Identifiability in Knowledge Space Theory: a survey of some of the recent results

Jean-Paul Doignon

Université Libre de Bruxelles
Definition

A **knowledge structure** \((Q, \mathcal{K})\) consists of
- a finite **domain** \(Q\) formed by **items**, and
- a collection \(\mathcal{K}\) of subsets of \(Q\) called **knowledge states**.

We assume \(\emptyset \in \mathcal{K}\) and \(Q \in \mathcal{K}\).

Motivation

An item is used in a questionnaire to test some notion;

a knowledge state consists of all items/notions
which are mastered by some hypothetic student.

See the on-line system **Aleks**.
Consider the following generalization of a relation.

**Definition**

An **attribution** $\sigma$ assigns to any element $q$ in $Q$ some nonempty collection $\sigma(q)$ of subsets of $Q$, called the **clauses** for $q$.

**From a knowledge structure $\mathcal{K}$, build an attribution $h(\mathcal{K})$**

$$h(\mathcal{K})(q) = \{K \in \mathcal{K} : K \text{ minimal for } q \in K\}.$$  

**From an attribution $\sigma$, build a knowledge structure $k(\sigma)$**

$$K \in k(\sigma)$$

when $\forall q \in K : \exists C \in \sigma(q) : C \subseteq K$. 
The above extends MONJARDET (1970) approach to Birkhoff Theorem (relations vs. attributions; quasi orders vs. “surmise systems”).
Proposition

The pair \((h, k)\) forms a Galois connection between
the set of knowledge structures ordered by inclusion
and
the set of all attributions ordered by \(\lesssim\), with
\[
\sigma' \lesssim \sigma \iff \forall q \in Q, \forall C \in \sigma(q), \exists C' \in \sigma'(q) : C \subseteq C'.
\]

I am cheating a bit here, because \(\lesssim\) is a quasi order and not an order.

Corollary

There is a 1-1 correspondence between the closed elements:
the knowledge structures closed under \(\cup\)
and
the surmise systems, that is the attributions \(\sigma\) satisfying
\begin{enumerate}[(i)]
  \item if \(C \in \sigma(q)\), then \(q \in C\); \hspace{1cm} \text{(reflexivity)}
  \item if \(p \in C \in \sigma(q)\), then \(D \subseteq C\) for some \(D\) in \(\sigma(p)\); \hspace{1cm} \text{(transitivity)}
  \item if \(C, D \in \sigma(q)\) and \(C \subseteq D\), then \(C = D\).
\end{enumerate}

This correspondence extends the one in Birkhoff Theorem (D&F, ’85).
To summarize:

knowledge structures which are relations which are

<table>
<thead>
<tr>
<th>closed under $\cup$ and $\cap$</th>
<th>$\leftrightarrow$</th>
<th>quasi orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>discriminative, $\cup$- and $\cap$-closed</td>
<td>$\leftrightarrow$</td>
<td>orders</td>
</tr>
<tr>
<td>closed under $\cup$</td>
<td>$\leftrightarrow$</td>
<td>surmise functions</td>
</tr>
</tbody>
</table>

Definition

We call the resulting structures respectively

- quasi ordinal knowledge space
- ordinal space
- knowledge space

For more details, see

Let’s now go stochastic!

Remark
From now on, we consider the most informative case in which all items appear in a questionnaire.
Motivation

A student answer to a questionnaire item is correct if either he masters the topic and flawlessly formulates his response, or he does not master the topic but makes a lucky guess.

Example

Consider $\mathcal{K} = \{\emptyset, \{a, b\}, \{a, c\}, Q\}$ with the probabilities

$\pi(\emptyset), \pi(\{a, b\}), \pi(\{a, c\}), \pi(Q)$ for the states (they add up to 1),

$\beta_a, \beta_b, \beta_c$ for careless errors,

$\eta_a, \eta_b, \eta_c$ for lucky guesses.

The probability of a correct answer to the item $b$ equals

$$\tau(b) = \sum_{K \in \mathcal{K}} \Pr(b | K) \cdot \pi(K)$$

$$\overset{\text{def}}{=} (1 - \beta_b) \cdot \pi(\{a, b\}) + (1 - \beta_b) \cdot \pi(Q) + \eta_b \cdot \pi(\emptyset) + \eta_b \cdot \pi(\{a, c\})$$

$$= (1 - \beta_b) \cdot \left( \pi(\{a, b\}) + \pi(Q) \right) + \eta_b \cdot \left( \pi(\emptyset) + \pi(\{a, c\}) \right)$$

$$= (1 - \beta_b) \cdot \pi(K_b) + \eta_b \cdot \pi(K_b) .$$
The Correct Response Model (CRM)
The **parameters** of the Correct Response Model CRM\(\mathcal{K}\) are
\[\pi(K), \quad \beta_q, \quad \eta_q\]
- the probability of the state \(K\) (where \(K \in \mathcal{K}\));
- the probability of a careless error for item \(q\) (where \(q \in Q\));
- the probability of a lucky guess for item \(q\) (where \(q \in Q\)).

Thus the **parameter domain** is the product of a simplex \(\Lambda_\mathcal{K}\) (having vertex set \(\mathcal{K}\)) with 2 \(|Q|\) intervals \([0, 1]\).

The **outcome space** is \([0, 1]^Q\).

The model predicts the probability \(\tau(q)\) of a correct answer to any item \(q\) in \(Q\). The **prediction function** \(f\) maps a parameter point \((\pi, \beta, \eta)\) to the prediction point \(\tau\), with
\[\tau(q) = (1 - \beta_q) \pi(K_q) + \eta_q \pi(K_\bar{q}).\]

Thus
\[f : \Lambda_\mathcal{K} \times [0, 1]^Q \times [0, 1]^Q \rightarrow [0, 1]^Q : (\pi, \beta, \eta) \mapsto \tau.\]

In the **straight case** \(\beta_q = \eta_q = 0, \quad \forall q \in Q,\)
\[f : \Lambda_\mathcal{K} \rightarrow [0, 1]^Q : \pi \mapsto \tau.\]
In the straight case \((\beta_q = \eta_q = 0, \forall q \in Q)\), the prediction function
\[
f : \Lambda_K \rightarrow [0, 1]^Q : \pi \mapsto \tau \quad \text{with} \quad \tau(q) = \pi(K_q) = \sum_{K \in \mathcal{K}, q \in K} \pi(K)
\]
is linear (that is: the restriction of a linear function from \(\mathbb{R}^K\) to \(\mathbb{R}^Q\)).

**Proposition (very easy)**

The Correct Response Model in the straight case

is **testable** \((f(D) \subset \mathcal{O})\) iff \(\mathcal{K} \neq 2^Q\);

is **numerically testable** \((\mu(f(D)) = 0 < \mu(\mathcal{O}))\) iff the rank of the incidence matrix of \(\mathcal{K}\) is strictly less than \(|Q|\);

has its collection \(\mathcal{K}\) of states recoverable from the prediction range.

So, a modification of \(\mathcal{K}\) entails a modification of the prediction range.
In the straight case, the prediction function

\[ f : \bigwedge_K^Q \to [0, 1]^Q : \pi \mapsto \tau \quad \text{with} \quad \tau(q) = \pi(K_q) = \sum_{K \in K, q \in K} \pi(K) \]

is linear, with matrix the incidence matrix \( M_K \) of \( K \).

The prediction range is the 0/1-polytope with vertices the columns of \( M_K \), that is the characteristic vectors of the states.

**Meta-Assertion**

Depending on \((Q, K)\), some of the models are easily characterized, other ones lead to hopeless problems.

See the huge literature on 0/1-polytopes in geometry; operations research; mathematical psychology; etc.
Non-Identifiability of the Correct Response Model
Example due to M. LANDY (±1987)

Assume $Q \supseteq \{a, b, c, d\}$ and $\mathcal{K} \supseteq \{\{a, b\}, \{c, d\}, \{a, c\}, \{b, d\}\}$.

The two probability distributions on $\mathcal{K}$, one with
$$\pi(\{a, b\}) = \pi(\{c, d\}) = 0.5$$
the other with
$$\pi(\{a, c\}) = \pi(\{b, d\}) = 0.5$$
give exactly the same correctness probabilities $\tau(q)$.

Proposition

A straight CRM $(\Lambda_{\mathcal{K}}, f, [0, 1]^Q)$ is identifiable
iff the incidence matrix $M_{\mathcal{K}}$ has rank $|\mathcal{K}| - 1$.

Necessary condition for identifiability: $|\mathcal{K}| \leq |Q| + 1$. 
In general, the correct response model is not identifiable.

Many possible reactions:

(0) don’t worry; !
(1) modify the structure (hoping to recover identifiability); No.
(2) restrict the parameter domain in a well-defined way; Yes.
(3) . . .
Restoring Identifiability for the Correct Response Model
Example

Let $Q = \{a, b\}$ and $K = \{\emptyset, \{a\}, \{a, b\}\}$. Then the mapping $f$ is bijective from a 2-dimensional simplex onto a half-square:

The other half-square arises by taking $K = \{\emptyset, \{b\}, \{a, b\}\}$. Hence, the prediction range of the structure $(\{a, b\}, 2^{\{a, b\}})$ is decomposed into those of two chains.

A generalization derives from a result of STANLEY (1986).
Proposition

If \((Q, \mathcal{K})\) is the ordinal space built from the partial order \(\leq\) on \(Q\), the prediction range of the CRM decomposes into the union of simplices \(S_{\succsim}\), one for each linear extension \(\succsim\) of \(\leq\) with

\[
S_{\succsim} \equiv \begin{cases} 
1 \geq x_q \geq 0, & \text{for } q \in Q, \\
x_q \geq x_r & \text{when } q \succsim r.
\end{cases}
\]

Restricting the parameter domain to

\[
\bigcup_{\mathcal{C}} \{ \lambda \in \Lambda_{\mathcal{K}} \mid \forall K \in \mathcal{K} \setminus \mathcal{C} : \lambda_K = 0 \},
\]

where the union is over all maximal chains \(\mathcal{C}\) in \(\mathcal{K}\), makes the Correct Response Model identifiable.

Notice that

the linear extensions of \(\leq\)

are in a one-to-one correspondence with

the maximal chains included in \(\mathcal{K}\).
The new domain makes the CRM identifiable without modifying the prediction range (and it has a natural interpretation).

The result easily extends to quasi-ordinal spaces, but to no other.

**Proposition**

The prediction range of the straight Correct Response Model \((\Lambda, f, [0, 1]^Q)\) is decomposed into simplices \(S_{\preceq}\), where \(\preceq\) is taken in some family of weak orders on \(Q\), if and only if the structure \((Q, \mathcal{K})\) is a quasi ordinal space.
Question 1. What about other knowledge structures?
(Use other triangulations of the polytope?)

Question 2. Investigate the general case, not necessarily straight but with restricted values of the $\beta_q$ and $\eta_q$.

Question 3. Are there similar notions for FCA? Yes! (future work)
1. Knowledge Structures

2. Knowledge Spaces vs. Surmise Systems

3. Let’s Now Go Stochastic!

4. The Correct Response Model

5. Non-Identifiability of the Correct Response Model

6. Restoring Identifiability for the CRM

7. Questions