Discovering Skill Prerequisite Structure through Bayesian Estimation and Nested Model Comparison

Soo-Yun Han
Dept. of Mathematics Education
Seoul National University
Seoul, South Korea
ssu1205@snu.ac.kr

Jiyoung Yoon
Dept. of Mathematics Education
Seoul National University
Seoul, South Korea
torol2@snu.ac.kr

Yun Joo Yoo
Dept. of Mathematics Education
Seoul National University
Seoul, South Korea
yyoo@snu.ac.kr

ABSTRACT

Identifying prerequisite relationships among skills is important for better student modeling in many educational systems. In this paper, we propose a new method to discover prerequisite structure from data using nested model comparisons in the context of Bayesian estimation. We evaluate our method with simulated data and real math test data.

Keywords

Prerequisite structure discovery, Bayesian Network, MCMC estimation, nested model comparison, pseudo-Bayes factor.

1. INTRODUCTION

In many educational systems, the process of learning usually proceeds sequentially according to a predetermined order that reflects cognitive theories about student learning. In this learning sequence some knowledge skills must be acquired prior to learning advanced skills. In this study, we refer to *prerequisite structure* as the relationships among skills that put strict constraints on the order in which these skills can be mastered.

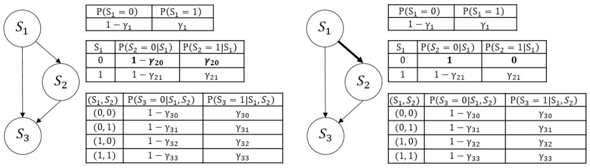
Identifying skill prerequisite structure is a crucial step to construct a valid and accurate student model in adaptive tutoring system or other educational system for estimation of student's skill mastery status and provision of appropriate remediation for them. Prerequisite structure can be specified by domain experts, but such process may be time-consuming and could produce subjective models lacking validity. Using large educational data and data mining techniques, several previous studies have tried to find prerequisite relationships among knowledge skills [1,2,3,7]. To derive prerequisite structure from student performance data is somewhat challenging in that a student's mastery status of skills cannot be directly observed, but can only be estimated, i.e, is latent in nature. Previous works mostly used Expectation-Maximization (EM) estimates for latent skill variables [1,2,3].

In this paper, we present a new method for discovering prerequisite structure from student performance data using Bayesian Markov Chain Monte Carlo (MCMC) estimation and nested model comparison. For nested model comparison, we use pseudo-Bayes factor (PsBF) [4], one of the Bayesian model selection criteria.

2. METHOD

In our method, it is assumed that student performance (item response) data at a certain point in time is given and skills related to items are specified. Skills and items are considered as binary random variables and the item-skill relationships are given by Q-matrix (a binary matrix that represents the mapping of items to skills) [9]. DINA model is used for modeling the probability of correct response to an item as a function of whether all the skills required are mastered and of slip and guess parameters [5]. To represent skill prerequisite structure, (static) Bayesian Network is used as student model. Bayesian network is a probabilistic graphical model representing the relationship of a set of random variables as a directed acyclic graph (DAG) with conditional probability tables (CPTs).

We now focus on the discovery of prerequisite relationship, that is, *strict hierarchical order* between mastery of two skills. To this end, we set two types of models: a *full model*, which parameterizes all possible dependencies between skills, and a *strict model*, which assumes prerequisite relationship between a pair of skills. For example, Figure 1 illustrates DAGs and CPTs of a full model consisting of three skills (S_1, S_2, S_3) and a strict model assuming prerequisite relationship between skill S_1 and S_2 (S_1 is a prerequisite for S_2). The difference between two models is that, while the full model contains the parameter γ_{20} related to the probability $P(S_2 = 1 \mid S_1 = 0)$, the strict model put a constraint that this probability is zero (that is, the strict model is nested within the full model). If skill S_1 is a true prerequisite for S_2, the parameter γ_{20} in the full model will be estimated to be closed to zero and there will be no significant difference in the degree to which the two models explain the data. The idea of nested model comparison is to statistically test the null hypothesis that the two models present the same likelihood on the data.



$P(S_1 = 0)$	$P(S_1 = 1)$
$1 - \gamma_1$	γ_1

S_1	$P(S_2 = 0 \mid S_1)$	$P(S_2 = 1 \mid S_1)$
0	$\mathbf{1 - \gamma_{20}}$	$\boldsymbol{\gamma_{20}}$
1	$1 - \gamma_{21}$	γ_{21}

(S_1, S_2)	$P(S_3 = 0 \mid S_1, S_2)$	$P(S_3 = 1 \mid S_1, S_2)$
$(0, 0)$	$1 - \gamma_{30}$	γ_{30}
$(0, 1)$	$1 - \gamma_{31}$	γ_{31}
$(1, 0)$	$1 - \gamma_{32}$	γ_{32}
$(1, 1)$	$1 - \gamma_{33}$	γ_{33}

(a) Full model

$P(S_1 = 0)$	$P(S_1 = 1)$
$1 - \gamma_1$	γ_1

S_1	$P(S_2 = 0 \mid S_1)$	$P(S_2 = 1 \mid S_1)$
0	**1**	**0**
1	$1 - \gamma_{21}$	γ_{21}

(S_1, S_2)	$P(S_3 = 0 \mid S_1, S_2)$	$P(S_3 = 1 \mid S_1, S_2)$
$(0, 0)$	$1 - \gamma_{30}$	γ_{30}
$(0, 1)$	$1 - \gamma_{31}$	γ_{31}
$(1, 0)$	$1 - \gamma_{32}$	γ_{32}
$(1, 1)$	$1 - \gamma_{33}$	γ_{33}

(b) Strict model

Figure 1. DAGs and CPTs of (a) a full model and (b) a strict model of skills S_1, S_2, S_3. The bolded directed edge from S_1 to S_2 in DAG of the strict model (b) means that S_1 is a prerequisite for mastery of S_2.

When two models are fitted to the data using maximum likelihood, the likelihood ratio test is used for hypothesis testing. In the context of Bayesian estimation, Bayes factor or its variants can be considered as the test method. We use *pseudo-Bayes factor*, which can be calculated by the MCMC estimation process, as the test statistic to contrast two models. The pseudo-Bayes factor for model M_1 relative to M_2 is the ratio of approximations of marginal likelihood based on predictive distributions and cross-validation strategies and defined as

$$\mathrm{PsBF}_{12} = \frac{\hat{p}(X \mid M_1)}{\hat{p}(X \mid M_2)} = \frac{\prod_{i=1}^{n} p(X_i \mid X_{-i},\ M_1)}{\prod_{i=1}^{n} p(X_i \mid X_{-i},\ M_2)} = \frac{\prod_{i=1}^{n} \int p(X_i \mid \Theta, M_1) p(\Theta \mid X_{-i}, M_1) d\Theta}{\prod_{i=1}^{n} \int p(X_i \mid \Theta, M_2) p(\Theta \mid X_{-i}, M_2) d\Theta}$$

where X_i is the response data of student i, X_{-i} is the complement of X_i in the data X, and Θ is the set of free parameters. The calculated PsBF value in MCMC estimation is compared to a critical value to decide whether to reject the null hypothesis or not. If the null hypothesis is not rejected, then the strict model is accepted, thus concluding that the prerequisite relationship exists.

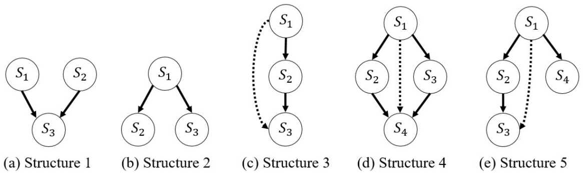
3. EVALUATIONS

To evaluate the efficiency of our method in discovering prerequisite structures, we first conducted a simulation study and then applied our method to a real dataset. In this process we faced a problem that PsBF values are dispersed from the known distribution of Bayes Factor [6]. To address this problem, we derived the critical value from the empirical distribution of PsBF values under the null hypothesis.

In our evaluation steps, all MCMC estimation algorithms were implemented using R package R2OpenBUGS [8]. For MCMC estimations, we set the priors as follows: a uniform prior $\mathrm{Unif}(0,1)$ on each structural parameters (γ_{ij}) and a beta prior $\mathrm{Beta}(6,21)$ on slip and guess parameters for each items.

3.1 Simulated Data

In this simulation part, we considered five prerequisite structures of latent skills (Figure 2). For each structure, we generated 500 datasets consisting of 1000 students' skill mastery status and their responses for test items using a balanced Q-matrix (each skills are measured with the same number and types of items) under the DINA model with low slip and guess probabilities randomly drawn from $\mathrm{Unif}(0,0.05)$.



(a) Structure 1 (b) Structure 2 (c) Structure 3 (d) Structure 4 (e) Structure 5

Figure 2. Five prerequisite structures of skills used in simulation study

We evaluate our method using two metrics: *true positive structure rate* (TPSR; # of correct structure recoveries in the output / # of true structures) and *true positive adjacency rate* (TPAR; # of correct adjacency recoveries in the output / # of adjacencies in true model).

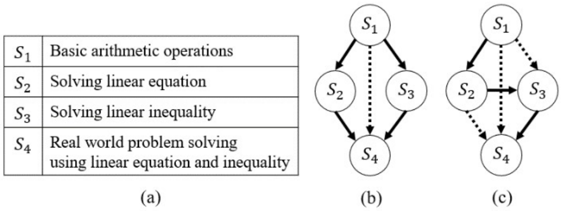
The results show that our method can efficiently discover prerequisite structure (Table 1). In all cases recovery rates of true structure are over 80% (the worst rate is 81.6% in structure 4). The recovery rates of true prerequisite relationship between two skills (edges) are even higher such as over 90%.

Table 1. TPSR and TPAR results for each structure

Structure	1	2	3	4	5
TPSR	0.926	0.840	0.872	0.816	0.874
TPAR	0.937	0.942	0.943	0.942	0.962

3.2 Real Data Application

We used mathematics cognitive diagnosis assessment data from 936 eighth grade students over a set of 16 items measuring four skills related to linear equation and linear inequality (Figure 3-a). The prerequisite structure of these skills (Figure 3-b) was initially set by knowledge experts.



S_1	Basic arithmetic operations
S_2	Solving linear equation
S_3	Solving linear inequality
S_4	Real world problem solving using linear equation and inequality

Figure 3. (a) Four skills in math test; (b) Prerequisite structure from knowledge experts; (c) Discovered prerequisite structure

Figure 3-c shows the prerequisite structure discovered by applying our method to the real data. All prerequisite relationships set by experts are well discovered, and one additional prerequisite relationship ($S_2 \rightarrow S_3$) is found. A possible explanation for this is that while knowledge experts judge that either linear equation or linear inequality can be learned first, students usually learn to solve linear equation first following the sequence in the curriculum.

4. CONCLUSION AND FUTURE WORK

We presented a method to discover skill prerequisite structure from data based on nested model comparison and evaluated the method using simulated data and real data. The performance of our prerequisite structure learning method was good within the settings used in our experiments. Since we used only low number of skills and certain assumptions for the evaluation, we need to further explore our method in various conditions.

In future work, we will investigate the idea of nested model comparison in the context of frequentist estimation (e.g., EM estimation) and compare with other previous methods. In this paper the focus is only on the prerequisite relationship between skills, but there may be other dependence relationships between them along with different types of response models. It would be interesting to study how to discover skill structures considering various dependency relationships in Bayesian Network modeling of skill mastery.

5. REFERENCES

[1] Brunskill, E. 2011. Estimating prerequisite structure from noisy data. In *Proceedings of the 4th International Conference on Educational Data Mining*.

[2] Chen, Y., González-Brenes, J. P., and Tian, J. 2016. Joint discovery of skill prerequisite graphs and student models. In *Proceedings of the 9th International Conference on Educational Data Mining*.

[3] Chen, Y., Wuillemin, P. H., and Labat, J. M. 2015. Discovering prerequisite structure of skills through probabilistic association rules mining. In *Proceedings of the 8th International Conference on Educational Data Mining*.

[4] Gelfand, A. E. 1996. Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 145-161). London: Chapman & Hall.

[5] Junker, B. W., and Sijtsma, K. 2001. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272.

[6] Kass, R. E., and Raftery, A. E. 1995. Bayes factors. *Journal of the American Statistical Association, 90*(430), 773-795.

[7] Scheines, R., Silver, E., and Goldin. I. 2014. Discovering prerequisite relationships among knowledge components. In *Proceedings of the 7th International Conference on Educational Data Mining*.

[8] Sturtz, S., Ligges, U., and Gelman, A. 2010. *R2OpenBUGS: a package for running OpenBUGS from R.* http://cran.r-project.org/web/packages/R2OpenBUGS/vignettes/R2OpenBUGS.pdf

[9] Tatsuoka, K. K. 1983. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345-354.