

Autoencoders for Amortized Joint Maximum Likelihood Estimation of Confirmatory Item Factor Models

Dylan Molenaar, Raoul P. P. P. Grasman & Mariana Cúri

To cite this article: Dylan Molenaar, Raoul P. P. P. Grasman & Mariana Cúri (2025) Autoencoders for Amortized Joint Maximum Likelihood Estimation of Confirmatory Item Factor Models, *Multivariate Behavioral Research*, 60:4, 657-677, DOI: [10.1080/00273171.2025.2456598](https://doi.org/10.1080/00273171.2025.2456598)

To link to this article: <https://doi.org/10.1080/00273171.2025.2456598>



© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC



[View supplementary material](#)



Published online: 12 Feb 2025.



[Submit your article to this journal](#)



Article views: 597



[View related articles](#)



[View Crossmark data](#)

Autoencoders for Amortized Joint Maximum Likelihood Estimation of Confirmatory Item Factor Models

Dylan Molenaar^a, Raoul P. P. P. Grasman^a, and Mariana Cúri^b

^aUniversity of Amsterdam, Amsterdam, The Netherlands; ^bUniversidade de São Paulo (USP), São Paulo, Brazil

ABSTRACT

Neural networks like variational autoencoders have been proposed as a statistical tool to fit item factor models to data. Advantages are that high dimensional models can be estimated more efficiently as compared to conventional approaches. In this study, we demonstrate advantages of a specific autoencoder as a tool for amortized joint maximum likelihood estimation of item factor models. Contrary to contemporary joint maximum likelihood estimation and marginal maximum likelihood estimation, no additional parameter constraints are necessary to ensure standard asymptotic theory to apply. In a simulation study, the performance of the autoencoder is compared to constrained joint maximum likelihood and various forms of marginal maximum likelihood under different distributions for the factor scores. Results show that the amortized joint maximum likelihood estimates of the factors scores are overall less biased as compared to the other approaches. We illustrate the use of the autoencoder in two real data examples.



KEYWORDS

Autoencoder; variational autoencoder; item factor models; parameter estimation; normal distribution

In psychometrics, confirmatory item factor models have many practical uses including the assessment of the psychometric properties of tests and questionnaires (e.g., Kline, 2013), accounting for individual differences and measurement error in inferences about theoretical constructs (e.g., Mellenbergh, 1994), test equating (e.g., Kolen & Brennan, 2004), computerized adaptive testing (e.g., Wainer et al., 2000), and modeling change over time (e.g., McArdle, 2009). Estimation of confirmatory item factor models has been dominated by variations of maximum likelihood estimation (e.g., Andersen, 1970; Bock & Aitkin, 1981; Cai, 2010; Kelderman & Rijkes, 1994; Klein & Moosbrugger, 2000; Lawley, 1943; Verhelst & Glas, 1995), least squares estimation (e.g., Browne, 1974; Li, 2016; Muthén, 1984), and Bayesian estimation (Albert, 1992; Edwards, 2010; Fox & Glas, 2001; Martin & McDonald, 1975). Although for continuous observed indicators, estimation is relatively fast and easily applied to high dimensional datasets, for discrete data, estimation is arguably more challenging. Therefore, recently a number of studies have focused on the development of estimation approaches for categorical

data that are computationally less demanding. Here we focus on developments with respect to joint maximum likelihood estimation and estimation based on models from the field of deep learning.

Joint maximum likelihood, originally considered for item response theory models by Birnbaum (1968), has recently been rediscovered as a fast and practical estimation approach. Specifically, in joint maximum likelihood both the person and the item parameters are assumed fixed effect parameters by which they are estimated simultaneously. As a result, no numerical integration is required which makes approaches like marginal maximum likelihood and Bayesian estimation computationally demanding. However, in traditional joint maximum likelihood, the number of free parameters increases linearly with the sample size which violates standard asymptotic theory causing the parameter estimates to be inconsistent (e.g., Haberman, 1977). Recently, solutions have been proposed involving either constraining (Chen et al., 2019) or regularizing (Bergner et al., 2022) the person and item parameters.

CONTACT Dylan Molenaar  D.Molenaar@uva.nl  Psychological Methods, Department of Psychology, University of Amsterdam, Postbus 15906, Amsterdam 1001 NK, The Netherlands

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Other recent studies have used models from the field of deep learning. In deep learning, latent variables are used in so-called deep neural networks to capture complex non-linear functions between observed dependent and observed independent variables to predicted future data (Goodfellow et al., 2016). Although the latent variables in neural networks and the latent variables in item factor models are used for different purposes, the models are intrinsically the same. For instance, Cúri et al. (2019, see also Converse et al., 2021) and Urban and Bauer (2021) focused on fitting item factor models for respectively binary items scores (Birnbaum, 1968; Lord, 1952) and ordinal item scores (Samejima, 1969; Takane & De Leeuw, 1987) by means of a variational autoencoder (Kingma & Welling, 2013). In these variational autoencoder item factor models, the factor scores are subject to a multivariate normal prior distribution, and the resulting posterior distribution of the factor scores is approximated by an (importance weighted) normal distribution where the mean and log-standard deviation depend non-linearly on the observed data. During the estimation of variational autoencoders, a value is drawn from the normal approximate posterior (Converse et al., 2021; Cúri et al., 2019) or from the importance weighted normal approximate posterior (Urban & Bauer, 2021) to replace the unknown factor scores during optimization. In doing so, fitting item factor models to categorical data is generally faster as compared to marginal maximum likelihood (Urban & Bauer, 2021).

Current work on joint maximum likelihood and variational autoencoders as an approach to item factor analysis has mainly focused on the computational advantages: shorter estimation time and increased flexibility with respect to the dimensionality of the models as compared to maximum likelihood estimation. This paper is motivated by a statistical advantage: We demonstrate how an fixed-effects autoencoder (Goodfellow et al., 2016) can be conceived as an amortized joint maximum likelihood estimator for item factor models. That is, we consider an autoencoder where the factor scores are considered fixed effects similar to joint maximum likelihood and contrary to variational autoencoders where the factors are random effects. Desirable properties of such a fixed-effects autoencoder are that (1) contrary to joint maximum likelihood and similar to variational autoencoders, the number of parameters does not depend on the sample size; and (2) contrary to joint maximum likelihood, marginal maximum likelihood, and the variational autoencoder, the fixed-effects autoencoder avoids direct constraints on, or regularizing of, the parameter space.

The fixed-effects autoencoder has been proposed before as a tool to fit item factor models by Guo et al. (2017) and Converse et al. (2019). Both Guo et al. and Converse et al. conducted small simulation studies to demonstrate the viability of the autoencoder to estimate the DINA model (Guo et al., 2017) and the two-parameter logistic item response theory model (Converse et al., 2019). Converse et al. found large bias and relatively small correlations between the true and estimated item response theory parameters for the autoencoder, but not for the variational autoencoder, which may indicate problems related to identification. Therefore, here, we expand on this work in several ways. First, we derive the autoencoder item factor model more formally as an amortized joint maximum likelihood approach of which the properties are known. Next, we study the identification of the model and propose a default configuration for the amortization part of the autoencoder. Furthermore, we study the performance of the autoencoder as compared to joint maximum likelihood, and various forms of marginal maximum likelihood under practical and (double approximate) asymptotic settings. Finally, we demonstrate that the autoencoder produces less biased factor score estimates as compared to the constrained approaches.

Theoretically, our incentive to study the fixed-effects autoencoder as an estimation approach next to variational autoencoders is to increase understanding of the theoretical relation between neural networks for deep learning and latent variable models for psychometric inference. It has already demonstrated that the variational autoencoder is the autoencoder counterpart of the marginal maximum likelihood factor model framework. In the present study we demonstrate how the fixed effects autoencoder is the autoencoder counterpart of the joint maximum likelihood factor model framework. A thorough understanding of such relations among models from psychometrics and deep learning will ideally benefit both fields of research. For instance, in the field of deep learning many well-established efficient and fast algorithms exist that can potentially be used for various psychometric purposes that are currently challenging due to numerical demands, while in psychometrics many powerful tools for statistical inferences and model fit exists from which deep learning applications can importantly benefit. In addition, the modeling frameworks developed within psychometrics can help the field of artificial intelligence in their work on more explainable and interpretable models (explainable artificial intelligence; e.g., Arrieta et al., 2020).

Practically, the fixed-effects autoencoders has a number of benefits, some of which are illustrated in the

present study. First, we will show that due to the non-parametric nature of the autoencoder, for normal population distributions and finite samples, there is no shrinkage effects in the factor score estimates, while these effects are common for marginal and joint maximum likelihood approaches. Related, for non-normal population distributions (e.g., due to heterogeneous subpopulations) and finite samples, there is less bias in the factor score estimates as compared to approaches assuming a normal distribution in the population (e.g., marginal maximum likelihood and the variational autoencoder) or impose similar constraints (as in regularized joint maximum likelihood and constrained joint maximum likelihood). Next, as opposed to variational autoencoders, fixed-effects autoencoders do not involve sampling during estimation which can be time consuming (especially in the case of multiple chains and many importance samples). Finally, after fitting an autoencoder, factor scores can be calculated for new, incoming, data from the same population without additional estimation, while for importance weighted variational autoencoders this would involve additional sampling.

The outline is as follows: First, we present the item factor model and marginal and joint maximum likelihood estimation of its parameters. Next, we present the fixed-effects autoencoder and demonstrate its relation to joint maximum likelihood and the variational autoencoder. We first focus on binary data, and show how this approach extension straightforwardly to polytomous data. Next, we present a simulation study in which we compare the performance of the fixed-effects autoencoder to the performance of constrained joint maximum likelihood estimation (Chen et al., 2019), and various forms of marginal maximum likelihood estimation (Cai, 2010; Chalmers, 2012) in a three dimensional item factor model with binary items. We consider different shapes of the factor score distribution and compare the parameter recovery of the different approaches. We then present two real data examples respectively illustrating the use of the autoencoder in recovering the factor score distribution and illustrating a robustness analysis of a 16 dimensional autoencoder for ordinal items to different configurations of the encoder. We end with a general discussion.

Item factor models

In the next sections we first focus on the item factor model for binary data (Birnbaum, 1968; Christofferson, 1975; Muthén, 1978; Takane & De Leeuw, 1987) but the principles discussed are applicable to all models in the generalized linear item

response theory framework (Mellenbergh, 1994; Moustaki & Knott, 2000) for continuous latent variables. For ordinal data we explicitly demonstrate this in a separate section. If \mathbf{X} denotes a matrix of stacked vectors $\mathbf{x}_p^T = [x_{p1}, \dots, x_{pn}]$ containing the item scores of person $p = 1, \dots, N$ on items $i = 1, \dots, n$, then the distribution of \mathbf{x}_p under the item factor model is:

$$f(x_{pi}|\boldsymbol{\eta}_p) = P(x_{pi} = 1|\boldsymbol{\eta}_p)^{x_{pi}} \times [1 - P(x_{pi} = 1|\boldsymbol{\eta}_p)]^{1-x_{pi}} \quad (1)$$

with

$$P(x_{pi} = 1|\boldsymbol{\eta}_p) = \Phi(-\tau_i + \boldsymbol{\lambda}_i^T \boldsymbol{\eta}_p) \quad (2)$$

in which $\boldsymbol{\eta}_p$ is the vector of factor scores with elements η_{pq} indexed by $q = 1, \dots, K$, τ_i is a threshold parameter for item i , $\Phi(\cdot)$ is the cumulative standard normal distribution function, and $\boldsymbol{\lambda}_i$ is a vector of discrimination parameters or factor loadings for item i with elements λ_{iq} . We assume throughout this paper that a sufficient number of elements of $\boldsymbol{\lambda}_i$ is set to zero to avoid rotational indeterminacy (we will return to this point later). In addition, we have omitted an item specific residual variance parameter for reasons of identification (although such a parameter can be identified in ordinal data, see e.g., Mehta et al., 2004; Millsap & Yun-Tein, 2004; Molenaar et al., 2012).

Marginal maximum likelihood

Marginal maximum likelihood estimation of the model above is conducted by maximizing the log-marginal likelihood of \mathbf{X} with respect to the unknown parameter vector

$$\boldsymbol{\theta}_{MML} = [\tau_1^{MML}, \dots, \tau_n^{MML}, \boldsymbol{\lambda}_1^{MML}, \dots, \boldsymbol{\lambda}_n^{MML}, \boldsymbol{\mu}_\eta, \text{vec}(\boldsymbol{\Sigma}_\eta)]$$

where $\boldsymbol{\mu}_\eta$ and $\boldsymbol{\Sigma}_\eta$ are respectively the mean vector and the covariance matrix of the factor scores in the marginal maximum likelihood specification of the item factor model, $\boldsymbol{\eta}_p^{MML}$. Specifically, the log-marginal likelihood function is given by

$$\ell(\boldsymbol{\theta}_{MML}; \mathbf{X}) = \sum_{p=1}^N \log \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \prod_{i=1}^n f(x_{pi}|\boldsymbol{\eta}_p^{MML}) g(\boldsymbol{\eta}_p^{MML}) d\boldsymbol{\eta}^{MML} \quad (3)$$

in which $g(\cdot)$ is a multivariate normal distribution. The model can be identified by fixing $\boldsymbol{\mu}_\eta$ to a vector of zeros and the diagonal elements of $\boldsymbol{\Sigma}_\eta$ to 1. To evaluate the log-marginal likelihood function above,

the integral needs to be approximated using, for instance, numerical quadrature (e.g., Bock & Aitkin, 1981) or stochastic imputation (Cai, 2010). As a result, the factor scores $\boldsymbol{\eta}_p^{JML}$ are no free model parameters but can be estimated in a second step by for instance empirical Bayes estimation or expected posterior scoring. Marginal maximum likelihood with numerical approximation of the integrals in the likelihood function is practically feasible but becomes numerically challenging for increasing the number of latent variable dimensions and increasing sample size.

Joint maximum likelihood

In joint maximum likelihood, the factor scores $\boldsymbol{\eta}_p^{JML}$ are considered fixed effects and estimated simultaneously with the thresholds and loadings. If the joint maximum likelihood parameter vector is given by $\boldsymbol{\theta}_{JML} = [\tau_1^{JML}, \dots, \tau_n^{JML}, \lambda_1^{JML}, \dots, \lambda_n^{JML}, \boldsymbol{\eta}_1^{JML}, \dots, \boldsymbol{\eta}_N^{JML}]$, the joint likelihood of the data matrix \mathbf{X} is given by

$$\ell(\boldsymbol{\theta}_{JML}; \mathbf{X}) = \sum_{p=1}^N \sum_{i=1}^n \log f(x_{pi} | \boldsymbol{\eta}_p^{JML}). \quad (4)$$

The advantage of joint maximum likelihood so is that no distributional assumption is needed for $\boldsymbol{\eta}_p^{JML}$ and that the procedure is generally much faster as compared to marginal maximum likelihood where the approximation of the integrals is time consuming. A disadvantage is that, for fixed n , the number of parameters in $\boldsymbol{\theta}_{JML}$ increases linearly with N (i.e., the number of elements in $\boldsymbol{\theta}_{JML}$ equals $2n + N$) by which standard asymptotic theory does not apply and the estimates are inconsistent as a result (Haberman, 1977). In addition, for subjects/items with strict 0 or 1 scores in the rows/columns of \mathbf{X} , no parameter estimates exist.

To solve the above, Chen et al. (2019) constrained the parameter space in the following way

$$\sqrt{1 + \|\boldsymbol{\eta}_p^{JML}\|^2} \leq S \quad \text{and} \quad \sqrt{\tau_i^{JML} 2 + \|\lambda_i^{JML}\|^2} \leq S \quad (5)$$

where $\|\cdot\|^2$ denotes the squared l^2 -norm, and S is commonly set to $5K^{\frac{1}{2}}$ as a default but can in principle be set to any sufficiently large number (see Chen et al., 2019). Chen et al. demonstrated that the parameter estimates are consistent in the double asymptotic case (both N and n approach infinity). To enforce the constraints in Equation 5, Chen et al. used a projected gradient descent algorithm to estimate the parameters. In each iteration of this algorithm, the parameters are transformed to a feasible parameter space using a

projection function. Bergner et al. (2022) achieved similar constraints by focusing on regularizing the parameters using the squared l^2 -norm, that is

$$\begin{aligned} \ell(\boldsymbol{\theta}_{JML}; \mathbf{X}) = & \sum_{p=1}^N \sum_{i=1}^n \log f(x_{pi} | \boldsymbol{\eta}_p^{JML}) \\ & - \psi \left(N + \sum_{p=1}^N \|\boldsymbol{\eta}_p^{JML}\|^2 + \sum_{i=1}^n \tau_i^{JML} 2 + \sum_{i=1}^n \|\lambda_i^{JML}\|^2 \right) \end{aligned} \quad (6)$$

where ψ is a tuning parameter that can be determined by cross-validation. Note that both approaches are similar in the restrictions that are imposed on the l^2 -norms of the parameters, although they are enforced in a different way (i.e., by projection versus regularization) and are person/item specific in Equation (5) but not in Equation (6) (i.e., ψ is not person or item specific). These restrictions effectively impose a normal prior on both the person and the item parameters in the likelihood function. This is well known for l^2 -regularization in Equation (6) and can be shown in a similar way for Equation (4) subject to the projection in Equation (5) by focusing on the Lagrangian function. Specifically, the restrictions in Equation (5) can be reformulated as $\|\boldsymbol{\eta}_p^{JML}\|^2 \leq S^2 - 1$. This adds a term $\zeta_p (\|\boldsymbol{\eta}_p^{JML}\|^2 - S^2 + 1)$ to the Lagrangian -where ζ_p is a person specific Lagrange multiplier- which, like regularized joint maximum likelihood in Equation (6), can be construed as a normal prior term. The main difference between constrained joint maximum likelihood and regularized joint maximum likelihood from this perspective is that regularized joint maximum likelihood is equivalent in form to having a uniform normal prior on all the parameters, where constrained joint maximum likelihood is equivalent in form to having separate normal priors, one for each set of parameters. Note that the constrained joint maximum likelihood approach is not assuming a normal prior, but the effect of the constraints is in form equivalent to imposing such priors. Thus both constrained joint maximum likelihood and regularized joint maximum likelihood are effectively not fully distribution free.

The effect of these constraints diminishes if $N \rightarrow \infty$ and $n \rightarrow \infty$ by which these joint maximum likelihood schemes obtain their asymptotically consistency. However, similar to prior distributions in a Bayesian sense, the effects of these constraints may be notable in finite samples and for finite items (which we also demonstrate in the simulation study). Therefore, below, we show how autoencoders can be used for amortized joint

maximum likelihood estimation without direct constraints on the person and item parameters.

Autoencoders

The key of autoencoders is that the data is first encoded into so-called hidden nodes. These nodes are organized in layers $l = 1, \dots, L$ and collected in $Q^{(l)}$ -dimensional vector $\mathbf{z}_p^{(l)}$. At layer l , $\mathbf{z}_p^{(l)}$ is given by

$$\mathbf{z}_p^{(l)} = h^{(l)}\left(\mathbf{b}^{(l)} + \mathbf{A}^{(l)}\mathbf{z}_p^{(l-1)}\right) \quad (7)$$

where $\mathbf{z}_p^{(0)} = \mathbf{x}_p = [x_{p1}, \dots, x_{pn}]^T$ and $Q^{(0)} = n$. In addition, $h^{(l)}(\cdot)$ is the encoding function at layer l , $\mathbf{b}^{(l)}$ is the $Q^{(l)}$ vector of encoding intercepts at layer l , and $\mathbf{A}^{(l)}$ is the $Q^{(l)}$ by $Q^{(l-1)}$ matrix of encoding slopes at layer l . After the final layer, L , predictions for the observed data in \mathbf{x}_p are obtained from hidden variables $\mathbf{z}_p^{(L)}$ by

$$\mathbf{x}'_p = k\left(\boldsymbol{\delta} + \boldsymbol{\Gamma}\mathbf{z}_p^{(L)}\right) \quad (8)$$

where \mathbf{x}'_p is a vector of model predictions with elements x'_{pi} , $k(\cdot)$ is the decoding function, $\boldsymbol{\delta}$ is a n -dimensional vector of decoding intercepts, and $\boldsymbol{\Gamma}$ is a n by $Q^{(L)}$ matrix of decoding slopes with rows γ_i . See Figure 1 for a graphical representation of the autoencoder including the encoding and decoding equations.

The autoencoder above can be configured in such a way that, under certain conditions, it is theoretically equivalent to the joint maximum likelihood specification in Eq., 1, 2, and 4. To demonstrate this, we first specify $Q^{(L)} = K$, $k(\cdot)$ to be a cumulative standard normal distribution function $\Phi(\cdot)$, and we denote $\mathbf{z}_p^{(L)}$ by $\boldsymbol{\eta}_p^{AE}$, δ_i by $-\tau_i^{AE}$ and γ_i by λ_i^{AE} . Then for a given item i , the decoder in Equation (8) simplifies to

$$x'_{pi} = \Phi\left(-\tau_i^{AE} + (\lambda_i^{AE})^T \boldsymbol{\eta}_p^{AE}\right) \quad (9)$$

with binary cross entropy

$$H(\mathbf{X}, \mathbf{X}') = -\sum_{p=1}^N \sum_{i=1}^n x_{pi} \log(x'_{pi}) + (1 - x_{pi}) \log(1 - x'_{pi}). \quad (10)$$

where \mathbf{X}' is the model predicted data matrix which consists of the stacked $\mathbf{x}'_1, \dots, \mathbf{x}'_N$ vectors which are a function of parameter vector

$$\boldsymbol{\theta}_{AE} = \left[\tau_1^{AE}, \dots, \tau_n^{AE}, \lambda_1^{AE}, \dots, \lambda_n^{AE}, \mathbf{b}^{(1)}, \dots, \mathbf{b}^{(L)}, \text{vec}(\mathbf{A}^{(1)}), \dots, \text{vec}(\mathbf{A}^{(L)}) \right]. \quad (11)$$

Note that, if $\boldsymbol{\eta}_p^{AE} = \boldsymbol{\eta}_p^{JML}$ for all p , the negative binary cross entropy in Equation (10) is equal to the joint likelihood function in Equation (4).

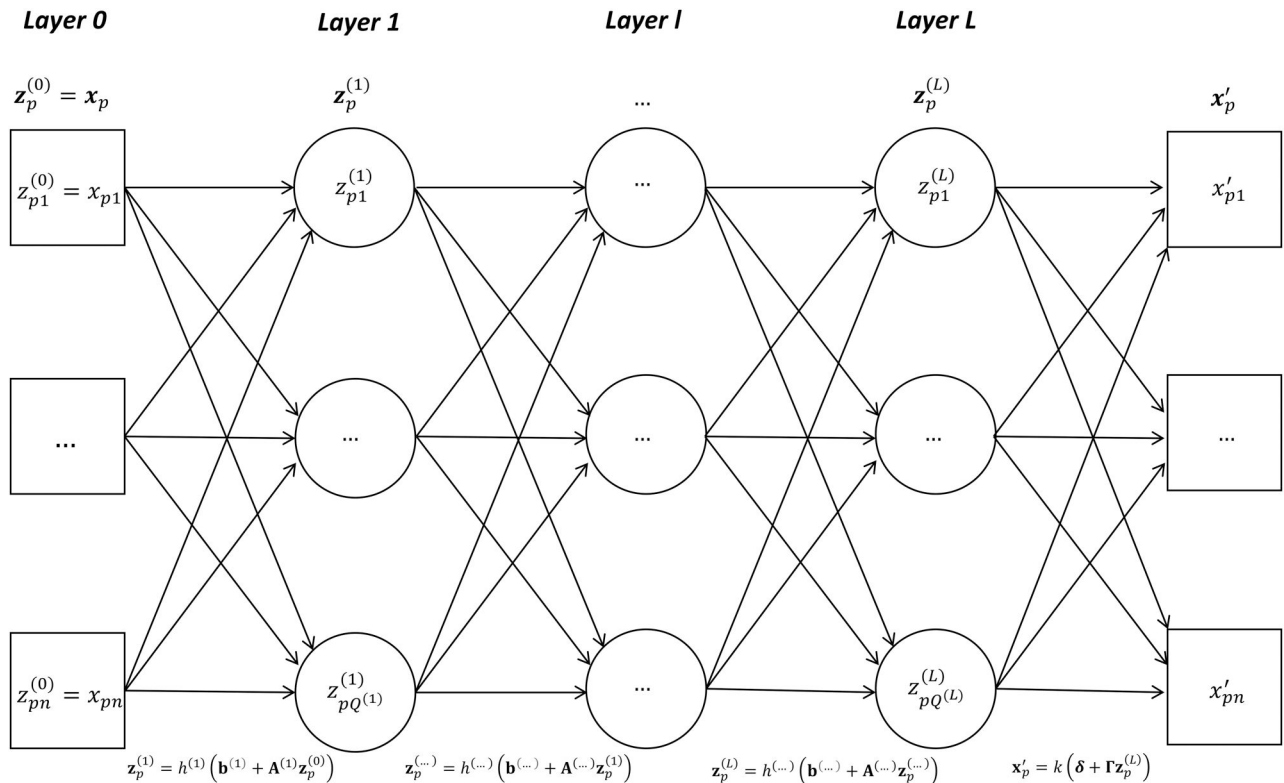


Figure 1. Graphical representation of the autoencoder including the encoder and decoder functions.

Amortized inference

Thus, the autoencoder and joint maximum likelihood item factor model coincide if $\boldsymbol{\eta}_p^{AE} = \boldsymbol{\eta}_p^{JML}$ for all p . To understand the condition for which this holds, it is important to note that in the autoencoder, $\boldsymbol{\eta}_p^{AE}$ is not a free model parameter. That is, $\boldsymbol{\eta}_p^{AE}$ is *amortized* (e.g., Marino et al., 2018) which means that $\boldsymbol{\eta}_p^{AE}$ is modeled as a (non-linear) function of the item scores with the parameters of this function common to all subjects p . Thus,

$$\boldsymbol{\eta}_p^{AE} = \nu(\mathbf{x}_p; \boldsymbol{\theta}_{\nu(\cdot)}) \quad (12)$$

where $\nu(\cdot)$ is referred to as the inference model and has a parameter vector $\boldsymbol{\theta}_{\nu(\cdot)}$. In the autoencoder, the inference model $\nu(\cdot)$ is the multilayer encoder given by Equation (7). That is, $\boldsymbol{\eta}_p^{AE}$ is amortized by

$$\begin{aligned} \boldsymbol{\eta}_p^{AE} &= \nu(\mathbf{x}_p; \boldsymbol{\theta}_{\nu(\cdot)}) \\ &= h^{(L)}\left(\mathbf{b}^{(L)} + \mathbf{A}^{(L)}h^{(L-1)}\left(\mathbf{b}^{(L-1)} + \mathbf{A}^{(L-1)}h^{(L-2)}(\dots\mathbf{x}_p)\right)\right). \end{aligned} \quad (13)$$

Consequently, the encoder in Equation (13) approximates $\boldsymbol{\eta}_p^{JML}$ up to a precision depending on the configuration of the nodes and layers $\mathbf{z}_p^{(1)}$ to $\mathbf{z}_p^{(L)}$ in Equation (7). That is, the approximation depends effectively on the number of nodes, $Q^{(l)}$, the number of layers L , and the nature of the encoding functions used in each layer, $h^{(l)}(\cdot)$.

Hence, the autoencoder above is a joint maximum likelihood approach with parameters $\boldsymbol{\eta}_p^{JML}$ amortized by $\nu(\cdot)$ in Equation (13). Key advantage of doing so is that instead of KN free $\boldsymbol{\eta}_p^{JML}$ -parameters to be estimated using (constrained) joint maximum likelihood, $\boldsymbol{\eta}_p^{AE}$ is parameterized by $\sum_{l=1}^L Q^{(l)}(Q^{(l-1)} + 1)$ free parameters in the autoencoder, which does not depend on N . Thus, for fixed n , $Q^{(1)}, \dots, Q^{(L)}$, sample size can grow to infinity without problems.

The function in Equation (13) needs to be specified so that it is sufficiently flexible to cover the parameter space of $\boldsymbol{\eta}_p$, but needs to be sufficiently parsimonious to avoid overfitting (which will generally occur if the number of parameters in $\boldsymbol{\theta}_{\nu(\cdot)}$ will be larger than N). Fortunately, Urban and Bauer (2021) provide useful recommendations which we will adopt in the simulation study and the real data applications, we elaborate on these recommendations in those sections.

As mentioned above, a common challenge to the joint maximum likelihood of binary data is that for subjects and items with strictly 0 or 1 scores, no parameter estimates exist. As discussed above, in their joint maximum likelihood approach Chen et al. (2019) solved this issue by directly constraining the

item and person parameter space by Equation (5). In the autoencoder, the parameter space of $\boldsymbol{\eta}_p^{AE}$ is also constrained but in an indirect way by the amortization in Equation (13). As a result, the effect on the parameter space is milder as compared to direct constraints like prior constraints and regularization (as we also demonstrate in the simulation study below). In addition, $\boldsymbol{\eta}_p^{AE}$ has a natural upper and lower bound that follow from plugging in a vector of respectively ones and zeros for \mathbf{x}_p . For instance, for the very simple model where $L = 1$ and $h^{(1)}(\cdot)$ is a linear function (i.e., $\boldsymbol{\eta}_p^{AE}$ is a linear transformation of \mathbf{x}_p), the lower bound of $\boldsymbol{\eta}_p^{AE}$ is equal to $\mathbf{b}^{(1)}$ and the upper bound is equal to $\mathbf{b}^{(1)} + \text{SUM}(\mathbf{A}^{(1)})$. As parameters $\mathbf{b}^{(1)}$ also occurs in the likelihood of other response patterns it is identified (given Equations (14) and (15)). Note that for items with strict zero or one scores no item parameter estimates exist (similarly to e.g., marginal maximum likelihood).

Identification

The autoencoder itself is not identified yet as a linear transformations of $\boldsymbol{\eta}_p^{AE}$ can be absorbed in τ_i^{AE} and λ_i^{AE} and produce the same likelihood. For constrained and regularized joint maximum likelihood, the model is identified by the constraints and regularization respectively. For the autoencoder, no identification restrictions have been proposed yet. As the autoencoder does not include scale and location parameters for the factor score distribution, the model can't be identified by fixing the mean and variance of the factor scores (as is common in item factor analysis). Therefore, we propose to identify the model using the following constraints on τ_i^{AE} and λ_i^{AE} :

$$\prod_{i \in F_q} |\lambda_{iq}^{AE}| = 1 \quad (14)$$

for each q , and

$$\sum_{i \in F_q} \tau_i^{AE} = 0 \quad (15)$$

for each q where F_q is the set of items that have a non-zero loading on factor q (assuming a simple structure see below).

In addition to these two constraints, as already mentioned before, we assume that sufficient elements in $\lambda_1^{AE}, \dots, \lambda_n^{AE}$ are constrained to 0 to avoid rotational indeterminacy. Specifically, similarly to traditional (item) factor analysis,

$K(K-1)$ factor loadings need to be constrained (in addition to the constraint in Equation (14)) to make

the factor structure just identified. The resulting model is an exploratory item factor model for which the matrix $\Lambda^{AE} \Sigma_{\eta}^{\frac{1}{2}}$ could be rotated to facilitate interpretation. The autoencoder presented here is equally amenable to such an exploratory use. However, in the present simulation and applications, we focus on a more confirmatory use in which Λ^{AE} follows a simple structure.

Note that Equation (14) makes sure that the geometric mean of the absolute factor loadings is equal to 1. Due to the absolute operator, the factor loadings λ_{iq}^{AE} are still allowed to be smaller than 0, which is for instance desirable in applications to personality items where contra-indicative items are used (see real data application 2). In addition, that Equation (15) ensures that the arithmetic mean of the thresholds is equal to 0. These identification constraints are not new and have been used before in item response theory modeling to identify a two parameter model (see e.g., Albert, 1992; De Jong et al., 2008). They are sufficient to identify parameters η_{pq}^{AE} , λ_{iq}^{AE} , and τ_i^{AE} from the decoder (given sufficient constraints on λ_{iq}^{AE} as discussed above). That is, with these constraints in place, there is only one set of decoder parameter estimates that optimizes the binary cross entropy in Equation (10). However, the parameters from the encoder are not unique so that there may be multiple functions $v(\cdot)$ in Equation (12) that result in the same η_p^{AE} . However, this is unproblematic for the present purpose as the encoder parameters underlying $v(\cdot)$ are not of direct interest. In fact, it can be proven using the asymptotic theory of concentrated likelihoods that given the encoder parameters in $v(\cdot)$, the decoder parameters λ_{iq}^{AE} , and τ_i^{AE} follow standard asymptotic theory (see Grasman, 2004, Appendix C). Note that the non-uniqueness of the encoder is not an exclusive property of the fixed effects autoencoder approach, the same applies to the (importance weighted) variational autoencoder as studied by Urban and Bauer (2021) and Cúri et al. (2019).

Alternative scales

The identification constraints above preclude direct comparison of the results from the autoencoder with results from other estimation techniques that use different identification constraints. To enable a comparison, the results need to be transformed to a common scale. One can either focus on the standardized parameter estimates, or on a transformation of the autoencoder results to the scale of the other estimation technique.

First, the standardized parameter estimates, are obtained largely in the same way as in traditional item

factor analysis, i.e. (note that we assume simple structure as discussed above),

$$\hat{\lambda}_{iq}^z = \frac{\hat{\lambda}_{iq} \hat{\sigma}_{\eta_q}}{\sqrt{\hat{\lambda}_{iq}^2 \hat{\sigma}_{\eta_q}^2 + \sigma_e^2}}$$

and

$$\hat{\tau}_i^z = \frac{\hat{\tau}_i}{\sqrt{\hat{\lambda}_{iq}^2 \hat{\sigma}_{\eta_q}^2 + \sigma_e^2}}$$

where superscript z indicates a standardized parameter, and a hat indicates a parameter estimate obtained using the estimation technique and identification constraints of choice. In these equations, σ_e^2 is the variance of the error which is commonly not estimated (as mentioned before) but depends on the link function for traditional item factor models, and on the activation function in the decoder for the autoencoder. For the probit link or normal ogive activation function $\sigma_e^2 = 1$ and for a logit link or logistic activation function $\sigma_e^2 = \frac{\pi^2}{3}$. Furthermore, $\hat{\sigma}_{\eta_q}^2$ denotes the variance of factor q which is an explicit parameter in the case of marginal maximum likelihood (i.e., in that case it is diagonal element q from Σ_{η} in the multivariate normal density $g(\cdot)$ in Equation (3)). In the case of the autoencoder, $\hat{\sigma}_{\eta_q}^2 = \text{VAR}(\hat{\eta}_{pq}^{AE})$. The standardized factor loading estimates above have the appealing property that they can be interpreted as the biserial correlation among the item score and the factor score.

Second, to transform the autoencoder results to the scale used in another estimation technique, it is useful to note that due to Equations (14) and (15), for the autoencoder it holds that $E(\hat{\eta}_{pq}^{AE}) = \frac{1}{n_q} \sum_{i \in F_q} \frac{\tau_i}{\sigma_{\eta_q} \lambda_{iq}}$ and $SD(\hat{\eta}_{pq}^{AE}) = \left(\prod_{i \in F_q} |\lambda_{iq}| \right)^{\frac{1}{n_q}}$ where n_q is the number of elements in F_q (i.e., the number of items loading on factor q). Using these results, the autoencoder parameter estimates can be transformed to a scale of choice. For instance, to transform the autoencoder results to the marginal maximum likelihood scale where $E(\eta_{pq}^{MML}) = 0$ and $\text{VAR}(\eta_{pq}^{MML}) = 1$ for all q , the marginal maximum likelihood estimates $(\hat{\tau}_i^{MML} \text{ and } \hat{\lambda}_{iq}^{MML})$ can be plugged in for τ_i and λ_{iq} (note that $\sigma_{\eta_q} = 1$) and the expressions for $E(\hat{\eta}_{pq}^{AE})$ and $\text{VAR}(\hat{\eta}_{pq}^{AE})$ given above can be used to transform $\hat{\lambda}_{iq}^{AE}$, $\hat{\tau}_i^{AE}$, and $\hat{\eta}_{pq}^{AE}$ to the marginal maximum likelihood scale (using the general admissible scale transformation $\tau'_i = \tau_i + \lambda_{iq} E(\eta_{pq})$, and $\lambda'_{iq} = \lambda_{iq} SD(\eta_{pq})$).

The above two transformations can be used to put the parameter estimates on the same scale across methods. As the different methods discussed above (constrained/regularized/amortized joint maximum likelihood, and marginal maximum likelihood) all involve the same likelihood function, $f(x_{pi}|\boldsymbol{\eta}_p)$, but different constraints on the parameters, the transformed parameters will show discrepancies across methods. Asymptotically these discrepancies are expected to diminish as the effects of the constraints diminish for increasing N and n . Therefore, the item factor model parameters enjoy the same theoretical interpretation across methods. For finite samples, there may be differences across the different methods, which should be interpreted in terms of the differences in the constraints adopted in the specific methods.

Estimation

We fit the final model in Equation (9) with $\boldsymbol{\eta}_p^{AE}$ given in Equation (13) by minimizing the binary cross entropy in Equation (10) with respect to the unknown parameters in vector $\boldsymbol{\theta}_{AE}$ (see Equation (11)). As the negative binary cross entropy is equivalent to the likelihood function, this procedure is the same as maximum likelihood estimation. There are multiple algorithms possible to optimize the likelihood function. Common iterative algorithms use starting values for the parameters, determine a step size and the direction of the optimum of the likelihood function, and update the parameters using this step size and direction. The step size can be determined using function evaluations only (e.g., Nelder-Mead algorithm), using the gradients of the likelihood function only (e.g., quasi-Newton algorithm), or using both the gradients and the Hessian of the likelihood function (e.g., Newton-Raphson algorithm). In this study, we use the AMSgrad algorithm (Reddi et al., 2019) which utilizes exponential moving averages of the gradients of the likelihood to determine the direction and step size in each iteration of the algorithm, that is:

$$\boldsymbol{\theta}^{AE(t)} = \boldsymbol{\theta}^{AE(t-1)} - \alpha \frac{\mathbf{m}_t}{\sqrt{\bar{\mathbf{v}}_t}} \quad (16)$$

where α is a sufficiently small constant referred to as the learning rate and t indexes the iterations of the algorithm. In addition:

$$\begin{aligned} \mathbf{m}_t &= \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t \\ \mathbf{v}_t &= \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2 \\ \bar{\mathbf{v}}_t &= \max(\bar{\mathbf{v}}_{t-1}, \mathbf{v}_t) \end{aligned} \quad (17)$$

where \mathbf{g}_t is the vector of gradients of the likelihood function, and where β_1 and β_2 are between 0 and 1 and

can be used to tune the algorithm. In the simulation section, we discuss default values for α , β_1 , and β_2 .

Gradient projection

In the AMSgrad algorithm above, we introduce the identification constraints from Equations (14) and (15) by rescaling the factor loadings and threshold parameters after each iteration (i.e., a gradient projection method; Nocedal & Wright, 2006, p. 485). For numerical feasibility, the factor loadings are rescaled on a log-scale to prevent overflow, that is:

$$\lambda'_{iq} = \text{sign}(\lambda_{iq}) \times \log \left\{ \exp \left(\log(|\lambda_{iq}|) - \frac{1}{n_q} \sum_{i=1}^{n_q} |\lambda_{iq}| \right) \right\}.$$

Note that rescaling of the factor loadings and thresholds only forces the algorithm to iterate to a specific maximum of the likelihood function where the geometric mean of the factor loadings equals 1 and the arithmetic mean of the thresholds equals 0. That is, it does not complicate the likelihood function further. Gradient project methods are known to converge slowly in the case of complex non-linear constraints, but as our constraints are relative simply, we did not encounter any issues with respect to convergence.

Ordinal data

The models above have been presented for binary data. If x_{pi} is an ordinal variable with C categories, we code the item scores using vector \mathbf{y}_{pic} with elements y_{pic} for $c = 1, \dots, C - 1$. Specifically

$$y_{pic} = 1 \text{ if } x_{pi} \geq c \text{ and } y_{pic} = 0 \text{ otherwise.} \quad (18)$$

for instance if $C = 5$ and $x_{pi} = 3$, $\mathbf{y}_{pic} = [1, 1, 1, 0]$. Then, the item factor model for \mathbf{y}_{pic} can be given by:

$$f(\mathbf{y}_{pic}|\boldsymbol{\eta}_p^{AE}) = P(x_{pic} = 1|\boldsymbol{\eta}_p^{AE})^{y_{pic}} [1 - P(x_{pic} = 1|\boldsymbol{\eta}_p^{AE})]^{1-y_{pic}} \quad (19)$$

with

$$P(x_{pic} = 1|\boldsymbol{\eta}_p^{AE}) = \Phi \left(-\tau_{ic}^{AE} + (\boldsymbol{\lambda}_i^{AE})^T \boldsymbol{\eta}_p^{AE} \right) \quad (20)$$

where τ_{ic}^{AE} is the c -th threshold parameter of item i . Due to the coding, it holds that $\tau_{i1}^{AE} < \dots < \tau_{i(C-1)}^{AE}$ which reflect the ordered nature of the data. It can be shown that the resulting model for x_{pi} is:

$$\begin{aligned} P(x_{pi} = c|\boldsymbol{\eta}_p^{AE}) &= \Phi \left(-\tau_{ic}^{AE} + (\boldsymbol{\lambda}_i^{AE})^T \boldsymbol{\eta}_p^{AE} \right) \\ &\quad - \Phi \left(-\tau_{i(c+1)}^{AE} + (\boldsymbol{\lambda}_i^{AE})^T \boldsymbol{\eta}_p^{AE} \right) \end{aligned}$$

with $\tau_{i0}^{AE} = -\infty$ and with $\tau_{iC}^{AE} = \infty$, which is the more familiar form of the ordinal item factor model

(Samejima, 1969; Takane & De Leeuw, 1987). Thus, the above amortized joint maximum likelihood approach is equally amenable to ordinal data, by applying the autoencoder in Equation (9) (10), and (13) on the y_{pic} variables. Note that this involves equating the factor loadings λ_i^{AE} to be equal across y_{pic} from the same item (i.e., $y_{pi1}, \dots, y_{pi(C-1)}$). Alternatively, one can directly plug the likelihood of \mathbf{x}_p based on Equation (19) into the loss function in Equation (10) which is an equivalent approach.

Relation to variational autoencoders

A variational autoencoder item factor model is obtained by replacing η_p^{AE} in the model above by η_p^{VAE} , a stochastic parameter for which its means and standard deviations are a deterministic functions of the nodes from the previous layers, i.e., $\mathbf{z}_p^{(0)}$ to $\mathbf{z}_p^{(L-1)}$. In the variational autoencoder, η_p^{VAE} is assumed to follow a multivariate standard normal prior distribution (Kingma & Welling, 2013), $p(\eta_p^{VAE})$. The model is fit to data using a variational inference framework (e.g., Gelman et al., 1995) by running W Monte Carlo chains in which R samples from an approximate posterior distribution of η_p^{VAE} , $q(\eta_p^{VAE}|\mathbf{x}_p)$ are combined into an importance weighted estimate of the log-likelihood. Distribution $q(\cdot)$ can be any continuous distribution but is commonly specified to be a multivariate normal distribution (see e.g., Converse et al., 2021; Cúri et al., 2019; Urban & Bauer, 2021). That is

$$\tilde{\eta}_{prw}^{VAE} \sim q(\eta_p^{VAE}|\mathbf{x}_p) = \text{MVN}(\boldsymbol{\mu}_p, \text{diag}(\boldsymbol{\sigma}_p)) \quad (21)$$

where $\tilde{\eta}_{prw}^{VAE}$ is used to denote importance sample $r = 1, \dots, R$ in chain $w = 1, \dots, W$. The means and log-standard deviations in distribution $q(\cdot)$ are functions of the previous layers comparable as in the encoder in Equation (13), that is:

$$\boldsymbol{\mu}_p = h_\mu(\mathbf{b}^{(L)} + \mathbf{A}^{(L)}\mathbf{z}_p^{(L-1)}) \quad (22)$$

$$\ln \boldsymbol{\sigma}_p = h_\sigma(\mathbf{d}^{(L)} + \mathbf{C}^{(L)}\mathbf{z}_p^{(L-1)}) \quad (23)$$

where $h_\mu(\cdot)$ is the encoding function for the means of η_p^{VAE} , and $h_\sigma(\cdot)$ is the encoding function for the log-transformed standard deviations of η_p^{VAE} . Note that in the fixed-effects autoencoder, it is the factor scores that are amortized, while for the variational autoencoder it is the mean and standard deviation of the approximate posterior of the factor scores being amortized.

Using the above, the observed data is decoded from $\tilde{\eta}_{prw}^{VAE}$ using

$$x'_{pirw} = \omega\left(-\tau_i^{VAE} + (\lambda_i^{VAE})^T \tilde{\eta}_{prw}^{VAE}\right). \quad (24)$$

where $\omega(\cdot)$ is commonly a logistic function. If the variational autoencoder parameter vector is given by

$$\boldsymbol{\theta}_{VAE} = \left[\tau_1^{VAE}, \dots, \tau_n^{VAE}, \lambda_1^{VAE}, \dots, \lambda_n^{VAE}, b^{(1)}, \dots, b^{(L)}, \mathbf{d}^{(1)}, \dots, \mathbf{d}^{(L)}, \text{vec}(\mathbf{A}^{(1)}), \dots, \text{vec}(\mathbf{A}^{(L)}), \text{vec}(\mathbf{C}^{(1)}), \dots, \text{vec}(\mathbf{C}^{(L)})\right]$$

then the estimates for these parameters are obtained by maximizing the importance weighted estimate of the likelihood, $p(\mathbf{x}_p)$, that is:

$$H = \sum_{p=1}^N \frac{1}{W} \sum_{w=1}^W \left(\log \frac{1}{R} \sum_{r=1}^R \frac{p(\tilde{\eta}_{prw}^{VAE}|\mathbf{x}_p)}{q(\tilde{\eta}_{prw}^{VAE}|\mathbf{x}_p)} p(\mathbf{x}_p) \right). \quad (25)$$

where $p(\tilde{\eta}_{prw}^{VAE}|\mathbf{x}_p)$ is the true posterior evaluated at samples $\tilde{\eta}_{prw}^{VAE}$ but which does not need to be evaluated explicitly as $p(\tilde{\eta}_{prw}^{VAE}|\mathbf{x}_p)p(\mathbf{x}_p) = p(\tilde{\eta}_{prw}^{VAE}, \mathbf{x}_p)$. For $R = 1$, the model above is referred to as variational autoencoder (Kingma & Welling, 2013) and is studied as an item factor model by Cúri et al. (2019), Converse et al. (2019), and Converse et al. (2021). For $R > 1$, the model is referred to as importance weighted variational autoencoder (Burda et al., 2015) and is studied as item factor model by Urban and Bauer (2021). If $R \rightarrow \infty$, Equation (25) is equivalent to marginal maximum log-likelihood estimation (see Appendix A from Burda et al., 2015).

The autoencoder as studied here can be obtained from the (importance weighted) variational autoencoder by two restrictions: First, by fixing $\boldsymbol{\sigma}_p$ to 0 for all p , the posterior of η_p^{VAE} is a discrete distribution of the $2^n \times K$ possible values in $\boldsymbol{\mu}_p$. As a result, we can use $R = 1$ and $W = 1$ as each sample from $q(\eta_p^{VAE}|\mathbf{x}_p)$ will be identical for each p . In such a case, Equation (25) reduces to:

$$H = \sum_{p=1}^N \log \left(p(\mathbf{x}_p|\tilde{\eta}_p^{VAE}) \right) - KL \left(q(\tilde{\eta}_p^{VAE}|\mathbf{x}_p) || p(\tilde{\eta}_p^{VAE}) \right) \quad (26)$$

where $KL(\cdot)$ is the Kullback–Leibler divergence. That is, $\tilde{\eta}_p^{VAE}$ is still constrained by a normal prior, $p(\eta_p^{VAE})$. That is, at this point, the model in Equation (26) is an autoencoder with a normal prior constraint on η_p^{VAE} . Therefore, a final step in obtaining the autoencoder as used in the present study, is to omit the prior distribution on η_p^{VAE} after which H is equivalent to $-H(\mathbf{X}, \mathbf{X}')$ in Equation (10).

Simulation study

A key advantage of the autoencoder proposed here is that it, contrary to constrained/regularized joint maximum likelihood, marginal maximum likelihood and the variational autoencoder, it precludes distributional assumptions about the model parameters. Therefore, in the simulation study below, we study the parameter recovery of the amortized joint maximum likelihood estimator as compared to the approaches above under normal and non-normal factor distributions.

Design

Following Chen et al. (2019) we incorporate conditions in our design that approximate the single and double asymptotic situation. That is, we use a confirmatory 3 factor model with a simple structure, binary item scores, and either $n_q = 30$ items per factor (reflecting a more practical setting) or $n_q = 100$ items per factor (reflecting an approximate asymptotic setting). The sample size is either $N = 1000$ (practical setting) or $N = 10,000$ (approximate asymptotic setting). The binary item scores are generated according to the item factor model in Equations (1) and (2) and where, for all q , η_{pq} follows either a multivariate normal distribution, a lognormal distribution with $\mu = 0$ and $\sigma = 0.4$, or a (bimodal) normal mixture distribution with $\mu_1 = -1$, $\mu_2 = 2$, $\sigma_1 = 1$, $\sigma_2 = 0.5$, and class probabilities $\pi_1 = 0.8$ and $\pi_2 = 0.2$. The correlations among the dimensions are 0.4. See Figure 2 for the resulting distribution of the first dimension in the different scenarios. In simulating the data, the true parameter values are fixed for both the item parameters and the person parameters. The factor loadings are fixed to $\lambda_{iq} = 1$ for all items loading on factor q , and the thresholds τ_i are fixed to equally spaced increasing values between -3 and 3 (normal condition), -5 and 2 (lognormal condition), or -2.5 and 2.5 (bimodal condition). The true η_p parameter values are obtained by the quantile functions of the factor distribution in the corresponding conditions (i.e., normal, lognormal, normal mixture) evaluated on a vector of increasing equally spaced numbers in the interval $(0, 1)$. The resulting true η_p are standardized to facilitate comparisons across different approaches (see below). We conducted 100 replications of the above design.

Models and estimation

Autoencoder

As discussed above, the configuration of the encoder should be done with care as the resulting function

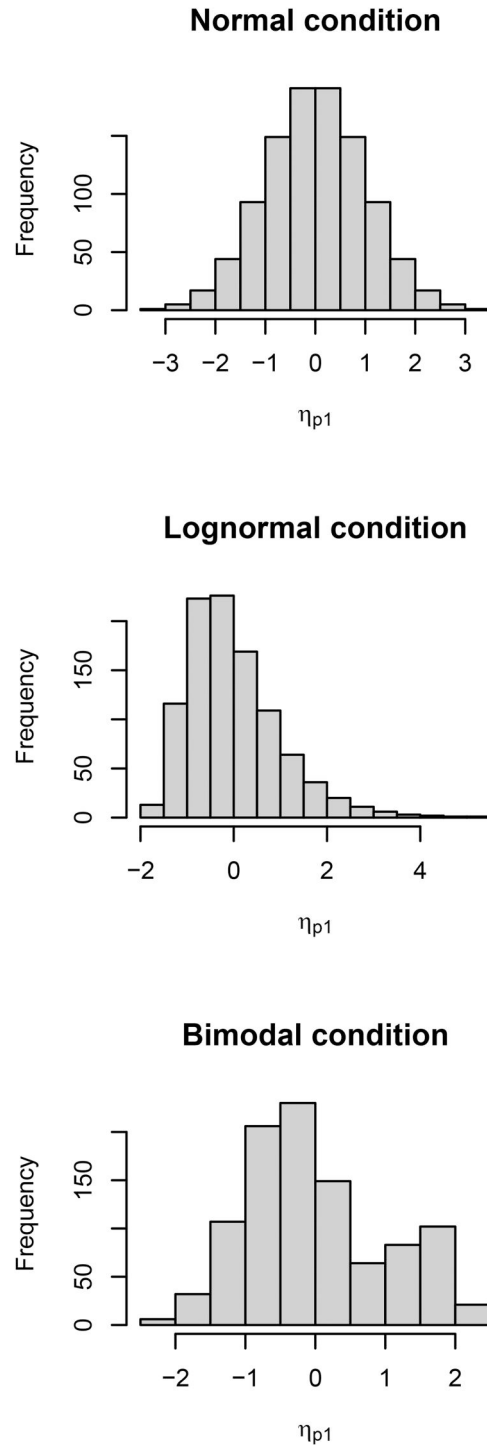


Figure 2. Latent variable distributions used in the different conditions in the simulation study. In this figure $N = 1000$.

for the amortization in Equation (13) needs to be flexible enough but should not result in overfitting. Urban and Bauer (2021) provide useful recommendations which they base on Heaton (2008) and that generally work well. Specifically, following the work by Urban and Bauer into the variational autoencoder item factor model (see above), we use $L = 2$ with an elu-function for $h^{(1)}(\cdot)$, a linear function for $h^{(2)}(\cdot)$,

and $Q^{(1)} = \text{floor}(\frac{1}{2}n + \frac{1}{2}Q^{(2)})$. From the simulation study results below, it turns out that in the conditions considered, these choices are satisfactory. In addition, in the real data example below, we illustrate how the robustness of these choices can be studied. We find that, for the data at hand, results are not affected by changes in the encoder. Note that $Q^{(2)} = K = 3$ which is the dimensionality of the factor model used in the present study.

In the decoder, we use a logistic activation function (instead of the normal ogive function in Equation 9) so that the final model becomes:

$$x'_{pi} = \omega \left(-\tau_i^{AE} + (\lambda_i^{AE})^T \eta_p^{AE} \right) \quad (27)$$

with η_p^{AE} being amortized by

$$\eta_p^{AE} = \mathbf{b}^{(2)} + \mathbf{A}^{(2)} \text{elu}(\mathbf{b}^{(1)} + \mathbf{A}^{(1)} \mathbf{x}_p) \quad (28)$$

where $\mathbf{A}^{(2)}$ is $n \times Q^{(1)}$ -dimensional, $\mathbf{b}^{(2)}$ is $Q^{(1)}$ -dimensional, $\mathbf{A}^{(1)}$ is $Q^{(1)} \times 3$ -dimensional, and $\mathbf{b}^{(1)}$ is 3-dimensional. See Figure 3 for a graphical representation of the above configuration of the amortized item factor model for $n_q = 30$.

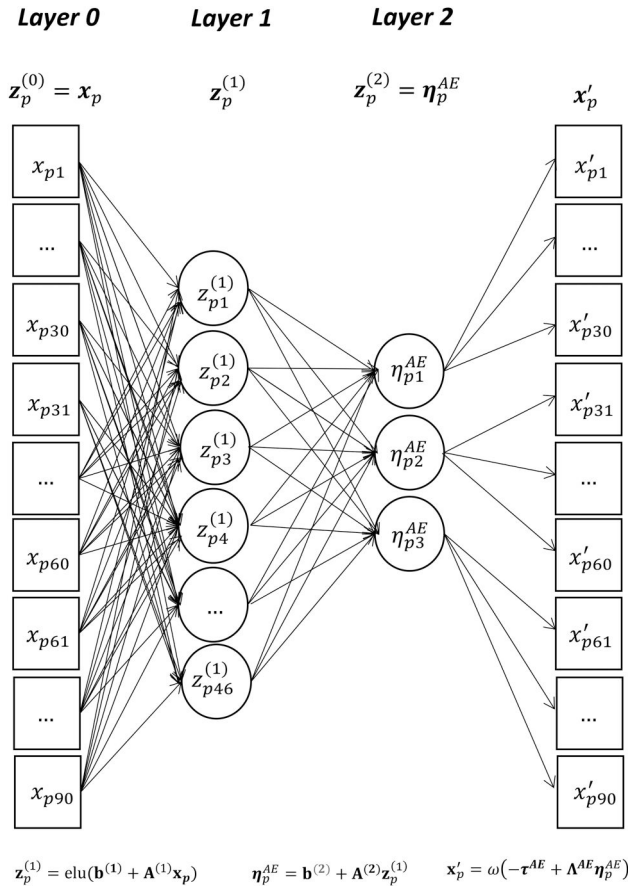


Figure 3. Graphical representation of the configuration of the autoencoder item factor model as used in the simulation study in the condition of $n_q = 30$.

We implemented this autoencoder (AE) using the Python package “Keras” (Chollet et al., 2015). The script to fit the model is available from the website of the first author. We used the AMSgrad algorithm (Reddi et al., 2019) in Equations (16) and (17) above to minimize Equation (10). We take a difference of $1e-8$ in H (Equation (10)) between consecutive iterations as an indication of convergence. In addition, we set $\beta_1 = 0.9$ and $\beta_2 = 0.999$ which can be considered default values (see Reddi et al., 2019; Urban & Bauer, 2021). In addition, we follow Urban and Bauer (2021) and set the learning rate $\alpha = 0.005$. We allowed for a maximum of 60,000 iterations, but all cases in all conditions converged far before this maximum without issues.

Other approaches

Besides the AE, we also fit the item factor model to the simulated data using: constrained joint maximum likelihood (cJML), marginal maximum likelihood (MML) with a normal prior distribution (nMML), MML with a prior distribution based on a Davidian curve of order 6 (dMML; Woods & Lin, 2009), and MML with the factor distribution approximated by a histogram as proposed by Bock and Aitkin (1981) and Woods (2007; wMML). We do not consider the importance weighted variational autoencoder as these estimates are close to the nMML estimates (for sufficiently large W and R , see above).

The cJML estimates are obtained using the R package “mirtjml” (Zhang et al., 2020) using $S = 5\sqrt{K} = 8.66$ in Equation 5 (which is the default value of the package) and with a tolerance of 0.001 (default of the package is 5 which may be too lenient for the present study). In addition, the nMML, dMML, and wMML estimates are obtained using the EM algorithm as implemented in the R package “mirt” (Chalmers, 2012), where dMML and wMML are applied to each dimension separately (as the full 3 dimensional factor model is numerically too demanding for these approaches).

Factor scores

An important aspect of the present study is to see how the factor scores, η_p , are being recovered by the estimates of the autoencoder as compared to the other approaches. For cJML, the factor scores are model parameters that are estimated directly during model estimation. For the MML approaches, the factor scores are estimated ad-hoc using expected a posteriori estimation with quasi Monte Carlo integration. For the autoencoder, factor scores can be obtained by the

non-linear transformation of the observed item scores in Equation (28).

In the “mirtjml” package, the cJML estimates are transformed to a scale comparable to that of MML. However, as also discussed above with respect to identification, the parameters from the AE are on a different scale due to Equations (14) and (15). We therefore used the transformations above (see paragraph ‘alternative scales’, but using the true parameter values to allow potential parameter bias to still be visible) to transform the AE estimates to that same scale.

Results

In the below, we discuss the results of the simulation study. We focus on the results with respect to the first factor. In tables, we present results with respect to the mean absolute bias and the root mean square error (RMSE) for all conditions. In addition, in figures, we visualize the results for the conditions with $N = 1000$ and $n_q = 30$ to demonstrate the effects of the prior constraints (MML) and the constraints on the parameters norms (cJML) on the estimates. Generally, these effects diminish for increasing N and n_q .

Factor scores

Table 1 depicts the mean absolute bias and RMSE across the conditions in the simulation study for the factors score estimates of the first factor. As can be seen the AE outperforms all other methods in terms of absolute bias at the expense of a larger RMSE. For the MML based approaches and cJML, the bias in the estimates is related to a shrinkage effect. See Figure 4 which displays the errors of the factor score estimates

of the first dimension in the condition $N = 1000$ and $n_q = 30$. The errors are ordered on the true factor score values. As can be seen, for the MML based approaches and for cJML, values in the lower tail of the factor score distribution are overestimated, and values in the upper tail are underestimated. This shrinkage effect is common in models like these and is due to the prior distribution (MML) and due to the constraints on the norm (cJML). This effect diminishes for both MML and cJML if the number of items increase. Most importantly, the AE does not suffer from such an effect. For cJML, there is bias in the log-normal condition which is related to the shrinkage effect of the thresholds (which are pulled to 0, see below).

Factor loadings

Table 2 depicts the mean absolute bias and the RMSE across the conditions in the simulation study for the factor loading estimates of the items loading on the first factor. For these estimates, the non-normal MML approaches outperform the other approaches in both the absolute bias and RMSE. The AE outperforms cJML and nMML in terms of both the absolute bias and RMSE. See Figure 5 which displays the errors of the factor loadings for the items that load on the first dimension in the condition $N = 1000$ and $n_q = 30$. The errors are ordered on the true item threshold value. As can be seen wMML and AE are relatively robust for the different distributions, while the nMML estimates are biased in the lognormal and bimodal conditions. The cJML estimates show some bias in all conditions, due to the constraint on the norm of the loadings which pulls the estimates toward 0.

Table 1. Mean absolute bias and Root Mean Squared Error (RMSE) for the factor score estimates over replications and over subjects for the first dimension in the Normal condition (N), the Lognormal condition (LN) and the bimodal normal condition (BN).

Data	n_q	N	Mean absolute bias					RMSE				
			AE	cJML	nMML	dMML	wMML	AE	cJML	nMML	dMML	wMML
N	30	1000	0.049	0.088	0.153	0.151	0.151	0.511	0.453	0.435	0.434	0.434
	30	10,000	0.055	0.089	0.153	0.152	0.152	0.509	0.450	0.434	0.433	0.433
	100	1000	0.022	0.033	0.031	0.053	0.054	0.269	0.260	0.273	0.256	0.256
	100	10,000	0.025	0.034	0.031	0.054	0.054	0.269	0.258	0.270	0.255	0.255
LN	30	1000	0.080	0.415	0.170	0.162	0.164	0.584	0.655	0.472	0.464	0.463
	30	10,000	0.085	0.538	0.172	0.162	0.163	0.577	0.732	0.472	0.463	0.462
	100	1000	0.031	0.302	0.046	0.055	0.062	0.300	0.422	0.303	0.283	0.281
	100	10,000	0.031	0.529	0.047	0.056	0.062	0.297	0.607	0.302	0.282	0.281
BN	30	1000	0.052	0.087	0.143	0.136	0.140	0.489	0.439	0.419	0.417	0.417
	30	10,000	0.053	0.086	0.144	0.137	0.139	0.481	0.435	0.417	0.416	0.416
	100	1000	0.024	0.033	0.032	0.036	0.049	0.258	0.249	0.266	0.246	0.245
	100	10,000	0.025	0.047	0.035	0.036	0.048	0.254	0.252	0.263	0.245	0.244

n_q is the number of items per factor. In addition, AE: amortized joint maximum likelihood using the autoencoder; cJML: constrained joint maximum likelihood; nMML, dMML, wMML: marginal maximum likelihood using respectively a normal prior, a prior based on a Davidian curve of order 6, and an empirical histogram. For dMML and wMML, the results are obtained by fitting unidimensional models to each dimension separately.

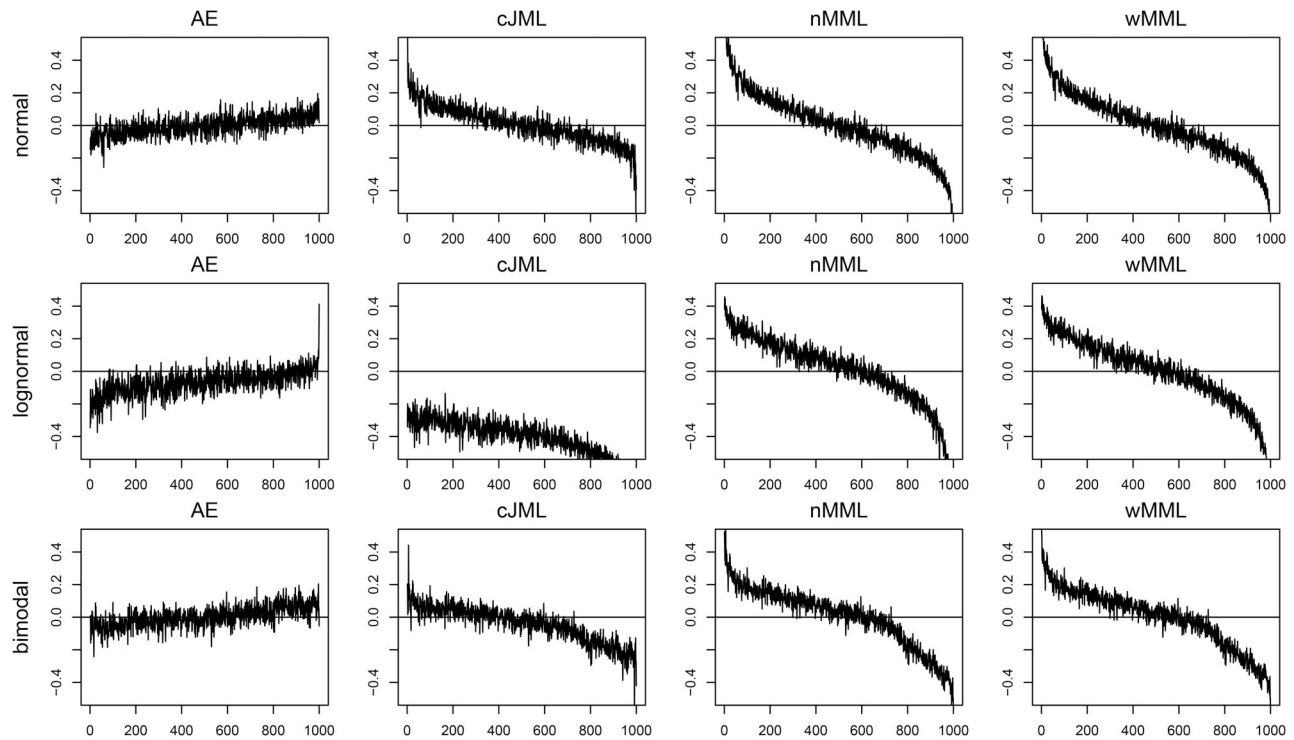


Figure 4. Plot of the errors of the factor score estimates of the first dimension for the condition with 30 items per factor and 1000 subjects. The errors on the y-axis are ordered according to their true factor scores on the x-axis. AE: amortized joint maximum likelihood using the autoencoder; cJML: constrained joint maximum likelihood; nMML, wMML: marginal maximum likelihood using respectively a normal prior and a prior based on a Davidian curve of order 6. Results for the histogram approach are not displayed as these are highly similar to the wMML results.

Table 2. Mean absolute bias and Root Mean Squared Error (RMSE) for the factor loadings estimates over replications and over subjects for the first dimension in the Normal condition (N), the Lognormal condition (LN) and the bimodal normal condition (BN).

Data	n_q	N	Mean absolute bias					RMSE				
			AE	cJML	nMML	dMML	wMML	AE	cJML	nMML	dMML	wMML
N	30	1,000	0.013	0.157	0.010	0.010	0.010	0.133	0.210	0.117	0.120	0.120
	30	10,000	0.009	0.148	0.003	0.003	0.003	0.043	0.157	0.037	0.038	0.038
	100	1,000	0.012	0.066	0.082	0.009	0.010	0.113	0.138	0.131	0.110	0.110
	100	10,000	0.004	0.050	0.086	0.003	0.003	0.035	0.062	0.092	0.034	0.034
LN	30	1,000	0.031	0.227	0.142	0.028	0.018	0.163	0.310	0.211	0.142	0.142
	30	10,000	0.020	0.196	0.134	0.022	0.004	0.053	0.206	0.143	0.048	0.042
	100	1,000	0.014	0.073	0.059	0.011	0.011	0.121	0.153	0.135	0.117	0.119
	100	10,000	0.008	0.058	0.062	0.012	0.003	0.041	0.072	0.077	0.040	0.038
BN	30	1,000	0.012	0.134	0.046	0.010	0.011	0.120	0.188	0.116	0.108	0.109
	30	10,000	0.005	0.143	0.044	0.009	0.004	0.040	0.151	0.057	0.037	0.036
	100	1,000	0.009	0.045	0.094	0.019	0.010	0.106	0.114	0.133	0.104	0.104
	100	10,000	0.003	0.046	0.102	0.023	0.002	0.032	0.057	0.106	0.039	0.031

n_q is the number of items per factor. In addition, AE: amortized joint maximum likelihood using the autoencoder; cJML: constrained joint maximum likelihood; nMML, dMML, wMML: marginal maximum likelihood using respectively a normal prior, a prior based on a Davidian curve of order 6, and an empirical histogram. For dMML and wMML, the results are obtained by fitting unidimensional models to each dimension separately.

Thresholds

Table 3 depict the mean absolute bias and the RMSE across the conditions in the simulation study for the threshold estimates of the items loading on the first factor. The non-normal MML approaches again outperform the other approaches in terms of both absolute bias and RMSE. In addition, the nMML outperforms the AE and cJML generally in terms of both absolute bias and RMSE, although the AE has a

slightly smaller absolute bias in the log-normal condition. The cJML approach in turn, has smaller absolute bias and RMSE as compared to the AE in the normal and bimodal conditions with a smaller number of items. The AE outperforms the cJML approach in the case of a larger number of items and in all cases with a lognormal distribution for the factor scores. The cJML estimates of the thresholds are systematically biased in the lognormal condition. This is due to the

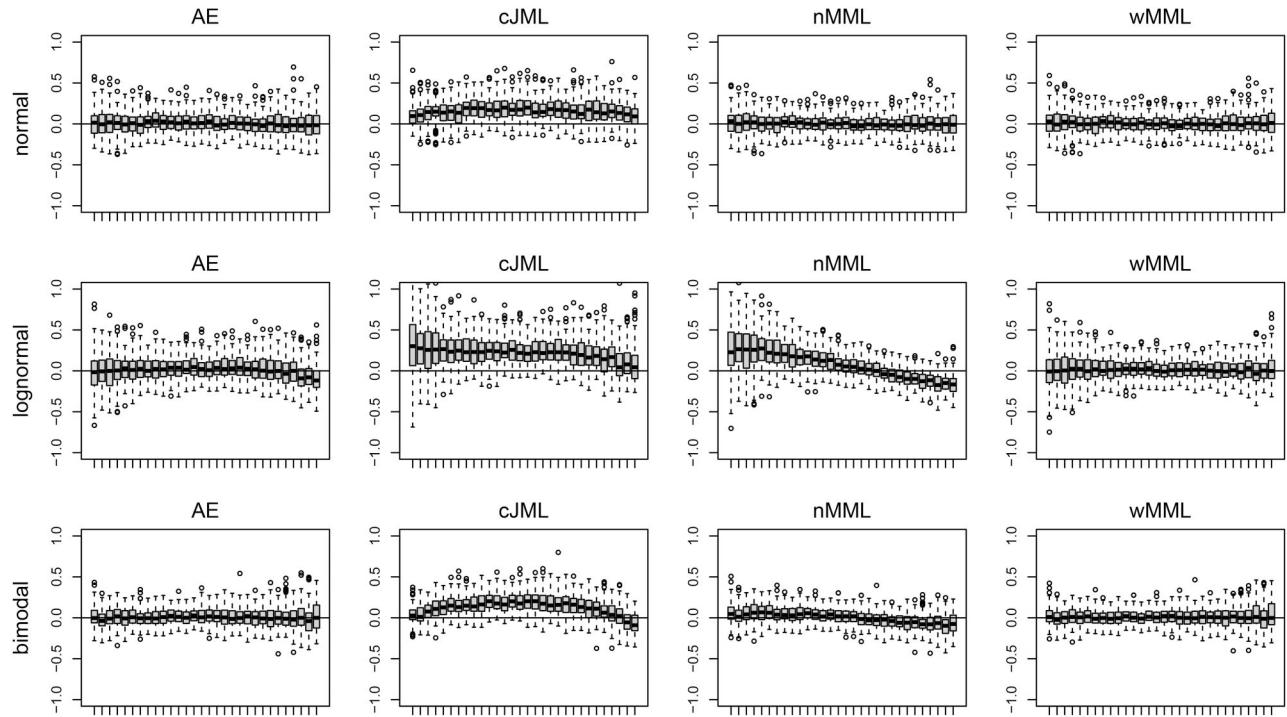


Figure 5. Boxplot of the errors of the factor loading estimates of the items loading on the first dimension for the condition with 30 items per factor and 1000 subjects. The errors on the y-axis are ordered according to the item difficulty (true threshold value) on the x-axis. AE: amortized joint maximum likelihood using the autoencoder; cJML: constrained joint maximum likelihood; nMML, wMML: marginal maximum likelihood using respectively a normal prior and a prior based on a Davidian curve of order 6. Results for the histogram approach are not displayed as these are highly similar to the wMML results.

Table 3. Mean absolute bias and Root Mean Squared Error (RMSE) for the threshold estimates over replications and over subjects for the first dimension in the Normal condition (N), the Lognormal condition (LN) and the bimodal normal condition (BN).

Data	n_q	N	Mean absolute bias					RMSE				
			AE	cJML	nMML	dMML	wMML	AE	cJML	nMML	dMML	wMML
N	30	1000	0.076	0.053	0.010	0.011	0.011	0.145	0.128	0.107	0.109	0.109
	30	10,000	0.072	0.047	0.002	0.002	0.002	0.082	0.063	0.033	0.033	0.033
	100	1000	0.026	0.050	0.007	0.009	0.009	0.110	0.131	0.110	0.104	0.104
	100	10,000	0.022	0.022	0.003	0.003	0.003	0.041	0.042	0.036	0.033	0.033
LN	30	1000	0.077	0.381	0.104	0.030	0.021	0.247	0.486	0.205	0.165	0.162
	30	10,000	0.071	0.562	0.088	0.016	0.004	0.103	0.574	0.102	0.050	0.047
	100	1000	0.035	0.270	0.054	0.023	0.018	0.195	0.355	0.173	0.158	0.156
	100	10,000	0.020	0.529	0.039	0.009	0.004	0.064	0.540	0.064	0.048	0.047
BN	30	1000	0.061	0.054	0.022	0.009	0.010	0.127	0.113	0.100	0.099	0.100
	30	10,000	0.058	0.050	0.022	0.003	0.003	0.068	0.064	0.037	0.030	0.030
	100	1000	0.022	0.016	0.014	0.009	0.009	0.101	0.094	0.107	0.096	0.097
	100	10,000	0.017	0.040	0.018	0.006	0.003	0.035	0.064	0.037	0.030	0.030

n_q is the number of items per factor. In addition, AE: amortized joint maximum likelihood using the autoencoder; cJML: constrained joint maximum likelihood; nMML, dMML, wMML: marginal maximum likelihood using respectively a normal prior, a prior based on a Davidian curve of order 6, and an empirical histogram. For dMML and wMML, the results are obtained by fitting unidimensional models to each dimension separately.

true thresholds having a mean smaller than 0 in this condition, while the cJML constraints on the norm of the threshold parameters pull the estimates to 0. See Figure 6 which displays the errors of the factor loadings for the items that load on the first dimension in the condition $N = 1000$ and $n_q = 30$. The errors are ordered on the true item threshold value.

It should be noted that for the AE, the recovery of the threshold parameters is somewhat worse for smaller N and n as compared to the recovery of the

discrimination parameters which is not typical for MML based estimation of item factor models. This is due to the slight bias in the factor scores (see Table 1) which is -due to the absence of a prior- compensated in the thresholds.

Additional simulations

To see if the pattern of results above holds for a smaller sample size and for smaller item numbers, we

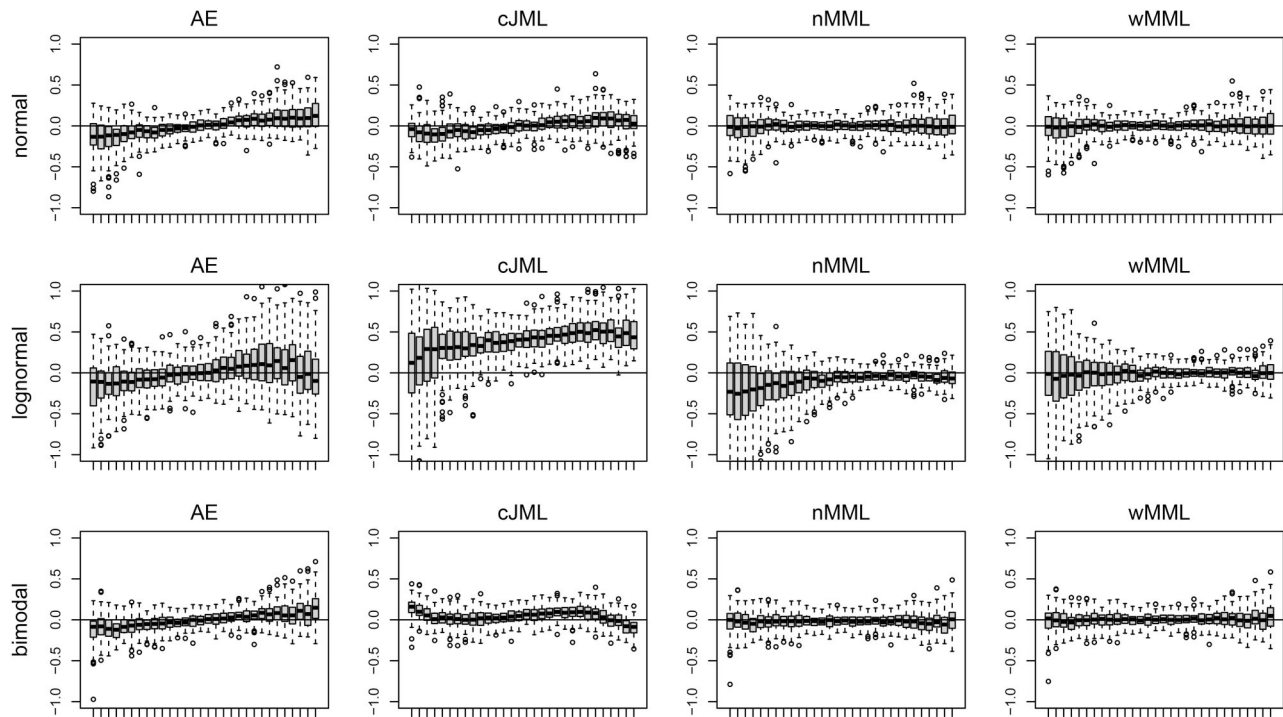


Figure 6. Boxplot of the errors of the threshold estimates of the items loading on the first dimension for the condition with 30 items per factor and 1000 subjects. The errors on the y-axis are ordered according to the item difficulty (true threshold value) on the x-axis. AE: amortized joint maximum likelihood using the autoencoder; cJML: constrained joint maximum likelihood; nMML, wMML: marginal maximum likelihood using respectively a normal prior and a prior based on a Davidian curve of order 6. Results for the histogram approach are not displayed as these are highly similar to the wMML results.

ran additional simulations using $N = 500$ and n_q equal to either 10, 15, 20, and 30. All other settings were the same as in the above, except that we decreased the number of nodes in the first hidden layer of the encoder, $Q^{(1)}$ to respectively 3, 5, and 8 for 10, 15, and 20 items respectively (for $n_q = 30$ we used the same setting as above) as for these settings the configuration above inflated parameter bias.

Results are in Table 4. As can be seen, for the factor scores $n_q = 10$ is already enough to outperform cJML and nMML in the normal and bimodal conditions in terms of absolute bias. However, for the lognormal condition, at least $n_q = 15$ is needed to reach a smaller absolute bias. For the factor loadings and thresholds, results indicate that the AE is generally associated with a smaller absolute bias for the factor loadings as compared to cJML, while cJML is generally associated with a smaller absolute bias for the thresholds as compared to the AE. However, both joint estimation approaches perform worse as compared to MML in the case of $N = 500$, except for $n_q = 30$, the AE has a smaller absolute bias in the non-normal conditions compared to nMML.

Finally, we also considered $N = 1000$ (results are on the website of the first author) which indicated that for this sample size, 10 items and 15/20 items are

sufficient for the AE to outperform cJML and nMML in terms of absolute bias of respectively the factor scores and factor loadings.

Conclusion/recommendation

The present results showed that, for factor score estimation, the AE generally has the smallest absolute bias, in particular in practical situation where n and N are modest. This is due to the other approaches suffering from shrinkage effects. This advantage of the AE comes at the expense of an increased RMSE.

For the item parameters, the non-normal MML approaches (dMML and wMML) generally perform best in all conditions in terms of absolute bias and RMSE. The AE produces less biased item parameter as compared to nMML and cJML in the lognormal condition and for sufficiently large sample sizes (at least 1000), but not in the normal and bimodal conditions. Thus, overall, it is recommendable that if the factor score estimation is the main aim of the study, the AE can be preferred over normal and non-normal MML or cJML approaches. If item parameter estimation is the main aim of the study, non-normal MML methods should be preferred. However, it should be noted that currently these approaches can only be

Table 4. Mean absolute bias and Root Mean Squared Error (RMSE) for the estimates over replications and over subjects for the first dimension in the Normal condition (N), the Lognormal condition (LN) and the bimodal normal condition (BN) for $N = 500$ and 10, 15, 20, or 30 items.

Parameter	n_q	Data	Mean absolute bias			RMSE		
			AE	cJML	nMML	AE	cJML	nMML
Factor scores	10	N	0.110	0.284	0.343	1.132	0.799	0.639
		LN	0.396	0.370	0.358	1.393	0.950	0.652
		BN	0.090	0.270	0.322	1.005	0.790	0.619
	15	N	0.085	0.197	0.258	0.796	0.811	0.561
		LN	0.242	0.288	0.280	1.383	0.904	0.589
		BN	0.087	0.228	0.243	0.735	0.739	0.544
	20	N	0.069	0.138	0.209	0.648	0.599	0.509
		LN	0.109	0.202	0.231	0.902	0.806	0.542
		BN	0.070	0.139	0.199	0.613	0.587	0.493
	30	N	0.050	0.093	0.152	0.515	0.457	0.435
		LN	0.074	0.280	0.171	0.587	0.697	0.473
		BN	0.049	0.087	0.144	0.490	0.442	0.422
Factor loadings	10	N	1.953	2.583	0.026	13.26	4.538	0.233
		LN	1.342	2.823	0.261	4.842	5.058	0.452
		BN	0.495	2.630	0.096	3.630	5.318	0.216
	15	N	0.070	1.227	0.019	0.570	4.081	0.197
		LN	1.575	1.476	0.193	8.575	4.240	0.345
		BN	0.070	1.136	0.075	0.513	3.922	0.199
	20	N	0.031	0.425	0.019	0.241	1.586	0.185
		LN	0.387	0.968	0.170	2.784	3.209	0.306
		BN	0.024	0.356	0.060	0.213	1.049	0.177
	30	N	0.022	0.179	0.016	0.193	0.293	0.169
		LN	0.038	0.244	0.137	0.234	0.701	0.268
		BN	0.020	0.170	0.047	0.180	0.273	0.164
Thresholds	10	N	0.465	0.141	0.034	1.091	0.414	0.187
		LN	1.237	0.307	0.252	2.415	0.721	0.534
		BN	0.375	0.108	0.045	0.908	0.418	0.159
	15	N	0.207	0.132	0.027	0.429	0.654	0.173
		LN	0.568	0.234	0.172	2.115	0.755	0.391
		BN	0.174	0.169	0.034	0.345	0.533	0.148
	20	N	0.140	0.118	0.017	0.259	0.360	0.159
		LN	0.320	0.163	0.145	1.396	0.711	0.327
		BN	0.114	0.106	0.031	0.222	0.353	0.142
	30	N	0.087	0.080	0.017	0.198	0.200	0.154
		LN	0.099	0.259	0.118	0.356	0.595	0.322
		BN	0.072	0.067	0.020	0.175	0.179	0.137

AE: amortized joint maximum likelihood using the autoencoder; cJML: constrained joint maximum likelihood; nMML: marginal maximum likelihood using a normal prior.

applied to unidimensional models, which may be undesirable in some situations. For the approaches that do take the full item factor model into account (nMML, AE, and cJML), the AE can be preferred in skewed settings if the sample size is large enough, and cJML and nMML can be preferred in normal settings (including bimodal).

Applications

Application 1

Unidimensional binary item factor model

Data and model. We analyze the scores on the 40 “choose a move A” items from the Amsterdam Chess Test (Van Der Maas & Wagenmakers, 2005). In each

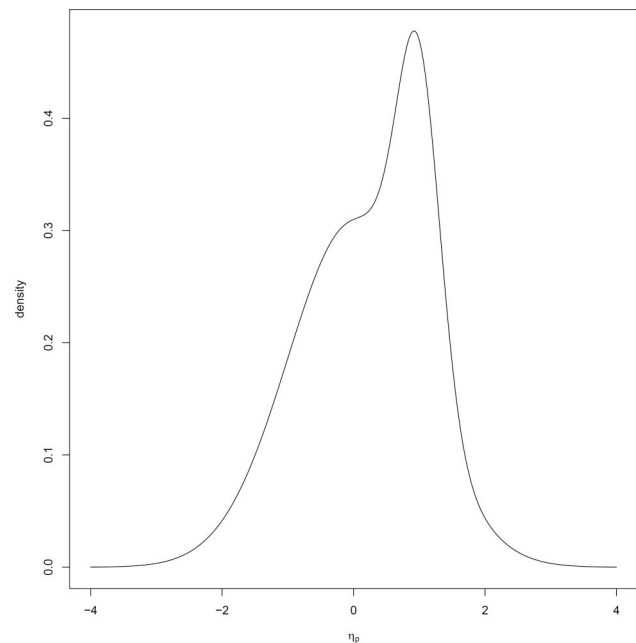


Figure 7. Model implied latent variable distribution by a two-class mixture latent variable model.

item, a configuration of chess pieces on the chess board is depicted. Subjects were instructed to select the next best move. Scores are correct (1) or incorrect (0). The sample consist of 234 amateur and professional chess players. As a result, due to this heterogeneity, the latent variable distribution is not expected to be normal. A mixture analysis confirms this: A two-class mixture latent variable model seems to fit somewhat better (AIC: 7885; BIC: 8172; sample size adjusted BIC: 7909) than a baseline latent variable model without mixtures (AIC: 7888; BIC: 8165; sample size adjusted BIC: 7911) at least according to the AIC and the sample size adjusted BIC. The latent variable mean and variance in the first class were fixed for identification reasons at a mean 0 and a variance of 1. The class size for this first class was estimated to be 0.233. For the second class, the class size was estimated to be 0.767 with a latent variable mean and variance of respectively 0.992 and 0.105. Thus, these two classes can represent a relatively homogeneous subgroup of professional chess players and a more heterogeneous subgroup of amateur chess players. See Figure 7 for a plot of the implied latent variable distribution by the mixture model. As can be seen, the distribution is negatively skewed.

To the data we fit a unidimensional item factor model using the AE, cJML, nMML, and the importance weighted autoencoder (iwVAE). We use the same procedure as in the simulation study with identical settings for the optimization algorithms. For the

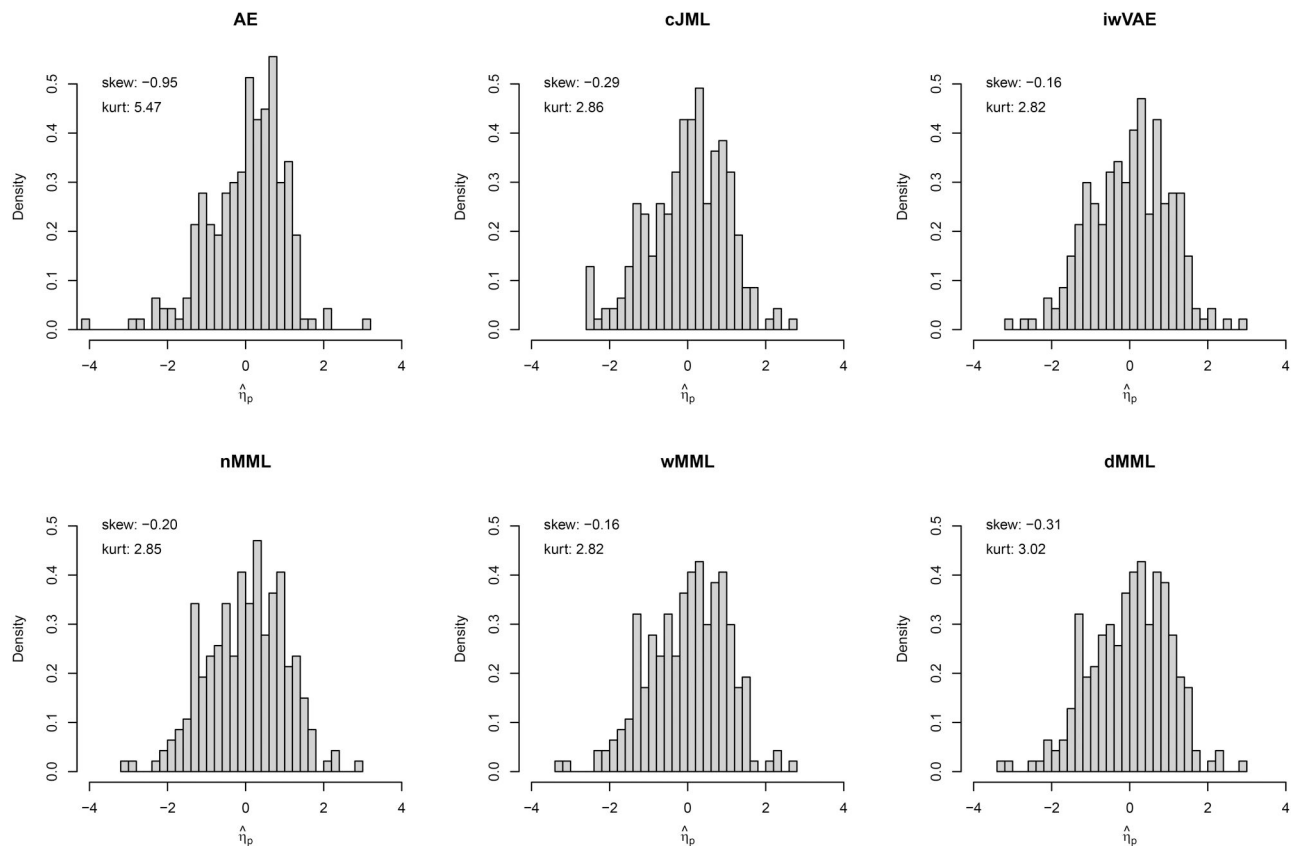


Figure 8. Histograms of the factor score estimates across the different approaches.

iwVAE we used the DeepIRTools Python package (Urban & He, 2022). We compare the results of the different approaches in the light of the simulation study results.

Results. First, the item parameters hardly show any differences between the different approaches with correlations between 0.95 and 0.99 for the factor loadings and with correlations of 0.99 and 1.00 for the threshold parameters. For the factor score estimates the results are more interesting: See Figure 8 for a histogram of the standardized factor score estimates in the different approaches. It seems that the estimates from the AE and cJML capture the heterogeneity in the chess data due to expertise better than MML and the iwVAE where the AE seems to outperform cJML in the sense that the factors scores for the AE are closer to the mixture distribution found in Figure 7. This is also confirmed by an actual mixture analysis of the standardized estimated factor scores, see Table 5. That is, for the AE estimates, a one-component normal mixture distribution is rejected in favor of a two-component normal mixture by both the AIC and BIC. For the factor scores obtained using MML and iwVAE, the one-component normal mixture model cannot be

rejected, while for cJML, the results are mixed as the AIC favors the two-component normal mixture and the BIC favors the one-component normal mixture. Thus, due to the normal prior used in MML and the iwVAE, the factor score estimates are pushed into a normal distribution, by which it is masked that the distribution is better represented by a mixture distribution. This is more clear from the AE factor score estimates.

Note that the procedure followed above is not one that we would recommend in practice (estimating the factor scores first, then conduct a mixture analysis on the estimates). In addition, there may be a better representation of the chess factor score distribution than a two-component normal mixture. However, these analyses are intended to illustrate the benefits of the AE in providing distribution free estimates of the factor scores underlying a psychometric dataset.

Application 2

Multidimensional ordinal item factor model

Data and model. Here we analyze a large scale personality assessment. Specifically, we use the data on Cattell's 16 factor personality test as retrieved from

Table 5. Model fit and parameter estimates of the one- and two-component normal mixture model on the standardized factor scores estimates of the different approaches.

	Estimates two-component model					One-component		Two-components	
	π_1	μ_1	μ_2	σ_1	σ_2	AIC	BIC	AIC	BIC
AE	0.555	-0.410	0.512	1.115	0.463	667.061	673.972	643.458	660.734
cJML	0.261	-1.180	0.416	0.667	0.725	667.061	673.972	664.320	681.597
iwVAE	0.507	-0.630	0.649	0.821	0.705	667.061	673.972	670.242	687.519
nMML	0.482	-0.670	0.625	0.820	0.699	667.061	673.972	669.520	686.797
wMML	0.804	-0.202	0.827	0.995	0.408	667.061	673.972	664.465	681.742
dMML	0.797	-0.203	0.797	0.999	0.434	667.061	673.972	665.277	682.553

π_1 probability of class one; μ_c and σ_c mean and standard deviation of class $c = 1, 2$. In addition, note that as we analyze the standardized estimates, the AIC and BIC are the same for all approaches in the one-component model.

openpsychometrics.org. These data consist of the responses of 49,159 subjects to 163 personality items on a 5 point Likert scale. According to Cattell's theory, these items measure 16 factors according to a simple structure.

To these data we fit a 16 dimensional ordinal item factor model using an AE for ordinal data as explained above. We use three different configurations of the encoder in Equation (13) to verify that the configuration that we have proposed in this paper is relatively robust for the choices made. That is, we focus on:

AE1: the autoencoder with 2 layers as described above and as studied in the simulation study;

AE2: similar as AE1 but with halve the number of nodes in layer 1;

AE3: similar as AE1 but with 3 layers, where layer 1 and layer 2 have the same configuration as layer 1 from AE1, and layer 3 has the same configuration as layer 2 from AE1.

For these autoencoders, all estimation settings are the same as discussed above for the simulation study, except that we use a learning rate of $\alpha = 0.001$ which is recommendable for models with many dimensions and many items (see also Urban & Bauer, 2021).

In addition, we fit unidimensional ordinal item factor models using MML in R-package ltm (Rizopoulos, 2006) to each dimension separately (referred to again as nMML). We do so because we failed to find a stable solution for the full model (at least using stochastic imputation). We also fit unidimensional ordinal item factor models with a prior based on a Davidian curve using MML (dMML, similar as above) but results are comparable to the normal MML results, so these are not considered in the below.

Results. Table 6 contains the correlations among the factor score estimates across the different models for each dimension. As can be seen, across the autoencoders, results are highly comparable indicating that

Table 6. Correlations among the factor score estimates obtained using three autoencoders and univariate nMML for each dimension q of Cattell's 16 dimensional personality test.

q	η_{pq} correlations					
	AE1-AE2	AE1-AE3	AE2-AE3	AE1-nMML	AE2-nMML	AE3-nMML
1	0.997	0.995	0.994	0.991	0.990	0.987
2	0.997	0.994	0.993	0.982	0.981	0.978
3	0.997	0.997	0.996	0.993	0.990	0.992
4	0.998	0.996	0.996	0.989	0.990	0.984
5	0.994	0.996	0.995	0.988	0.980	0.985
6	0.998	0.997	0.996	0.986	0.987	0.983
7	0.998	0.998	0.997	0.995	0.995	0.994
8	0.841	0.908	0.955	0.988	0.843	0.899
9	0.999	0.997	0.997	0.992	0.994	0.991
10	0.996	0.996	0.995	0.971	0.965	0.970
11	0.995	0.997	0.995	0.992	0.992	0.990
12	0.998	0.997	0.997	0.993	0.993	0.992
13	0.992	0.995	0.991	0.979	0.970	0.974
14	0.998	0.996	0.993	0.985	0.983	0.985
15	0.957	0.996	0.952	0.989	0.976	0.986
16	0.998	0.997	0.995	0.990	0.992	0.985

AE1: The autoencoder with 2 layers as proposed in this study and as studied in the simulation study. AE2: An autoencoder with 2 layers, but with halve the nodes in layer 1 as compared to AE1. AE3: An autoencoder with 3 layers, see text.

the results are robust to the exact configuration of the encoder used.

Discussion

Methods from deep learning in general and the fixed-effects autoencoder and (importance weighted) variational autoencoder in particular, are promising tools for the field of psychometrics. Previous work on the variational autoencoder has already demonstrated the benefits of these methods in terms of less computation time and less numerical challenges for increasing dimensionality of the models (e.g., Urban & Bauer, 2021, Cúri et al., 2019). In this paper we illustrated how the fixed-effects autoencoder can provide distribution free amortized joint maximum likelihood estimates of item factor models. We showed that this approach is less biased in estimating factor scores as compared to other approaches. These benefits, however, come with the cost of more parameter variability.

The increased parameter variability is due to the fixed-effects autoencoder being free of direct restrictions of the parameter space, for instance by means of a normal prior distribution. Some existing work has been concerned with incorporating more flexible prior distributions in the item factor model including the skew-normal distribution (e.g., Azevedo et al., 2011; Molenaar et al., 2010; Smits et al., 2016), the Johnson distribution (van den Oord, 2005), and the log-beta distribution (Andersen & Madsen, 1977). In addition, semi-parametric approaches have been proposed based on mixture distributions (e.g., Haberman, 2005; Schmitt et al., 2006; Vermunt, 2004). Although these distributions are more flexible than the normal distribution used in marginal maximum likelihood and variational autoencoders, these existing approaches still impose restrictions on the factor score distribution which may result in shrinkage in the factor score estimates (as we will demonstrate in this study) or parameter bias if an incorrect shape is used (e.g., Swaminathan & Gifford, 1983; Zwinderman & van den Wollenberg, 1990). In addition, most of these approaches are computationally more demanding as compared to the present approach. Thus, in practice, it should thus be considered how comfortable one is with a given (non-)normal prior and how complex the resulting model becomes. If in doubt, the amortized methodology from this study can be of help.

In the simulation study, we explored the number of items for which the autoencoder still performed adequately. It turned out that, depending on the exact purpose and setting of the study, 10 or 15 items may be sufficient. However, for these smaller item numbers, we downsized the configuration of the encoder as the configuration based on the general recommendation by Urban and Bauer (2021) -which worked well in the current study for 30 items and 1000 subjects- turned out to inflate bias in the case of fewer items and subjects. Future research should thus focus on general rules for the setup of the encoder for different samples sizes, item numbers, and factor numbers, especially for smaller datasets.

In general, the practical benefits of using the fixed-effects autoencoder for estimating item factor models lie in its computational speed, its non-parametric nature, and its flexibility. Thus, the autoencoder seems an appealing choice for fitting high dimensional models to very large datasets (as illustrated in the second real data example where a 16 dimensional item factor model was fit on a dataset with over 50,000 subjects and 163 items), and/or for datasets for which heterogeneity is expected (as was illustrated in the first real

data example where two subgroups are more evident in the autoencoder factor scores as compared to the constrained approaches). In addition, in computerized adaptive settings, factor scores can be estimated by a straightforward evaluation of the encoder, without the need of additional algorithmic computations (e.g., maximizing a likelihood or sampling from the approximate posterior). In addition, the (variational) autoencoder framework naturally facilitates fitting more complex nonstandard item factor models with very flexible item characteristic curves by adding layers to the decoder part of the (variational) autoencoder.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The scripts to fit the autoencoder and the data of illustration 1 are available from www.dylanmolenaar.nl. The data from illustration 2 are available from <http://openpsychometrics.org>. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the author's institutions is not intended and should not be inferred.

References

- Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, 17(3), 251–269. <https://doi.org/10.3102/10769986017003251>
- Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society Series B*, 32(2), 283–301. <https://doi.org/10.1111/j.2517-6161.1970.tb00842.x>

- Andersen, E., & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika*, 42(3), 357–374. <https://doi.org/10.1007/BF02293656>
- Azevedo, C. L. N., Bolfarine, H., & Andrade, D. F. (2011). Bayesian inference for a skew-normal IRT model under the centred parameterization. *Computational Statistics & Data Analysis*, 55(1), 353–365. <https://doi.org/10.1016/j.csda.2010.05.003>
- Arrieta, B. A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bergner, Y., Halpin, P., & Vie, J. J. (2022). Multidimensional Item Response Theory in the Style of Collaborative Filtering. *Psychometrika*, 87(1), 266–288. <https://doi.org/10.1007/s11336-021-09788-9>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In E. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 17–20). Addison Wesley.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4), 443–459. <https://doi.org/10.1007/BF02293801>
- Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, 8(1), 1–24.
- Burda, Y., Grosse, R., & Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv:1509.00519*.
- Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins Monro algorithm. *Psychometrika*, 75(1), 33–57. <https://doi.org/10.1007/s11336-009-9136-x>
- Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Chen, Y., Li, X., & Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84(1), 124–146. <https://doi.org/10.1007/s11336-018-9646-5>
- Chollet, F., et al. (2015). Keras [computer software]. <https://keras.io>
- Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40(1), 5–32. <https://doi.org/10.1007/BF02291477>
- Converse, G., Curi, M., & Oliveira, S. (2019). Autoencoders for educational assessment. In: S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, R. Luckin (Eds.) *Artificial intelligence in education. AIED 2019. Lecture notes in computer science* (vol 11626, pp. 41–45). Springer.
- Converse, G., Curi, M., Oliveira, S., & Templin, J. (2021). Estimation of multidimensional item response theory models with correlated latent variables using variational autoencoders. *Machine Learning*, 110(6), 1463–1480. <https://doi.org/10.1007/s10994-021-06005-7>
- Cúri, M., Converse, G., Hajewski, J., & Oliveira, S. (2019). *Interpretable Variational Autoencoders for Cognitive Models* [Paper presentation]. 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1–8). <https://doi.org/10.1109/IJCNN.2019.8852333>
- De Jong, M. G., Steenkamp, J. B. E., Fox, J. P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, 45(1), 104–115. <https://doi.org/10.1509/jmkr.45.1.104>
- Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, 75(3), 474–497. <https://doi.org/10.1007/s11336-010-9161-9>
- Fox, J. P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66(2), 271–288. <https://doi.org/10.1007/BF02294839>
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
- Grasman, R. P. P. (2004). *Sensor array signal processing and the neuro-electromagnetic inverse problem in functional connectivity analysis of the brain* [Unpublished doctoral thesis]. University of Amsterdam.
- Guo, Q., Cutumisu, M., Cui, Y. (2017). A neural network approach to estimate student skill mastery in cognitive diagnostic assessments. *10th International Conference on Educational Data Mining*, 370–371.
- Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, 5(5), 815–841.
- Haberman, S. J. (2005). Latent–class item response models. *ETS Research Report Series*, 2005(2), 1–7. <https://doi.org/10.1002/j.2333-8504.2005.tb02005.x>
- Heaton, J. (2008). *Introduction to neural networks for Java* (2nd ed.). Heaton Research Inc.
- Kelderman, H., & Rijkes, C. P. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, 59(2), 149–176. <https://doi.org/10.1007/BF02295181>
- Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, 65(4), 457–474. <https://doi.org/10.1007/BF02296338>
- Kline, P. (2013). *Handbook of psychological testing*. Routledge.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer.
- Lawley, D. N. (1943). The application of the maximum likelihood method to factor analysis. *British Journal of Psychology*, 33(3), 172–175.
- Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 48(3), 936–949. <https://doi.org/10.3758/s13428-015-0619-7>
- Lord, F. M. (1952). *A theory of test scores*. Psychometric Society.
- Marino, J., Yue, Y., & Mandt, S. (2018). *Iterative Amortized Inference*. Proceedings of the 35th International

- Conference on Machine Learning (vol. 80, pp. 3403–3412).
- Martin, J. K., & McDonald, R. R. (1975). Bayesian estimation in unrestricted factor analysis; a treatment for Heywood cases. *Psychometrika*, 40(4), 505–517. <https://doi.org/10.1007/BF02291552>
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, 60(1), 577–605. <https://doi.org/10.1146/annurev.psych.60.110707.163612>
- Mehta, P. D., Neale, M. C., & Flay, B. R. (2004). Squeezing interval change from ordinal panel data: Latent growth curves with ordinal outcomes. *Psychological Methods*, 9(3), 301–333. <https://doi.org/10.1037/1082-989X.9.3.301>
- Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, 115(2), 300–307. <https://doi.org/10.1037/0033-2909.115.2.300>
- Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39(3), 479–515. https://doi.org/10.1207/S15327906MBR3903_4
- Molenaar, D., Dolan, C. V., & De Boeck, P. (2012). The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika*, 77(3), 455–478. <https://doi.org/10.1007/s11336-012-9273-5>
- Molenaar, D., Dolan, C. V., & Verhelst, N. D. (2010). Testing and modeling non-normality within the one factor model. *British Journal of Mathematical and Statistical Psychology*, 63(2), 293–317. <https://doi.org/10.1348/000711009X456935>
- Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65(3), 391–411. <https://doi.org/10.1007/BF02296153>
- Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43(4), 551–560. <https://doi.org/10.1007/BF02293813>
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115–132. <https://doi.org/10.1007/BF02294210>
- Nocedal, J., & Wright, S. J. (2006). *Numerical optimization* (2nd ed.). Springer.
- Reddi, S. J., Kale, S., & Kumar, S. (2019). On the convergence of adam and beyond. arXiv preprint arXiv: 1904.09237. <https://doi.org/10.48550/arXiv.1904.09237>
- Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25. <http://www.jstatsoft.org/v17/i05/> <https://doi.org/10.18637/jss.v017.i05>
- Samejima, F. (1969). *Estimation of ability using a response pattern of graded scores: Psychometric monograph*. (Vol. 17). The Psychometric Society.
- Schmitt, J. E., Mehta, P. D., Aggen, S. H., Kubarych, T. S., & Neale, M. C. (2006). Semi-nonparametric methods for detecting latent non-normality: A fusion of latent trait and ordered latent class modeling. *Multivariate Behavioral Research*, 41(4), 427–443. https://doi.org/10.1207/s15327906mbr4104_1
- Smits, I. A., Timmerman, M. E., & Stegeman, A. (2016). Modelling non-normal data: The relationship between the skew-normal factor model and the quadratic factor model. *The British Journal of Mathematical and Statistical Psychology*, 69(2), 105–121. <https://doi.org/10.1111/bmsp.12062>
- Swaminathan, H., & Gifford, J. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 13–30). Academic Press.
- Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3), 393–408. <https://doi.org/10.1007/BF02294363>
- Urban, C. J., & Bauer, D. J. (2021). A deep learning algorithm for high-dimensional exploratory item factor analysis. *Psychometrika*, 86(1), 1–29. <https://doi.org/10.1007/s11336-021-09748-3>
- Urban, C. J., He, S. (2022). DeepIRTools: Deep learning-based estimation and inference for item response theory models. Python package. <https://github.com/cjurban/deepirtools>
- van den Oord, E. J. (2005). Estimating Johnson curve population distributions in MULTILOG. *Applied Psychological Measurement*, 29(1), 45–64. <https://doi.org/10.1177/0146621604269791>
- Van Der Maas, H. L., & Wagenmakers, E. J. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology*, 118(1), 29–60.
- Verhelst, N. D., & Glas, C. A. (1995). The one parameter logistic model. In *Rasch models* (pp. 215–237). Springer.
- Vermunt, J. K. (2004). An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica*, 58(2), 220–233. <https://doi.org/10.1046/j.0039-0402.2003.00257.x>
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge. <https://doi.org/10.4324/9781410605931>
- Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, 33(2), 102–117. <https://doi.org/10.1177/0146621608319512>
- Zhang, S., Chen, Y., Li, X. (2020). mirtjml: Joint maximum likelihood estimation for high-dimensional item factor analysis. R package version 1.4.0. <https://CRAN.R-project.org/package=mirtjml>
- Zwinderman, A. H., & van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurement*, 14(1), 73–81. <https://doi.org/10.1177/014662169001400107>