Identifiability in Knowledge Space Theory: a survey of recent results

Jean-Paul Doignon

Université Libre de Bruxelles,
Bd du Triomphe, c.p. 216,
B–1050 Bruxelles.
Belgium.
doignon@ulb.ac.be

Abstract. Knowledge Space Theory (KST) links in several ways to Formal Concept Analysis (FCA). Recently, the probabilistic and statistical aspects of KST have been further developed by several authors. We review part of the recent results, and describe some of the open problems. The question of whether the outcomes can be useful in FCA remains to be investigated.

Keywords: knowledge space, Basic Local Independence Model, Correct Response Model, model identifiability

In Knowledge Space Theory (KST, see Doignon & Falmagne, 1999; Falmagne & Doignon, 2011), a body of knowledge is represented by a finite set, say Q, of test items. The knowledge state of a student is identified with the collection of items he masters. Because of dependencies among the items, not any subset of Q can be a knowledge state; for instance, if Q is structured by a prerequisite relation, the states should be taken as the ideals of the transitive closure of the prerequisite relation. In general, the collection $\mathcal{K}$ of all possible *knowledge states* forms a *knowledge structure* $(Q, \mathcal{K})$; it is assumed $\varnothing$, $Q \in \mathcal{K}$. The correctness of the answer provided at a certain time by a student to any item is granted to depend only on his knowledge state, except for careless errors and lucky guesses.

Because variations are routinely observed in such answers, a probabilistic extension of KST was designed. So, assume the knowledge state of a student may vary (around a certain time point of his apprenticeship) in $\mathcal{K}$ according to a probability distribution π on $\mathcal{K}$. Moreover, for any item q in Q, let β_q be the probability of a careless error in answering q, and η_q be the probability of a lucky guess in answering q. All the numbers $\pi(K)$ (for K in $\mathcal{K}$), β_q and η_q (for q in Q) will be considered as parameters with (unknown) latent values. (Of course, the $\pi(K)$'s are not independent parameters, because they add up to 1.) The *straight case* obtains when $\beta_q = \eta_q = 0$, for any q in Q. We now propose two models for the probabilities of correctness of student answers (considered as the observables). Both models are based on the latent knowledge structure together with the various parameters we have just introduced. The second model

is described in Doignon & Falmagne (1999) (see also Falmagne & Doignon, 2011), while the first one is only implicit there.

The first model, the Correct Response Model (CRM), defines the probability $\tau(q)$ of a correct answer to any isolated item q. It first conditions the probability of a correct answer to item q on the state of the student:

$$\tau(q) \;=\; \sum_{K\in\mathcal{K}} Pr(q \,|\, K)\cdot\pi(K).$$

Then, it specifies each conditional probability $Pr(q \,|\, K)$ by taking into account the careless error probabilities β_q and the lucky guess probabilities η_q:

$$Pr(q \,|\, K) = \begin{cases} 1-\beta_q & \text{if } q\in K,\\ \eta_q & \text{if } q\notin K. \end{cases} \tag{1}$$

A second model, the Basic Local Independence Model (BLIM), defines the probability of a pattern of responses. Here, a pattern is a subset of Q meant to contain all items to which a student (at a given time) produces a correct answer. Exactly as the CRM, the BLIM conditions the pattern probability on the state of the student. Thus, the probability of a given pattern R of responses (with $R\subseteq Q$) equals

$$\rho(R) \;=\; \sum_{K\in\mathcal{K}} r(R,K)\cdot\pi(K),$$

where $r(R,K)$ is specified as follows:

$$r(R,K) = \Big(\prod_{q\in K\setminus R}\beta_q\Big)\Big(\prod_{q\in K\cap R}(1-\beta_q)\Big)\Big(\prod_{q\in R\setminus K}\eta_q\Big)\Big(\prod_{q\in Q\setminus(R\cup K)}(1-\eta_q)\Big).$$

Both of our models, the CRM and the BLIM, are instances of probabilistic models. On the basis of a fixed knowledge structure, they predict from any parameter point (that is, any list of values for all parameters) some definite probability values for the observables (in our case, the observables are either individual correct responses, or whole patterns of correct responses). We use the term *predicted distribution* to designate "any distribution of probability values for the observables that are predicted by the model". The questions we will consider are as follows (the first two are clearly stated in Bamber & van Santen, 2000 for probabilistic models in general).

1. Model testability: is there some distribution of probability values for the observables that the model does not predict?
2. Model identifiability: is each predicted distribution produced from at most one parameter point?
3. Model characterizability: are the predicted distributions susceptible of an effective characterization (without reference to the underlying parameter values)?

Recently Spoto, Stefanutti & Vidotto (2012) have investigated the first two questions for the BLIM, the model for pattern probabilities. Moreover, Stefanutti, Heller, Anselmi & Robusto (2012) have produced additional, nice results about identifiability of the BLIM, especially in its local version: *local identifiability* means identifiability when the model is restricted to some neighborhood of any given parameter point.

On our part, we consider the three types of questions for the CRM, the correct response model, however working mainly in the straight case. First, we are able to characterize testability of the model using a simple criterion (and also to reformulate a variant of it, numerical testability, in a manageable, technical way). Second, about characterizability, we point out unavoidable difficulties in recognizing when it holds. Third, as regards identifiability, we give a tractable equivalent, concluding that identifiability is not often met. On the positive side, we indicate how to modify the parameter domain (consisting of the knowledge state probabilities) in order to restore identifiability while keeping the same prediction range; nevertheless, we show that the construction works well only for the knowledge structures $(Q, \mathcal{K})$ which are derived from a quasi order on Q (as it is the case in the presence of a prerequisite relation). As a matter of fact, the construction heavily relies on a theorem of Stanley (1986) for a convex polytope he associates to a partial order.

The results presented during the talk are taken from a manuscript under preparation (Doignon, 2013).

Bibliography

Bamber, D., & van Santen, J. P. H. (2000). How to assess a model's testability and identifiability. *J. Math. Psych.*, *44*, 20–40.

Doignon, J.-P. (2013). A correct response model in knowledge space theory. Manuscript in preparation.

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Berlin: Springer-Verlag.

Falmagne, J.-C., & Doignon, J.-P. (2011). *Learning Spaces*. Berlin: Springer-Verlag.

Spoto, A., Stefanutti, L., & Vidotto, G. (2012). Considerations about the identification of forward- and backward-graded knowledge structures. Submitted.

Stanley, R. (1986). Two poset polytopes. *Discrete Comput. Geom.*, *1*, 9–23.

Stefanutti, L., Heller, J., Anselmi, P., & Robusto, E. (2012). Assessing the local identifiability of probabilistic knowledge structures. *Behavior Research Methods*, *44*, 1197–1211.