An extension of the irt to a network model

Article *in* Behaviormetrika · January 2002
DOI: 10.2333/bhmk.29.59

1 author:

Maomi Ueno
The University of Electro-Communications

Behaviormetrika
Vol.29, No.1, 2002, 59-79

AN EXTENSION OF THE *IRT* TO A NETWORK MODEL

Maomi Ueno*

The traditional Item Response Theory (IRT) models assume local independence, which is equivalent to the assumption of unidimensionality. This assumption states that a subject's responses to different items in a test are statistically independent. For the assumption to be true, a subject's performance on one item must not affect, either for better or for worse, his or her responses to any other items in the test. The main purpose of this paper is to relax the local independence assumption in the traditional IRT models by extending to a network model. A new IRT model is defined which assumes probabilistic network structures for the assumption of local independence.

Another unique feature of the model proposed is that it is a new probabilistic network model with the conditional probability parameters depending on a latent trait variable.

Information criteria AIC and BIC are used to evaluate the performance of the model proposed, using actual test data. It shows that the proposed model provides better results than the traditional model.

In addition, this paper proposes an item selection criterion from the decision theoretic approach. The amount of test information is defined as the amount of mutual information between a variable for the item and all variables over the test, to maximize the prediction efficiency of the subject's responses. The new item selection method is used to compare the prediction efficiency between the proposed model and the traditional IRT model. The proposed model is shown to be more efficient.

1. Introduction

Since Lord and Novick (1968) took a modern mathematical statistical approach to formulate the basic constructs of the Item Response Theory (IRT), a great deal of research effort has been spent in developing their idea from different perspectives (e.g., statistical theory, parameter estimation algorithms). Many possible IRT models exist, differing in the mathematical form of the item characteristic function and/or the number of parameters specified in the model, for example, the Rasch model (Rasch, 1960; Rasch, 1961; Rasch, 1966a; Rasch, 1966b), the normal ogive model (Lord & Novick, 1968), the two parameters logistic model (Birnbaum, 1957), and the three parameters logistic model (Birnbaum, 1968). There are more general well known IRT models such as the graded response model (Samejima, 1969), the free response model (Samejima, 1972), the partial credit model (Masters, 1982), and the multi-nominal model (Bock, 1972). All IRT models contain one or more parameters describing the subject.

Key Words and Phrases: Item Response Model, probabilistic network, Bayesian networks, graphical model, multinomial model, test information.

* Nagaoka University of Technology, Department of Management & Information Systems Science 1603-1, Kamitomioka, Nagaoka, Niigata, 940-2188, Japan. E-mail: ueno@kjs.nagaokaut.ac.jp

The IRT rests on the following three basic postulates: (a) The performance of a subject on a test item can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; (b) the relationship between the subject's item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an *item characteristic function* or *item characteristic curve* (ICC); and (c) when the abilities influencing test performance are held constant, the subject's responses to any pair of items are statistically independent, which is called *local independence*. It should be particularly noted that the assumption (c) states that a subject's responses to different items in a test are statistically independent. For this assumption to be true, a subject's performance on one item must not affect, either for better or for worse, his or her responses to any other items in the test. Concerning this local independence assumption, Junker (1991) proposed an index to represent deviation from the local independence assumption, and showed that the parameter estimation often fails when the local independence assumption is violated. It could be deduced from his result that the true probability structures of actual data are more complex than the local independence structure, and that if a more flexible structure in the IRT model is developed, the prediction efficiency of the model is improved.

In this paper we shall introduce a network structure to relax the local independence assumption in the traditional IRT model. That is, we shall propose a new IRT model with a probabilistic network structure instead of assumption (c) of the local independence structure. More concretely, the features of the model to be proposed are as follows:

1. The proposed IRT model is regarded as one of the probabilistic network models (Pearl, 1988);

2. The proposed model is a new probabilistic network model with conditional probability parameters depending on a latent trait variable.

In Section 2 we review probabilistic networks and develop a new network IRT model in Section 3. Section 4 used the probability propagation to discuss the prediction efficiency of IRT models. Section 5 defines the EVSI (expected sample information) for Bayesian networks. In Section 5, we discuss a method of the marginal maximum likelihood estimation to estimate the item parameter vector. Numerical examples are provided in Section 7. We end with conclusion and discussion.

2. Probabilistic Network

This paper proposes a new Item Response Theory (IRT model) from the probabilistic network approach. First, this section introduces the probabilistic network(see, e.g., Pearl, 1988; Lauritzen, 1996) as follows:

Let $X = \{X_1, X_2, \ldots, X_N\}$ be a set of N discrete variables; each can take values in the set $\{0, \ldots, r_i - 1\}$. We write $x_i = K$ when we observe that variable

x_i is state K. We use $p(x_i = K \mid y = K', \xi)$ to denote the probability of $x_i = K$ given the observation $y = K'$ with background knowledge ξ. When we observe the state for every variable in set X, we call this set of observations an instance of X. We use $p(X \mid Y, \xi)$ to denote the set of probabilities for all possible observations of X, given all possible observations of Y.

A probabilistic network represents a joint probability distribution over domain U by encoding assertions of conditional independence as well as a collection of probability distributions. From the chain rule of probability, we know

$$p(x_1, x_2, \ldots, x_N \mid \xi) = \prod_{i=1}^{N} p(x_i \mid x_1, x_2, \ldots, x_{i-1}, \xi). \tag{1}$$

For each variable x_i, let $\Pi_i \subseteq \{x_1, \ldots, x_{i-1}\}$ be a set of variables that renders x_i and $\{x_1, \ldots, x_{i-1}\}$ conditionally independent. Here, $x_1, \ldots, x_{i-1}$ are called parent nodes. That is,

$$p(x_i \mid x_1, x_2, \ldots, x_{i-1}, \xi) = p(x_i \mid \Pi_i, \xi). \tag{2}$$

The probabilistic network model is represented as a pair (B_S, B_P) of network structures where B_S encodes the assertions of conditional independence in the above equation and B_P is a set of conditional probability parameters. In particular B_S is a directed acyclic graph such that (1) each variable in U corresponds to a node in B_S, and (2) the parent nodes of the node corresponding to x_i are the nodes corresponding to the variables in Π_i. (In the remainder of this paper, we use x_i to refer to both the variables and their corresponding nodes in the graph.) Associated with nodes x_i in B_S are the probability distributions $p(x_i \mid \Pi_i, \xi)$. Combining (1) and (2), we see that any network for U uniquely determines a joint probability distribution for U. That is,

$$p(x_1, x_2, \ldots, x_n \mid B_S) = \prod_{i=1}^{N} p(x_i \mid \Pi_i, B_S). \tag{3}$$

Figure 1 shows an example of the probabilistic network model with variables $\{x_1, \ldots, x_5\}$. In this case, the joint probability distribution of the probabilistic network model is given by

$$\begin{aligned} &p(x_1, x_2, \ldots, x_5 \mid B_S) \\ &\quad = p(x_1 \mid x_2, x_5) p(x_2 \mid x_5) p(x_3 \mid x_2, x_4, x_5) \\ &\quad\quad \times p(x_4 \mid x_5) p(x_5). \end{aligned}$$

The probabilistic network model is usually parameterized as the multinomial model(see, e.g., Suzuki, 1993). This approach assumes that the conditional probabilities $\{p(x_i \mid \Pi_i, B_S)\}$ are constant over the subject, and represent the conditional probabilities as a set of parameters. This paper will propose a probabilistic network model where the conditional parameters depend on latent variables which reflect individual characteristics.

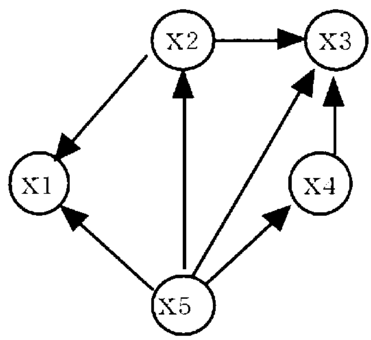

Figure 1: An example of the probabilistic networks

3. Network IRT Model

In this section, the IRT model is extended to the network model.

Let the subject j's response pattern to N items be

$$\mathbf{u}_j = (u_{j1}, u_{j2}, \cdots, u_{ji}, \cdots, u_{jN})^t, \tag{4}$$

where

$$u_{ij} = \begin{cases} 1 & \text{for a correct response} \\ 0 & \text{for an incorrect response} \end{cases}$$

Let the subject j 's latent ability parameter be θ_j, then the joint probability of a score matrix $U = (u_{ij})$ for n subjects is given by

$$p(U|\theta, \boldsymbol{\xi}) = \prod_{j=1}^{n} p(\mathbf{u}_j|\theta_j, \boldsymbol{\xi}), \tag{5}$$

where

$\boldsymbol{\xi}_i$: a parameter vector for an item i,

$\boldsymbol{\xi} = (\boldsymbol{\xi}_1^t, \boldsymbol{\xi}_2^t, \ldots, \boldsymbol{\xi}_i^t, \cdots \boldsymbol{\xi}_N^t)^t$,

and

$$\theta = (\theta_1, \theta_2, \ldots, \theta_j, \ldots, \theta_n).$$

For the traditional IRT model, from the assumption of local independence, the joint probability distribution $p(\mathbf{u}_j|\theta_j, \boldsymbol{\xi})$ is given by

$$p(\mathbf{u}_j|\theta_j, \boldsymbol{\xi}) = \prod_{i=1}^{N} p(u_{ij} = 1|\theta_j, \boldsymbol{\xi}_i)^{u_{ij}} [1 - p(u_{ij} = 1|\theta_j, \boldsymbol{\xi}_i)]^{1-u_{ij}}. \tag{6}$$

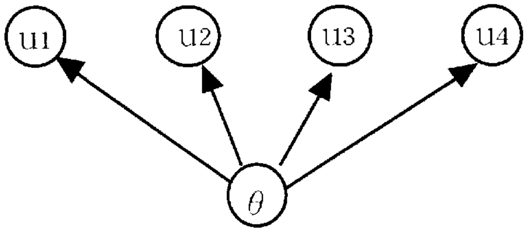

Figure 2: The structure of the local independence model

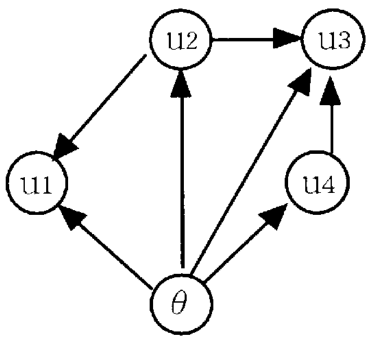

Figure 3: The structure of the network model

Although there are some IRT models, we will consider the following two-parameter logistic model here.

That is, $p(u_{ij} = 1|\theta_j, \boldsymbol{\xi}_i)$ is given by

$$p(u_{ij} = 1|\theta_j, \boldsymbol{\xi}_i) = \frac{1}{1 + \exp(-1.7a_i(\theta_j - b_i))} \tag{7}$$

with $\boldsymbol{\xi}_i = (a_i, b_i)^t$.

Here, the probabilistic structure of the traditional IRT model is shown in Figure 2 from the probabilistic network approach.

From the model structure B_S shown in Figure 3, we obtain

$$\begin{aligned} p(\mathbf{u}_j|\theta_j, \boldsymbol{\xi}, B_S) &= \prod_{i=1}^{N} \prod_{k=1}^{2^{m_i}} p(u_{ij}|\widetilde{u_{ijk}}, \theta_j, \boldsymbol{\xi}_i)^{u_{ijk}} \\ &= \prod_{i=1}^{N} \prod_{k=1}^{2^{m_i}} p(u_{ij} = 1|\widetilde{u_{ijk}}, \theta_j, \boldsymbol{\xi}_i)^{u_{ij} \cdot u_{ijk}} \\ &\quad \times [1 - p(u_{ij} = 1|\widetilde{u_{ijk}}, \theta_j, \boldsymbol{\xi}_i)]^{u_{ij} \cdot (1 - u_{ijk})}, \end{aligned} \tag{8}$$

where

$$u_{ijk} = \begin{cases} 1 & \text{for the subject } j\text{'s } k\text{-th response pattern} \\ & \text{to parent node items of item } i \\ 0 & \text{for the other patterns} \end{cases}$$

m_i : the number of parent node items of item i.
$\widetilde{u_{ijk}}$: a subject j's k-th response pattern to parent node items of item i,
$\boldsymbol{\xi}_i$: a parameter vector for $\widetilde{u_{ijk}}$,
$\boldsymbol{\xi}$ $=$ $(\boldsymbol{\xi}_1^t, \boldsymbol{\xi}_2^t, \ldots, \boldsymbol{\xi}_i^t, \ldots, \boldsymbol{\xi}_{2^{m_i}}^t)^t$.

For the probabilistic structure shown in Figure 3, the joint probability distribution is given by

$$p(u_1, u_2, \ldots, u_4, \theta_j \mid \boldsymbol{\xi}, B_S)$$
$$= p(u_1 \mid u_2, \boldsymbol{\xi}_1, \theta_j)p(u_2 \mid \boldsymbol{\xi}_2, \theta_j)p(u_3 \mid u_2, u_4, \boldsymbol{\xi}_3, \theta_j)p(u_4 \mid \boldsymbol{\xi}_4, \theta_j)p(\theta_j),$$

where

$$\theta_j \sim N(0, 1^2).$$

The proposed model is a new IRT model that assumes probabilistic network structures. This model integrates the probabilistic structure into the IRT model. Moreover, it should be noted that this model is also a new probabilistic network model with a latent trait. Whereas the traditional probabilistic network model implicitly assumes that the probabilistic structure is invariable for the subjects, the model proposed assumes that the conditional probabilities of the probabilistic network model can change corresponding to the subject's trait.

4. Probability Propagation

When the true probability structure of the IRT model is a network structure, the traditional IRT model loses much information, and therefore, can not provide a reliable prediction for a subject's unknown responses. This paper evaluates the proposed model from the aspect of prediction efficiency. The remaining problem is how to evaluate the prediction efficiency of the IRT models. This section provides a method which predicts a subject's unknown responses from observed data. For this purpose, it is well known that the probability propagation technique (Pearl, 1988) is efficient. Probability propagation is a technique to predict unknown events from observed data by using probabilistic networks. That is, when the observation $u_{i'}$, which is a child node of node i, is given, the probability $p(u_i \mid B_S)$ is propagated to $p(u_i \mid u_{i'}, B_S)$ using Bayes' theorem.

$$p(u_i \mid u_{i'}, B_S) = \frac{p(u_{i'} \mid u_i, B_S)p(u_i \mid B_S)}{\sum_{u_i} p(u_{i'} \mid u_i, B_S)p(u_i \mid B_S)}. \quad (9)$$

When the observation u_i, which is a parent node of node i', is given, the prior probability $p(u_{i'} \mid B_S)$ is replaced with the conditional probability $p(u_{i'} \mid u_i, B_S)$.

These procedures are propagated over the network. However, this algorithm can not be applied to the proposed method because the proposed IRT model has a latent variable. Therefore, we propose a probability propagation method for the IRT model in the following algorithm that combines the Monte Carlo integration with the Newton-Raphson method.

1. Generate $\theta^{(q)} \sim g(\theta \mid \boldsymbol{\tau}^{(q-1)})$, where $g(\theta \mid \boldsymbol{\tau}^{(q-1)})$ is the normal density function with a parameter vector $\boldsymbol{\tau}^{(q-1)} = (\mu^{(q-1)}, \sigma^{2(q-1)})^t$. The initial value of $\boldsymbol{\tau}^{(q-1)}$ is $\boldsymbol{\tau}^{(0)} = (0, 1^2)^t$.

2. Generate $u_i^{(q)}(=0,1) \sim P(u_j \mid u_{ip}^{(q-1)}, u_{ic}^{(q-1)}, \theta^{(q-1)})$,

   where
   $u_i^{(q)}$: the q-th binary data generated from $P(u_j \mid u_{ip}^{(q-1)}, u_{ic}^{(q-1)}, \theta^{(q-1)})$
   i_p : a set of the parent nodes of node i,
   i_c : a set of the child nodes of node i,

   and $P(u_j \mid u_{ip}^{(q-1)}, u_{ic}^{(q-1)}, \theta^{(q-1)})$ is calculated by

$$\begin{aligned} P(u_i \mid u_{ip}^{(q-1)}, u_{ic}^{(q-1)}, \theta^{(q)}) &= P(u_i = 1 \mid u_{ip}^{(q-1)}, \theta^{(q)}, \boldsymbol{\xi}_i, B_S) \\ &\quad \times \prod_{i' \in i_c} P(u_i = 1 \mid u_{ip}^{(q-1)}, \theta^{(q)}, \boldsymbol{\xi}_i, B_S) \\ &= P(u_i = 1 \mid u_{ip}^{(q-1)}, \theta^{(q)}, \boldsymbol{\xi}_i, B_S) \\ &\quad \times \prod_{i' \in i_c} \frac{P(u_{i'}^{(q-1)} \mid u_i = 1, \theta^{(q)}, \boldsymbol{\xi}_i, B_S) P(u_i = 1)}{\sum_{u_i} P(u_{i'}^{(q-1)} \mid u_i, \theta^{(q)}, \boldsymbol{\xi}_i, B_S) P(u_i)} \end{aligned} \quad (10)$$

   We can obtain the q-th Monte Calro random pattern $\boldsymbol{u}^{(q)} = (u_1^{(q)}, u_2^{(q)}, \ldots, u_i^{(q)}, \ldots, u_N^{(q)})$ by calculating the above procedure for i.

3. Estimation of $\boldsymbol{\tau}^{(q)}$
   We can obtain the MAP estimates $\mu^{(q)}$ which maximize the following posterior $g(\theta^{(q)} \mid \boldsymbol{u}^{(q)}, \boldsymbol{\xi})$:

$$g(\theta^{(q)} \mid \boldsymbol{u}^{(q)}, \boldsymbol{\xi}) \propto p(\boldsymbol{u}^{(q)} \mid \theta^{(q)}, \boldsymbol{\xi}) g(\theta \mid \boldsymbol{\tau}). \quad (11)$$

   Finally, the estimation procedure of the $\hat{\theta}$ is shown as follows:

$$\hat{\theta}_{T+1} = \hat{\theta}_T - \left[\frac{\partial^2 \{\log g(\hat{\theta}_T \mid \boldsymbol{u}^{(q)}, \boldsymbol{\xi})\}}{\partial^2 \hat{\theta}_T} \right]^{-1} \frac{\partial \{\log g(\hat{\theta}_T \mid \boldsymbol{u}^{(q)}, \boldsymbol{\xi})\}}{\partial \hat{\theta}_T}.$$

   By iterating these procedures T times until the gradient is smaller than 10^{-5}, the obtained final value of $\hat{\theta}_{T+1}$ is the estimated value of $\mu^{(q)}$.

Although $\sigma^{2(q)}$ should be estimated as usual, the asymptotic variance of the mode of θ is used here to avoid the convergence problem. That is,

$$\sigma^{2(q)} = \left[\frac{\partial \{\log g(\hat{\theta}_T \mid \boldsymbol{u}^{(q)}, \boldsymbol{\xi})\}^2}{\partial^2 \hat{\theta}_T} \right]^{-1}. \tag{12}$$

The procedure from 1 to 3 is iterated t times. In addition, the marginal posterior probabilities are obtained as follows by considering only $s(< t)$ times results:

$$P(u_i = 1 \mid u_{i'}) = \frac{1}{s} \sum_{q=t-s+1} u_i^{(q)}, \tag{13}$$

$$\mu = \frac{1}{s} \sum_{q=t-s+1} \mu^{(q)}, \tag{14}$$

$$\sigma^2 = \frac{1}{s} \sum_{q=t-s+1} \sigma^{2(q)}. \tag{15}$$

Consequently, we can obtain the propagated posterior $P(u_i = 1 \mid u_{i'}, B_S)$. It should be noted that the proposed method can provide the propagation in the traditional probabilistic network structure and the traditional IRT procedure.

5. Amount of Test Information from Information Theory

The amount of test information with Bayesian networks has been already proposed in the Test Theory with probabilistic networks (Ueno, 1994). In Bayesian decision theory, the value of the data is evaluated by the expected value of sample information (*EVSI*). *EVSI* is the difference between the expected value given data x and the expected value given no data (Shigemasu, 1985). It is defined as follows:

$$\begin{aligned} EVSI = & \int_x \{\max_a \int_\Theta \boldsymbol{utility}(a, \theta) p(\theta|x) d\Theta\} dx \\ & - \max_a \int_\Theta \boldsymbol{utility}(a, \theta) p(\theta) d\theta. \end{aligned} \tag{16}$$

where a indicates the action, $\boldsymbol{utility}(a, \theta)$ is the utility function and Θ indicates the parameter space. The integral is the Stielties integral. Shigemasu (1988) has already proposed some educational evaluation by using $EVSI$. However, now we will consider the $EVSI$ for Bayesian networks, $EVTIN$(Expected Value of Test Information with Networks); the problem is how to define the utility $\boldsymbol{utility}(a, \theta)$. A joint probability over the network is shown in (8).

When we obtain observations $u_1, u_2, \ldots, u_{N'}, N'$ $(< N)$, the predictive distribution for $u_{N'+1}, u_{N'+2}, \ldots, u_{N-N'}$ is given by

$$\begin{aligned} & p(u_{N'+1}, u_{N'+2}, \ldots, u_{N-N'} | u_1, u_2, \ldots, u_{N'}, B_S) \\ = & \int p(u_{N'+1}, u_{N'+2}, \ldots, u_{N-N'} | u_1, u_2, \ldots, u_{N'}, \boldsymbol{\tau}, B_S) p(\boldsymbol{\tau} | \mathbf{U}, B_S) d\boldsymbol{\tau}. \end{aligned} \tag{17}$$

Now, we consider the log-predictive joint distribution

$$\log p(u_{N'+1}, u_{N'+2}, \ldots, u_{N-N'} | u_1, u_2, \ldots, u_{N'}, B_S)$$

as the utility function. We then have

$$\begin{aligned} & EVTIN(u_{N'N}, u_{1N'}, B_S) \\ & = \int_u \left\{ \max_a \int_\Theta p(u_{N'N} | u_{1N'}, B_S) \right. \\ & \qquad \left. \times \log p(u_{N'N}0 | u_{1N'}, B_S) d\, u_{N'N} \right\} p(u_{1N'} | B_S) du_{1N'} \\ & \qquad - \max_a \int_\Theta p(u_{N'N} | B_S) \log p(u_{N'N} | B_S) du_{N'N}, \end{aligned} \tag{18}$$

where $u_{1N'} = \{u_1, u_2, \ldots, u_{N'}\}$ and $u_{N'N} = \{u_{N'+1}, u_{N'+2}, \ldots, u_{N-N'}\}'$.

Thus, we see that $EVTIN$ is a measure that selects the unknown node set, i.e., the test items that are the best predictor for future observations.

We should note that the expression of $EVTIN$ is similar to the amount of mutual information. However, the $EVTIN$ has some unique features, among them, non-symmetricity and consistency (Ueno, 1996a,b).

6. Parameter Estimation

The parameter vector $\boldsymbol{\xi}_i$ can be estimated by Marginal Maximum Likelihood Estimation(MMLE) as follows: The following marginal log likelihood can be maximized by the method proposed by Bock (1981):

$$\phi = \log \int_{-\infty}^{+\infty} P(\boldsymbol{u}_i \mid \theta_j, \boldsymbol{\xi}_i) g(\theta_j \mid \boldsymbol{\tau}) d\theta, \tag{19}$$

where $g(\theta_j \mid \boldsymbol{\tau})$ is the normal distribution with the hyper parameters $\boldsymbol{\tau} = (\mu, \sigma^2)$ $(\mu = 0, \sigma^2 = 1)$.

The marginal maximum likelihood equation

$$\partial\phi / \partial\boldsymbol{\xi}_i = 0$$

can be solved as follows. Here, we use the following simplified notation to avoid complicated formulae:

$$P_{ijkl} = P(u_i = 1 \mid \tilde{u}_{ijk}, X_l, \boldsymbol{\xi}_i, B_S),$$

$$Q_{ijkl} = 1 - P(u_i = 1 \mid \tilde{u}_{ijk}, X_l, \boldsymbol{\xi}_i, B_S).$$

E-Step

$$\overline{f_{jkl}} = \sum_{i}^{N} \left[\frac{\prod_{j=1}^{n} \prod_{k=1}^{2^{m_j}} (P_{ijkl})^{u_{ij} \cdot u_{ijk}}}{\sum_{l=1}^{q} \prod_{j=1}^{n} \prod_{k=1}^{2^{m_j}} (Q_{ijkl})^{(1-u_{ij}) \cdot u_{ijk}}} \times \frac{(Q_{ijkl})^{(1-u_{ij}) \cdot u_{ijk}} A(X_l)}{(P_{ijkl})^{u_{ij} \cdot u_{ijk}} \cdot u_{ijk} A(X_l)} \right] \tag{20}$$

and

$$\overline{r_{jkl}} = \sum_{i}^{N} \left[\frac{\prod_{j=1}^{n} \prod_{k=1}^{2^{m_j}} u_{ij} \cdot u_{ijk} (P_{ijkl})^{u_{ij} \cdot u_{ijk}}}{\sum_{l=1}^{q} \prod_{j=1}^{n} \prod_{k=1}^{2^{m_j}} (Q_{ijkl})^{(1-u_{ij}) \cdot u_{ijk}}} \times \frac{(Q_{ijkl})^{(1-u_{ij}) \cdot u_{ijk}} A(X_l)}{(P_{ijkl})^{u_{ij} \cdot u_{ijk}} \cdot u_{ijk} A(X_l)} \right], \tag{21}$$

where $A(X_l)$, $(l = 1, \ldots, q)$ indicates the weights of the Hermite-Gauss quadrature.

M-Step Choose $\boldsymbol{\xi}_i$ such that the following posterior expectation is maximized:

$$\log(L) = \sum_{l=1}^{q} \sum_{j=1}^{n} \sum_{k=1}^{m_j} [\overline{r_{jkl}} \log P_{ijkl} + \overline{f_{jkl}} \log Q_{ijkl}] \tag{22}$$

Consequently, we can obtain the conditional probability parameters.

7. Numerical Examples

7.1 Parameter estimation

In this study, we conducted a mathematical test of the structure described in Figure 4 to 428 junior high school students. The content of the test items, corresponding to the nodes in Figure 4, are shown as follows:

1. Represent the following numbers with positive or negative signs.
   - The number which is 7 less than 0.
   - The number which is 8 greater than 0.
2. Calculate the following: $(+6) + (-11) + (-21)$
3. Calculate the following: $(+8) - (+6) + (-12) - (-6)$
4. Calculate the following: $(-4) \times (-7) \times 25$
5. Calculate the following: $(-4) \div (-\frac{2}{3}) \times 5 \div (-3)$
6. A person bought eight notebooks, each of which costs a yen, and paid 1000 yen. Represent the amount of change by a literal formula.

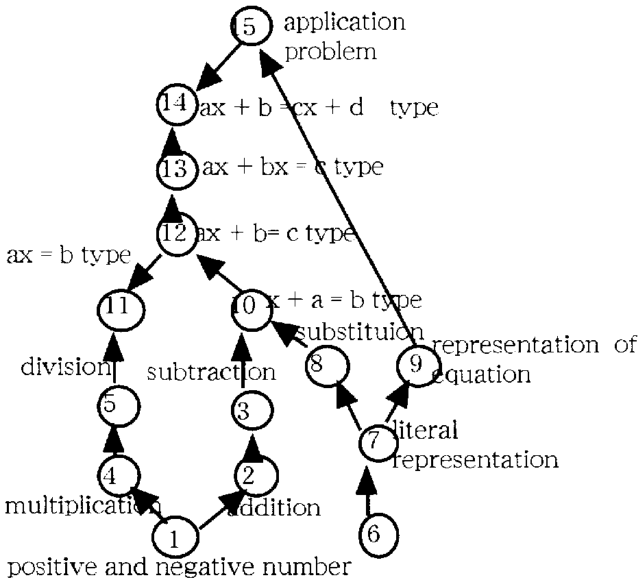

Figure 4: The structure of the expanded IRT model

7. Represent the following expression without using the symbols *times* and $\div$: $a \times b \div c$

8. When $a = -3$, calculate $-a + 8$.

9. Subtracting 2 from 6 times a certain number x, the result is equal to 4 times the result of x plus 3. Represent the foregoing relation by an equation.

10. Solve the equation $x + 8 = -2$.

11. Solve the equation $-4x = -6$.

12. Solve the equation $4x + 7 = -5$.

13. Solve the equation $5x = 7x - 4$.

14. Solve the equation $5x - 6 = -4x + 3$.

15. Adding 2 to 5 times a certain number x, the result is equal to 3 times the result of subtracting 6 from x. Determine the some of x.

Based on the data, we estimated the parameters ξ_i, whose results are presented in Tables 1 and 2. Here SE denotes the standard error of an estimator. It is known that the likelihood of a model with more parameters is greater than that of a model with less parameters. According to AIC and BIC, the extended IRT model is seen to be a more efficient model than the traditional two-parameter IRT model.

Table 1: The estimated parameters of the extended IRT model

i	$\tilde{u}_{ijk}$	$P(u_i \mid \tilde{u}_{ijk})$	a_i	$SE\ a_i$	b_i	$SE\ b_i$
1	—	0.964029	0.51133	0.10390	- 3.24276	1.34563
2	$u_1 = 1$	0.809353	0.84651	0.11480	- 2.15350	0.21017
2	$u_1 = 0$	0.025180			3.46324	1.53166
3	$u_2 = 1$	0.701439	0.90482	0.11008	- 1.13939	0.25537
3	$u_2 = 0$	0.079137			3.12432	1.49373
4	$u_1 = 1$	0.737410	0.53674	0.11225	- 1.32179	0.26622
4	$u_1 = 0$	0.025180			3,46324	1.53164
5	$u_4 = 1$	0.579137	0.82620	0.13784	- 0.27707	0.11130
5	$u_4 = 0$	0.125899			2.67896	1.47120
6	—	0.845324	1.28102	0.27558	- 1.32562	0.15585
7	$u_6 = 1$	0.780576	1.17644	0.21784	- 1.02232	0.13446
7	$u_6 = 0$	0.089928			3.26943	1.61479
8	$u_7 = 1$	0.780526	1.12663	0.24588	-1.00962	0.16842
8	$u_7 = 0$	0.014362			3.24636	1.14386
9	$u_8 = 1$	0.381295	0.98065	0.12200	0.69706	0.17160
9	$u_8 = 0$	0.014388			3.96860	1.16577
10	$u_3 = 1, u_8 = 1$	0.636691	1.12884	0.18917	- 0.43988	0.98986
10	$u_3 = 0, u_8 = 1$	0.122302			2.21654	0.67628
10	$u_3 = 1, u_8 = 0$	0.636691			- 0.43988	0.98983
10	$u_3 = 0, u_8 = 0$	0.122302			2.21654	0.67352
11	$u_5 = 1$	0.449640	1.31344	0.21619	0.21172	0.80277
11	$u_5 = 0$	0.079137			2.64713	1.14795
12	$u_{10} = 1, u_{11} = 1$	0.410072	1.53359	0.25587	0.33037	0.77086
12	$u_{10} = 0, u_{11} = 1$	0.035971			3.36564	1.14989
12	$u_{10} = 1, u_{11} = 0$	0.190647			2.21645	0.14760
12	$u_{10} = 0, u_{11} = 0$	0.035971			3.3656	1.98740
13	$u_{12} = 1$	0.543165	1.22571	0.13376	- 0.13000	0.10665
13	$u_{12} = 0$	0.107914			2.89598	0.98741
14	$u_{13} = 1$	0.943165	1.37188	0.11820	- 0.15267	0.12706
14	$u_{13} = 0$	0.100719			3.00416	0.94653
15	$u_9 = 1, u_{14} = 1$	0.965468	1.22680	0.21961	2.64915	1.09911
15	$u_9 = 0, u_{14} = 1$	0.053957			3.36954	1.02341
15	$u_9 = 1, u_{14} = 0$	0.010791			3.38436	1.02416
15	$u_9 = 0, u_{14} = 0$	0.010891			3.84362	1.02416

log-likelihood : -1983.72699
AIC : 4034.434
BIC : 4380.8

Item characteristic curves (ICC) can be drawn in the proposed model as in the traditional IRT model. Figure 5 shows the ICCs of the proposed IRT model for item 2 and item 11. The upper figures indicate the ICCs in the traditional model, the middle figures indicate the ICCs in the proposed model, and the lower figures indicate the marginal ICCs reflecting the network structure $\sum^{q_i} = jp(u_{ij}|\widetilde{u_{ijk}}, \theta_j, \boldsymbol{\xi}_i)p(\widetilde{u_{ijk}})$, corresponding to the upper figures in the traditional IRT model. It is seen that this marginal ICC can reflect the true ICC better. The figures on the left indicate the ICCs given the correct answer to the parent node item, and the figures on the right indicate the ICCs given the wrong answer to the parent node item.

Table 2: The estimated parameters for the 2-parameter IRT model

i	$P(u_i = 1)$	a_i	$SE\ a_i$	b_i	$SE\ b_i$	$I_i(\theta)$
1	0.964029	0.54393	0.17690	-4.00553	1.09237	34.53975807
2	0.834532	0.46289	0.10603	-2.30348	0.49513	3.783406272
3	0.780576	0.52828	0.11415	-1.63480	0.33144	3.486375704
4	0.762590	0.43701	0.09611	-1.74735	0.38001	2.010399946
5	0.705036	0.77297	0.12933	-0.85970	0.16192	5.338765845
6	0.845324	0.91440	0.17814	-1.52182	0.21139	25.70399042
7	0.870504	0.97900	0.17697	-1.64310	0.21694	42.65967648
8	0.604317	0.67715	0.11826	-0.44706	0.14079	2.216873472
9	0.395683	0.70553	0.13805	0.46844	0.12791	0.82024279
10	0.758993	1.58024	0.25675	-0.81003	0.09575	63.58645339
11	0.728777	0.79756	0.13801	-0.09176	0.11184	2.076921518
12	0.672662	1.34714	0.23438	-0.52678	0.08882	17.49362876
13	0.651079	0.87393	0.13829	-0.56454	0.12298	5.106288696
14	0.643885	0.95784	0.15018	-0.50647	0.10609	6.048478333
15	0.230216	0.90547	0.22574	1.10354	0.17192	0.433430516

log-likelihood : -3390.0912
AIC : 6810.1824
BIC : 6963

From the tables, it can be seen that most of the estimated values of parameter a_i for the proposed model are larger than the corresponding values of the traditional IRT model. This indicates that the standard error of the estimated θ can be estimated to be smaller than the standard error of the traditional IRT model by adding the structure knowledge to the traditional IRT model. However, it should be noted that since there are more parameters b_i in the proposed model than in the traditional IRT model, the standard errors of the estimators for b_i in the proposed model are estimated to be larger than those in the traditional model.

On the other hand, the marginal ICCs of the proposed model shown in the lower figures are more precisely estimated than the traditional ICCs shown in the upper figures, although the ICCs for the proposed model shown in the lower figures are similar to the ICCs of the traditional model shown in the upper figures.

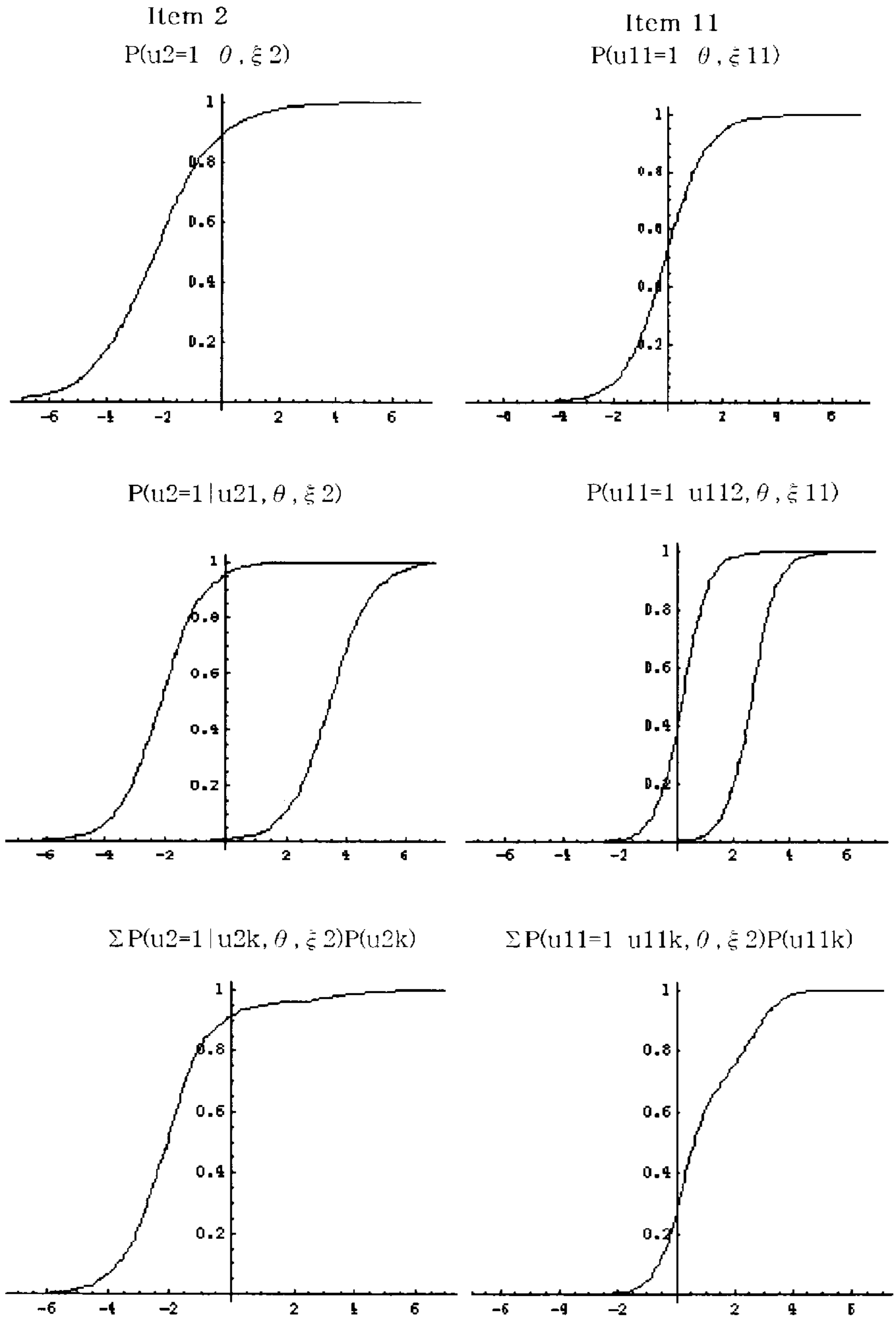

Figure 5: Examples of ICCs

Next, we demonstrate an example of the probability propagation shown in the Section 3. Figure 6 shows the probability propagation when we observe the subject's response for item 5. It is shown that the prior probabilities for the other items are propagated to the posterior probabilities. Thus, we can estimate the subject's knowledge structure.

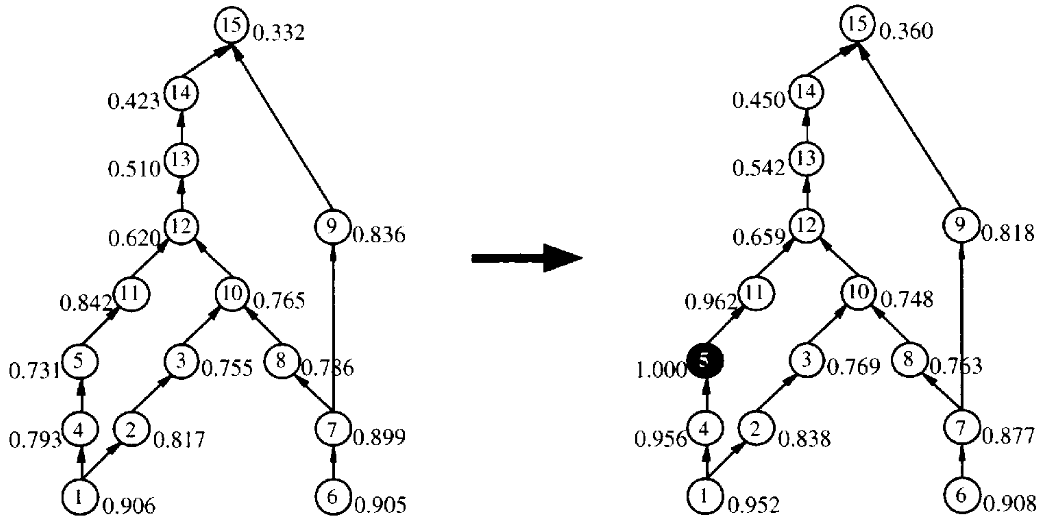

Figure 6: An example of the probability propagation

7.2 Some comparisons of the test construction processes

In this section, we use $EVTIN$ to propose the test construction method, and make comparisons with other test construction methods. Let $EVTIN^{(n'+1)}$ be the value of $EVTIN$ for the number of items N. Ueno (1994) proved that $EVTIN^{(n')}$ is monotonically increasing, i.e.,

$$EVTIN^{(n')} \leq EVTIN^{(n'+1)}, (n' = 1, \ldots, n-1). \tag{23}$$

Using this consequence, we propose the following test construction procedure in this study:

1. Given n', select an item set that maximizes the value of $EVTIN^{(n'+1)}$.

2. Iterate the procedure 1 for $n' = 1, 2, \ldots$ until the following convergence criterion is met:.

$$EVTIN^{(n'+1)} - EVTIN^{(n')} \leq \varepsilon. \tag{24}$$

For the following model, we compare the test construction process:

CASE 1 the extended IRT model,

CASE 2 the traditional probabilistic network model, and

CASE 3 the traditional IRT model.

Figure 7 shows the values of $EVTIN$ for each case. The horizontal axis denotes the number of items, and the vertical axis denotes the maximum values of $EVTIN$. The figure shows that the values of $EVTIN$ increase monotonically as does the number of items and finally converge. It is interesting to note that each slope of the curves is in order CASE 1, CASE 2, CASE 3, the values of $EVTIN$ converge at 9 items for CASE 1, converge at 10 items for CASE 2, and converge at 13 items for CASE 3 in the case of $\varepsilon = 0.01$. It is seen from these figures that the original 15 items can be reduced to 9 items in CASE 1, to 10 items in the CASE 2, and to 13 items in CASE 3, which demonstrates the efficiency of a network model including domain structure knowledge.

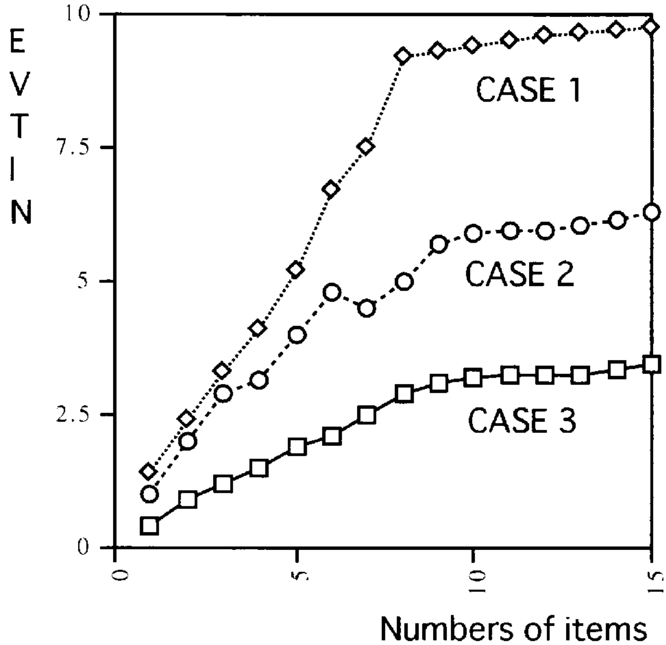

Figure 7: The values of EVTIN

Next we consider the Fisher information $I_i(\theta_j)$ in the test theory. We have

$$I_i(\theta_j) = \frac{\{P'(u_i = 1 \mid \theta_j)\}^2}{P(u_i = 1 \mid \theta_j)\{1 - P(u_i = 1 \mid \theta_j)\}}, \tag{25}$$

where $P'(u_i = 1 \mid \theta_j)$ denotes the first-order partial derivatives of $P(u_i = 1 \mid \theta_j)$. The amount of test information with N items is then given as

$$I(\theta_j) = \sum_{i=1}^{N} \frac{\{P'(u_i = 1 \mid \theta_j)\}^2}{P(u_i = 1 \mid \theta_j)\{1 - P(u_i = 1 \mid \theta_j)\}}. \tag{26}$$

When we have no knowledge of the values of θ_j, we estimate it as follows:

$$I = \sum_{i=1}^{N} \int I_i(\theta_j) g(\theta_j \mid \boldsymbol{\tau}) d\theta_j. \tag{27}$$

We define this case as CASE4. The values of $I_i(\theta_j)$ are shown in Table 2. The item patterns that are selected in CASE1 to CASE4 are shown in Tables 3-6, respectively.

Table 3: CASE1

the number of items	*the selected items*								
1	12								
2	5	14							
3	5	12	14						
4	5	12	14	15					
5	2	5	12	14	15				
6	2	5	9	12	14	15			
7	2	5	6	9	12	14	15		
8	2	5	6	7	9	12	14	15	
9	2	5	6	7	8	9	12	14	15

Table 4: CASE2

the numbers of items	*the selected items*								
1	12								
2	5	15							
3	5	9	15						
4	5	9	12	15					
5	2	5	9	12	15				
6	2	5	6	9	12	15			
7	2	5	6	9	12	14	15		
8	2	5	6	7	9	12	14	15	
9	2	5	6	7	8	9	12	14	15

It can be seen from these tables that while the $EVTIN$ selects different item patterns for each of the test item numbers, the Fisher information selects the same item pattern for each of the test item numbers.

Table 5: CASE3

the numbers of items	*the selected items*								
1	10								
2	10	12							
3	10	12	14						
4	10	11	12	14					
5	10	11	12	14	15				
6	10	11	12	13	14	15			
7	9	10	11	12	13	14	15		
8	8	9	10	11	12	13	14	15	
9	5	8	9	10	11	12	13	14	15

Table 6: CASE4

the numbers of items	*the numbers of items*								
1	10								
2	10	7							
3	10	7	1						
4	10	7	1	6					
5	10	7	1	6	12				
6	10	7	1	6	12	14			
7	10	7	1	6	12	14	5		
8	10	7	1	6	12	14	5	13	
9	10	7	1	6	12	14	5	13	2

7.3 Prediction efficiency

In this section, we will evaluate the prediction efficiency for the network model. That is, we evaluate how the model can predict the response for the unknown items by using data.

Let $\underline{y_j} = \{y_{1j}, \ldots, y_{kj}, \ldots, y_{m'j}\}$ be a subject j's response data for the test constructed by $EVTIN$, then we can obtain the parameters ξ, θ and $\{\underline{P_{ij}}\} = \{P(u_{ij} = 1 \mid \underline{y_j})\}$, $(i = 1, \ldots, m', j = 1, \ldots, n)$ for n subjects, where

$$y_{ij} = \begin{cases} 1 & \text{if the subject } j\text{'s response for } i\text{th item is correct} \\ 0 & \text{if the subject } j\text{'s response for } i\text{th item is wrong} \end{cases}$$

Let $\underline{\overline{y}_j} = \{\overline{y}_{1j}, \ldots, \overline{y}_{ij}, \ldots, \overline{y}_{(m-m')j}\}$ be the subject j's response except for $\underline{y_i}$, and then we define the square error E_i between the data and the posterior probability $P(u_{ij} \mid \theta_j, \underline{y_i})$ as follows:

$$E_j = \sum_{i=1}^{m-m'} \{\overline{y_{ij}} - P(u_{ij} \mid \underline{y_j})\}^2 / (m - m')$$

$$= (\underline{\overline{y}_j} - \underline{P_j})(\underline{\overline{y}_j} - \underline{P_j})^t/(m - m'). \tag{28}$$

The whole distance between data and a model can be defined as

$$E = \sum_{j=1}^{n} E_j/n. \tag{29}$$

We also evaluate the IRT model in the same way and the results are shown in Table 7. From the results, we can say that the prediction efficiency of the extended model with a domain structure is better than the traditional model with a simple independence structure.

Table 7: The comparisons of prediction efficiency

CASE	E
CASE1	0.039
CASE2	0.046
CASE3	0.148
CASE4	0.152

8. Conclusion and Discussion

In this paper, we extended the IRT model to the network model to relax the conditional independence assumption. The numerical experiments showed that the extension is successful. Here we did not employ the traditional evaluation approach based on the estimated latent variable θ. The latent variable as an ability measure has a certain problem in evaluating model adequacy as pointed out by Fischer (1995) and Ramsay (1996). Thus, this paper mainly discussed prediction of a subject's unknown responses. This approach is considered to be efficient for cognitive assessment. It can infer the items which the subject understands correctly or misunderstands.

The new model proposed here, however, has a drawback because it does not work for a large number of items for NP-hard problems. Concerning this problem, Ueno (1998b) derived the following analytical result: When $N \to \infty$ and $n \to \infty$, it holds that

$$\begin{aligned} &p(u_1, u_2, \ldots, u_N \mid \theta_j, \xi) \\ &= \prod_{i=1}^{N} \left(\frac{1}{1 + \exp(-a_i\theta_j + b_i)} \right)^{u_{ij}} \left(1 - \frac{1}{1 + \exp(-a_i\theta_j + b_i)} \right)^{1-u_{ij}} \end{aligned}$$

From this result, we would say that the traditional IRT is better for a large number of items, and the network model is better for a small number of items. It would be meaningful to investigate for how many items the IRT model would be preferable. In addition, construction of a structure B_S derived by BIC or MDL, which enjoys a strong consistency property, is an area for further investigation.

Acknowledgements

The author would like to thank Professor Yutaka Kano of Osaka University for his comments to improve the presentation of this paper.

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker, Inc.

Birnbaum, A. (1957). Efficient design and use of tests of a mental ability for various decision-making problems. (Series Report 58-16, No. 7755-23). *USAF School of Aviation Medicine, Randolph Air Force Base*, Texas.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp.395-479). Reading, MA: Addison Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29-51.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters. An application of an EM algorithm. *Psychometrika*, **46**, 443-459.

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika*, **60**, 459-487.

Fischer, G. H. (1924). On a distribution yielding the error function of several well-known statistics. *Proc. inter. cong. math. Vol.2* (pp.805-13). Toronto: University of Toronto Press.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, **56**, 255-278.

Lauritzen, S. L. (1996). *Graphical models.* Oxford Science Publications.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* MA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **35**, 43-50.

Mislevy, R. J. & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, **55**, 195-215.

Pearl, J. (1988). *Probabilistic reasoning in intelligent system.* California: Morgan Kaufman Publishers.

Ramsay, J. O. (1996). A geometrical approach to item response theory. *Behaviormetrika*, **23**, 3-16.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute of Educational Research.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the 4th Berkeley symposium on mathematical statistics, Vol.4* (pp.321-334). Berkeley: University of California Press.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, **19**, 49-57.

Rasch, G. (1966). *An individualistic approach to item analysis.* In P. Lazarsfeld and N. V. Henry (Eds.), *Readings in mathematical social science* (pp.89-107). Chicago: Science Research Association.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph, No.17.*

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph, No.18.*

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika*, **38**, 221-223.

Samejima, F. (1997). Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. *Psychometrika*, **62**, 471-493.

Shigemasu, K. (1985). *Introduction to Bayes statistics.* Tokyo: University of Tokyo Press (in Japanese).

Shigemasu, K. (1988). Educational evaluation from the perspective of normative decision theory. *Japan Journal of Educational Technology*, **12(3)**, 77-89. (in Japanese).

Suzuki, J. (1993). A construction of Bayesian networks from databases based on an MDL scheme. *Proceedings of ninth conference on uncertainty in artificial intelligence* (pp.266-273).

Wilson, M. R. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, **105**, 276-289.

Ueno, M., Ohnishi, H., & Shigemasu, K. (1994). A test theory using probabilistic networks, *Electronics and Communications*, **77**, 14-23.

Ueno, M. (1996). Construction of Bayesian belief networks by decision theoretic approach. *Journal of Japanese Society for Artificial Intelligence*, **13(6)**, 49-58. (in Japanese).

Ueno, M. (1996). *The test theory including the probabilistic networks.* Unpublished Doctorial Dissertation. Tokyo Institute of Technology. (in Japanese).

Ueno, M. (1998a). Potentialities of Bayesian networks. *Proceedings of the 12th annual conference of JSAI* (pp.106-109).

Ueno, M. (1998b). A derivation of IRT model from a joint binary probability distribution and it's parameter estimation. *Proceedings of the 26th annual meeting of the Behaviormetric Society of Japan* (pp.383-386).

Received March 1999. Final version received October 2001.