Test Construction

Robert Henson

The University of North Carolina at Greensboro

Introduction

- So now that you have a feel for how to develop your own model based on Q-matrix construction.
  - We will assume that the Q-matrix is known.
- We will now get to test construction and scale refinement.
- Specifically, we will answer the question:

Outline

- In order to think about test refinement or test construction (of fixed length) we are going to:
  - Briefly discuss test construction when the latent variable is continuous.
  - Discuss what a “good” item is in cognitive diagnosis.
  - Define a set of indices that can be used to measure a good item.
  - Discuss their use in application.

Traditional Test Construction

- We have methods of defining a good item and a good test in traditional settings where we assume that the latent variable is a continuous latent trait.

- A good test is one that measures the latent variable well.

  – Minimizes the standard error of estimate.

Traditional Test Construction

- For example:
  - Classical Test Theory may define a good item as one that correlates highly with test score.
  - Item Response Theory defines a good item as one that maximizes the Fisher's information for the ability score of interest.

Traditional Test Construction

- Based on this definition (and especially in IRT), methods of test construction have been developed.

- In IRT, one of the characteristics that is focused on is that Fisher's information is additive across items.

Cognitive Diagnosis

- One problem with trying to generalize the concepts of test construction using traditional methods to cognitive diagnosis models is that cognitive diagnosis models have classes.

- Values such as Fisher's information are not defined in cognitive diagnosis.

- In addition, the association of an item with test score does not mean quite the same thing.

Cognitive Diagnosis

- So, our definition of a “good” test must be slightly changed.
- We will need to define what is meant by a good test because measurement error does not mean quite the same thing with latent classes.
- A “good” test is one that correctly classifies examinees.
  - Correctly estimates examinees’ profiles.

Objective

- It is our goal to define an index or set of indices that:
  - Relate to correct classification rates.
  - Have similar properties as in IRT.
    - Uses all of the relevant information.
    - Have a meaningful interpretation.
    - Defined for the item and the test.
  - Are additive (the test index should equal the sum of item index).

Discrimination Indices

- We will define a set of indices that have these characteristics.
  - Kullback-Leibler Information.
  - Test discrimination index for CDMs (C_{j}).
  - Attribute discrimination indices for CDMs.
    - Indices $d_{(\mathrm{A})\mathrm{jk}}$ and $d_{(\mathrm{B})\mathrm{jk}}$.

Kullback-Leibler Information

- The Kullback-Leibler information, $\delta[f, g]$, is most commonly described as a distance measure.
  - Specifically, the "distance" between the two probability distributions $f(X)$ and $g(X)$.

$$\delta(f, g) = E_f\left[\log\left[\frac{f(X)}{g(X)}\right]\right]$$

Kullback-Leibler Characteristics

- Not quite a distance.
  - It is not symmetric.
  - Does not satisfy the triangle inequality.

- But, the higher the value the easier it is to discriminate between the two distributions.

- If the distributions are the same then $\delta(f, g)=0$.

K-L for CDMs

- For CDMs we can start by thinking about any two skill patterns $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$.

- We define: $f(x) = P(X_j \mid \alpha_u)$

$$g(x) = P(X_j \mid \alpha_v)$$

- So: $$\delta_j[\alpha_u, \alpha_v] = E_{\alpha_u}\left[\log\left[\frac{P(X_j \mid \alpha_u)}{P(X_j \mid \alpha_v)}\right]\right]$$

K-L for CDMs

- The K-L defined in this way will measure the degree to which the distributions differ.
  - This also is an indication of how well we can discriminate between skill pattern u and skill pattern v.
- Also, based on its definition, the test K-L comparing u to v is simply the sum of all item K-L for these two skill patterns.

K-L for CDMs

- However, there are $2^k(2^k-1)$ possible pairs of comparison.

- For simplicity we will keep all comparison for the *j*th item in a matrix $\mathbf{D}_j$

$$D_{juv} = E_{\alpha_u}\left[\log\left[\frac{P(X_j \mid \alpha_u)}{P(X_j \mid \alpha_v)}\right]\right]$$

Test Discrimination (C_j)

- $\mathbf{D}_j$ contains all possible comparisons of skill patterns for item *j*.
- One possible measure of discrimination is the whole matrix.
- For simplicity, we should summarize the matrix.
- The summary will describe the discrimination power between skill *patterns.*

Test Discrimination (C_j)

- So we will define C_j as

$$C_j = \frac{1}{\sum_{u \neq v} h(\alpha_u, \alpha_v)^{-1}} \sum_{u \neq v} h(\alpha_u, \alpha_v)^{-1} D_{juv}$$

- where

$$h(\alpha_u, \alpha_v) = \sum_{k=1}^{K} (\alpha_{uk} - \alpha_{jk})$$

A count of the number of skills that differ between the two skill patterns.

- Note that test index $C_.$ equals the sum of the C_j

Attribute Discrimination

- One limitation is that C_j describes all possible comparisons.
- To reduce the loss of information, indices are desired at the attribute level.
- Separate indices for classification of masters and classification of nonmasters
- Two will be discussed.

Attribute Discrimination A

- The average discriminating power of the j^{th} item for attribute k holding all other attributes constant.

$$d_{(A)jk1} = \frac{1}{2^{K-1}} \sum_{\{\alpha_u, \alpha_v\} \in \Omega_{k1}} D_{juv}$$

- where $\Omega_{k1} \equiv \left\{ \alpha_{uk} = 1 \cap \alpha_{vk} = 0 \cap \alpha_{um} = \alpha_{vm} \forall\, m \neq k \right\}$

Attribute Discrimination B

- The weighted average of the discriminating power of the j^{th} item for attribute k holding all other attributes constant.

$$d_{(A)jk1} = \sum_{\{\alpha_u, \alpha_v\} \in \Omega_{k1}} w_{uk} D_{juv}$$

- Where

$$w_{uk} = p(\alpha_u \mid \alpha_{uk} = 1)$$

Attribute level indices

- In general, there is an attribute level index for the masters and the nonmasters.
- However, for simplification I will discuss the case where we are interested in both equally.
- In that case, we can simply average the index for masters and for nonmasters.

Attribute level indices

- Specifically, for the remainder of the talk we will define the discrimination of the j^{th} item for the k^{th} attribute using Attribute index A as:

$$d_{(A)jk} = \frac{1}{2} d_{(A)jk1} + \frac{1}{2} d_{(A)jk0}$$

- We could do the same computation for Index B.

A Small Example

- For this small example we will assume that we have a single item for a test that measures two attributes and is parameterized using the RUM.
  - Q=(1, 0)
  - π^*=.8
  - r_{11}^*=*.125*
- To compute any of the indices we will need the matrix $\mathbf{D}_j$.

A Small Example

Definition of any element in $\mathbf{D}_j$

$$D_{juv} = E_{\alpha_u}\left[\log\left[\frac{P(X_j \mid \alpha_u)}{P(X_j \mid \alpha_v)}\right]\right]$$

Write out the expectation

$$D_{juv} = P(X_j = 1 \mid \alpha_u)\left[\log\left[\frac{P(X_j = 1 \mid \alpha_u)}{P(X_j = 1 \mid \alpha_v)}\right]\right]$$

$$+ P(X_j = 0 \mid \alpha_u)\left[\log\left[\frac{P(X_j = 0 \mid \alpha_u)}{P(X_j = 0 \mid \alpha_v)}\right]\right]$$

A Small Example

- First we will compute the probability of a correct response for all possible attribute patterns.

Attribute Pattern	P(X|α)
(0,0)	.1
(0,1)	.1
(1,0)	.8
(1,1)	.8

If (0,1) was α_u

If (1,0) was α_v

$$D_{juv} = .1\left[\log\left[\frac{.1}{.8}\right]\right] + .9\left[\log\left[\frac{.9}{.2}\right]\right]$$

A Small Example

- So in this case, the final D_j would be:

Zeros such as this are the K-L for attribute patterns that differ only by the second attribute, which was not required by this item

$$\mathbf{D}_j = \begin{pmatrix} 0 & 0 & 1.36 & 1.36 \\ 0 & 0 & 1.36 & 1.36 \\ 1.14 & 1.14 & 0 & 0 \\ 1.14 & 1.14 & 0 & 0 \end{pmatrix}$$

- By simply computing a weighted sum of all of these elements we would get C_j

A Small Example (Attribute 1)

Attribute pattern
(0,0)
(0,1)
(1,0)
(1,1)

$$\mathbf{D}_j = \begin{pmatrix} 0 & 0 & 1.36 & 1.36 \\ 0 & 0 & 1.36 & 1.36 \\ 1.14 & 1.14 & 0 & 0 \\ 1.14 & 1.14 & 0 & 0 \end{pmatrix}$$

To compute the attribute discrimination for attribute 1 (when mastery is the concern) we use these.

To compute the attribute discrimination for attribute 1 (when nonmastery is the concern) we use these.

A Small Example (Attribute 1)

Attribute pattern
(0,0)
(0,1)
(1,0)
(1,1)

$$\mathbf{D}_j = \begin{pmatrix} 0 & 0 & 1.36 & 1.36 \\ 0 & 0 & 1.36 & 1.36 \\ 1.14 & 1.14 & 0 & 0 \\ 1.14 & 1.14 & 0 & 0 \end{pmatrix}$$

To compute the attribute discrimination for attribute 2 (when mastery is the concern) we use these

To compute the attribute discrimination for attribute 2 (when nonmastery is the concern) we use these

A Small Example

- So in this case $C_j = .75$

- Attribute Index A

	Skill 1	Skill 2
Masters	1.14	0
NonMast	1.36	0

A Small Example

- As was discussed previously, if we did not want to concern ourselves with Masters and Nonmasters we can simply average the two to get:

	Skill 1	Skill 2
Attribute Discrimination	1.25	0

Index B

- Finally, to compute index B we need to assume that we know something about the probability distribution of attribute patterns.

- Then, instead of simply averaging the values from $\mathbf{D}_{j,,}$ we use a weighted average.

$$w_{uk} = p(\alpha_u \mid \alpha_{uk} = 1)$$

Summary of Indices

- So to summarize, for each item we could compute:
  - The item discrimination C_j.
  - The attribute level discrimination for determining mastery.
  - The attribute level discrimination for determining nonmastery.

Summary of Indices

- These indices are general indices that can be used for any model.

- All one must do is compute the matrix $\mathbf{D}_j$ which is only a function of the probability of a correct response given each attribute pattern.

Summaries of Simulation Studies

- Simulation studies show that by selecting items with large C_j, correct classification rates are also high.

- By selecting items so that attribute level indices are high for the test across all attributes, tests are generated with high correct classification rates.

- Therefore, the results suggest that for a fixed test length, the defined discrimination indices are related to correct classification rates and should be used for test construction and refinement.

Test Refinement (Using C_j)

- In a typical testing situation more items have been generated than are necessary.
  - For example, 40 are written with the intent of keeping 30 for the test.
- In refining a test, it is assumed that the item parameters have been obtained.

Test Refinement (Using C_j)

- Compute C_j for every item.
- Eliminate the items with the lowest values until desired test size has been reached.
- One concern is that eliminating items in this way may eliminate items that all measure the same attribute.
- So may simply make note of which items are being eliminated.
  - May incorporate constraints.

Test Refinement (Using d_j)

- Compute the attribute level index for each attribute for each item.
- Eliminate the set of items that are minimally discriminate for the attributes they measure.
- Also, consider the test attribute level discrimination when that item is eliminated.
- Eliminate those items that keep all test level attribute discrimination indices high.

Example

- Assume that I have a test of 6 items measuring 2 attributes.
- First, I will calibrate the items using the RUM, although I could use whatever model is appropriate.
- Given the item parameters, I can compute the probability of a correct response for each item for all attribute patterns.
- And then compute $\mathbf{D}_j$ for j=1 to 6, which is used to compute the item discrimination C_j and the attribute discrimination.

Example

- So lets assume that my indices for the 6 items are as follows.
- Here I will only give Index A, because I am assuming no prior knowledge of the population.

Item	π^*	r_1^*	r_2^*
1	.81	.6	.6
2	.84	.5	
3	.76		.4
4	.95	.4	
5	.87	.6	.8
6	.83		.7

Example

	C_j	$d_{(A)j1}$	$d_{(A)j1}$
1	.19	.16	.16
2	.25	.42	
3	.27		.45
4	.58	.98	
5	.17	.24	.06
6	.09		.16

To eliminate one item based on C_j we would pick item 6

However, to eliminate an item based on attribute index we would want to consider eliminating item number 2

Note that we do real well with attribute 1 and not with attribute 2.

1.80 0.83

Other Methods

- I have presented a set of methods of test construction and model refinement, but they are not the only methods.

- However, any test construction of refinement of a fixed length test should be based on the K-L information and or $\mathbf{D}_j$

- Because the discrimination indices (both item and attribute) are additive, any method used to construct tests base on IRT and Fisher's information can be used.

Quick Measures

- While item discrimination has meaning and seems to be one of the best measures of attribute discrimination there are simpler indices that can help in test refinement of construction.

- I will quickly talk about two

Quick Measures of Item Value

- The most basic idea behind this alternative set of indices is that the most informative items are those that perfectly determine the response.
  - Given attribute pattern the response is known.
- Therefore, by defining indices that indicate the extent that an item is determinant, we also indicate the value of an item.
- Note these are also related to some degree to the K-L information.

Quick Measures of Item Value

- The two example are:

- Dina Index:

$$c_j^{DINA} = (1 - s_j) - g_j$$

- RUM Attribute Index:

$$d_{jk}^{RUM} = \pi_j^* - \pi_j^* r_{jk}^* = \pi_j^* (1 - r_{jk}^*)$$

Summary

- So, we have defined a “good” test as a test that correctly classifies individuals (i.e., correctly assigns attribute profiles).

- We also defined discrimination based on K-L as an intuitive measure of the value of an item.

  – It is a summary of how easily response distributions given attribute patterns can be differentiated.

  – It is also additive across the items, allowing any method of test construction used in IRT to apply.

Summary

- In defining these new indices, we are able to determine the value of each item relative to all items being considered.

- In using this, we can refine, and construct “good” tests from a prespecified set of items.