IRTest: An R Package for Item Response Theory with Estimation of Latent Distribution

by Seewoo Li

Abstract Item response theory (IRT) models the relationship between respondents' latent traits and their responses to specific items. One key aspect of IRT is the assumption about the distribution of the latent variables, which can influence parameter estimation accuracy. While a normal distribution has been conventionally assumed, this may not always be appropriate. When the assumption of normality is violated, latent distribution estimation (LDE) can enhance parameter estimation accuracy by accommodating non-normal characteristics. Despite there being several methods proposed for LDE in IRT, there is a lack of software designed to handle their implementations. This paper introduces IRTest, a software program developed for IRT analysis that incorporates LDE procedures. It outlines the statistical foundation of LDE, details the functionalities of IRTest, and provides examples of IRT analyses to demonstrate the software's applications.

1 Introduction

Item response theory (IRT) is a widely used statistical framework for modeling the probabilistic relationship between examinees' latent traits (i.e., ability parameters) and their responses to specific items (de Ayala, 2009; Hambleton et al., 1991; van der Linden, 2016). These latent traits, an essential component of IRT, typically represent unobservable human characteristics in educational and psychological assessments, such as academic ability, depression severity, or extroversion levels.

In the estimation process of IRT models, particularly when using marginal maximum likelihood (MML), the latent trait's distribution can impact parameter estimation (Woods, 2015). MML is obtained by marginalizing a joint likelihood with respect to the latent variable, and any misspecification in the latent distribution can lead to biased parameter estimates. It is also worth noting that IRT necessarily assumes that the latent variables of interest are continuous. In some cases, a discrete latent distribution may better reflect observed data, such as in located latent class (LLC) models (see Clogg, 1981; Follmann, 1988; Haberman, 2005; McCutcheon, 1987; Xu and von Davier, 2008).

The conventional assumption of normality in latent distributions has been questioned, and empirical evidence and potential drawbacks of violating this assumption have been addressed (Dudley-Marling, 2020; Li, 2022; Mislevy, 1984; Sass et al., 2008; Seong, 1990; Woods and Lin, 2009). Previous studies have identified factors that can result in a skewed and/or bimodal latent distribution (Harvey and Murry, 1994; Ho and Yu, 2015; Woods, 2015; Yadin, 2013): disparities between high-achieving and low-achieving groups, the presence of an extreme group, difficulties of test items, and an innate human tendency to be inclined to one side of a latent ability scale. Potential problems of this assumption being violated include biases in parameter estimates and errors in ensuing decision-making processes. In this case, latent distribution estimation (LDE) can effectively reduce the biases in parameter estimates by capturing non-normal characteristics of the latent distribution.

Several methods have been proposed for LDE: the empirical histogram method (EHM: Bock and Aitkin, 1981; Mislevy, 1984), a mixture of two normal components (2NM: Li, 2021; Mislevy, 1984), the Ramsay-curve method (RCM: Woods, 2006b), the Davidian-curve method (DCM: Woods and Lin, 2009), the log-linear smoothing method (LLS: Casabianca and Lewis, 2015; Xu and von Davier, 2008), and the kernel density estimation method (KDM: Li, 2022).

IRTest is an R package for unidimensional IRT analyses which aims to handle LDE in IRT. **IRTest** is available from the Comprehensive R Archive Network (CRAN) at `https://cran.r-project.org/package=IRTest`. In **IRTest**, the model-fitting functions estimate the latent distribution through MML estimation using the EM algorithm (MML-EM: Bock and Aitkin, 1981). Along with the conventional method of assuming normality, five LDE methods (EHM, 2NM, DCM, LLS, and KDM) are currently available as model-fitting procedures. Extensions of these methods to multidimensional settings will be updated in accordance with advancements in theoretical research, as the current lack of research in this area necessitates ongoing development.

Currently, there are not many software programs to implement LDE, and their choices of LDE methods are somewhat limited since LDE may not be one of their main concerns. Most of them offer only EHM which may be the most straightforward way to carry out LDE (e.g., BILOG-MG, Zimowski et al., 2003; flexMIRT, Cai, 2022), some software programs such as RCLOG (Woods, 2006a), LLSEM

(Casabianca and Lewis, 2011), and **sirt** (Robitzsch, 2024) focus on one particular LDE method, and **mirt** (Chalmers, 2012) is equipped with EHM and DCM.

This paper details the statistical foundation of **IRTest** and its implementation. The remainder of this paper is organized as follows: Section 2 explains the basic statistical concepts of the IRT parameter estimation and the LDE procedure within the MML-EM framework. Section 3 discusses the LDE methods. Section 4 validates **IRTest** by comparing it with existing packages. Section 5 demonstrates the **IRTest** implementations. Lastly, Section 6 presents a discussion on the package.

2 LDE in IRT

This section provides a brief overview of the statistical aspects of LDE in IRT, and specific LDE methods will be discussed in detail in the next section. The objectives of this section are 1) to introduce the basic concepts of IRT, 2) to summarize the role of the latent distribution within the estimation process, and 3) to explain how LDE can enhance parameter estimation accuracy.

IRT is a statistical framework widely used in educational and psychological measurement to analyze how examinees' responses to test items reflect their underlying traits. IRT models offer mathematical functions to estimate the probability of examinees' responses based on item parameters and their ability levels. Among the estimation methods, MML is often preferred over conditional maximum likelihood (CML) and joint maximum likelihood (JML) for its versatility in model selection and statistical consistency of estimation. MML incorporates the latent distribution when integrating out the latent variable from the joint likelihood.

Statistically, LDE extends the normality assumption models by adding distribution parameters. While the marginal likelihood of the normality assumption models is dependent only on item parameters, LDE incorporates both item and distribution parameters in the marginal likelihood. It is worth noting that some LDE methods may not follow the maximum likelihood estimation (MLE) approach. Instead, they may utilize another index to estimate distribution parameters.

2.1 IRT models

Most IRT models follow a monotonically increasing probabilistic form and specify a functional relationship between item parameters and ability parameters. This section presents five well-known and widely-used IRT models, all of which are available in **IRTest**. For brevity, the discussion is mainly focused on three-parameter logistic model (3PLM: Birnbaum, 1968) and generalized partial credit model (GPCM: Muraki, 1992), since 3PLM can be reduced to one-parameter logistic model (1PLM: Rasch, 1960) or two-parameter logistic model (2PLM: Birnbaum, 1968) and GPCM can be reduced to partial credit model (PCM: Masters, 1982). The former handles dichotomous item responses and the latter handles polytomous item responses.

More than one model can be applied to item response data. For example, when analyzing a test of dichotomous items, the 2PLM can be applied to short-answer items, while the 3PLM can be applied to multiple-choice items. This differentiation allows researchers to reflect the differences in guessing behaviors observed between the two types of items. Also, the pair of one dichotomous response model and one polytomous response model could be used for a mixed-format test that comprises both dichotomous and polytomous items (Baker and Kim, 2004).

Three-parameter logistic model (3PLM) Let $u \in \{0, 1\}$ be a dichotomous item response, θ be the ability parameter of an examinee, and a, b, and c $(0 < c < 1)$ be the item discrimination, difficulty, and guessing parameters, respectively. The item response function of the 3PLM can be expressed as,

$$\Pr(u = 1 \mid \theta, a, b, c) = c + (1 - c)\frac{\exp\left(a(\theta - b)\right)}{1 + \exp\left(a(\theta - b)\right)}, \tag{1}$$

where $u = 1$ indicates the correct response of the examinee. The probability ranges from c to 1 because of the guessing parameter determining the lower bound. Given the nature of dichotomous items, $1 - \Pr(u = 1 \mid \theta, a, b, c)$ represents the probability of an incorrect response. The model reduces to the 2PLM when $c = 0$, and to the 1PLM when $c = 0$ and the same a value is assumed across all items.

Generalized partial credit model (GPCM) The GPCM can be regarded as a polytomous form of the 2PLM. Let $u \in \{0, 1, \ldots, M\}$ $(M \geq 2)$ be a polytomous item response, b_v $(v = 1, 2, \ldots, M)$ be the boundary parameters, and the rest be the same as previously defined. The item response function of

the GPCM can be expressed as,

$$\Pr(u = k \mid \theta, a, b_1, \ldots, b_M) = \frac{\exp \sum_{v=0}^{k} (a(\theta - b_v))}{\sum_{m=0}^{M} \exp \sum_{v=0}^{m} (a(\theta - b_v))}, \tag{2}$$

where $\exp \sum_{v=0}^{0} (a(\theta - b_v)) = 1$ for notational convenience. Equation (2) represents the probability of providing a response of $u = k$. The GPCM reduces to the PCM when the same a value is assumed across all items. When $M = 1$, they are reduced to their dichotomous counterparts: the 1PLM and the 2PLM, respectively.

2.2 Role of the latent distribution

In the implementation of MML-EM, quadrature schemes are used to numerically approximate the integral with respect to the latent variable (Baker and Kim, 2004; Bock and Aitkin, 1981). A quadrature scheme transforms a continuous latent variable θ into a discrete variable θ^*; the domain of θ is divided into non-overlapping Q grids, each of which is assigned to a certain value of θ^* called a quadrature point. Typically, quadrature points are set to the middle of the grids. The default option of **IRTest** is to set quadrature points from -6 to 6 with an increment of 0.1, resulting in 121 quadrature points. Within the estimation functions of **IRTest**, the `range` and `q` arguments determine the range and the number of quadrature points, respectively. The corresponding probability mass function (PMF) of the latent variable can be expressed as,

$$A\left(\theta_q^*\right) = \frac{g\left(\theta_q^*\right)}{\sum_{q=1}^{Q} g\left(\theta_q^*\right)}, \tag{3}$$

where $g(\theta)$ is the probability density function (PDF) of the latent variable and θ_q^* is the qth quadrature point ($q = 1, 2, \ldots, Q$) (Baker and Kim, 2004). In this paper, the term latent distribution indicates either $g(\theta)$ or $A(\theta^*)$ depending on the context.

The marginal log-likelihood of the model is the quantity to be maximized in the estimation procedure, which can be expressed as follows (Baker and Kim, 2004):

$$\begin{aligned} \log L &= \sum_{j=1}^{N} \log \int_{\theta} L_j(\theta) g(\theta)\, d\theta \\ &\approx \sum_{j=1}^{N} \log \sum_{q=1}^{Q} L_j\left(\theta_q^*\right) A\left(\theta_q^*\right). \end{aligned} \tag{4}$$

In the equation above, the integral is approximated by the summation to facilitate the EM algorithm. The quantity $L_j\left(\theta_q^*\right)$ is the jth examinee's ($j = 1, 2, \ldots, N$) likelihood for their item responses given that their ability parameter is θ_q^*.

Equation (4) shows that the latent distribution is one of the components of the marginal log-likelihood. Consequently, the specification of the latent distribution influences the value of the marginal log-likelihood of a model, potentially impacting the accuracy of parameter estimates.

Additionally, the latent distribution plays a role in model identification and affects the convergence of the MML-EM procedure. In **IRTest**, a scale is assigned to the latent variable by setting the mean and standard deviation of the latent distribution to 0 and 1, respectively. Meanwhile, for LDE methods such as LLS and DCM, where a hyperparameter determines the number of distributional parameters, the MML-EM procedure may not converge with a small sample size and a large number of distributional parameters.

2.3 LDE in the MML-EM procedure

The decomposition of the marginal log-likelihood would help explicate the separate estimation of the item and distribution parameters, which can be expressed as follows (Li, 2021):

$$\begin{aligned} \log L &\approx \sum_{q=1}^{Q} \sum_{j=1}^{N} \gamma_{jq} \log L_j\left(\theta_q^*\right) + \sum_{q=1}^{Q} \sum_{j=1}^{N} \gamma_{jq} \log A\left(\theta_q^*\right) - \sum_{q=1}^{Q} \sum_{j=1}^{N} \gamma_{jq} \log \gamma_{jq} \\ &= \log L_{\text{item}} + \log L_{\text{distribution}} - (\text{constant}). \end{aligned} \tag{5}$$

The quantity $\gamma_{jq} = E\left(\Pr_j\left(\theta_q^*\right)\right)$, calculated through Bayes' theorem in the expectation-step (E-step) of the EM algorithm, represents the expected probability of jth examinee's ability parameter belonging to the qth grid (see Baker and Kim, 2004). Then, in the maximization-step (M-step), the item and distribution parameters are estimated. Since γ_{jq} is a function of the latent distribution, precise specification of the latent distribution would enhance the accuracy of γ_{jq}. Thus, the parameter estimates are implicitly affected by the latent distribution through γ_{jq}.

In equation (5), regarding γ_{jq} as a constant, the $L_j\left(\theta_q^*\right)$ in the first term depends only on item parameters, while the $A\left(\theta_q^*\right)$ in the second term depends only on distribution parameters. This probabilistic independence allows the separate estimation of the item and distribution parameters, from which a selection of estimation methods of the distribution parameters may emerge.

To elaborate more on the second term of equation (5), it can be rewritten and simplified as,

$$\begin{aligned} \log L_{\text{distribution}} &= \sum_{q=1}^{Q}\sum_{j=1}^{N}\gamma_{jq}\log A\left(\theta_q^*\right) \\ &= \sum_{q=1}^{Q}\hat{f}_q\log A\left(\theta_q^*\right), \end{aligned} \tag{6}$$

where f_q is an unknown true frequency at the qth grid and $\hat{f}_q = E(f_q) = \sum_{j=1}^{N}\gamma_{jq}$ is the expected frequency at the qth grid by the definition of γ_{jq}. In the E-step, the latent distribution is involved in calculating $\hat{f}_q$, then, in the M-step, the distribution parameters are estimated and updated by using the quantity $\hat{f}_q$. This E and M cycle iterates until the algorithm converges. The estimated parameters in the last iteration would be the final output, and the corresponding distribution of the final output becomes the estimated latent distribution.

Unlike the item parameter estimation being aligned with the MLE approach of the MML-EM procedure, the distribution parameters are not always estimated by maximizing equation (6). With the MLE approach still being the dominant choice, different approaches, such as minimizing the *approximate mean integrated squared error* (Li, 2022), can be applied to estimate distribution parameters, which is discussed in the next section. In all case, every strategy for the estimation of distribution parameters utilizes γ_{jq} in its estimation procedure.

3 LDE methods

In principle, almost every density estimation method can be used for LDE in IRT. However, existing studies have selectively inspected and developed some methods that would enhance the effectiveness of practical applications of IRT and/or benefit the researchers working on IRT. This section focuses on four LDE methods to highlight their methodological diversity. The choice and order of the methods in this section are not intended to imply any superiority of one method over the other.

At the time of writing this article, there is a lack of research comparing LDE methods. Given that comparing models in meaningful ways is crucial for practical applications, further studies examining the evaluation criteria of LDE methods would be beneficial for practitioners to select the most suitable method for their analyses.

3.1 Empirical histogram method (EHM)

One simple LDE strategy would be to directly employ the outputs obtained from the E-step. EHM does this by simply calculating the expected probabilities for each grid of the quadrature scheme, which can be considered either as the normalized expected sample size or nonparametric maximum likelihood estimates (Bock and Aitkin, 1981; Laird, 1978; Mislevy, 1984). The entire estimation process can be easily portrayed in the form of an equation as follows:

$$\hat{A}_q = E(A_q) = \frac{\sum_{j=1}^{N} E\left(\Pr_j\left(\theta_q^*\right)\right)}{N} = \frac{\sum_{j=1}^{N}\gamma_j q}{N} = \frac{\hat{f}_q}{N}, \tag{7}$$

where A_q denotes $A\left(\theta_q^*\right)$ for notational brevity. It can be seen that the estimates are simply the expected frequencies $\hat{f}_q$'s divided by the total population N. EHM can be implemented in the estimation functions of **IRTest** by specifying the argument as `latent_dist="EHM"`.

Alternatively, the MLE solution can be derived in the following manner using Lagrangian multi-

pliers. With Lagrangian multipliers, the quantity to be maximized becomes,

$$\mathcal{L} = \sum_{q=1}^{Q} \hat{f}_q \log A_q - \lambda \left(\sum_{q=1}^{Q} A_q - 1 \right), \tag{8}$$

where the second term is introduced from the constraint for a proper distribution (i.e., $\sum_q A_q = 1$). Differentiating $\mathcal{L}$ with respect to A_q and equating it to zero yields

$$\frac{\partial \mathcal{L}}{\partial A_q} = \frac{\hat{f}_q}{A_q} - \lambda = 0. \tag{9}$$

Then, $A_q = \frac{\hat{f}_q}{\lambda}$ for all $q = 1, 2, \ldots, Q$, which results in $\lambda = N$ by the constraint. This shows that $\hat{A}_q = \frac{\hat{f}_q}{N}$ maximizes the likelihood $\mathcal{L}$.

In addition to its simplicity and expediency, EHM has been shown to be effective in reducing biases in parameter estimates when the normality assumption is violated. However, the performance of EHM could be limited to some extent when it fails to screen out the random noise of the data, thereby producing less accurate parameter estimates (Li, 2021; Woods, 2015; Woods and Lin, 2009). To address this issue, some methods incorporate smoothing procedures to alleviate the impacts of the random noise, which are addressed later in this section.

3.2 Two-component normal mixture distribution (2NM)

The 2NM is made up of two normal components added up together to form a single distribution. As a natural extension of the normality assumption, the 2NM could be thought of as a non-normal distribution caused by two different latent groups where each group is assumed to follow a normal distribution. The addition of a normal component imparts flexibility to the 2NM to reflect bimodality and skewness of the latent distribution. The 2NM method can be implemented in the estimation functions of **IRTest** by specifying the argument as `latent_dist="2NM"`.

Letting $\tau = [\pi, \mu_1, \mu_2, \sigma_1, \sigma_2]'$ be the vector of five original parameters of the 2NM, the PDF of the 2NM can be expressed as follows (Li, 2021):

$$g(\theta \mid \tau) = \pi \times \phi(\theta \mid \mu_1, \sigma_1) + (1 - \pi) \times \phi(\theta \mid \mu_2, \sigma_2), \tag{10}$$

where $\phi(\theta)$ is a normal component.

Proportionality holds when $A\left(\theta_q^*\right)$ in equation (6) is substituted with $g(\theta_q^* \mid \tau)$, resulting in $\log L_{\text{distribution}} \propto \sum_{q=1}^{Q} \hat{f}_q \log g(\theta_q^* \mid \tau)$. The MLE results for the 2NM parameters can be obtained by introducing another EM algorithm. In this paper, this additional optimization algorithm would be referred to as *the secondary EM algorithm* nested in the M-step of the primary EM algorithm.

To estimate the 2NM parameters, another quantity η_q is calculated in the E-step of the secondary EM algorithm, which represents the expected probability of θ_q belonging to the first normal component. With the quantity η_q, the likelihood of the M-step of the secondary EM algorithm can be expressed as follows (Li, 2021):

$$\begin{aligned} \log L_{\text{distribution}} \propto & \sum_{q=1}^{Q} \hat{f}_q \eta_q \log \left[\pi \phi(\theta_q \mid \mu_1, \sigma_1) \right] \\ & + \sum_{q=1}^{Q} \hat{f}_q (1 - \eta_q) \log \left[(1 - \pi) \phi(\theta_q \mid \mu_2, \sigma_2) \right]. \end{aligned} \tag{11}$$

The closed-form solution for the 2NM parameters can be obtained by differentiating the likelihood with respect to each parameter and setting the first derivatives equal to zero (Li, 2021):

$$\hat{\pi} = \frac{\sum_{q=1}^{Q} \hat{f}_q \eta_q}{\sum_{q=1}^{Q} \hat{f}_q} = \frac{\sum_{q=1}^{Q} \hat{f}_q \eta_q}{N}, \tag{12}$$

$$\hat{\mu_1} = \frac{\sum_{q=1}^{Q} \hat{f}_q \eta_q \theta_q^*}{\sum_{q=1}^{Q} \hat{f}_q \eta_q}, \tag{13}$$

$$\hat{\mu_2} = \frac{\sum_{q=1}^{Q} \hat{f}_q (1-\eta_q)\theta_q^*}{\sum_{q=1}^{Q} \hat{f}_q (1-\eta_q)}, \tag{14}$$

$$\hat{\sigma_1^2} = \frac{\sum_{q=1}^{Q} \hat{f}_q \eta_q \left(\theta_q^* - \hat{\mu_1}\right)^2}{\sum_{q=1}^{Q} \hat{f}_q \eta_q}, \tag{15}$$

and

$$\hat{\sigma_2^2} = \frac{\sum_{q=1}^{Q} \hat{f}_q (1-\eta_q) \left(\theta_q^* - \hat{\mu_2}\right)^2}{\sum_{q=1}^{Q} \hat{f}_q (1-\eta_q)}. \tag{16}$$

Both advantages and disadvantages of the 2NM method stem from the parametric nature of the 2NM. The parameters of the 2NM render the estimated latent distribution interpretable. Also, the reparameterization of the 2NM parameters offers an inherent way to fix the mean and variance of the latent distribution to constants (see Li, 2021), which is a typical way to assign a scale to the latent variable in the MML-EM procedures. On the other hand, compared with its nonparametric counterparts, the flexibility of the 2NM is limited to some extent. For example, the 2NM is incapable of forming a wiggly-shaped distribution.

3.3 Davidian-curve method (DCM)

DCM uses a semi-nonparametric distribution, where the hyperparameter $h = 1, 2, \ldots, 10$ determines the complexity of the density (Woods and Lin, 2009). When $h = 1$, the distribution reduces to the normal distribution. In general, the search for an optimal hyperparameter involves balancing the flexibility of the latent distribution with the prevention of overfitting. DCM can be implemented in the estimation functions of **IRTest** by specifying the argument as `latent_dist="DC"`.

In DCM, the latent distribution can be expressed as follows:

$$g(\theta \mid h, \mathbf{m}) = \{P_h(\theta)\}^2 \varphi(\theta) = \left\{ \sum_{k=0}^{h} m_k \theta^k \right\}^2 \varphi(\theta), \tag{17}$$

where $\mathbf{m} = [m_0, m_1, \ldots, m_h]^{'}$ is a vector of coefficients, P_h is a polynomial of order h, φ is the standard normal distribution, and $m_h \neq 0$ (Woods and Lin, 2009; Zhang and Davidian, 2001). The following constraint guarantees that the function is a proper distribution (Zhang and Davidian, 2001):

$$\begin{aligned} E\left(P_h(Z)^2\right) &= E\left((\mathbf{mZ})^2\right) \\ &= \mathbf{m}^{'} E\left(\mathbf{ZZ}^{'}\right)\mathbf{m} \\ &= \mathbf{m}^{'}\mathbf{Mm} \\ &= \mathbf{m}^{'}\mathbf{B}^{'}\mathbf{Bm} \\ &= \mathbf{c}^{'}\mathbf{c} \\ &= 1. \end{aligned} \tag{18}$$

In the constraint above, $Z \sim N(0,1)$, $\mathbf{Z} = \left[1, Z^1, Z^1, \ldots, Z^h\right]^{'}$, $\mathbf{M} = E\left(\mathbf{ZZ}^{'}\right)$, $\mathbf{B}^{'}\mathbf{B} = \mathbf{M}$ by eigenvalue decomposition, and $\mathbf{c} = \mathbf{Bm}$ is a $h+1$ dimensional vector. By applying a polar coordinate transformation of $\mathbf{c}$, the constraint is always satisfied (see Woods and Lin, 2009; Zhang and Davidian, 2001).

As DCM follows the MLE approach, the quantity to be maximized for LDE can be expressed as,

$$\log L_{\text{distribution}} \propto \sum_{q=1}^{Q} \hat{f}_q \log \left[\left\{ \left[\mathbf{B}^{-1}\mathbf{c}\right]^{'} \begin{bmatrix} \left(\theta_q^*\right)^0 \\ \left(\theta_q^*\right)^1 \\ \vdots \\ \left(\theta_q^*\right)^h \end{bmatrix} \right\}^2 \varphi\left(\theta_q^*\right) \right]. \tag{19}$$

Since the elements of $\mathbf{B}^{-1}$ are constants, the latent distribution is estimated by finding $\mathbf{c}$ that maximizes the likelihood above.

In the implementation of DCM, ten models are typically fitted according to each value of the hyperparameter ($h = 1, 2, \ldots, 10$). Then, the best model is selected by Hannan-Quinn (HQ) criterion (Hannan and Quinn, 1979):

$$HQ = -2 \log L + 2p \left(\log \left(\log N \right) \right), \tag{20}$$

where N is the total number of examinees and p is the number of parameters to be estimated. Focusing on whether HQ criterion selects $h = 1$ or not, this model selection procedure in DCM can be used to examine the normality of the latent distribution (Woods and Lin, 2009).

This paper does not go into details of LLS (Casabianca and Lewis, 2015; Xu and von Davier, 2008) because of the similarities between DCM and LLS: both of them take the MLE approach for parameter estimation, and the hyperparameter h determines the number of distribution parameters to control the degree of smoothing.

3.4 Kernel density estimation method (KDM)

KDM is a nonparametric method for conducting the LDE procedure, and it can be implemented in the estimation functions of **IRTest** by specifying the argument as `latent_dist="KDE"`. In general, a kernel function is assigned to every observation to be stacked up all together and form a density function. In the context of LDE in IRT, the former statement means that $\hat{f}_q$ kernels are assigned to θ_q^*, which can be expressed as,

$$g(\theta \,|\, h) = \frac{1}{Nh} \sum_{q=1}^{Q} \hat{f}_q K\left(\frac{\theta - \theta_q^*}{h} \right), \tag{21}$$

where $K(\cdot)$ is a kernel function and h is a hyperparameter often referred to as the *bandwidth*. The following discussion assumes the Gaussian kernel for $K(\cdot)$ as a general default choice (Gramacki, 2018; Silverman, 1986).

KDM takes a different approach from the previously discussed methods in carrying out the LDE procedure. Instead of the log-likelihood ($\log L_{\text{distribution}}$), the approximate mean integrated squared error (AMISE) is used, which is the Tayor-series approximation of the mean integrated squared error:

$$AMISE(\hat{g_h}) = \frac{1}{2Nh\sqrt{\pi}} + \frac{h^4}{4} R(g''). \tag{22}$$

In the equation above, $\hat{g_h}$ is the estimated latent distribution using the bandwidth h and $R(g'') = \int g''(x)\, dx$. Note that this is a simplified version of AMISE by the adoption of the Gaussian kernel. To find h that minimizes equation (22), another equation is obtained by differentiating AMISE with respect to h and equating it to 0 (Gramacki, 2018; Silverman, 1986; Wand and Jones, 1995):

$$h_{AMISE} = \left[2N\sqrt{\pi} R(g'') \right]^{-\frac{1}{5}}. \tag{23}$$

Unfortunately, this solution cannot be immediately employed in estimating the bandwidth, because $R(g'')$ still depends on the unknown density $g(\theta)$. For the details of the methods that deal with this situation, refer to Silverman (1986), Gramacki (2018), Sheather (2004), and Wand and Jones (1995). Utilizing the built-in R function `stats::density()` for the KDM procedure, the default option of **IRTest** is `bandwidth="SJ-ste"`, a recommended method for bandwidth estimation (Jones et al., 1996; Sheather and Jones, 1991). Other available options for the `bw` argument of the `stats::density()` can also be passed to the `bandwidth` argument of **IRTest**.

On the one hand, KDM and DCM are similar in that their hyperparameters determine the degree of smoothing. On the other hand, once the γ_{jq} is calculated and treated as a constant, the hyperparameter of KDM is the only parameter to influence the LDE results, whereas the LDE results of DCM depends on both the hyperparameter and the corresponding $h + 1$ density parameters. The absence of distribution parameter in KDM, except for the hyperparameter (bandwidth) itself, allows KDM to estimate the hyperparameter in a single model-fitting procedure, thereby obviating the need for a model selection process. Compared with other methods having a model selection step, this advantage may decrease computation time and expedite the analysis (Li, 2022).

4 Package validation

This section examines and validates the estimation performance of **IRTest** by comparing its results with those from **mirt** (Chalmers, 2012) and **ltm** (Rizopoulos, 2006). **mirt** and **ltm** are among the widely used R packages for IRT analyses. To this end, a dataset of ten dichotomous items and 1,000 examinees is generated using the 2PLM and under a right-skewed latent distribution, which is used throughout

Table 1: Item parameters and their estimates

	parameter	IRTest::IRTest_Dich()	IRTest::IRTest_Poly()	mirt::mirt()	ltm::ltm()
Item discrimination parameter (a)					
Item1	0.82	0.9264	0.9264	0.9263	0.9263
Item2	1.26	1.0174	1.0174	1.0174	1.0174
Item3	1.88	1.7614	1.7614	1.7613	1.7614
Item4	1.41	1.3730	1.3730	1.3729	1.3729
Item5	2.33	1.7300	1.7300	1.7296	1.7301
Item6	1.34	1.2904	1.2904	1.2903	1.2903
Item7	1.19	1.2593	1.2593	1.2592	1.2593
Item8	1.15	1.3579	1.3579	1.3578	1.3579
Item9	2.42	2.1221	2.1221	2.1218	2.1221
Item10	2.31	1.7667	1.7667	1.7665	1.7668
Item difficulty parameter (b)					
Item1	1.23	1.1484	1.1484	1.1482	1.1481
Item2	-0.72	-0.7642	-0.7642	-0.7645	-0.7645
Item3	-0.19	-0.1788	-0.1788	-0.1791	-0.1791
Item4	-0.14	-0.1565	-0.1565	-0.1568	-0.1568
Item5	-1.16	-1.2970	-1.2970	-1.2975	-1.2973
Item6	0.14	0.2154	0.2154	0.2151	0.2151
Item7	0.29	0.2619	0.2619	0.2616	0.2616
Item8	1.28	1.1198	1.1198	1.1196	1.1195
Item9	0.12	0.2919	0.2919	0.2916	0.2916
Item10	-0.96	-0.9906	-0.9906	-0.9909	-0.9909

this section.

The following functions are used to fit the models: `IRTest::IRTest_Dich()` and `IRTest::IRTest_Poly()` from **IRTest**, `mirt::mirt()` from **mirt**, and `ltm::ltm()` from **ltm**. The standard normal latent distribution, the 2PLM, and 61 quadrature points are applied to all functions, and the rest of the options are set to the default. Note that `IRTest::IRTest_Poly()` can also be applied to binary data, but with lower computational efficiency compared to `IRTest::IRTest_Dich()`.

Table 1 shows the item parameters and their estimates. The result shows that the parameter estimates from **IRTest** are almost identical to those from **mirt** and **ltm**. For example, the root mean square error (RMSE) of the a (item discrimination) parameter estimates between `IRTest::IRTest_Dich()` and `mirt::mirt()` is 2.04×10^{-4}, and that between `IRTest::IRTest_Dich()` and `ltm::ltm` is 0.29×10^{-4}. Similarly, the RMSE of the b (item difficulty) parameter estimates between `IRTest::IRTest_Dich()` and `mirt::mirt()` is 3.23×10^{-4}, and that between `IRTest::IRTest_Dich()` and `ltm::ltm()` is 3.21×10^{-4}. This shows that the estimates from the four functions are practically identical.

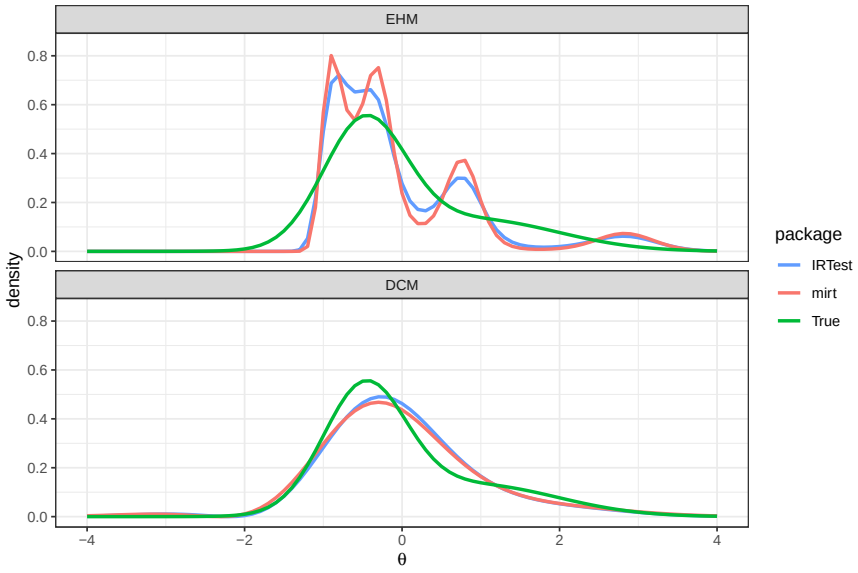

Figure 1: Estimated latent distributions from IRTest and mirt

To validate the LDE procedures of **IRTest**, the estimated latent distributions of **IRTest** are compared with those estimated from **mirt**, using EHM and DCM. The hyperparameter h of DCM is set to $h = 3$ for an illustrative purpose by considering the complexity of the true distribution and the size of the

data. Total 121 quadrature points are used for the MML-EM procedure. Integrated squared error (ISE), one of the widely used criterion in comparing two distributions, is used as an index to measure the closeness of the estimated distributions of **IRTest** and **mirt** (for the usages of ISE, see Jones, 1991):

$$\begin{aligned} ISE\left(\hat{g}^{\text{IRTest}}, \hat{g}^{\text{mirt}}\right) &= \int_{-\infty}^{\infty} \left(\hat{g}(\theta)^{\text{IRTest}} - \hat{g}(\theta)^{\text{mirt}}\right)^2 d\theta \\ &\approx \sum_{q=1}^{Q} \left(\hat{g}\left(\theta_q^*\right)^{\text{IRTest}} - \hat{g}\left(\theta_q^*\right)^{\text{mirt}}\right)^2 \Delta\theta^*, \end{aligned} \tag{24}$$

where $\Delta\theta^*$ is the distance between quadrature points, and $\hat{g}^{\text{IRTest}}$ and $\hat{g}^{\text{mirt}}$ are the estimated latent distributions of **IRTest** and **mirt**, respectively. In equation (24), the integral is approximated by the summation using the quadrature scheme of the MML-EM procedure.

Figure 1 illustrates the estimated latent distributions from both packages and the true latent distribution. Note that values outside of $|\theta| \leq 4$ are truncated for visualization, but they are included in the ISE calculation. The figure shows that the estimated distribution of **IRTest** (blue line) and the one from **mirt** (red line) closely resemble each other. The ISE values for EHM and DCM are 0.0107 and 0.0009, respectively, which numerically validate the similarities of the estimated distributions of the two packages.

5 Implementations of IRTest

The primary purpose of this section is to demonstrate the usages of **IRTest** with examples. Section 5.1 provides an example using a simulated dataset. In doing so, it illustrates the effect of LDE when the normality assumption is violated. Section 5.2 performs an IRT analysis using the Generic Conspiracist Beliefs Scale (GCBS) data (Brotherton et al., 2013) available from the Open-Source Psychometric Project at `http://openpsychometrics.org/_rawdata/GCBS.zip`.

5.1 The effect of LDE

This section illustrates the effect of LDE by showing an improvement in estimation accuracy. To this end, an artificial item response dataset is generated from a non-normal latent distribution. The parameters of this dataset are used as the true values in evaluating errors.

Data generation Using the function `IRTest::DataGeneration()`, a dataset of 40 dichotomous items and 2,000 respondents is generated by

```
Alldata <- IRTest::DataGeneration(N = 2000,
                                  nitem_D = 40,
                                  latent_dist = "2NM",
                                  d = 1.414,
                                  sd_ratio = 2,
                                  prob = 2/3)
simulated_data <- Alldata$data_D
true_item <- Alldata$item_D
true_theta <- Alldata$theta
```

where `simulated_data` is the item response matrix, `true_item` is the item parameter matrix, and `true_theta` is the vector of ability parameters. The 2PLM is employed in generating the dataset, and a 2NM distribution is employed to simulate a non-normal latent distribution (`latent_dist = "2NM"`) with its parameters being `d = 1.414`, `sd_ratio = 2`, and `prob = 2/3` (for the reparameterization of the 2NM, see Li, 2021, 2024).

The highly skewed distribution is chosen to clearly demonstrate the effect of LDE, which is likely to be rare in practice (see Figure 2). Therefore, cautions are needed in interpreting and understanding the magnitude of the effect.

Model fitting The function `IRTest::IRTest_Dich()` is applied to the dichotomous data for the model-fitting, and an LDE method is specified by `latent_dist` argument. Two types of ability parameter estimates are illustrated: expected *a posteriori* (EAP) and maximum likelihood estimate (MLE). Note that *MLE* is also used to abbreviate maximum likelihood estimation. In **IRTest**, estimation methods of

ability parameters are determined by specifying the argument `ability_method`, where the default option is EAP. Either a model-fitting function (e.g., `IRTest::IRTest_Dich()`) or `IRTest::factor_score()` can be used to estimate ability parameters.

```
model_normal <- IRTest::IRTest_Dich(data = simulated_data,
                                    latent_dist = "Normal")
model_KDM <- IRTest::IRTest_Dich(data = simulated_data,
                                 latent_dist = "KDE")
theta_eap_normal <- IRTest::factor_score(model_normal)
theta_mle_normal <- IRTest::factor_score(model_normal, ability_method = "MLE")
theta_eap_KDM <- IRTest::factor_score(model_KDM)
theta_mle_KDM <- IRTest::factor_score(model_KDM, ability_method = "MLE")
```

Among these two models, `model_normal` assumes normality, whereas `model_KDM` estimates the latent distribution using KDM during the MML-EM procedure. KDM is arbitrarily selected for illustrative purposes.

Estimated latent distribution **IRTest** provides two ways to draw a density curve of the estimated latent distribution. The first is to use `plot()`, and the second is to use `IRTest::latent_distribution()`. The `plot()` can be considered as a shortcut for using `IRTest::latent_distribution()`. Note that `IRTest::latent_distribution()` is a PDF, and, thus, can only be applied to LDE methods that estimate a PDF. For example, since EHM (or LLS) estimates a *PMF*, a message will be printed without evaluated density values if an EHM-based (or LLS-based) object is passed to `IRTest::latent_distribution()`. The `plot()` can be utilized in all cases.

The following code can be an example for utilizing `IRTest::latent_distribution()`, where `IRTest::dist2()` is a density function of the 2NM and `stats::dnorm()` is the built-in function for a normal distribution.

```
density_plot <- ggplot2::ggplot() +
  ggplot2::stat_function(fun = IRTest::dist2,
                         args = list(d = 1.414, sd_ratio = 2, prob = 2/3),
                         linewidth = 1,
                         mapping = aes(color = "True")) +
  ggplot2::stat_function(fun = stats::dnorm,
                         linewidth = 1,
                         mapping = aes(color = "Normal")) +
  ggplot2::stat_function(fun = IRTest::latent_distribution,
                         args = list(model_KDM),
                         linewidth = 1,
                         mapping = aes(color = "KDM"))
```

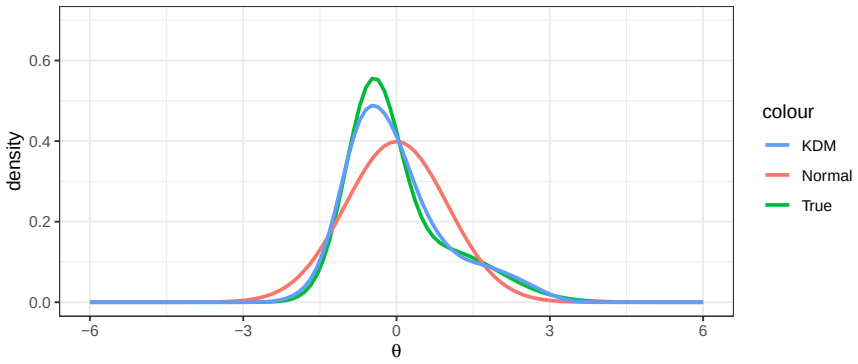

Figure 2: The true, normal, and estimated latent distributions

Figure 2 is drawn with a slight addition of aesthetic codes to the object `density_plot`. It shows the density curves of the true latent distribution, the standard normal distribution, and the estimated latent distribution. Notably, even with the discrepancies between the true distribution (green line) and the standard normal distribution (red line), the estimated distribution (blue line) almost recovered the shape of the true latent distribution.

Table 2: RMSEs from the normality assumption model and KDM

	Normal	KDM
a	0.181	0.123
b	0.124	0.079
EAP	0.272	0.252
MLE	0.285	0.272

Parameter estimates Differences in parameter estimates caused by the LDE procedure are portrayed in Figure 3, where the x-axis represents the true values (parameters) and the y-axis shows the corresponding errors. It shows that the parameter estimates from KDM (blue dots) are generally located closer to the "Error $= 0$" line than those from the normality assumption model (red dots), which indicates that the estimates from KDM are more accurate on average. RMSE is used as an evaluation criterion to assess the accuracy of the parameter estimates, where the lower RMSE indicates higher accuracy. Table 2 shows that, for all types of parameters, estimates from the KDM model are more accurate with lower RMSE values than those from the normality assumption model. Especially, the considerable amounts of decreases in RMSEs for $\hat{a}$ and $\hat{b}$ substantiates the effectiveness of LDE in enhancing the estimation accuracy of item parameters, where the magnitudes in the RMSE reduction are originated from the highly skewed latent distribution. As stated in Section 2.2, the results show that the appropriate specification of the latent distribution can have a positive impact on the accuracy of parameter estimates.

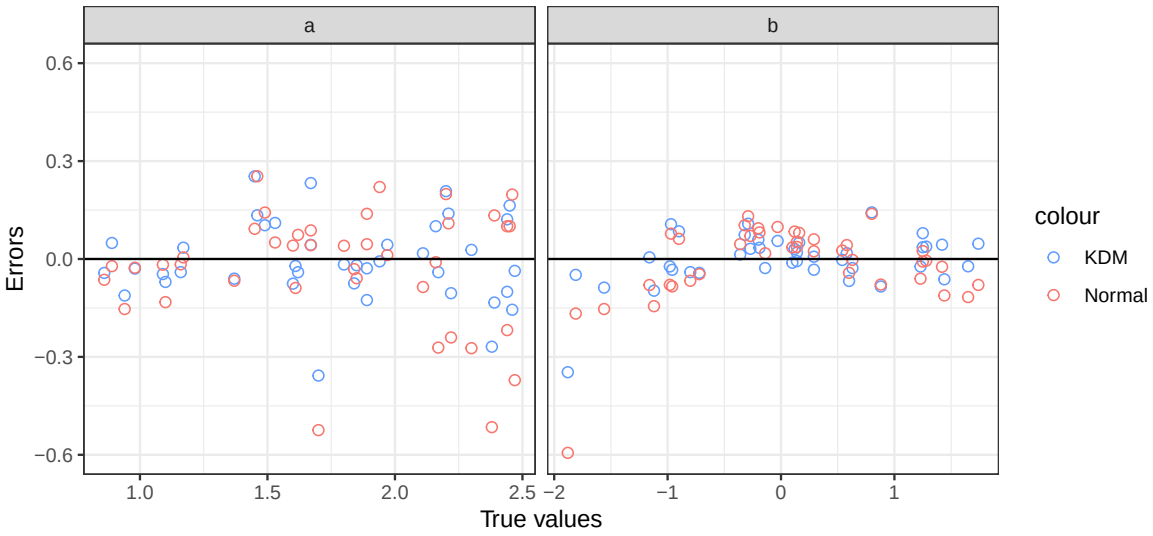

Figure 3: Differences in the errors of the item parameter estimates caused by LDE

5.2 An empirical example

This section performs an IRT analysis using the GCBS data. Along with the data analysis, one of the objectives of the section is to illustrate the usages of the functions of **IRTest**, such as those for calculating reliability coefficients and item/test information.

Data The GCBS data contains responses from 15 polytomous items and 2,391 respondents, where each item has five categories scored from one to five. The data can be loaded in the following manner.

```
data_GCBS <- read.csv("data/data_GCBS.csv")
```

There are 108 missing values in the data. **IRTest** uses a full-information maximum likelihood (FIML) approach in handling missing data.

Model selection DCM and KDM are used in performing LDE for illustrative purposes, where the DCM's h is the hyperparameter indicating the complexity of the latent distribution. The PCM, GPCM,

and graded response model (GRM: Samejima, 1969) are available IRT models of `IRTest::IRTest_Poly()`, where the default is the GPCM. Models are fitted as follows:

```
# Davidian-curve method
DCMs <- list()
for(i in 1:10){
    DCMs[[i]] <- IRTest::IRTest_Poly(data = data_GCBS, latent_dist = "DC", h = i)
}
names(DCMs) <- paste0("DC", 1:10)

# kernel density estimation method
KDM <- IRTest::IRTest_Poly(data = data_GCBS, latent_dist = "KDE")
```

A model-selection step is required for DCM: the "best-DCM" can be selected by the HQ criterion (Hannan and Quinn, 1979) presented in equation (20). After ten models are fitted, the best model can be selected with the function `IRTest::best_model()` (`stats::anova()` can also be used). The `IRTest::best_model()` function uses a specific criterion to determine the best model, with options including `"logLik"`, `"deviance"`, `"AIC"`, `"BIC"`, and `"HQ"`. The default is `criterion = "HQ"`.

```
do.call(what = IRTest::best_model, args = DCMs)
```

```
#> The best model: DC8
#>
#>            HQ
#> DC1  91232.30
#> DC2  91236.40
#> DC3  91237.18
#> DC4  91220.93
#> DC5  91217.55
#> DC6  91210.21
#> DC7  91209.33
#> DC8  91207.85
#> DC9  91209.87
#> DC10 91213.24
```

The result indicates that `DC8` is the best DCM.

The `DC8` and `KDM` can also be compared by the function `IRTest::best_model()`.

```
IRTest::best_model(DCMs$DC8, KDM, criterion = "logLik")
```

```
#> The best model: KDM
#>
#>             logLik
#> DCMs$DC8 -45433.63
#> KDM      -45433.02
```

The rest of this section looks into the details of `KDM` by following the model-comparison result.

Summary of the model A brief summary of the model-fitting results can be printed with `summary()`. It presents the convergence status, model-fit indices, and the number of parameters and items. Also, it displays a cursory shape of the estimated latent distribution.

```
summary(KDM)
```

```
#> Convergence:
#> Successfully converged below the threshold of 1e-04 on 63rd iterations.
#>
#> Model Fit:
#>  log-likeli   -45433.02
#>    deviance   90866.05
#>         AIC   91018.05
#>         BIC   91457.48
```

```
#>          HQ   91177.92
#>
#> The Number of Parameters:
#>        item   75
#>        dist   1
#>       total   76
#>
#> The Number of Items:  15
#>
#> The Estimated Latent Distribution:
#> method - KDE
#> ----------------------------------------
#>
#>                     @ @ .
#>                   @ @ @ @
#>                 . @ @ @ @ @
#>               . @ @ @ @ @ @ .
#>               @ @ @ @ @ @ @ @
#>           . @ @ @ @ @ @ @ @ @ @
#>         . @ @ @ @ @ @ @ @ @ @ @ .
#>     . @ @ @ @ @ @ @ @ @ @ @ @ @ @ .
#> . @ @ @ @ @ @ @ @ @ @ @ @ @ @ @ @ @ .
#> +---------+---------+---------+---------+
#> -2        -1        0         1         2
```

Alternatively, the shape of the estimated latent distribution can be easily visualized by the `plot()` function which produces a `ggplot`-class object. Figure 4 shows the estimated latent distribution of the GCBS data which is left-skewed. Figure 4 is produced by adding aesthetic codes to `plot()` and the standard normal distribution. If a PDF is estimated, additional arguments in `plot()` are passed to `ggplot2::stat_function()` of **ggplot2** (Wickham, 2016). Otherwise, if a PMF is estimated (e.g., EHM or LLS), they are passed to `ggplot2::geom_line()` of **ggplot2**.

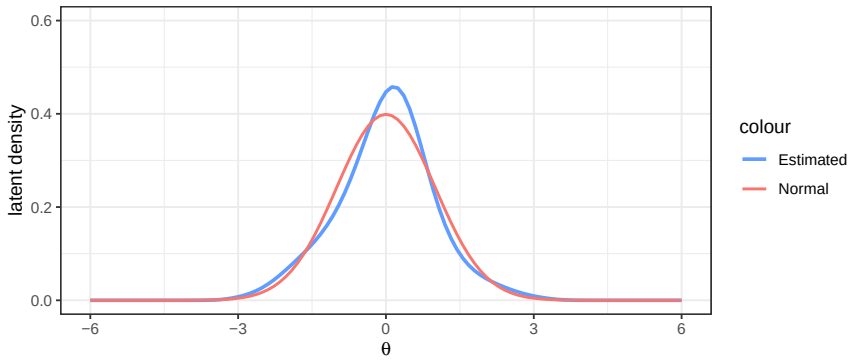

Figure 4: Estimated latent distribution for the GCBS data

Parameter estimates Users can access to item parameter estimates and corresponding standard errors with `stats::coef()` and `IRTest::coef_se()`, respectively.

```
stats::coef(KDM)
```

```
#>                a         b_1          b_2          b_3        b_4
#> Item1  0.9689201 -0.76302183 -0.391283148 -0.855328244  0.3535741
#> Item2  1.0789839 -0.53792382  0.009783414  0.005704736  0.7242562
#> Item3  0.7387147  1.67238829  0.366676394  0.966121767  1.3765605
#> Item4  1.3101921 -0.13464756  0.161325353  0.204687878  1.3519716
#> Item5  0.7816865 -0.54657035 -0.191682964 -0.831793749  0.7254415
#> Item6  1.1236494 -0.46965441 -0.187144001 -0.214049525  0.6453753
#> Item7  1.1752710 -0.07562134  0.301472014  0.146375779  0.9623761
#> Item8  0.6667551  1.54879257 -0.087408217  0.644569978  0.3019026
#> Item9  1.0243926  0.53049910  0.622129709  0.614778365  1.3307188
#> Item10 0.5422634 -0.67411059 -0.728342737 -1.374793623  0.4975367
```

```
#> Item11 1.0969859 -0.85223166 -0.647107727 -0.193551333  0.7969608
#> Item12 1.6043792 -0.13515994  0.138427510  0.313100514  0.9981246
#> Item13 0.8859117  1.15644479  0.243555061  1.098831006  1.3876873
#> Item14 1.0693468 -0.36458700 -0.054602648 -0.080358875  0.8726474
#> Item15 0.8699800 -2.01564618 -1.551317019 -1.872535217 -0.6573178

IRTest::coef_se(KDM)

#>                 a        b_1        b_2        b_3        b_4
#> Item1  0.03678565 0.08584960 0.08654712 0.08019167 0.05779205
#> Item2  0.03966144 0.06329931 0.06584207 0.06508447 0.06232969
#> Item3  0.03208946 0.11778055 0.10831653 0.11904930 0.13613987
#> Item4  0.04714700 0.05119141 0.05476176 0.05471713 0.06194283
#> Item5  0.03074447 0.09686522 0.09959919 0.09509046 0.07449349
#> Item6  0.04106021 0.06583399 0.06718341 0.06322656 0.05630556
#> Item7  0.04324230 0.05617174 0.06338652 0.06467895 0.06362815
#> Item8  0.02800753 0.12930642 0.12010949 0.12179173 0.12290590
#> Item9  0.04025272 0.06541014 0.07425989 0.08161815 0.09079149
#> Item10 0.02420022 0.15487167 0.14764212 0.13523579 0.09904333
#> Item11 0.04013028 0.07370413 0.06734274 0.05764279 0.05617594
#> Item12 0.05686980 0.04318188 0.04603551 0.04652548 0.04914336
#> Item13 0.03648830 0.09029916 0.08542261 0.09660685 0.11458529
#> Item14 0.03930729 0.06541102 0.06780640 0.06541428 0.06253578
#> Item15 0.03830840 0.16985780 0.13629512 0.10927582 0.06689059
```

Likewise, `IRTest::factor_score()` returns ability parameter estimates (`IRTest::factor_score()$theta`) and their standard errors (`IRTest::factor_score()$theta_se`) which are not printed here for their lengths being 2397.

```
IRTest::factor_score(KDM, ability_method = "EAP")
```

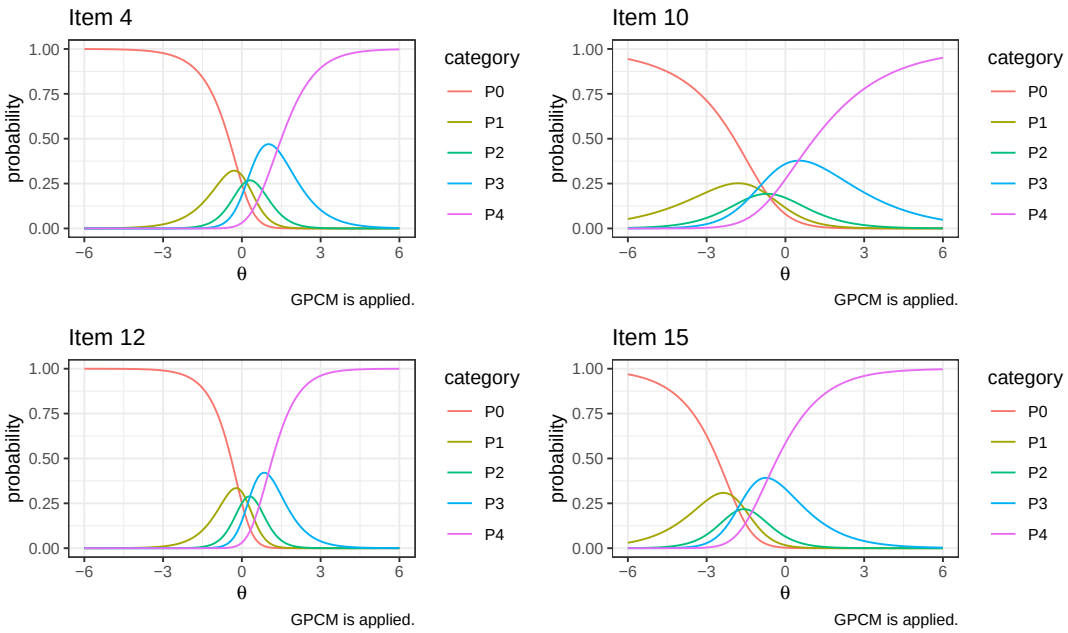

Figure 5: Item response functions of Item 4, 10, 12, and 15

Plotting item response functions Figure 5 shows item response functions of the four items (Item 4, 10, 12, and 15) by using `IRTest::plot_item()`. For example, a plot of the item response function of Item 11 can be drawn with `IRTest::plot_item(x = KDM, item.number = 11)`.

Reliability coefficient Among various types of IRT reliability coefficients, **IRTest** applies those discussed by Green et al. (1984) and May and Nicewander (1994). The coefficient from Green et al.

(1984) is calculated on the latent variable θ scale, and the one from May and Nicewander (1994) is calculated on the summed-score scale. May and Nicewander (1994)'s approach has an advantage of providing reliability coefficients for individual items. The function `IRTest::reliability()` returns the coefficients mentioned above: item reliability coefficients, and test reliability coefficients on the θ scale and the summed-score scale.

```
IRTest::reliability(KDM)

#> $summed.score.scale
#> $summed.score.scale$test
#> test reliability
#>        0.9293983
#>
#> $summed.score.scale$item
#>     Item1     Item2     Item3     Item4     Item5     Item6     Item7     Item8
#> 0.5039295 0.5248032 0.3818564 0.5740997 0.4227400 0.5468987 0.5526063 0.3795133
#>     Item9    Item10    Item11    Item12    Item13    Item14    Item15
#> 0.4929443 0.2898281 0.5196563 0.6437762 0.4386300 0.5232466 0.4021941
#>
#>
#> $theta.scale
#> test reliability
#>        0.9159476
```

Item and test information functions In **IRTest**, `IRTest::inform_f_item()` and `IRTest::inform_f_test()` evaluate item and test information, respectively. Figure 6 visually illustrates how each item contributes to the test information and how item information functions are added up all together to form the test information function. The test information function is drawn with a black line and the item information functions are colored.

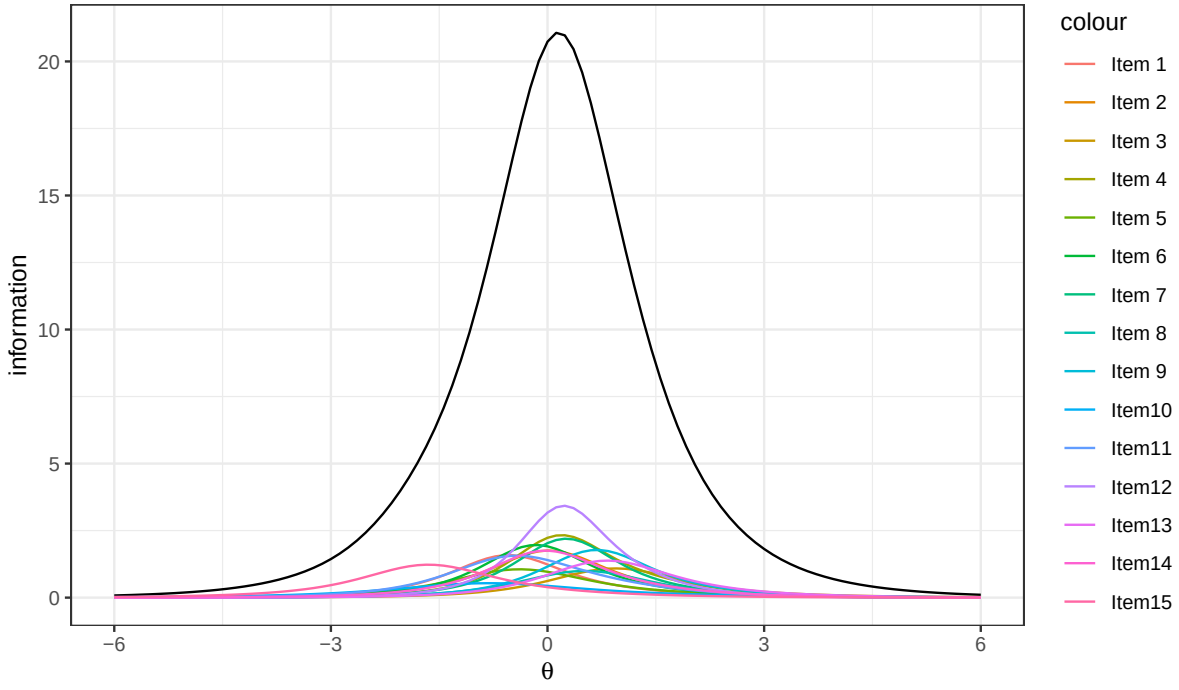

Figure 6: Item and test information functions

6 Discussion

While some may not find the magnitude of biases resulting from the deviations from the normality assumption particularly significant in the context of IRT, LDE remains an appealing option. Within IRT, increasing the number of items and respondents may require additional time and resources to improve parameter estimation accuracy. Meanwhile, an appropriate usage of LDE would increase

the estimation accuracy with the cost of adding a few distribution parameters. Therefore, unless the addition of distribution parameters compromises the model-data fit significantly, LDE can be a reasonable strategy for reducing estimation biases.

Although the current literature on LDE in IRT lacks clear guidelines on how and when to implement the LDE procedure, practitioners may choose to start with either KDM or DCM, even when there is limited or no information available about the shape of the latent distribution. Both approaches not only perform well in estimating non-normal latent distributions, but also effectively recover normal distributions when the normality assumption holds (Li, 2022; Woods and Lin, 2009).

The **IRTest** package offers the choice of LDE methods for conducting IRT analyses, and its functions utilize the estimated latent distribution for the subsequent analyses. **IRTest** is actively being updated, and some potential enhancements can be made for a better application of the package, which may include expanding the range of available IRT models, decreasing computation time, and imposing constraints on parameters. User feedback would also guide a way for the package maintenance and enhancement. Yet, there is much more to explore in the field of LDE beyond what has already been studied. Thus, further research on LDE is expected to provide valuable guidance to both package users and the developer.

Acknowledgments

Special thanks to Nagap Park, Hyesung Shin, Ryan Lerch, and anonymous reviewers for their insightful suggestions that improved the article.

Additional features

Analysis of continuous response data Continuous item responses, ranging from 0 to 1, are often encountered in test datasets. Such item responses can also be handled by the **IRTest** package with `IRTest::IRTest_Cont()`. All features of **IRTest** presented in this paper are applicable to continuous item responses. An example code is shown below:

```
# Generating a continuous item response data
data <- IRTest::DataGeneration(N = 1000, nitem_C = 10)$data_C

# Analysis
model_continuous <- IRTest::IRTest_Cont(data)
```

Analysis of mixed-format data An IRT analysis of mixed-format data can also be conducted with `IRTest::IRTest_Mix()`. The difference between `IRTest::IRTest_Mix()` and `IRTest::IRTest_Dich()` (or `IRTest::IRTest_Poly()`) is that `IRTest::IRTest_Mix()` requires two separate data: one for dichotomous items and the other for polytomous items. An example code is shown below:

```
model_mixed.format <- IRTest::IRTest_Mix(data_D = dichotomous_data,
                                         data_P = polytomous_data,
                                         model_D = rep(c("2PL", "3PL"), each = 5),
                                         model_P = "GPCM")
```

In this case, the 2PLM is applied to the first five dichotomous items, the 3PLM is applied to the rest of the dichotomous items, and the GPCM is applied to the polytomous items. The rest are the same with `IRTest::IRTest_Dich()` and `IRTest::IRTest_Poly()`.

The `IRTest::IRTest_Dich()` can also take a vector for the argument `model` to apply multiple IRT models to different types of items.

Item-fit statistic An item-fit statistic can be calculated with `IRTest::item_fit()`. Currently, Bock (1960)'s χ^2 and Yen (1981)'s Q_1 are available.

References

F. B. Baker and S.-H. Kim. *Item Response Theory: Parameter Estimation Techniques*. CRC press, 2nd edition, 2004. [p24, 25, 26]

A. Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, editors, *Statistical Theories of Mental test Scores*, chapter 17, pages 397–479. Addison-Wesley, 1968. [p24]

D. R. Bock. Methods and applications of optimal scaling. University of North Carolina Psychometric Laboratory Research Memorandum, No. 25, 1960. [p38]

D. R. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459, 1981. doi: 10.1007/BF02293801. [p23, 25, 26]

R. Brotherton, C. C. French, and A. D. Pickering. Measuring belief in conspiracy theories: The generic conspiracist beliefs scale. *Frontiers in Psychology*, 4, 2013. doi: 10.3389/fpsyg.2013.00279. [p31]

L. Cai. *flexMIRT version 3.65: Flexible Multilevel Multidimensional Item Analysis and Test Scoring [Computer software]*. Vector Psychometric Group, 2022. Chapel Hill, NC. [p23]

J. M. Casabianca and C. Lewis. *LLSEM 1.0: LogLinear Smoothing in an Expectation Maximization Algorithm for Item Response Theory Item Parameter Estimation [Computer software]*. Washington University in St. Louis, 2011. Bronx, NY. [p24]

J. M. Casabianca and C. Lewis. Irt item parameter recovery with marginal maximum likelihood estimation using loglinear smoothing models. *Journal of Educational and Behavioral Statistics*, 40(6): 547–578, 2015. doi: 10.3102/1076998615606112. [p23, 29]

P. R. Chalmers. mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6):1—-29, 2012. doi: 10.18637/jss.v048.i06. URL https://www.jstatsoft.org/index.php/jss/article/view/v048i06. [p24, 29]

C. C. Clogg. New developments in latent structure analysis. In D. J. Jackson and E. F. Borgatta, editors, *Factor analysis and measurement in sociological research*. Sage, 1981. [p23]

R. J. de Ayala. *The Theory and Practice of Item Response Theory*. The Guilford Press, New York, 2009. [p23]

C. Dudley-Marling. The tyranny of the normal curve: How the "bell curve" corrupts educational research and practice. In D. M. Allen and J. W. Howell, editors, *Groupthink in Science: Greed, Pathological Altruism, Ideology, Competition, and Culture*, chapter 17, pages 201–210. Springer, Cham, 2020. doi: 10.1007/978-3-030-36822-7_17. [p23]

D. Follmann. Consistent estimation in the rasch model based on nonparametric margins. *Psychometrika*, 53(4):553–562, 1988. [p23]

A. Gramacki. *Nonparametric Kernel Density Estimation and Its Computational Aspects*. Springer, 2018. [p29]

B. F. Green, D. R. Bock, L. G. Humphreys, R. L. Linn, and M. D. Reckase. Technical guidelines for assessing computerized adaptive tests. *Journal of Educational measurement*, 21(4):347–360, 1984. doi: 10.1111/j.1745-3984.1984.tb01039.x. [p36]

S. J. Haberman. Latent-class item response models (RR-05-28), 2005. [p23]

R. K. Hambleton, H. Swaminathan, and H. J. Rogers. *Fundamentals of Item Response Theory*. Sage, 1991. [p23]

E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society B*, 41(2):190–195, 1979. doi: 10.1111/j.2517-6161.1979.tb01072.x. [p29, 34]

R. J. Harvey and W. D. Murry. Scoring the myers-briggs type indicator: Empirical comparison of preference score versus latent-trait methods. *Journal of Personality Assessment*, 62(1):116–129, 1994. doi: 10.1207/s15327752jpa6201_11. [p23]

A. D. Ho and C. C. Yu. Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, 75(3):365–388, 2015. doi: 10.1177/0013164414548576. [p23]

M. C. Jones. The roles of ISE and MISE in density estimation. *Statistics & Probability Letters*, 12(1):51–56, 1991. doi: 10.1016/0167-7152(91)90163-L. [p31]

M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996. doi: 10.1080/01621459.1996.10476701. [p29]

N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978. doi: 10.1080/01621459.1978.10480103. [p26]

S. Li. Using a two-component normal mixture distribution as a latent distribution in estimating parameters of item response models. *Journal of Educational Evaluation*, 34(4):759–789, 2021. doi: 10.31158/JEEV.2021.34.4.759. [p23, 25, 27, 28, 31]

S. Li. The effect of estimating latent distribution using kernel density estimation method on the accuracy and efficiency of parameter estimation of item response models, 2022. [p23, 26, 29, 38]

S. Li. *IRTest: Parameter Estimation of Item Response Theory with Estimation of Latent Distribution*, 2024. URL https://CRAN.R-project.org/package=IRTest. R package version 2.0.0. [p31]

G. N. Masters. A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, 1982. doi: 10.1007/BF02296272. [p24]

K. May and A. W. Nicewander. Reliability and information functions for percentile ranks. *Journal of Educational Measurement*, 31(4):313–325, 1994. doi: 10.1111/j.1745-3984.1994.tb00449.x. [p36, 37]

A. L. McCutcheon. *Latent class analysis*. Sage, 1987. [p23]

R. J. Mislevy. Estimating latent distributions. *Psychometrika*, 49(3):359–381, 1984. doi: 10.1007/BF02306026. [p23, 26]

E. Muraki. A generalized partial credit model: Application of an em algorithm. *Applied Psychological Measurement*, 16(2):159–176, 1992. doi: 10.1177/014662169201600206. [p24]

G. Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests.* Danish Institute for Educational Research, Copenhagen, 1960. [p24]

D. Rizopoulos. ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software*, 17(5):1—25, 2006. doi: 10.18637/jss.v017.i05. [p29]

A. Robitzsch. *sirt: Supplementary Item Response Theory Models*, 2024. URL https://CRAN.R-project.org/package=sirt. R package version 4.1-15. [p24]

F. Samejima. *Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17)*. Psychometric Society, Richmond, VA, 1969. URL https://www.psychometrika.org/journal/online/MN17.pdf. [p34]

D. A. Sass, T. A. Schmitt, and C. M. Walker. Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education*, 21(1):65–88, 2008. doi: 10.1080/08957340701796415. [p23]

T.-J. Seong. Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, 14(3):299–311, 1990. doi: 10.1177/014662169001400307. [p23]

S. J. Sheather. Density estimation. *Statistical Science*, 19(4):588–597, 2004. doi: 10.1214/088342304000000297. [p29]

S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society B*, 53(3):683–690, 1991. doi: 10.1111/j.2517-6161.1991.tb01857.x. [p29]

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, Boca Raton, FL, 1986. [p29]

W. J. van der Linden. *Handbook of Item Response Theory: Three Volume Set*. CRC Press, 2016. [p23]

M. P. Wand and C. M. Jones. *Kernel Smoothing*. Chapman & Hall, Boca Raton, FL, 1995. [p29]

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org. [p35]

C. M. Woods. *RCLOG v.2: Software for Item Response Theory Parameter Estimation with the Latent Population Distribution Represented Using Spline-based Densities [Computer software]*. Washington University in St. Louis, 2006a. St. Louis, MO. [p23]

C. M. Woods. Ramsay-curve item response theory (rc-irt) to detect and correct for nonnormal latent variables. *Psychological Methods*, 11(3):253–270, 2006b. doi: 10.1037/1082-989X.11.3.253. [p23]

C. M. Woods. Estimating the latent density in unidimensional irt to permit non-normality. In S. P. Reise and D. A. Revicki, editors, *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*, chapter 4, pages 60–84. Routledge, New York, 2015. doi: 10.4324/9781315736013-4. [p23, 27]

C. M. Woods and N. Lin. Item response theory with estimation of the latent density using davidian curves. *Applied Psychological Measurement*, 33(2):102–117, 2009. doi: 10.1177/0146621608319512. [p23, 27, 28, 29, 38]

X. Xu and M. von Davier. Fitting the structured general diagnostic model to naep data (RR-08-27), 2008. [p23, 29]

A. Yadin. Using unique assignments for reducing the bimodal grade distribution. *ACM Inroads*, 4(1): 38–42, 2013. doi: 10.1145/2432596.2432612. [p23]

W. M. Yen. Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5 (2):245–262, 1981. doi: 10.1177/014662168100500212. [p38]

D. Zhang and M. Davidian. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57(3):795–802, 2001. doi: 10.1111/j.0006-341X.2001.00795.x. [p28]

M. F. Zimowski, E. Muraki, R. J. Mislevy, and D. R. Bock. *BILOG-MG: Multiple-group IRT Analysis and Test Maintenance for Binary Items [Computer software]*. Scientific Software International, 2003. [p23]

Seewoo Li
University of California, Los Angeles
Social Research Methodology, Department of Education
ORCiD: 0000-0002-6290-2777
seewooli@g.ucla.edu