情報処理学会研究報告
IPSJ SIG Technical Report

マトリクス分解から見た項目反応理論の精度評価と実試験データへの適用

廣瀬 英雄[1,a)]

概要：マトリクス分解から見た項目反応理論の精度評価を実試験データによって確認した．項目反応理論から推定された応答マトリクスと同等なマトリクス分解による低ランクマトリクスのランクは低いが項目反応理論の予測能力は高い．

キーワード：項目反応理論, マトリクス分解, 特異値分解, フローベニウスノルム, 不完全マトリクス．

Matrix Decomposition Perspective for Accuracy Assessment of Item Response Theory with Applications to Actual Examination Data

Hirose Hideo[1,a)]

Abstract: This paper investigates the predictive effectiveness of item response theory from matrix decomposition perspective. Comparing the difference in terms of matrix norm between the observed item response matrix and the estimated item response matrix with that between the observed item response matrix and the low-rank approximation matrix generated by the matrix decomposition method, it is found that the rank of the generated low-rank approximation matrix that is equivalent to the estimated item response matrix is very low. However, the predictive ability of item response theory still seems to be high enough.

Keywords: Item response theory, matrix decomposition, singular value decomposition, Frobenius matrix norm, incomplete matrix treatment.

1. Introduction

Item response theory (IRT) ([1], [2], [6], [23]) is a theory based on a statistical parametric model that simultaneously assesses abilities of examinees and difficulties of problems. Because of its versatility and reliability, IRT has been regarded as one of the standard methods in assessing the performance of examinees. For this reason, IRT is used in various official examinations, including the TOFLE. To an *observed item response matrix* consisting of examinee user rows and problem item columns, estimates of IRT parameters and their confidence intervals can be obtained using the maximum likelihood estimation method.

The maximum likelihood estimators are known to be consistent and asymptotically efficient under certain conditions ([10]); that is, no consistent estimators have lower asymptotic mean squared errors other than the maximum likelihood estimators. This means that under the assumed mathematical model and its parameter space, the estimators are the best. However, there might be other models that are better than IRT. While the AIC (Akaike's information criterion) is often used to compare the superiority of parametric models, we use another criterion, the root mean squared error (RMSE), in order to accomplish such a purpose. This is a primary challenge to see the prediction effectiveness of item response theory from a new

[1] Bioinformatics Center, Kurume University Graduate School of Medicine, Fukuoka Tokyo 830–0011, Japan
[a)] hirose_hideo@kurume-u.ac.jp

perspective; we look at the item response matrix itself directly from the matrix decomposition perspective. The typical data cases corresponding to this new look appear in the examination data in education.

Using the estimates for parameters in IRT, the item response matrix can be reconstructed; we call this the *estimated item response matrix*. Then, the difference between the observed and estimated item response matrices can be computed using an appropriate matrix norm such as the Frobenius matrix norm. Thus, it is possible to measure how close the observed item response matrix is to the estimated item response matrix. This is the criterion to measure the difference between two matrices.

Many researchers have proposed new methods to achieve superiority over the standard IRT performance. For example, multidimensional item response theory (MIRT) (see [15]) and knowledge tracing (KT) (see [13], [24]) have been proposed to find examinee proficiency using parametric models. [25] describe the results of a performance comparison among those parametric models. In the evaluation of parametric models, the log-likelihood values are primarily used. In a nonparametric approach, [21] show how to predict student performance using a recommender system. In addition, [20] use a recommender system to predict student performance. Since recommender systems often use a matrix factorization algorithm, the RMSE is used to evaluate the closeness of the two matrices. More complex cases have also been proposed, with [11] integrating KT and IRT, and [26] comparing deep learning approaches to simple IRT.

However, unlike papers that provide such new mathematical models, this paper intends to examine the effectiveness of IRT itself from a different perspective as slightly mentioned above using actual educational data cases. To accomplish this, we use matrix decomposition (MD) perspective. [7] introduced matrix completion (MC) and low-rank singular value decomposition (SVD) to evaluate the difference between two matrices. By using SVD, we can obtain a low-rank matrix that is close to the original matrix in the sense of Frobenius matrix norm. The second challenge of this paper is to apply the method to a number of actual examination data cases performed at universities. By applying the matrix decomposition and singular value decomposition methods to more than 40 examination data cases, ranging from small to large matrix sizes, we were able to derive a very clear conclusion about the effective approximated low-rank matrix that is equivalent to the estimated item response matrix. Therefore, the objective of the paper is to clarify the position of the IRT performance in the sense of *low-rank approximation matrix* equivalent to the estimated item response matrix using a number of actual examination data sets.

2. Item response theory

2.1 Mathematical model

The standard IRT estimates proficiency parameters θ_i $(i = 1, \ldots, n)$ and problem parameters a_j, b_j, c_j $(j = 1, \ldots, m)$ simultaneously by using the observed item response matrix. Usually, this item response matrix (matrix size is $n \times m$) consists of 1/0 valued elements δ_{ij}, with the value 1 for the (i, j) element corresponding to the case where examinee i solved question j correctly and the value 0 for the case where he/she solved it incorrectly.

Assume that the logistic probability function p_{ij} of examinee i correctly answering question j is expressed such that

$$\begin{aligned} p_{ij}(\theta_i; a_j, b_j, c_j) &= c_j + \frac{1 - c_j}{1 + \exp\{-1.7a_j(\theta_i - b_j)\}}, \\ &= 1 - q_{ij}(\theta_i; a_j, b_j, c_j), \end{aligned} \tag{1}$$

where θ_i is called the ability for examinee i and a_j, b_j, c_j are called the discrimination parameter, difficulty parameter, and pseudo-guessing parameter, respectively; q_{ij} is the probability that examinee i answers question j incorrectly.

2.2 Parameter estimation

Using the maximum likelihood estimation (MLE) method, the maximum likelihood estimates $\hat{\theta}_i$ and $\hat{a}_j, \hat{b}_j, \hat{c}_j$ for parameters θ_i and a_j, b_j, c_j can be obtained by maximizing the likelihood function,

$$L = \prod_{i=1}^{n} \prod_{j=1}^{m} \left(p_{ij}^{\delta_{ij}} \times q_{ij}^{1-\delta_{ij}} \right). \tag{2}$$

When only difficulty parameter b_j is considered, such the model is called the Rasch model. Usually, the two-parameter model ($c_j = 0$) is the standard, and we will deal with this case below. Also, in terms of recommender systems, we will refer to examinees as users and questions as items.

If we denote parameters θ_i and a_j, b_j, c_j together by Θ, and the observed matrix by $\Delta = (\delta_{ij})$, then the estimation process is expressed as follows.

$$\Delta \rightsquigarrow \hat{\Theta}. \tag{3}$$

2.3 Estimated item response matrix

Applying the delta method to p_{ij} in equation (1) as a

function of $\hat{\Theta}$, we can obtain $\hat{\delta}_{ij}$ which is a continuous value in $[0,1]$. This estimation process can be expressed such that

$$\hat{\Theta} \rightsquigarrow \hat{\Delta}, \tag{4}$$

and we call $\hat{\Delta}$ the *estimated item response matrix*. The value $\hat{\delta}_{ij}$ is corresponding to the probability of correctly answering the question using equation (1).

Considering such treatment, the δ_{ij} value is extended from a discrete value of $1/0$ to a continuous value of $[0,1]$, although δ_{ij} takes values $\delta_{ij}=1$ if the question is successfully answered and $\delta_{ij}=0$ if it is not. We also deal with the null value of the element (i,j), corresponding to the case where the examinee i has not tackle problem j, or the case that the response is unknown. How to deal with such cases is explained in [9], [16], [17].

Once, estimates $\hat{a}_j$ and $\hat{b}_j$ for a_j and b_j are obtained, we can perform estimation procedure for θ_i to each i independently. If there is a sequence of random variables X_l satisfying

$$\sqrt{l}[X_l-\theta_i] \xrightarrow{d} \mathcal{N}(0,\sigma_i^2), \tag{5}$$

where θ_i and σ_i^2 are finite valued constants, $\mathcal{N}$ is a normal distribution, and $\xrightarrow{d}$ denotes convergence in distribution, then

$$\sqrt{l}[g(X_l)-g(\theta_i)] \xrightarrow{d} \mathcal{N}(0,\sigma_i^2\cdot[g'(\theta_i)]^2), \tag{6}$$

for any function g satisfying the property that $g'(\theta)$ exists and is non-zero valued. According to this, and regarding g as p_{ij} in equation (1), $\hat{\delta}_{ij}$ becomes optimal in the likelihood sense. Therefore, $\hat{\Delta}$ is optimal in the mathematical model assuming equation (1) and the parameter space of Θ.

3. Singular value decomposition

3.1 Singular value decomposition procedure

Assuming that $A=(a_{ij})$ is a $m\times n$ matrix. Then, $A^{\mathsf{T}}A$ becomes a $n\times n$ symmetric matrix, and AA^{T} becomes a $m\times m$ symmetric matrix, where A^{T} denotes the transpose of A. Eigen values and eigen vectors to these two matrices $A^{\mathsf{T}}A$ and AA^{T} are the same if they exist. We denote the eigen values and eigen vectors to matrix $A^{\mathsf{T}}A$ as $\{\xi_1,\xi_2,\cdots,\xi_n\}$ and $\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}$. That is,

$$A^{\mathsf{T}}A\boldsymbol{v}_i=\xi_i\boldsymbol{v}_i. \tag{7}$$

Eigen values can be reordered such that $\xi_1\geq\xi_2\geq\cdots\geq\xi_r>0,\xi_{r+1}=\cdots=\xi_n=0$, where r is the rank of $A^{\mathsf{T}}A$. Since $A^{\mathsf{T}}A$ is symmetric, eigen vectors can be made as orthonormal system. That is, $\boldsymbol{v}_i\cdot\boldsymbol{v}_j=I_{ij}$, where I_{ij} is the indicator function; i.e., $I_{ii}=1$, and $I_{ij}=0$ $(i\neq j)$. We make vector $\boldsymbol{u}_i$ by $\boldsymbol{u}_i=A\boldsymbol{v}_i/\sigma_i$, $(i\geq r)$, where $\sigma_i=\sqrt{\xi_i}$. In addition, if we produce matrices $U=(\boldsymbol{u}_i)$ and $V=(\boldsymbol{v}_j)$, then A can be expressed as $A=U\Sigma V^{\mathsf{T}}$, or equivalently, $A=\sum_{l=1}^{r}\sigma_l\boldsymbol{u}_l\boldsymbol{v}_l^{\mathsf{T}}$. Here, Σ is a diagonal matrix using σ_i. This is the typical singular value decomposition (SVD) (see [5], [18], [19]).

3.2 Generating the low-rank matrix

We define A_k such that

$$A_k=\sum_{l=1}^{k}\sigma_l\boldsymbol{u}_l\boldsymbol{v}_l^{\mathsf{T}}, \tag{8}$$

using the first k columns in the matrices of U and V. This procedure generates the *low-rank matrix* A_k for A as shown below.

It is interesting to remind the following theorem ([3]).

Theorem1 (Eckart-Young)

1) $rank(A_k)=k$

2) For any $m\times n$ matrix B, $(rank(B)\leq k)$,

$$||A-A_k||_F=\min_{B,rank(B)\leq k}||A-B||_F=(\sum_{l=k+1}^{n}\sigma_l^2)^{1/2},$$

where $||\cdot||_F$ means the Frobenius matrix norm, i.e., $||(a_{ij})||_F=(\sum_{i,j}|a_{ij}|^2)^{1/2}$.

The theorem claims that A_k is best approximated to A among all the matrices with rank of less than $k+1$ in the sense of matrix norm.

3.3 Construction of the low-rank item response matrix

When we regard A as the observed item response matrix Δ, and we regard A_k as Δ_k, we can construct the low-rank item response matrix Δ_k from Δ. When it is desired to emphasize that Δ_k is derived from SVD, it is denoted as Δ_k^{SVD} if necessary.

4. Matrix decomposition

4.1 Matrix decomposition procedure

SVD is a promising method to find an approximate matrix that is close to a certain matrix in the sense of the matrix norm. However, it requires that all the elements must be occupied. Occasionally, one encounters cases where not all elements of the observed response matrix are occupied. In such cases, the matrix decomposition (MD) method ([12]) can be used; other methods, such as the imputation method ([22]) or the matrix completion method ([7]) are also used.

MD is similar to SVD. The target matrix $R\in\mathbb{R}^{m\times n}$ can be constructed from two matrices $U\in\mathbb{R}^{m\times k}$ and

$V \in \mathbb{R}^{n\times k}$, but the decomposed form has not the diagonal singular value matrix found in SVD. MD is described as

$$R = UV^{\mathsf{T}}. \tag{9}$$

Using the values of the non-null elements of A, we find U and V so that

$$E = \sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(a_{ij} - r_{ij})^2 \tag{10}$$

becomes small, where $r_{ij} = \sum_{l=1}^{k} u_{il}v_{jl}$, and I_{ij} is the indicator function such that $I_{ij} = 1$ if a_{ij} is non-null and $I_{ij} = 0$ if a_{ij} is null. For stable computation, we use another function with penalty terms such that

$$W = \sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(a_{ij} - r_{ij})^2 + k_u \sum_{i=1}^{m}\sum_{l=1}^{k} u_{il}^2 + k_v \sum_{j=1}^{n}\sum_{l=1}^{k} v_{jl}^2,$$

where, k_u and k_v are regularization factors to prevent overfitting. To find the optimum value, we use the descent method ([4], [14]). From appropriately set initial values of $u_{il}^{(0)}$ and $v_{jl}^{(0)}$, we proceed the following iterations until $|u_{il}^{(t+1)} - u_{il}^{(t)}|$ and $|v_{jl}^{(t+1)} - v_{jl}^{(t)}|$ are sufficiently small.

$$\begin{aligned} u_{il}^{(t+1)} &\leftarrow u_{il}^{(t)} - \lambda \frac{\partial W}{\partial u_{il}}|^{(t)} \\ v_{jl}^{(t+1)} &\leftarrow v_{jl}^{(t)} - \lambda \frac{\partial W}{\partial v_{jl}}|^{(t)}, \end{aligned} \tag{11}$$

where, λ is the learning coefficient. This is a typical MD procedure ([12]).

4.2 Construction of the low-rank item response matrix

When only k vectors are used for the matrices U and V, we denote R in such a case as R_k. As in the SVD case, when we regard R as the observed item response matrix Δ, and we regard R_k as Δ_k, we can construct the *low-rank item response matrix* Δ_k from Δ. When it is desired to emphasize that Δ_k is derived from MD, it is denoted as Δ_k^{MD} if necessary.

5. Examination data analysis (complete matrix treatment)

5.1 A typical example case of the observed item response matrix

As a typical case for complete matrix treatment, we use an observed item response matrix obtained from a mathematics midterm examination given at a certain university. The number of examinees n is 216 and the number of questions m is 31. There are no missing data in this matrix. We name this example case A.

The figure on the left in Figure 1 shows the observed item response matrix. In the figure, only the responses of 28 users are shown for clarity. This matrix is composed of binary elements, with 1 for correct answers and 0 for incorrect answers. The observed item response matrix is denoted as $\Delta = (\delta_{ij})$.

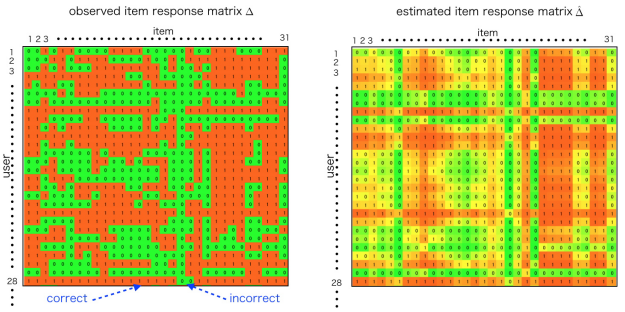

図 1 Observed item response matrix and estimated item response matrix.

5.2 Estimated item response matrix by using IRT

Applying the maximum likelihood estimation method to this observed item response matrix Δ yields the maximum likelihood estimate $\hat{\Theta}$ for the parameter Θ. Using this estimated value $\hat{\Theta}$, the estimated item response matrix $\hat{\Delta}$ can be reconstructed. As explained earlier, this $\hat{\Delta}$ is optimal in the sense of the likelihood principle. The figure on the right in Figure 1 shows this $\hat{\Delta}$. Comparing $\hat{\Delta}$ and Δ in Figure 1, we can roughly imagine the original observed item response matrix Δ from $\hat{\Delta}$. However, this approximation appears to be inaccurate.

To see if this is correct, we will now use the Frobenius matrix norm. Using this matrix norm, the proximity of two equal-sized matrices $A = (a_{ij})$ and $B = (b_{ij})$ can be expressed by the RMSE(A, B) such that

$$\begin{aligned} \mathrm{RMSE}(A,B) &= \sqrt{\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(a_{ij} - b_{ij})^2} \\ &= \sqrt{\frac{1}{nm}(||A - B||_F)^2}. \end{aligned} \tag{12}$$

In this case, the RMSE$(\hat{\Delta}, \Delta)$ of the difference between the observed item response matrix Δ and the estimated item response matrix $\hat{\Delta}$ is computed to be 0.3915. This indicates that the distance between an observed δ_{ij} and its estimated value $\hat{\delta}_{ij}$ lies on average around 0.3915. Intuitively, this value does not seem small.

5.3 Low-rank item response matrix

As described above, the estimated matrix $\hat{\Delta}$ by IRT does not seem to accurately reproduce the observed response matrix, although each estimate $\hat{\delta}_{ij}$ becomes optimal in the likelihood sense under the defined parameter space and likelihood function. Therefore, we investigate how accurate $\hat{\Delta}$ is in terms of MD and SVD.

Applying the methods described in sections 3 and 4, the low-rank item response matrices Δ_k^{SVD} and Δ_k^{MD} can be computed. Table 1 shows the RMSE $(\Delta_k^{\mathrm{SVD}}, \Delta)$ and $\mathrm{RMSE}(\Delta_k^{\mathrm{MD}}, \Delta)$ in the cases of $k = 1, \ldots, 5, 10, 20, 31$. The table also shows the $\mathrm{RMSE}(\hat{\Delta}, \Delta)$ corresponding to IRT estimation.

The table shows, first, that the performance of SVD and MD are almost the same. In other words, MD catches up well with SVD. This means that MD can be applied as an alternative method when SVD cannot be used directly. Such a case may occur in the case of incomplete matrix, especially when the matrices are sparse.

Next, we see that the $\mathrm{RMSE}(\hat{\Delta}, \Delta)$ obtained by IRT lies between the RMSE (Δ_1, Δ) and $\mathrm{RMSE}(\Delta_2, \Delta)$ in both the SVD and MD cases. This is amazing in terms of matrix approximation. The estimated response matrix $\hat{\Delta}$ using IRT would not exceed the accuracy obtained from a $k = 2$ low-rank response matrix generated from the observed item response Δ. In this example, this value $(k = 2)$ would be very small given that the rank of observed matrix Δ is 31. In other words, the reproducibility of the observed item response matrix appears to be low for IRT.

Such properties can also be seen in the pictures of the low-rank response matrices. The figure on the left in Figure 2 shows Δ_2^{SVD} and that on the right Δ_2^{MD}. These are very similar to $\hat{\Delta}$ in Figure 1.

表 1 RMSE of the difference of the two matrices between the low-rank response matrix and the observed response matrix

k	$\mathrm{RMSE}(\Delta_k^{\mathrm{SVD}}, \Delta)$	$\mathrm{RMSE}(\Delta_k^{\mathrm{MD}}, \Delta)$	$\mathrm{RMSE}(\hat{\Delta}, \Delta)$
			0.3915
1	0.4066	0.4067	
2	0.3851	0.3854	
3	0.3652	0.3656	
4	0.3479	0.3485	
5	0.3306	0.3314	
10	0.2562	0.2583	
20	0.1325	0.1400	
31	0	0.0570	

However, this is only the result of one case study. It would be necessary to collect other examination cases to see if these properties hold true in other cases.

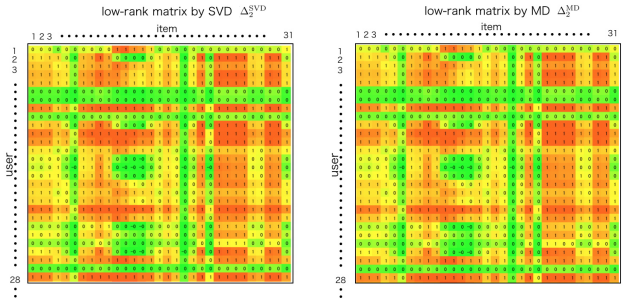

図 2 Low-rank matrices Δ_2^{SVD} and Δ_2^{MD} reproduced from the oserved item response matrix.

5.4 42 examination cases

To make sure that the above mentioned properties hold true for other examination cases, 42 examination cases were collected, including case A. For all examinations, answers were given as discrete values of 1/0 (1 for correct answers and 0 for incorrect answers).

Figure 3 shows the $\mathrm{RMSE}(\hat{\Delta}, \Delta)$ for the 42 examination cases. In the figure, the case *id* shown on the horizontal axis are arranged in ascending order of the magnitude of the $\mathrm{RMSE}(\hat{\Delta}, \Delta)$ shown on the vertical axis for easy understanding. Also shown are the $\mathrm{RMSE}(\Delta_k^{\mathrm{SVD}}, \Delta)$ $(k = 1, 2, 3)$ for each case *id*. Looking at the figure, we see that $\mathrm{RMSE}(\Delta_2^{\mathrm{SVD}}, \Delta) < \mathrm{RMSE}(\hat{\Delta}, \Delta)$ is obtained in all cases. This suggests that the effectiveness of IRT is similar to that of a very low-rank approximation matrix.

図 3 $\mathrm{RMSE}(\hat{\Delta}, \Delta)$ and $\mathrm{RMSE}(\Delta_k^{\mathrm{SVD}}, \Delta)$ $(k = 1, 2, 3)$ for 42 complete matrix using full element data.

However, this is the result when all the element data of Δ are used, i.e., when the estimation is performed with training data only. In this case, the estimated values $\hat{\delta}_{ij}$ may be overfitting values. In order to clarify whether the estimates obtained with IRT are really inaccurate, we

next investigate the prediction accuracy in the case of IRT, SVD and MD using the training data and test data.

6. RMSE for test data using incomplete matrix treatment

We, here, create two matrices S and T for the training and test data from the observed item response matrix Δ, respectively. We denote the corresponding estimated matrices for S and T as $\tilde{S}$ and $\tilde{T}$ in IRT. Similarly, we denote the low-rank matrices with rank k for S and T as $\tilde{S}_k^{\mathrm{SVD}}$ and $\tilde{T}_k^{\mathrm{SVD}}$ in SVD, and as $\tilde{S}_k^{\mathrm{MD}}$ and $\tilde{T}_k^{\mathrm{MD}}$ in MD.

6.1 Case A

Using examination example case A in section 5, we compute the RMSE of the training and test data. To do this, 10% of the elements are randomly selected from the original matrix as test data, and the remaining 90% elements are used as the training data. Then, the RMSE$(\tilde{S}, S)$ for the training data and RMSE$(\tilde{T}, T)$ for the test data are obtained. Since there may be fluctuations of the RMSE due to the selection of test data, this is repeated (bootstrapped) 10 times and the average RMSE is expressed as μ(RMSE).

Figure 4 shows the RMSE for the training test data using MD and SVD. Circles denote the RMSE$(\tilde{S}_k^{\mathrm{SVD}}, S)$ and RMSE$(\tilde{T}_k^{\mathrm{SVD}}, T)$ for each data selection, and similarly triangles denote the RMSE$(\tilde{S}_k^{\mathrm{MD}}, S)$ and RMSE$(\tilde{T}_k^{\mathrm{MD}}, T)$. For reference, the μ(RMSE$(\tilde{S}, S)$) (mean value of the RMSE$(\tilde{S}, S)$) and the μ(RMSE$(\tilde{T}, T)$) (mean value of the RMSE$(\tilde{T}, T)$) using 10 bootstrapped cases reproduced by IRT are shown with horizontal dotted lines. In the figure, the cases of $k = 1, \ldots, 10$ for MD and SVD are presented.

Looking at the figure, as k increases, the RMSE of the training data shows monotonically decreasing characteristics as expected. Moreover, the range of variation in RMSE among the 10 bootstrapped data sets is also small. As explained earlier, the RMSE$(\tilde{S}_k, S)$ with training data is found to lie between RMSE$(\tilde{S}_1^{\mathrm{SVD}}, S)$ and RMSE$(\tilde{S}_2^{\mathrm{SVD}}, S)$. This property is also found in the MD case.

However, for the test data, the RMSE$(\tilde{T}_k^{\mathrm{SVD}}, T)$ curves as a function of k were found to exhibit very different curve shape patterns. The RMSE follows a V-shaped curve (first decreasing and later increasing with increasing k). Such V-shaped curves due to model complexity are well known (e.g., [8]). In addition, the RMSE varies to some extent among 10 bootstrapped data cases. Surprisingly, near the bottom of the V-shaped curve, the values of the RMSE$(\tilde{T}_k^{\mathrm{SVD}}, T)$ and RMSE$(\tilde{T}_k^{\mathrm{MD}}, T)$ for the bootstrapped 10 cases are located near the mean value of the RMSE$(\tilde{T}, T)$.

Table 2 shows the mean values μ(RMSE$(\tilde{T}_k^{\mathrm{SVD}}, T)$) and μ(RMSE$(\tilde{T}_k^{\mathrm{MD}}, T)$) as well as the mean values of the μ(RMSE$(\tilde{T}, T)$ using 10 bootstrapped cases, where $k = 1, \ldots, 10$. In the table, boldface type denotes the smallest value among the various k values. For MD, the smallest RMSE$(\tilde{T}_k^{\mathrm{MD}}, T)$ is obtained when $k = 3$, and for SVD, the smallest RMSE$(\tilde{T}_k^{\mathrm{SVD}}, T)$ is obtained when $k = 5$. It is considered that the RMSE$(\tilde{T}_k^{\mathrm{SVD}}, T)$ or RMSE$(\tilde{T}_k^{\mathrm{MD}}, T)$ would not be smaller than a certain value, unlike the monotonically decreasing behavior of the RMSE$(\tilde{S}_k^{\mathrm{SVD}}, S)$ or RMSE$(\tilde{S}_k^{\mathrm{MD}}, S)$. In this case, this lower limit is located in the neighborhood of the RMSE$(\tilde{S}, S)$. This means that the accuracy (reliability) of mimicking the original data Δ of IRT in terms of prediction can be explained by a very low-rank matrix decomposed by SVD or MD.

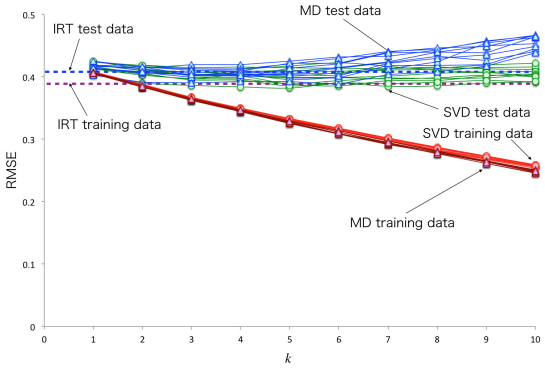

図 4 RMSE for the test data and the training data via MD and SVD (case A).

表 2 Smallest mean values of 10 bootstrapped RMSE for the test data to each Δ_k

k	μ(RMSE$(\tilde{T}_k^{\mathrm{SVD}}, T)$)	μ(RMSE$(\tilde{T}_k^{\mathrm{MD}}, T)$)	RMSE$(\hat{\Delta}, \Delta)$
			0.4056
1	0.4168	0.4160	
2	0.4085	0.4095	
3	0.4026	**0.4053**	
4	0.4025	0.4059	
5	**0.3985**	0.4096	
6	0.4009	0.4126	
7	0.4019	0.4256	
8	0.4025	0.4326	
9	0.4048	0.4403	
10	0.4052	0.4556	

Recall that the RMSE for IRT lies between the $k = 1$ RMSE and the $k = 2$ RMSE for MD or SVD with full

matrix data, as indicated in Table 1. This superficial result seems to indicate that IRT's ability to represent the observed item response matrix is weak. This is because the prediction ability using $\hat{\Delta}$ is similar to that using a very low-rank approximation matrix, although the rank of matrix Δ is the same as the minimum number of users and items, which in this case is 31.

However, as seen above, the RMSE in the test case is also close to the RMSE of the best reconstructed low-rank matrix (rank lower than 5) using the matrix decomposition. This tells us that the mathematical model of IRT (two-parameter logistic model) is well defined to represent the actual examination case (case A). However, this characteristic could be found only in certain cases. Therefore, we next investigate whether such a characteristic applies to other examination cases, using various data cases.

6.2 8 examination cases among 42 cases

From 42 data cases, we picked 8 cases, including case A, to verify whether the RMSE of the test cases in IRT is close to the RMSE of the low-rank matrix in SVD. Table 3 shows $\mu(\mathrm{RMSE}(\tilde{T}^{\mathrm{SVD}}_{k\_opt}, T))$ and $\mu(\mathrm{RMSE}(\tilde{T}, T))$, where k_opt means k for which $\mu(\mathrm{RMSE}(\tilde{T}^{\mathrm{SVD}}_{k\_opt}, T))$ is minimum when $k = 1, \ldots, k_{max}$.

Looking at the table, in almost all cases, we see that $\mu(\mathrm{RMSE}(\tilde{T}^{\mathrm{SVD}}_{k\_opt}, T))$ is close to $\mu(\mathrm{RMSE}(\tilde{T}, T))$ and k_opt is very small. The exception is case B (case id is 30), where the matrix size is $n = 1101, m = 84$, and $k\_opt = 16$. In this case, the number of examinee is large to some extent and $\mu(\mathrm{RMSE}(\tilde{T}^{\mathrm{SVD}}_{k\_opt}, T))$ is clearly smaller than $\mu(\mathrm{RMSE}(\tilde{T}, T))$. Thus, we next consider the relation between the $\mathrm{RMSE}(\tilde{T}^{\mathrm{SVD}}_{k\_opt}, T)$ and $\mathrm{RMSE}(\tilde{T}, T)$ in case B.

表 3 Mean values of 10 bootstrapped RMSE for the test data to 8 data cases

case id	$\mu(\mathrm{RMSE}(\tilde{T}^{\mathrm{SVD}}_{k\_opt}, T))$	k_opt	$\mu(\mathrm{RMSE}(\tilde{T}, T))$
5	0.2935	1	0.2908
10	0.3547	1	0.3591
15	0.3457	1	0.3344
20	0.3801	1	0.3726
25	0.3701	5	0.3789
30	0.3442	16	0.3771
35	0.3982	2	0.3964
40	0.4053	3	0.4068

Figure 5 shows the RMSE for the training data and the test data using SVD. Circles indicate the $\mathrm{RMSE}(\tilde{S}^{\mathrm{SVD}}_k, S)$ and $\mathrm{RMSE}(\tilde{T}^{\mathrm{SVD}}_k, T)$ for each data selection. For reference, the $\mu(\mathrm{RMSE}(\tilde{S}, S))$ (mean value of the $\mathrm{RMSE}(\tilde{S}, S)$) and the $\mu(\mathrm{RMSE}(\tilde{T}, T))$ (mean value of the $\mathrm{RMSE}(\tilde{T}, T)$) using 10 cases reproduced by IRT are shown using horizontal dotted lines. In the figure, cases of $k = 1, \ldots, 30$ for SVD are illustrated.

As shown in the figure, the mean value $\mu(\mathrm{RMSE}(\tilde{T}, T))$ using 10 bootstrapped cases by IRT lies between the $\mathrm{RMSE}(\tilde{T}^{\mathrm{SVD}}_1, T)$ and $\mathrm{RMSE}(\tilde{T}^{\mathrm{SVD}}_2, T)$. In addition, $\mu(\mathrm{RMSE}(\tilde{S}, S))$ lies between the $\mathrm{RMSE}(\tilde{S}^{\mathrm{SVD}}_1, S)$ and RMSE $(\tilde{S}^{\mathrm{SVD}}_2, S)$. This property is the same as the fact that the $\mathrm{RMSE}(\tilde{S}, S)$ using the training data lies between the values of $\mathrm{RMSE}(\tilde{S}^{\mathrm{SVD}}_1, S)$ and $\mathrm{RMSE}(\tilde{S}^{\mathrm{SVD}}_2, S)$ for one case. Such a result may be realized by chance, but the property that the $\mu(\mathrm{RMSE}(\tilde{T}^{\mathrm{SVD}}_k, T))$ is not much different from the $\mu(\mathrm{RMSE}(\tilde{T}, T))$ remains the same as in other 7 cases.

In other words, the potential ability of IRT to mimic the observed response matrix Δ is found to be equivalent to the ability of the low-rank approximation matrix generated by matrix decomposition to mimic the observed response matrix, not only in the full training data use study but also in the training and test data use study. Although the rank of the low-rank approximation response matrix by matrix decomposition corresponding to $\hat{\Delta}$ is extremely smaller than the rank of the observed response matrix Δ, the predictive ability of IRT seems to be high enough since the $\mathrm{RMSE}(\tilde{T}^{\mathrm{SVD}}_{k\_opt}, T)$ is almost equal to $\mu(\mathrm{RMSE}(\tilde{T}, T))$. In other words, if the size of the item response matrix is moderate, i.e., less than 1000 users and less than 100 items, it would be difficult to obtain more information than IRT produces from the observed item response matrix alone using the matrix decomposition method.

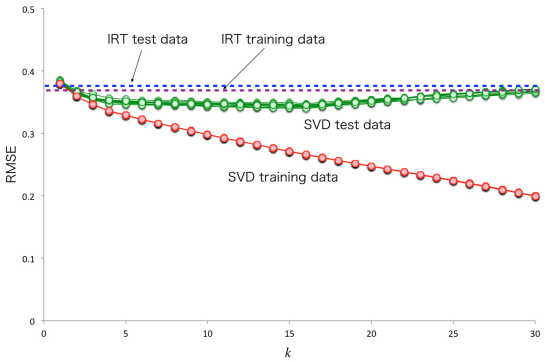

図 5 RMSE for the test data and the training data via MD and SVD (case30).

7. Concluding remarks

This paper has investigated the predictive effectiveness of item response theory itself from matrix decomposition perspective.

If the observed item response matrix is given, the maximum likelihood estimates for parameters in item response theory can be estimated, and the estimated item response matrix is also reconstructed using the estimates. Then, the difference between the observed and estimated item response matrices can be determined in the sense of matrix norm. Matrix decomposition and singular value decomposition methods can generate a low-rank approximation matrix from the observed item response matrix, and the difference between the observed and the generated low-rank matrix can be measured in the sense matrix norm.

The effectiveness of item response theory may be evaluated by comparing the two matrices between the estimated item response matrix obtained from the item response theory and the low-rank approximation matrix obtained from matrix decomposition or singular value decomposition.

Applying item response theory, matrix decomposition, and singular value decomposition to many actual examination data, it is found that the rank of the generated low-rank approximation matrix that is equivalent to the estimated item response matrix is very low. However, the predictive ability of IRT seems to be high enough since the minimum root mean squared errors for test data using matrix decomposition and singular value decomposition methods are almost equal to the root mean squared error from item response theory.

参考文献

[1] Baker, F. B. and Kim, S.-H.: *Item Response Theory: Parameter Estimation Technique, 2nd edn.*, Marcel Dekker (2004).
[2] de Ayala, R.: *The Theory and Practice of Item Response Theory*, Guilford Press (2009).
[3] Eckart, C. and Young, G.: The Approximation of One Matrix by Another of Lower Rank, *Psychometrika*, Vol. 1, pp. 211–218 (1936).
[4] Fletcher, R.: *Practical Methods of Optimization*, Wiley (2000).
[5] Golub, G. H. and Van Loan, C. F.: *Matrix Computations*, Johns Hopkins Univ. Press (2012).
[6] Hambleton, R., Swaminathan, H. and Rogers, H. J.: *Fundamentals of Item Response Theory*, Sage Publications (1991).
[7] Hastie, T., Mazumder, R., Lee, J. D. and Zadeh, R.: Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares, *Journal of Machine Learning Research*, Vol. 16, pp. 3367–3402 (2015).
[8] Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning*, Springer (2009).
[9] Hirose, H. and Sakumura, T.: Item Response Prediction for Incomplete Response Matrix Using the EM-type Item Response Theory with Application to Adaptive Online Ability Evaluation System, *IEEE International Conference on Teaching, Assessment, and Learning for Engineering*, pp. 8–12 (2012).
[10] Kendall, M. G. and Stuart, A.: *Advanced Theory of Statistics*, Macmillan Pub Co (1983).
[11] Khajah, M., Huang, Y., Gonzlez-Brenes, J. P., Mozer, M. C. and Brusilovsk, P.: Integrating knowledge tracing and item response theory: A tale of two frameworks, *4th International Workshop on Personalization Approaches in Learning Environments*, pp. 7–15 (2014).
[12] Koren, Y., Bell, R. M. and Volinsky, C.: Matrix Faxctorization Techniques for Recommender Systems, *Computer*, Vol. 42, pp. 30–37 (2009).
[13] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J. and Sohl-Dickstein, J.: Deep knowledge tracing, *Advances in Neural Information Processing Systems*, pp. 505–513 (2015).
[14] Polak, E.: *Optimization : Algorithms and Consistent Approximations*, Springer (1997).
[15] Reckase, D.: *Multidimensional Item Response Theory*, Springer (2011).
[16] Sakumura, T. and Hirose, H.: Making up the Complete Matrix from the Incomplete Matrix Using the EM-type IRT and Its Application, *Transactions on Information Processing Society of Japan (TOM)*, Vol. 72, pp. 17–26 (2014).
[17] Sakumura, T., Kuwahata, T. and Hirose, H.: An Adaptive Online Ability Evaluation System Using the Item Response Theory, *Education & e-Learning*, pp. 51–54 (2011).
[18] Strang, G.: Multiplying and Factoring Matrices, *The American Mathematical Monthly*, Vol. 125, pp. 223–230 (2018).
[19] Strang, G.: *Introduction to Linear Algebra*, Wellesley-Cambridge Press (2021).
[20] Sweeney, M., Lester, J., Rangwala, H. and Johri, A.: Next-term student performance prediction: A recommender systems approach, *Journal of Educational Data Mining*, Vol. 8, pp. 22–51 (2016).
[21] Thai-Nghe, N., Drumond, L., Horva, T., Krohn-Grimberghe, A., Nanopoulos, A. and Schmidt-Thieme, L.: Factorization techniques for predicting student performance, *Educational recommender systems and technologies: Practices and challenges*, pp. 129–153 (2011).
[22] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B.: Missing Value Estimation Methods for DNA Microarrays, *Bioinformatics*, pp. 520–525 (2001).
[23] van der Linden, W. J.: *Handbook of Item Response Theory*, Chapman and Hall/CRC (2016).
[24] Vie, J.-J.: Deep factorization machines for knowledge tracing, *Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 370–373 (2018).
[25] Vie, J.-J. and Kashima, H.: Knowledge Tracing Machines: Factorization Machines for Knowledge Tracing, *The Thirty-Third AAAI Conference on Artificial Intelligence*, pp. 750–757 (2019).
[26] Wilson, K. H., Xiong, X., Khajah, M., Lindsey, R. V., Zhao, S., Karklin, Y., Van Inwegen, E. G., Han, B., Ekanadham, C., Beck, J. E., Heffernan, N. and Mozer, M. C.: Estimating student proficiency: Deep learning is not the panacea, *Workshop on Machine Learning for Education, Neural Information Processing Systems*, pp. 1–8 (2016).