Discovering Prerequisite Relationships among Knowledge Components

Richard Scheines, Elizabeth Silver
Department of Philosophy
Carnegie Mellon University
{scheines,silver}@cmu.edu

Ilya Goldin
Center for Digital Data, Analytics, and Adaptive Learning
Pearson Education
ilya.goldin@pearson.com

ABSTRACT
Knowing the prerequisite structure among the knowledge components in a domain is crucial for instruction and assessment. Treating Knowledge Components as latent variables, we investigate how data on the items that test these KCs can be used to discover the prerequisite structure among the KCs. By modeling the pre-requisite relations as a causal graph, we can then search for the causal structure among the latents via an extension of an algorithm introduced by Spirtes, Glymour, and Scheines in 2000. We validate the algorithm using simulated data.

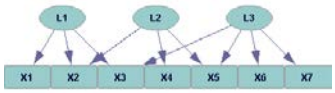
Keywords
Domain models, knowledge components, q-matrix, prerequisites, causal discovery

1. INTRODUCTION
In general, we need to determine the prerequisite structure of a domain. [3,4] Instead of relying on expert knowledge, which is subject to an "expert blind spot," in this paper we explore using causal model search to discover prerequisite structures from data.

As prerequisite relations are a form of causal relations, and as skills can be modeled as "latent" (unmeasured) variables, our approach is a generalization of causal structure discovery algorithms involving latent variables (Build Pure Clusters (BPC) [6] and MIMbuild [2, page 319]). In these algorithms, however, items are assumed to be "pure," that is, direct measures of only a single latent skill. In education this assumption is unreasonable, so we need to generalize the algorithms to handle models with impure measures. Further, BPC was written for continuous items. It would need to be extended to work on binary data before it could be applied to "correct/incorrect" test items.

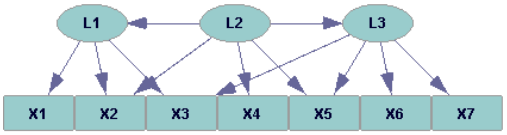
We begin with a simplifying assumption that we hope to eventually relax: that the Q-matrix (the matrix that specifies which item measures which skills) is known. We know of no current method for learning the prerequisite structure among skills in cases where there are very few pure items; so although the method we propose here is limited to cases where the Q-matrix is known, our method solves a novel problem. There are existing techniques for discovering and refining a Q-matrix, so there will be many cases where the Q-matrix is known or can be estimated to some approximation.

2. PREREQUISITE DISCOVERY
We model skills as continuous variables that represent the degree to which a student has mastered or has knowledge of a particular skill. We treat items as continuous variables that reflect the degree to which a student completed a task correctly. In practice, the measure of task completion is often a binary variable with values = correct/incorrect. A binary item can, however, be considered as a projection of a continuous item, and correlations among idealized continuous items can be estimated by computing the tetrachoric correlation matrix among the measured binary items.



(a: Measurement Model)

(b: Structural Model)



(c: Full Structural Equation Model)

Figure 1: Structural Equation Models

The Q-matrix typically defines which items "load" on which latent skills. We can define a "measurement model" that relates latent skills to measured items (Fig 1-a). By modeling the relations among the skills as a path analytic causal model among the latent variables (Fig 1-b), called the "structural model," we can then combine the "measurement model" and the "structural model" to form a full *linear structural equation model* [1].

By assuming that the measurement model is known, we can search for the structural model with the PC causal discovery algorithm [2], in which the inputs are the independence and conditional independence relations that hold among the latent variables. We compute or test the independence relations among the latents by constructing a distinct structural model and fitting it to the data for each particular independence test required. Our model construction method produces a provably consistent test of each conditional independence relation. [7]

3. VALIDATION ON SIMULATED DATA
To measure the method's ability to recover prerequisite structure, we conducted a large simulation study in which we varied (i) the structural model, (ii) the purity of the measurement model, (iii) the sample size, and (iv) whether the observed data were continuous or binary. In each of these conditions, we performed 100 simulations with different parameterizations. We used three structural models representing different causal relations between the latent skills, and varying degrees of impure measurement models (complicated Q-matrices). We ran the PC algorithm using the new test for independence involving a constructed structural equation model, and produced an equivalence class for the structural model in which we assumed no additional latent confounding, called a *pattern* [5].

We then scored each graph on the following metrics:

1. *True positive adjacency rate* (# correct adjacencies in output / # adjacencies in true graph), a.k.a. recall

2. *True positive orientations* or *orientation recall* (# correctly oriented edges in output / # orientable edges in true equivalence class). Defined to be 1 if none of the edges in the true equivalence class are orientable.

4. Results

Our results show that the algorithm performs well for discovering adjacencies (Figure 2). Even in the most difficult (and most realistic) case, where the sample size is 150, the measurement model is impure, and the data are binary, we still recover 74%, 76%, and 89.5% respectively for the three generating models.

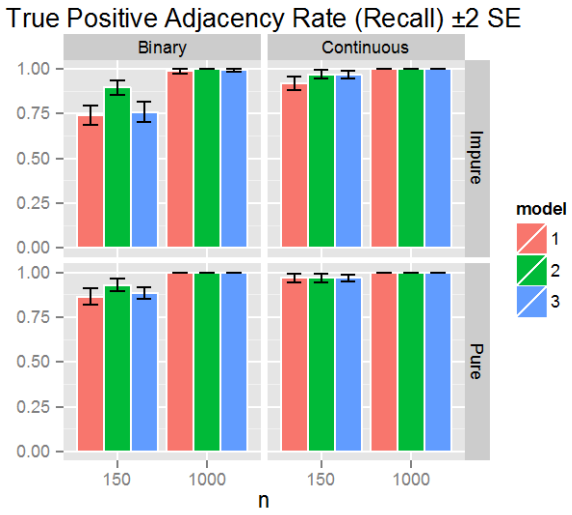


Figure 2: True positive adjacency rate

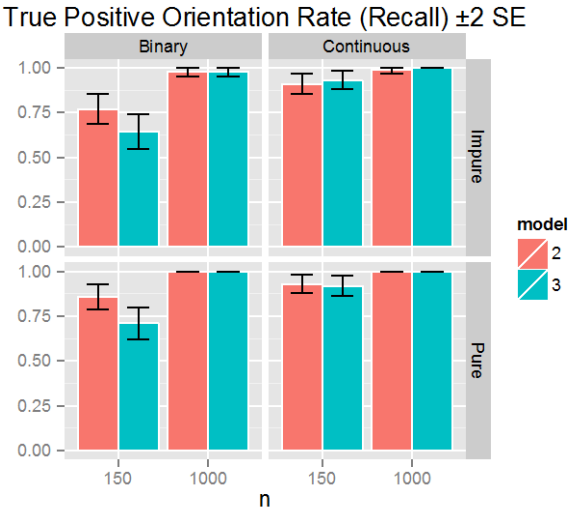


Figure 3: True positive orientation rate

The true positive orientation rate (recall) is shown in Figure 3. The worst score is 64.5% (for Model 3, with binary data, an impure measurement model and sample of 150), which is still quite good. Results for other metrics (omitted for lack of space) are also very good [7], including *false positive adjacency rate*, *true adjacency discovery rate*, *false positive orientation rate*, *true orientation discovery rate*, and f*alse negative orientation rate*.

5. CONCLUSIONS

The prerequisite graph is an important pedagogical artifact in itself, because we can use it to examine the structure of a domain, and it is furthermore a critical element of adaptive learning environments, where it can be used to create personalized and efficient learning trajectories for students. We expect that our algorithm can be used to discover fine-grained prerequisite structures to make student learning more efficient and more effective.

Unlike prior work [8], our method of prerequisite discovery only requires a single assessment from a point in time, and it applies to an assessment of any scope, regardless of whether it covers multiple problem-solving strategies on a skill, or multiple skills on a single learning objective, or multiple objectives in a syllabus, or multiple courses in a multi-year curricular sequence (e.g., a standardized test). Our algorithm is the only method currently available for inferring latent structure when the measurement model contains few pure items (i.e. items that load on only one latent). It performed well in our simulations, but has several important limitations including the assumption of linear relations, that the Q-matrix is known, and that the models are identified.

We intend to extend the work by expanding the range of models that can be identified, by investigating the robustness of the procedure to errors in the Q-matrix specified, and by including steps for Q-matrix discovery.

6. REFERENCES

[1] Bollen, K. (1989). *Structural Equation Models with Latent Variables*. Wiley.

[2] Spirtes, P, Glymour, G., Scheines, R. (2000). *Causation, Prediction and Search*, 2nd Edition, MIT Press.

[3] Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J.-P., & Johannesen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review*, 97(2), 201.

[4] Tatsuoka, K. K. (2012). *Cognitive Assessment: An Introduction to the Rule Space Method*. Routledge.

[5] Pearl, J. (2009). *Causality: Models and Reasoning, and Inference*, 2nd Edition. Cambridge University Press.

[6] Silva, R., Scheines, R., Glymour, C., Spirtes, P. (2006). "Learning the structure of linear latent variable models."*The Journal of Machine Learning Research* 7: 191-246.

[7] Scheines, R., Silver, E., Goldin, I. (in preparation) Discovering Prerequisite Relationships among Knowledge Components. Pearson Research Report.

[8] Vuong, A., Nixon, T., & Towle, B. (2011). A Method for Finding Prerequisites Within a Curriculum. In *Proceedings of 4th International Conference on Educational Data Mining* (pp. 211–216).