Review of AOS1010-046R0A0

General Comments

Conjunctive cognitive diagnosis models (CDM's), a.k.a. diagnostic classification models (DCM's), are two-layer bayes nets / graphical models defined by bipartite graphs connecting binary latent attribute variables $A^1, \ldots, A^k$ to binary observable response variables $R^1, \ldots, R^m$. These models are of great interest in cognitive/educational testing, and have applications in other geontype/phenotype-style probabilistic classification problems as well. The edges at each node R^i indicate the relevant attributes A^j which must be present together (in conjunction) in order to change (usually increase) the probability of a positive response for R^i. The Q matrix is the incidence matrix for the bipartite graph. Specific CDM's specify probability models for the A's and R's subject to these constraints. An open problem in the CDM literature is inference about the structure of the Q matrix, given only the observable response data (iid observations of the R's).

The present ms. is the first, to my knowledge, to provide general, principled, and potentially useful sufficient conditions for consistent estimation of the identifiable parts of Q from just the response data, under one common CDM, the DINA model. (The DINA model specifies response probabilities $P[R_i = 1] = 1 - s_i$ when all the relevant attributes are present, and $P[R_i = 1] = g_i$ when one or more relevant attributes is missing.)

The key insight is to formulate the problem as a kind of variable selection problem in a least-MAD linear model predicting margins of the 2^m table cross-classifying R's, and then choose Q (selecting variables) to minimize the residuals in this linear model. Mean absolute deviation (MAD) is chosen over likelihood-based calculation, with an eye toward computational efficiency in applications. Further results allow for joint estimation of Q and some of the response probabilities under the DINA model.

This is a very nice paper. It provides insights that have been absent until now. Someone reading this paper could write code immediately to do the estimation suggested here. Moreover the linear algebra machinery in the latter parts of the paper may be useful for other conjunctive CDM's. As a first paper breaking this ground there is much more to do, but this is a great start.

I have made several suggestions and corrections below, but all are addressible quickly by the authors. In addition to these comments I would like to add that the paper could use some light copy-editing to clean up English usage.

Specific Suggestions and Corrections

p3, two lines after (2.1), and throughout ms. "conjunctive" would be a better word than "non-compensatory", here and throughout. A conjunction ("AND") of the relevant attributes is needed to produce a positive response on each item. Other noncompensatory schemes are not fully conjunctive.

p 7, Remark 2.5, and throughout ms. The applicability of these models is broader than cognitive testing; see for example

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6,* 219–262.

p 7 conditions C1–C4 since you often refer to C1 and C2 by english names (completeness, saturation) and english labels exist for C3 and C4 also (iid, nondegenerate attribute distribution), it would be mnemonically better to refer to all by name and drop the abstract labesl C1–C4.

p 8, Remark 2.8 The situation is reminiscent of

Reckase, M. D. (1990). *Unidimensional data from multidimensional test and multidimensional data from unidimensional tests.* Paper presented at the Annual Meeting of the American Educatioal Research Association, Boston MA, April 16–20, 1990. Obtained from http://eric.ed.gov/PDFS/ED318758.pdf.

and in subsequent work. A less trivial example might help drive home the point more convincingly.

p 8, section 3 and beyond The notation here for s_i is reversed from what is common in the literature for the DINA model. That is, what is referred to as s_i here is referred to as $(1 - s_i)$ in the rest of the literature. Since a single letter is convenient for this paper I suggest defining, e.g., c_i to be the probability of a correct response, given that the relevant skills are known, and remarking that $c_i = 1 - s_i$ elsewhere in the literature.

p 9 bottom the ability to deal with $c_i < g_i$ is useful.

p 11, remark 3.2 This is really important for scaling as I imagine will be increasingly relevant as the use of CDM's grows.

p 11, sect 4, treatment of g_i's There are a couple of aspects of this discussion of assuming the g_i's are known that strike me as naive.

- Practical empirical experience with an item response theory (IRT) model known as the 3PL model (basically mixed effects logistic regression with a nonzero lower asymptote that functions analogously to g_i suggests that for multiple choice questions it is often the case that a data-based estimate $\hat{g}_i$ is not equal to $1/n_c$ where there are n_c choices. This has a lot to do with how the incorrect choices are designed/written, for example.

Similarly, while open-ended questions often arguably have $g_i = 0$, this need not be the case (indeed, you allow that chance of guessing "$(3 + 2) \times 2 =?$" is very small, rather than 0).

- What you are able to estimate with $\overline{s}_i$ is really the difference $s_i - g_i$. Sometimes we will know g_i but sometimes perhaps we will know s_i – E.g. if the student knows all the skills they are sure to the the item right, but if they are missing some they may still be able to guess with some probability that is interesting (perhaps associated with the prevalence of an alternate solution strategy not involving the skills specified for item i in the present Q matrix).

p 11, eq. (4.1) since consistency of $\tilde{s}$ is not dealt with in the paper (but only assumed when needed later on), I am not sure why not to include g in the argmin as well. Or is there a more general argument that I missed that s and g are not separately identifiable (even if p^* is suitably constrained)?

p 13, remark 4.3, first sentence It's remarkable that consistency of $\hat{Q}_{\hat{s}}(g)$ doesn't depend on consistency of $\hat{s}$. This seems to hinge on Prop. 6.6, which maybe should be played up a bit more.

p 13, remark 4.3, remainder Two related questions:

- what does the rest of this remark say about existing MCMC and E-M algorithms in the CDM literature, that purport to estimate s, g and p^*, for the DINA and related conjunctive models?
- In many applications, p^* has special structure: the components of A may be modeled as independent bernoulli's, or multivariate curved exponential family (IRT structure), or as probits with underlying correlated multidimensional normals. All of these reduce the dimensionality of the parameter space from $2^k - 1 + k$ to something that is basically linear or quadratic in k and m. This seems to greatly reduce the practical impact of your counter-example, doesn't it?

p 13 bottom I am not sure that the results in this paper can be applied to general *noncompensatory* models but the basic supporting results in Section 6 seem potentially relevant to other *conjunctive* models in the CDM/DCM world.

p 14 top I have the following comments about the to-do list:

- Your ideas about constraints on the Q matrix (or really the distribution of A, I think) may provide better examples for Remark 2.8 above.
- penalized optimization to incorporate expert knowledge is a nice idea.
- estimation of number of skills is important. I am not sure how serious the BIC suggestion is, though, since it will require something like a likelihood calculation, in contradiction to the scaling benefits suggested in remark 3.2.

- If I had one wish to weaken your sufficient conditions, it would be to weaken the completeness requirement on Q.
- It's surprising not to see joint estimation of s and g on the to-do list. This is certainly of practical/applied importance.
- Two other important practical to-do item would be convergence rates for the limit theorems, and some practical computational costs, as a function of problem size, for implementing these methods.