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

by Seewoo Li

Abstract Item response theory (IRT) provides a statistical explanation on the relationship between respondents' latent trait and their responses to specific items. Latent variable is a key feature of IRT, where latent distribution is the population distribution of the latent variable. While latent distribution has been conventionally assumed to be normal, this assumption may not always hold true. In cases where the assumption is violated, latent distribution estimation (LDE) can improve parameter estimation accuracy by accounting for the non-normal characteristics of the latent distribution.
Although several studies have been conducted on LDE of IRT, there is a lack of available software programs to address this issue. This paper introduces IRTest, a software program developed for IRT analyses that incorporate LDE. This paper addresses the statistical background of LDE and examines the functionalities of IRTest. Examples of IRT analyses are provided to demonstrate the package's usages.

1 Introduction

Item response theory (IRT) is a widely used statistical framework for modeling the probabilistic relationship between examinees' levels of an underlying latent variable (i.e., ability parameters) and their responses to specific items (de Ayala 2009). As with some other psychometric methodologies, latent variable is a key feature of IRT, which typically represents invisible human traits in educational measurement or psychometric settings, such as students' academic ability, severity of depression, and degree of extrovertedness. In IRT, latent distribution — distribution of latent variable — can play an important role in the estimation process, particularly when using marginal maximum likelihood (MML). This is because the MML is obtained by marginalizing a joint likelihood with respect to the latent variable. Therefore, misspecification of the latent distribution can cause biases in parameter estimates of IRT.

While normal distribution has been conventionally assumed as a form of latent distribution of IRT, empirical evidence and potential drawbacks of violating this assumption have been addressed (Dudley-Marling 2020; Li 2022; Mislevy 1984; Sass, Schmitt, and Walker 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 violation of the normality assumption are 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 with their own characteristics: empirical histogram method (EHM: Bock and Aitkin 1981; Mislevy 1984), a mixture of two normal components (2NM: Mislevy 1984; Li 2021), Ramsay-curve method (RCM: Woods 2006a), Davidian-curve method (DCM: Woods and Lin 2009), log-linear smoothing method (LLS: Casabianca and Lewis 2015), and kernel density estimation method (KDM: Li 2022).

IRTest is an R package for unidimensional IRT analyses which aims to deal extensively with the issue of the LDE when the normality assumption is dubious. The **IRTest** is available from the Comprehensive R Archive Network (CRAN) at `https://cran.r-project.org/package=IRTest`. In **IRTest**, model-fitting functions estimate the latent distribution within MMLE-EM (MML estimation using EM algorithm) framework. Along with the conventional method of using normal distribution, five LDE methods (i.e., EHM, 2NM, DCM, LLS, and KDM) are currently available for the LDE implementation. Then, other functions of the **IRTest** employ the estimated latent distribution for accompanying IRT analyses.

Currently, there are not many software programs to perform the LDE, and their choices of the LDE methods are somewhat limited since the LDE may not be one of their main concerns: most of them offer only the EHM which may be the most straightforward way to perform the LDE (e.g., BILOG-MG Zimowski et al. 2003; flexMIRT Cai 2022); some software programs such as RCLOG (Woods 2006b) and LLSEM (Casabianca and Lewis 2011) focus on one particular LDE method; and **mirt** (Chalmers 2012) is equipped with the EHM and the DCM.

In this paper, implementations and a brief statistical background of **IRTest** is discussed. The remainder of this paper is organized as follows: Section 2 explains the basic statistical concept of the IRT parameter estimation and the LDE within the MMLE-EM framework. Section 3 discusses LDE methods. Section 4 examines the performance of the **IRTest** by a comparison of parameter estimates with those obtained from existing packages. Section 5 demonstrates the **IRTest** implementations. Lastly, Section 6 presents a discussion on the package.

2 LDE in IRT

This section presents a brief introduction of the statistical aspects of the LDE before delving into the details of an individual LDE method. The objectives of this section are 1) to explain basic concepts of IRT, 2) to provide insights into the role of the latent distribution in the estimation algorithm, and 3) to explain how the LDE can enhance parameter estimation accuracy.

Statistically, the overall LDE procedure can be conceptualized as an extension of Bock and Aitkin (1981)'s MMLE-EM procedure. Bock and Aitkin (1981) used the marginal likelihood which is dependent only on item parameters, whereas the LDE procedure includes both item and distribution parameters in the marginal likelihood. Note that not all LDE methods maximize the likelihood for the estimation of distribution parameters, even though they are nested in the MMLE-EM procedures.

2.1 IRT models

Most of the 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 the **IRTest**. And, 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 is applied to dichotomous item responses and the latter is applied to polytomous item responses.

More than one model can be applied to an 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, a pair of a dichotomous response model and a 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, a be the item discrimination parameter, b be the item difficulty parameter, and c $(0 < c < 1)$ be the guessing parameter. The item response function of the 3PLM for an item can be expressed as,

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

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, $1 - \Pr(u = 1 \,|\, \theta, a, b, c)$ represents the probability of the incorrect response. The model reduces to a 2PLM if $c = 0$, and to a 1PLM if $a = 1$ and $c = 0$.

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 for an item can be expressed as,

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

where $\exp\sum_{v=0}^{0}\left(a(\theta - b_v)\right) = 1$ for notational convenience. The equation (2.2.1) represents the probability of providing a response of $u = k$. The GPCM reduces to a PCM if $a = 1$. The dichotomous counterparts of the PCM and the GPCM would be the 1PLM and the 2PLM, respectively. They are reduced to their dichotomous counterparts when $M = 1$.

2.2 Role of latent distribution

In the MMLE-EM implementation, quadrature schemes have been 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^* named as a quadrature point. Typically, quadrature points are set to the middle of the grids. The default option of the **IRTest** is to set quadrature points from -6 to 6 with an increment of 0.1, resulting 121 quadrature points. In estimation functions of the **IRTest**, arguments `range` and `q` determine the range and the number of quadrature points, respectively. The corresponding probability mass function (PMF) of the latent variable can be expressed as,

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

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 is a conceptual terminology capable of indicating either a PDF $g(\theta)$ or a PMF $A(\theta^*)$ 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):

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

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

The equation (2.2.2) shows that the latent distribution is one of the elements of the marginal log-likelihood. Therefore, the specification of the latent distribution affects the value of the marginal log-likelihood of a model, thereby having a potential impact on the accuracy of the parameter estimates.

2.3 LDE in the MMLE-EM procedure

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

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

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 the equation (2.2.3), regarding γ_{jq} as a constant value, the $L_j\left(\theta_q^*\right)$ in the first term depends only on the item parameters, while the $A\left(\theta_q^*\right)$ in the second term depends only on the distribution parameters. This property of independence enables the separate estimation of the item and distribution parameters, from which a selection of alternative approaches for the distribution parameter estimation may emerge.

To elaborate more on the second term of the equation (2.2.3), it can be rewritten and simplified as,

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

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 aligned with the maximum likelihood estimation (MLE) approach of the MMLE-EM procedure, the distribution parameters are not always estimated by maximizing the equation (2.2.3). With the MLE approach still being a dominant choice, a different approach (i.e., KDM: Li 2022) has also been adopted for the distribution parameter estimation, which is discussed in the next section. In any case, every strategy for the estimation of distribution parameters makes use of the quantity γ_{jq} in its estimation procedure.

3 LDE methods

Theoretically, almost every density estimation method would become a candidate for an LDE method of 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 mainly discusses four LDE methods to illustrate the variation of the LDE strategies. The choice and order of methods in this section are not intended to imply any superiority of one method over the other.

3.1 Empirical histogram method (EHM)

One simple LDE strategy would be to directly employ the information obtained from the E-step. The EHM does this by simply calculating the expected probability 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:

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

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

Alternatively, an MLE solution can be derived in the following manner by using a Lagrangian multiplier. With the Lagrangian multiplier, the quantity to be maximized is

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

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

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

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, the EHM has been shown to be effective in reducing

biases in parameter estimates when the normality assumption is violated. However, when compared with other methods, the performance of the EHM could be limited to some extent because it may fail to screen out the random noise from the given information, thus producing less accurate parameter estimates (Li 2021; Woods 2015; Woods and Lin 2009). Some methods incorporate smoothing procedures to alleviate the impact 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/or skewness of the latent distribution. The 2NM method can be implemented in the estimation functions of the **IRTest** by specifying the argument as `latent_dist="2NM"`.

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

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

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

The $\log L_{\text{distribution}}$ is proportional to the one obtained by substituting $A\left(\theta_q^*\right)$ in the equation (2.2.3) with $g(\theta_q^* \mid \tau)$, resulting $\sum_{q=1}^{Q} \hat{f}_q \log g(\theta_q^* \mid \tau)$. 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.

For the estimation of 2NM parameters in the LDE, 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. This would make the likelihood in the M-step of the secondary EM algorithm be expressed as follows (Li 2021):

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

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):

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

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

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

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

and

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

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 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 MMLE-EM procedures. On the other hand, compared with nonparametric counterparts,

the flexibility of 2NM is limited to some extent. For example, the 2NM is incapable of forming a wiggly-shaped distribution.

3.3 Davidian-curve method (DCM)

The DCM uses a semi-nonparametric distribution where a hyperparameter $h = 1, 2, \ldots, 10$ determines the complexity of the density (Woods and Lin 2009). If $h = 1$, the distribution reduces to the normal distribution. Typically, hyperparameters in nonparametric density estimation methods not only add enough flexibility to the shape of the latent distribution but also prevent the latent distribution from overfitting the data. The DCM can be implemented in the estimation functions of the **IRTest** by specifying the argument as `latent_dist="DC"`.

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

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

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):

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

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

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

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

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

In DCM implementation, ten models are typically fitted according to each value of the hyperparameter. Then, the best model can be selected by Hannan-Quinn (HQ) criterion (Hannan and Quinn 1979):

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

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 a latent distribution (Woods and Lin 2009).

This paper does not go into details of the LLS (Casabianca and Lewis 2015) for its similarity to the DCM in the context of the paper: both of them take the MLE approach for the estimation, and their hyperparameter h determines the degree of smoothing and the number of distribution parameters.

3.4 Kernel density estimation method (KDM)

The KDM is a nonparametric method for conducting LDE, and it can be implemented in the estimation functions of the **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 the LDE of IRT, this means that $\hat{f}_q$ kernels are assigned to θ_q^*, which can be expressed as,

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

where $K(\cdot)$ is a kernel function and h is a hyperparameter often referred to as a bandwidth. This paper adopts Gaussian kernel as a general default choice (Li 2022).

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

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

In the equation above, $\hat{g_h}$ is the estimate of the latent distribution using the bandwidth h and $R(g'') = \int g''(x)\, dx$. It should be noted that this is a simplified version of AMISE since the Gaussian kernel is used. Then, an equation can be obtained by differentiating AMISE with respect to h and equating it to 0 (Gramacki 2018; Silverman 1986; Wand and Jones 1995):

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

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

On the one hand, the KDM and the DCM are similar in that a hyperparameter determines 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 result, whereas the LDE result of DCM depends on both the hyperparameter and the corresponding $h + 1$ density parameters. The absence of distribution parameter in the KDM, except for the hyperparameter (i.e., bandwidth) itself, allows the 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 would decrease computation time and expedite the analysis (Li 2022).

4 Package validation

This section examines and validates the estimation performance of the **IRTest** by comparing parameter estimates with those estimated from **mirt** (Chalmers 2012) and **ltm** (Rizopoulos 2006). The **mirt** and **ltm** are among the widely used R packages for IRT analyses. For the purpose, a dataset of ten dichotomous items and 1000 examinees is generated using the 2PLM and under the normality assumption.

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

Table 1 and 2 show 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 mean absolute difference of the a (item discrimination) parameter estimates between `IRTest_Dich()` and `mirt()` is 1.31×10^{-4}, and that between `IRTest_Dich()` and `ltm` is 0.5×10^{-4}. Similarly, the mean absolute difference of the b (item difficulty) parameter estimates between `IRTest_Dich()` and `mirt()` is 4.38×10^{-4}, and that between `IRTest_Dich()` and `ltm()` is 3.47×10^{-4}. This shows that the estimation accuracy of the four functions are practically identical.

Table 1: Item discrimination parameters and their estimates

	parameter	IRTest_Dich()	IRTest_Poly()	mirt()	ltm()
Item1	0.82	0.8630	0.8630	0.8629	0.8630
Item2	1.26	1.3884	1.3884	1.3884	1.3884
Item3	1.88	1.8243	1.8243	1.8242	1.8243
Item4	1.41	1.5255	1.5255	1.5254	1.5254
Item5	2.33	2.0115	2.0115	2.0118	2.0115
Item6	1.34	1.2502	1.2502	1.2502	1.2502
Item7	1.19	1.1518	1.1518	1.1517	1.1518
Item8	1.15	0.9963	0.9963	0.9962	0.9963
Item9	2.42	2.0414	2.0414	2.0411	2.0413
Item10	2.31	1.9970	1.9970	1.9972	1.9970

Table 2: Item difficulty parameters and their estimates

	parameter	IRTest_Dich()	IRTest_Poly()	mirt()	ltm()
Item1	1.23	1.0468	1.0468	1.0464	1.0465
Item2	-0.72	-0.6358	-0.6358	-0.6363	-0.6362
Item3	-0.19	-0.1633	-0.1633	-0.1637	-0.1636
Item4	-0.14	-0.0899	-0.0899	-0.0904	-0.0903
Item5	-1.16	-1.1787	-1.1787	-1.1792	-1.1791
Item6	0.14	0.1441	0.1441	0.1437	0.1438
Item7	0.29	0.2053	0.2053	0.2049	0.2050
Item8	1.28	1.4411	1.4411	1.4408	1.4408
Item9	0.12	0.1275	0.1275	0.1270	0.1272
Item10	-0.96	-0.9272	-0.9272	-0.9277	-0.9276

5 Implementations of the IRTest

The primary purpose of this section is to demonstrate the usages of the **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 Generic Conspiracist Beliefs Scale (GCBS) data (Brotherton, French, and Pickering 2013) available from the Open-Source Psychometric Project at `http://openpsychometrics.org/_rawdata/GCBS.zip`.

5.1 The effect of the LDE

This section illustrates the effect of the LDE by showing an improvement in estimation accuracy. For this purpose, I generate an artificial item response dataset to use true values as evaluation criteria, and ability parameters are generated from a non-normal distribution.

Data generation Using the function `DataGeneration()`, a dataset of 40 dichotomous items and 2000 respondents is generated by

```
Alldata <- 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 random generating the dataset, and the 2NM distribution is employed to simulate a non-normal latent distribution (`latent_dist = "2NM"`) with its parameters `d = 1.414`, `sd_ratio = 2`, and `prob = 2/3` (for the reparameterization of 2NM, see Li 2021, 2023).

The shape of the distribution is highly skewed and likely to be rare in a real situation (see Figure 1). However, the shape of the distribution is chosen to clearly demonstrate the effect of the LDE. Therefore, cautions are needed in interpreting and understanding the magnitude of the effect.

Model fitting The function `IRTest_Dich()` is used for the model-fitting because all items are dichotomously scored, and the LDE method is specified by `latent_dist` argument. Two types of ability parameter estimates to be illustrated are expected *a posteriori* (EAP) and maximum likelihood estimate (MLE; note that *MLE* is also used as an abbreviation of maximum likelihood estimation strategy). In the **IRTest**, estimation method of ability parameter is determined by specifying the argument `ability_method`, where the default option is the EAP. Either an estimation function (e.g., `IRTest_Dich()`) or `factor_score()` can be used to calculate ability parameter estimates.

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

Among these two models, `model_normal` assumes a normal distribution on the latent distribution, whereas `model_KDM` estimates the latent distribution using KDM during its MMLE-EM procedure. The KDM is arbitrarily selected for an illustrative purpose.

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

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

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

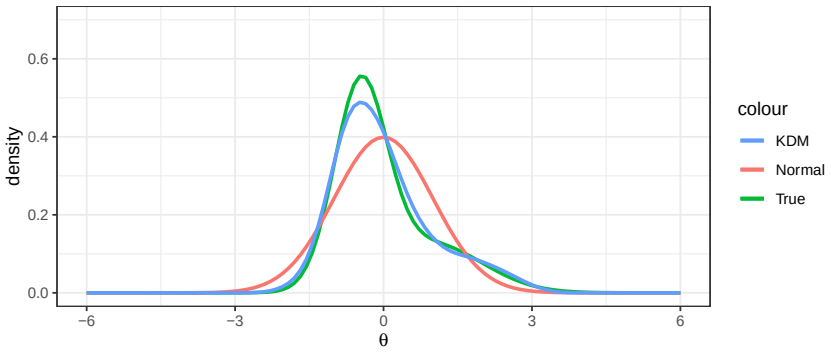

Figure 1: The true, normal, and estimated latent distributions

Figure 1 is drawn with a slight addition of aesthetic codes to the object `density_plot`. The Figure 1 shows the density curves of the true latent distribution, the normal distribution, and the estimated

Table 3: RMSEs from the normality assumption model and the KDM model

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

latent distribution. Even with the discrepancy between the true latent distribution (green line) and the normal distribution (red line), the estimated distribution (blue line) almost recovered the shape of the true latent distribution.

Parameter estimates Differences in parameter estimates caused by the LDE are portrayed in Figure 2, where x-axis is for true values and y-axis is for estimated values. It shows that the parameter estimates from the KDM model (blue dots) are located closer to the $y = x$ line than those from the normality assumption model (red dots), which indicates that the estimates from the KDM model are more accurate. Root mean squared error (RMSE) is used as an evaluation criterion for the accuracy of the parameter estimates, where lower RMSE indicates higher accuracy. Table 3 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 the RMSEs for $\hat{a}$ and $\hat{b}$ substantiate the effectiveness of the LDE procedure in enhancing the estimation accuracy of item parameters, where the magnitudes of the RMSE decreases are originated from the highly skewed latent distribution. As stated in Section 2.2, this shows that appropriate specification of a latent distribution can have a positive impact on the accuracy of parameter estimates.

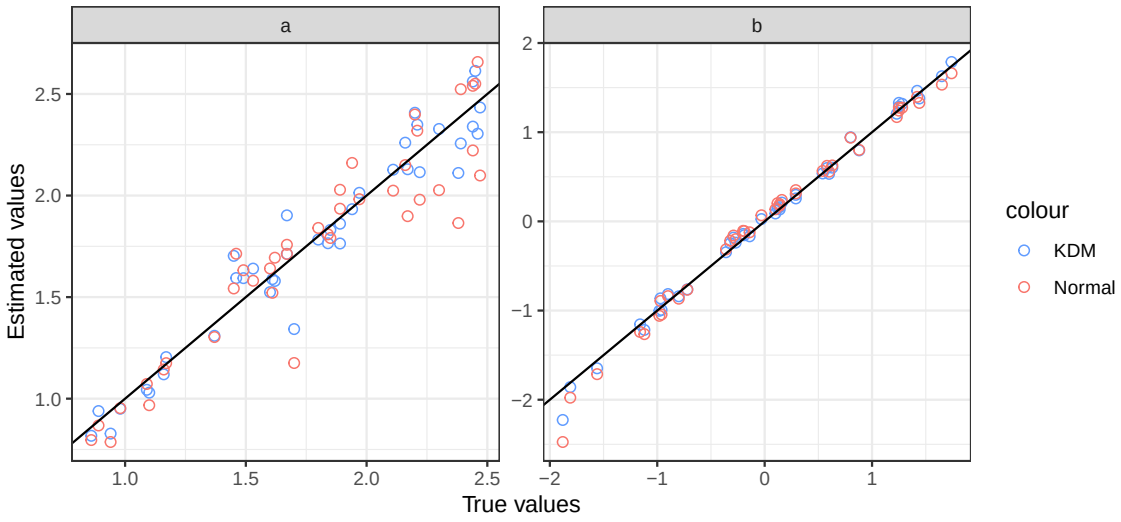

Figure 2: Differences in item parameter estimates caused by the LDE

5.2 An empirical example

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

Data The GCBS data contains responses from 15 polytomous items and 2391 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. The **IRTest** uses a full-information maximum likelihood (FIML) approach in handling missing data.

Model selection The DCM and the KDM are used in performing the LDE for an illustrative purpose. PCM, GPCM, and graded response model (GRM: Samejima 1969) are available IRT models of `IRTest_Poly()`, where the default is the GPCM. Models are fitted as follows:

```
# Davidian-curve method
DC1 <- IRTest_Poly(data = data_GCBS, latent_dist = "DC", h = 1)
DC2 <- IRTest_Poly(data = data_GCBS, latent_dist = "DC", h = 2)
DC3 <- IRTest_Poly(data = data_GCBS, latent_dist = "DC", h = 3)
DC4 <- IRTest_Poly(data = data_GCBS, latent_dist = "DC", h = 4)
DC5 <- IRTest_Poly(data = data_GCBS, latent_dist = "DC", h = 5)
DC6 <- IRTest_Poly(data = data_GCBS, latent_dist = "DC", h = 6)
DC7 <- IRTest_Poly(data = data_GCBS, latent_dist = "DC", h = 7)
DC8 <- IRTest_Poly(data = data_GCBS, latent_dist = "DC", h = 8)
DC9 <- IRTest_Poly(data = data_GCBS, latent_dist = "DC", h = 9)
DC10 <- IRTest_Poly(data = data_GCBS, latent_dist = "DC", h = 10)

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

A model-selection step is required for the DCM: the "best-DCM" can be selected by the HQ criterion (Hannan and Quinn 1979) presented in the equation (2.3.3). After ten models are fitted, the best model can be selected with the function `best_model()` (`anova()` can also be used). The `best_model()` employs a certain criterion to determine the best model, where available options are `"logLik"`, `"deviance"`, `"AIC"`, `"BIC"`, and `"HQ"`. The default is `criterion = "HQ"`.

```
best_model(DC1, DC2, DC3, DC4, DC5, DC6, DC7, DC8, DC9, DC10)
```

```
#> The best model: DC8
#>
#>            HQ
#> DC1  91232.30
#> DC2  91236.40
#> DC3  91237.18
#> DC4  91220.93
#> DC5  91217.55
#> DC6  91211.66
#> DC7  91209.33
#> DC8  91207.92
#> DC9  91209.94
#> DC10 91213.25
```

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

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

```
best_model(DC8, KDM)
```

```
#> The best model: KDM
#>
#>           HQ
#> DC8 91207.92
#> KDM 91177.92
```

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

Summary of the model A brief summary of the model-fitting result can be printed with `summary()`. It presents 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:
#> dichotomous   0
#> polyotomous   15
#>
#> The Estimated Latent Distribution:
#> method - KDE
#> ----------------------------------------
#>
#>                     @ @ .
#>                   @ @ @ @
#>                 . @ @ @ @ @
#>               . @ @ @ @ @ @ .
#>               @ @ @ @ @ @ @ @
#>           . @ @ @ @ @ @ @ @ @ @
#>         . @ @ @ @ @ @ @ @ @ @ @ .
#>     . @ @ @ @ @ @ @ @ @ @ @ @ @ @ .
#> . @ @ @ @ @ @ @ @ @ @ @ @ @ @ @ @ @ .
#> +---------+---------+---------+---------+
#> -2        -1        0         1         2
```

Alternatively, the shape of the estimated latent distribution can be easily examined by `plot()` which produces a ggplot-class object. Figure 3 shows the estimated latent distribution of the GCBS data which is left-skewed. The Figure 3 is produced by adding aesthetic codes to `plot()` and the curve of normal distribution. If a PDF is estimated from an estimation function, additional arguments in `plot()` are passed on to `stat_function()` of **ggplot2** (Wickham 2016). Otherwise, if a PMF is estimated (e.g., EHM or LLS), they are passed on to `geom_line()` of **ggplot2**.

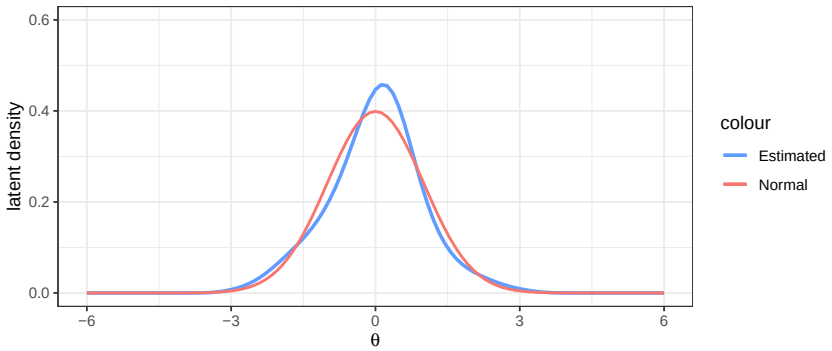

Figure 3: Estimated latent distribution for the GCBS data

Parameter estimates Users can have access to item parameter estimates and their standard errors with `coef()` and `coef_se()`, respectively.

```
coef(KDM)
```

```
#>                a         b_1          b_2          b_3        b_4
#> Item1  0.9689175 -0.76302060 -0.391284939 -0.855329989  0.3535797
#> Item2  1.0789773 -0.53792319  0.009783417  0.005707348  0.7242646
```

```
#> Item3  0.7387106  1.67240035  0.366682299  0.966132283  1.3765691
#> Item4  1.3101835 -0.13464636  0.161326651  0.204693609  1.3519835
#> Item5  0.7816844 -0.54656956 -0.191683411 -0.831793841  0.7254484
#> Item6  1.1236432 -0.46965320 -0.187145629 -0.214048841  0.6453831
#> Item7  1.1752611 -0.07561864  0.301474839  0.146380253  0.9623862
#> Item8  0.6667503  1.54880707 -0.087406668  0.644577920  0.3019064
#> Item9  1.0243844  0.53050596  0.622137122  0.614787964  1.3307289
#> Item10 0.5422616 -0.67410890 -0.728343182 -1.374795950  0.4975419
#> Item11 1.0969804 -0.85223020 -0.647110890 -0.193550801  0.7969690
#> Item12 1.6043650 -0.13515863  0.138428327  0.313106850  0.9981359
#> Item13 0.8859066  1.15645384  0.243560168  1.098843084  1.3876961
#> Item14 1.0693400 -0.36458551 -0.054603047 -0.080356862  0.8726566
#> Item15 0.8699796 -2.01564440 -1.551315231 -1.872536010 -0.6573154
```

```
coef_se(KDM)
```

```
#>                a        b_1        b_2        b_3        b_4
#> Item1  0.03678552 0.08584978 0.08654735 0.08019185 0.05779224
#> Item2  0.03966117 0.06329963 0.06584249 0.06508489 0.06233005
#> Item3  0.03208922 0.11778115 0.10831715 0.11904995 0.13614052
#> Item4  0.04714667 0.05119173 0.05476215 0.05471753 0.06194313
#> Item5  0.03074436 0.09686544 0.09959946 0.09509069 0.07449369
#> Item6  0.04105996 0.06583429 0.06718378 0.06322692 0.05630586
#> Item7  0.04324191 0.05617220 0.06338708 0.06467953 0.06362862
#> Item8  0.02800727 0.12930732 0.12011036 0.12179261 0.12290677
#> Item9  0.04025235 0.06541069 0.07426051 0.08161882 0.09079209
#> Item10 0.02420014 0.15487216 0.14764260 0.13523620 0.09904366
#> Item11 0.04013006 0.07370442 0.06734305 0.05764310 0.05617621
#> Item12 0.05686928 0.04318224 0.04603594 0.04652593 0.04914371
#> Item13 0.03648803 0.09029968 0.08542312 0.09660738 0.11458583
#> Item14 0.03930702 0.06541139 0.06780684 0.06541472 0.06253614
#> Item15 0.03830836 0.16985798 0.13629516 0.10927581 0.06689061
```

Likewise, `factor_score()` returns ability parameter estimates (`factor_score()$theta`) and their standard errors (`factor_score()$theta_se`) which are not printed here because both of their lengths are 2397.

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

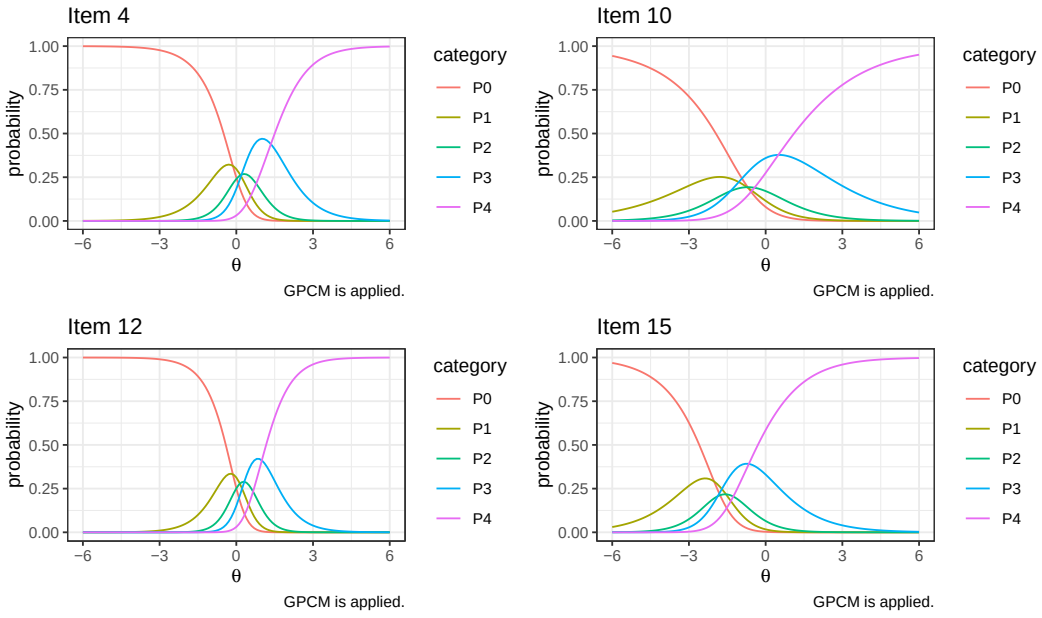

Figure 4: Item response functions of Item 4, 10, 12, and 15

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

Reliability coefficient Among various types of IRT reliability coefficients, the **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 in terms of obtaining reliability coefficients for individual items. The function `reliability()` returns all of the coefficients mentioned above: item reliability coefficients, and test reliability coefficients on the θ scale and the summed-score scale.

```
reliability(KDM)
```

```
#> $summed.score.scale
#> $summed.score.scale$test
#> test reliability
#>        0.9293986
#>
#> $summed.score.scale$item
#>     Item1     Item2     Item3     Item4     Item5     Item6     Item7     Item8
#> 0.5039288 0.5248048 0.3818573 0.5741023 0.4227403 0.5468998 0.5526087 0.3795150
#>     Item9    Item10    Item11    Item12    Item13    Item14    Item15
#> 0.4929462 0.2898269 0.5196557 0.6437788 0.4386315 0.5232480 0.4021903
#>
#>
#> $theta.scale
#> test reliability
#>        0.9159472
```

Item and test information functions In the **IRTest**, `inform_f_item()` and `inform_f_test()` evaluate item and test information, respectively. Figure 5 visually illustrates how each item contributes to the test information and how item information functions add up altogether to form the test information function. The test information function is drawn with black line and the item information functions are colored.

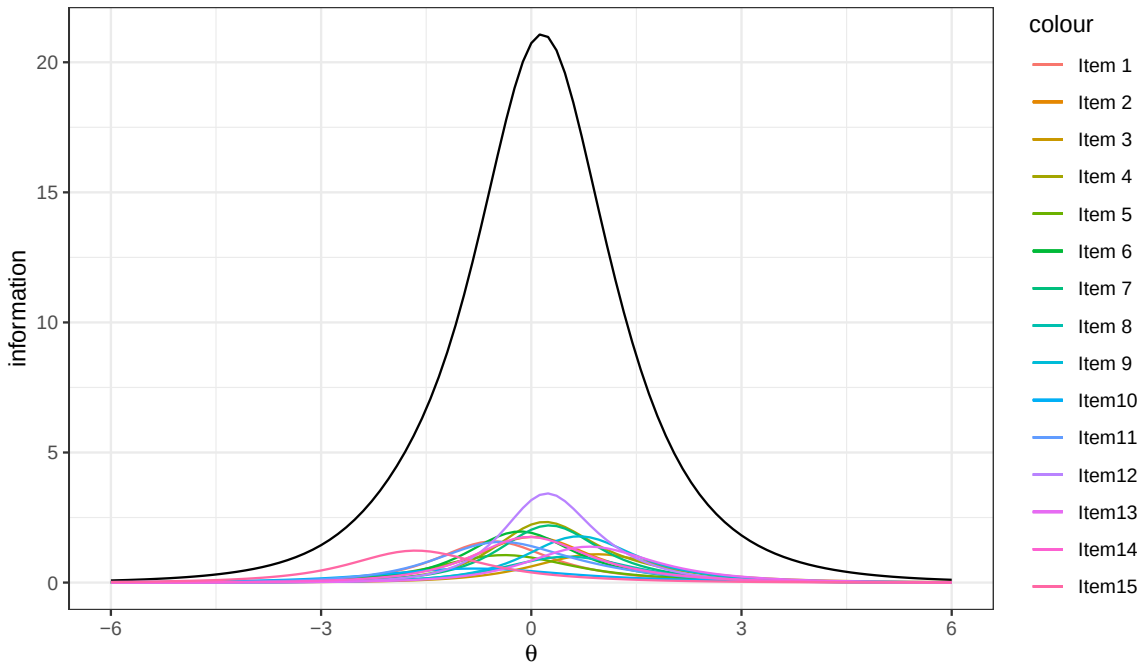

Figure 5: Item and test information functions

6 Discussion

While some may not find the extent of biases resulting from deviations from the normality assumption particularly significant in the context of IRT, the LDE remains an appealing option. Within IRT, an increase in the number of items and respondents can improve parameter estimation accuracy, which may cost a certain amount of time and resources. Meanwhile, an appropriate usage of the LDE would increase the estimation accuracy at almost no cost.

The R package **IRTest** offers users the choice of LDE methods for conducting IRT analyses, and its functions utilize the estimated latent distribution throughout the analysis. The **IRTest** is actively being updated, and some potential enhancements can be made for the 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 be explored in the field of LDE than what has been explored. Therefore, further studies on LDE are anticipated to provide valuable guidance for both package users and the developer.

Acknowledgments

Special thanks to Nagap Park and Hyesung Shin for their insightful suggestions that improved the article.

Additional features

An analysis of mixed-format data An IRT analysis of mixed-format data can also be conducted in the **IRTest**. The function `IRTest_Mix()` takes charge of it. The difference between `IRTest_Mix()` and `IRTest_Dich()` (or `IRTest_Poly()`) is that `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_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_Dich()` and `IRTest_Poly()`.

The `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 `item_fit()`. Currently, Bock (1960)'s χ^2 and Yen (1981)'s Q_1 are available.

References

Baker, Frank B., and Seock-Ho Kim. 2004. *Item Response Theory: Parameter Estimation Techniques*. 2nd ed. CRC press.

Birnbaum, Allan. 1968. "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In *Statistical Theories of Mental Test Scores*, edited by Frederic M. Lord and Melvin R. Novick, 397--479. Addison-Wesley.

Bock, Darrell R. 1960. "Methods and Applications of Optimal Scaling." Chapel Hill: NC: University of North Carolina Psychometric Laboratory Research Memorandum, No. 25.

Bock, Darrell R., and Murray Aitkin. 1981. "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm." *Psychometrika* 46 (4): 443–59. https://doi.org/10.1007/BF02293801.

Brotherton, Robert, Christopher C. French, and Alan D. Pickering. 2013. "Measuring Belief in Conspiracy Theories: The Generic Conspiracist Beliefs Scale." *Frontiers in Psychology* 4. https://doi.org/10.3389/fpsyg.2013.00279.

Cai, Li. 2022. *flexMIRT Version 3.65: Flexible Multilevel Multidimensional Item Analysis and Test Scoring [Computer Software]*. Vector Psychometric Group.

Casabianca, Jodi M., and Charles Lewis. 2011. *LLSEM 1.0: LogLinear Smoothing in an Expectation Maximization Algorithm for Item Response Theory Item Parameter Estimation [Computer Software]*. Washington University in St. Louis.

———. 2015. "IRT Item Parameter Recovery with Marginal Maximum Likelihood Estimation Using Loglinear Smoothing Models." *Journal of Educational and Behavioral Statistics* 40 (6): 547–78. https://doi.org/10.3102/1076998615606112.

Chalmers, Philip R. 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.

de Ayala, Rafael J. 2009. *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.

Dudley-Marling, Curt. 2020. "The Tyranny of the Normal Curve: How the "Bell Curve" Corrupts Educational Research and Practice." In *Groupthink in Science: Greed, Pathological Altruism, Ideology, Competition, and Culture*, edited by David M. Allen and James W.Howell, 201–10. Cham: Springer. https://doi.org/10.1007/978-3-030-36822-7_17.

Gramacki, Artur. 2018. *Nonparametric Kernel Density Estimation and Its Computational Aspects*. Springer.

Green, Bert F., Darrell R. Bock, Lloyd G. Humphreys, Robert L. Linn, and Mark D. Reckase. 1984. "Technical Guidelines for Assessing Computerized Adaptive Tests." *Journal of Educational Measurement* 21 (4): 347–60. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x.

Hannan, Edward J., and Barry G. Quinn. 1979. "The Determination of the Order of an Autoregression." *Journal of the Royal Statistical Society B* 41 (2): 190–95. https://doi.org/10.1111/j.2517-6161.1979.tb01072.x.

Harvey, Robert J., and William D. Murry. 1994. "Scoring the Myers-Briggs Type Indicator: Empirical Comparison of Preference Score Versus Latent-Trait Methods." *Journal of Personality Assessment* 62 (1): 116–29. https://doi.org/10.1207/s15327752jpa6201_11.

Ho, Andrew D., and Carol C. Yu. 2015. "Descriptive Statistics for Modern Test Score Distributions: Skewness, Kurtosis, Discreteness, and Ceiling Effects." *Educational and Psychological Measurement* 75 (3): 365–88. https://doi.org/10.1177/0013164414548576.

Jones, Michael C., J. S. Marron, and Simon J. Sheather. 1996. "A Brief Survey of Bandwidth Selection for Density Estimation." *Journal of the American Statistical Association* 91 (433): 401–7. https://doi.org/10.1080/01621459.1996.10476701.

Laird, Nan. 1978. "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution." *Journal of the American Statistical Association* 73 (364): 805–11. https://doi.org/10.1080/01621459.1978.10480103.

Li, Seewoo. 2021. "Using a Two-Component Normal Mixture Distribution as a Latent Distribution in Estimating Parameters of Item Response Models." *Journal of Educational Evaluation* 34 (4): 759–89. https://doi.org/10.31158/JEEV.2021.34.4.759.

———. 2022. "The Effect of Estimating Latent Distribution Using Kernel Density Estimation Method on the Accuracy and Efficiency of Parameter Estimation of Item Response Models." Master's thesis. Seoul: Yonsei University.

———. 2023. *IRTest: Parameter Estimation of Item Response Theory with Estimation of Latent Distribution*. https://CRAN.R-project.org/package=IRTest.

Masters, Geoff N. 1982. "A Rasch Model for Partial Credit Scoring." *Psychometrika* 47 (2): 149–74. https://doi.org/10.1007/BF02296272.

May, Kim, and Alan W. Nicewander. 1994. "Reliability and Information Functions for Percentile Ranks." *Journal of Educational Measurement* 31 (4): 313–25. https://doi.org/10.1111/j.1745-3984.1994.tb00449.x.

Mislevy, Robert J. 1984. "Estimating Latent Distributions." *Psychometrika* 49 (3): 359–81. https://doi.org/10.1007/BF02306026.

Muraki, Eiji. 1992. "A Generalized Partial Credit Model: Application of an EM Algorithm." *Applied Psychological Measurement* 16 (2): 159–76. https://doi.org/10.1177/014662169201600206.

Rasch, Georg. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Rizopoulos, Dimitris. 2006. "Ltm: An r Package for Latent Variable Modeling and Item Response Analysis." *Journal of Statistical Software* 17 (5): 1--25. https://doi.org/10.18637/jss.v017.i05.

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

Sass, Daniel A., Thomas A. Schmitt, and Cindy M. Walker. 2008. "Estimating Non-Normal Latent Trait Distributions Within Item Response Theory Using True and Estimated Item Parameters." *Applied Measurement in Education* 21 (1): 65–88. https://doi.org/10.1080/08957340701796415.

Seong, Tae-Je. 1990. "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. https://doi.org/10.1177/014662169001400307.

Sheather, Simon J. 2004. "Density Estimation." *Statistical Science* 19 (4): 588–97. https://doi.org/10.

1214/088342304000000297.

Sheather, Simon J., and Michael C. Jones. 1991. "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation." *Journal of the Royal Statistical Society B* 53 (3): 683–90. https://doi.org/10.1111/j.2517-6161.1991.tb01857.x.

Silverman, Bernard W. 1986. *Density Estimation for Statistics and Data Analysis*. Boca Raton, FL: Chapman & Hall.

Wand, Matt P., and Chris M. Jones. 1995. *Kernel Smoothing*. Boca Raton, FL: Chapman & Hall.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Woods, Carol M. 2006a. "Ramsay-Curve Item Response Theory (RC-IRT) to Detect and Correct for Nonnormal Latent Variables." *Psychological Methods* 11 (3): 253–70. https://doi.org/10.1037/1082-989X.11.3.253.

———. 2006b. *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.

———. 2015. "Estimating the Latent Density in Unidimensional IRT to Permit Non-Normality." In *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*, edited by Steven P. Reise and Dennis A. Revicki, 60–84. New York: Routledge. https://doi.org/10.4324/9781315736013-4.

Woods, Carol M., and Nan Lin. 2009. "Item Response Theory with Estimation of the Latent Density Using Davidian Curves." *Applied Psychological Measurement* 33 (2): 102–17. https://doi.org/10.1177/0146621608319512.

Yadin, Aharon. 2013. "Using Unique Assignments for Reducing the Bimodal Grade Distribution." *ACM Inroads* 4 (1): 38–42. https://doi.org/10.1145/2432596.2432612.

Yen, Wendy M. 1981. "Using Simulation Results to Choose a Latent Trait Model." *Applied Psychological Measurement* 5 (2): 245–62. https://doi.org/10.1177/014662168100500212.

Zhang, Daowen, and Marie Davidian. 2001. "Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data." *Biometrics* 57 (3): 795–802. https://doi.org/10.1111/j.0006-341X.2001.00795.x.

Zimowski, Michele F., Eiji Muraki, Robert J. Mislevy, and Darrell R. Bock. 2003. *BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items [Computer Software]*. Scientific Software International.

Seewoo Li
Yonsei University
Educational Measurement and Statistics
Department of Education
Seoul, Korea
cu@yonsei.ac.kr