayes. Frontiers in Psychology, 7. doi:ARTN 1422 10.3389/fpsyg.2016.01422

Ormerod, J. T., & Wand, M. P. (2010). Explaining Variational Approximations. the american statistician, 64(2), 140-153. doi:10.1198/TAST.2010.09058

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In.

Ranganath, R., Gerrish, S., & Blei, D. M. (2014). Black Box Variational Inference. Paper presented at the International Conference on Artificial Intelligence and Statistics.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic Programming in Python using PyMC3. peerj, 2. doi:10.7717/PEERJ-CS.55

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. psychometrika, 70(3), 533-555. doi:10.1007/S11336-003-1141-X

StanDevelopmentTeam. (2019). Cholesky Factors of Correlation Matrices. In Stan Reference Manual (v2.18 ed.).

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. psychometrika, 47(2), 175-186. doi:10.1007/BF02296273

Torre, J. d. l. (2009). DINA Model and Parameter Estimation: A Didactic. journal of educational and behavioral statistics, 34(1), 115-130. doi:10.3102/1076998607309474

Torre, J. d. l. (2011). The Generalized DINA Model Framework. psychometrika, 76(2), 179-199. doi:10.1007/S11336-011-9207-7

Torre, J. D. L., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. psychometrika, 69(3), 333-353. doi:10.1007/BF02295640

Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M. R., Liang, D., & Blei, D. M. (2016). Edward: A library for probabilistic modeling, inference, and criticism. In.

Williams, R. J. (1992). Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. machine learning, 8(3), 229-256. doi:10.1007/BF00992696

Wingate, D., & Weber, T. (2013). Automated Variational Inference in Probabilistic Programming. In.

Wu, M., Davis, R. L., Domingue, B. W., Piech, C., & Goodman, N. D. (2020). Variational Item Response Theory: Fast, Accurate, and Expressive. In.

Yamaguchi, K. (2020). Variational Bayesian inference for the multiple-choice DINA model. behaviormetrika, 47(1), 159-187. doi:10.1007/S41237-020-00104-W

Yamaguchi, K., & Okada, K. (2020). Variational Bayes Inference for the DINA Model. Journal of Educational and Behavioral Statistics, 45(5), 569-597. doi:doi: 10.3102/1076998620911934

Zhan, P., Jiao, H., Man, K., & Wang, L. (2019). Using JAGS for Bayesian Cognitive Diagnosis Modeling: A Tutorial. Journal of Educational and Behavioral Statistics, Online First. doi:10.3102/1076998619826040

Zhang, C., Bütepage, J., Kjellström, H., & Mandt, S. (2019). Advances in Variational Inference. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8), 2008-2026. doi:10.1109/TPAMI.2018.2889774

Zhang, S. L., Chen, Y. X., & Liu, Y. (2020). An improved stochastic EM algorithm for large-scale full-information item factor analysis. British Journal of Mathematical & Statistical Psychology, 73(1), 44-71. Retrieved from ://WOS:000509696500003

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv — Machine translation. Verify with original.