The Korean Communications in Statistics
Vol. 13 No. 3, 2006, pp.467-477

A Refinement on DETECT for Polytomous Test Data[1)]

Hae Rim Kim[2)]

Abstract

A multidimensionality detecting procedure DETECT, based on conditional covariances between items, is extended and refined to deal with polytomous item data as well as binary one. A large body of simulation study shows extraordinary performance of DETECT in both enumerating degrees of multidimensionality in a test and discovering dimensionally distinctive item clusters. Real data study also provides very meaningful results, making DETECT a strong dimensionality assessment tool for the test data analysis.

Keywords : DETECT; polytomous; test dimensionality; dimensionally distinctive.

1. Introduction

An educational/psychometrical test is called unidimensional if there is only one trait needed to solve items in the test. In a case of unidimension, ability measurement means locating a corresponding point to one examinee on one-dimensional trait continuum. There exist, however, other trait(s) involved in a 'supposed to be' unidimensional test intentionally or inadvertently. Otherwise a test is designed to be multidimensional in the first place. When multidimensionality is suspected, there often needs to figure out how many traits involved and how much.

One method to detect test multidimensionality introduced by Kim (1994, 1997), using an index of multidimensionality, called DETECT, for dichotomously 0/1 scored items. It has shown that a conditional covariance based DETECT performs excellently to enumerate the amount of multidimensionality and to count the number of dimensions existed in the test. Moreover a preliminary study of DETECT for polytomously scored items has been conducted by Kim (2000a) in a small scale, showing that DETECT preserves the important characteristics, successfully detecting degrees of multidimensionality and counting number of dimensions, even in the case of polytomous items. However, results were

---

1) This work is supported by Sangji University Fund in 2002.
2) Associated Professor, Department of Data Information, Sangji University, 660 Woosan -dong, Wonju, Gangwon-do, 220-702, Korea.
E-mail: hrkim@sangji.ac.kr

incompatible with those of dichotomous items; DETECT ranges 0-4 for dichotomous cases while it exceeds 17 for polytomous cases with 4 categories. It naturally necessitates to develop a unified procedure of DETECT, which fits both dichotomous and polytomous cases equally well. This article is aiming at refining upon DETECT so that it is versatile in assessing test dimensionality under any circumstances. A revised version of DETECT is suggested and sizable simulation results are shown. Simulation presents very promising performance of DETECT. Also the result using a real data reveals well enough the DETECT's role of exploring dimensional structure involved in the test.

The following section briefly introduces item response theory (IRT) models and notions of DETECT. It also describes newly developed version of DETECT. The third section shows methods and results of simulations. The following section deals with findings in a real data study. The last section discusses the relevant future works.

2. DETECT

2.1 Item Response Theory Modeling

A binary scoring scheme of items is giving score 1 for the correct answer and 0 for the incorrect. For a given ability θ the probability $P_i(\theta)$ of answering an item X_i correct is denoted $P_i(\theta) = P(X_i = 1 | \theta)$ and called the item characteristic curve. A way of denoting this probability is three-parameter logistic (3PL) model defined below;

$$P_i(\theta) = c_i + \frac{1 - c_i}{\exp[D\, a_i(\theta - b_i)]} ,$$

where $D = -1.7$ is the scale factor to put the ability and item parameters together in the same scale. The parameter c is represented for guessing, a for item discriminating, and b for item difficulty, respectively. Here the ability θ is a possibly vector valued latent variable.

A set of items $\{X_i ; 1 \leqq i \leqq n\}$ is called a test of length n. In order to make the model statistically illuminating and psychologically meaningful, definitions of local independence and monotonicity are assumed, and dimension of a test is defined. See Kim (1994, 2000b) for details.

More than 2 categories are employed often to locate an examinee's performance level in a graded manner. For instance, scores 0, 1, and 2 are given, respectively, for an examinee answering wholly incorrectly, for an examinee answering partially correctly, and for an examinee answering correctly. Let m denote the number of categories in an item. In the generalized partial credit model (Masters, 1982; Masters & Wright, 1996; Muraki, 1992) when $m > 2$ categories are involved the probability of choosing the h_{th} $(1 \leqq h \leqq m)$ category in an item X_i, $P_{ih}(\theta)$, is defined as below;

$$P_{ih}(\theta) = \frac{\exp[\sum_{v=1}^{h} D a_i(\theta - b_{iv})]}{\sum_{c=1}^{m} \exp[\sum_{v=1}^{c} D a_i(\theta - b_{iv})]},$$

with a restriction $\sum_{h=1}^{m} P_{ih}(\theta) = 1$. Here D and a are same as in dichotomous case, but b_{ih} is the difficulty-category parameter concerning item i and category h simultaneously. Notions of local independence, monotonicity, and dimension remain intact while we use the polytomous model.

2.2 Representation of DETECT

One straightforward and theoretically sound approach to investigate the possibility of multidimensionality is to consider covariances of disjoint subsets of items after conditioning on a reliable substitute for the latent variable. Item pairs are found to be sufficient for disjoint item subsets and total score is a natural candidate for a substitute of Θ in examining conditional covariances.

Let's assume that a test is composed of disjoint clusters that are dimensionally distinctive from each other and each cluster is dimensionally homogeneous. Then it is conjectured that the conditional covariance estimate of an item pair given score of the remaining test is positive or negative subject to whether two items in the pair belong to the same cluster or not, respectively. The index DETECT by Kim (1994; 1997) combines non-zero second order conditional covariances of item pairs evidencing violation of unidimensionality by adding conditional covariances when two items come from the same cluster and subtracting conditional covariances when two items come from different clusters. In the first version of DETECT the average covariance over all pairs are subtracted before summing up the covariances and the index was divided by the number of item pairs after summing. The subtraction was made to extract off the bias occurred possibly in the unidimensional case (Kim, 1994). That is, for an (i,j) item pair

$$DETECT = \frac{1}{\binom{n}{2}} \sum_{(i,j)} (-1)^{\delta_{ij}} (\widehat{Cov}_{ij} - \overline{Cov})$$

where summation extends over all item pairs and δ_{ij} is determined to be 0 when two items belong to the same cluster or 1 otherwise. Obviously the maximum

DETECT occurs if the correct cluster formation is utilized.

A primary objective of dimensionality assessment is to identify dimensionally homogeneous item clusters if exist in a test. DETECT makes an attempt to identify such clusters and also to quantify the amount of the lack of unidimensionality present in the test. A sizable body of research has shown excellent performance of DETECT using binary items. For details, see Kim (1994; 1997). Instead of subtracting the average Zhang and Stout (1999) modified DETECT by taking the mean out of two DETECTs, one uses covariances conditioning on the remaining test while the other uses covariances conditioning on total test including all items, in order to balance out positive or negative associations between conditional item scores. They showed modified DETECT serves better. To figure out the item clustering corresponding to the maximum DETECT they also introduced the genetic algorithm. Details can be found in Zhang and Stout (1999).

A preliminary study of DETECT extending to polytomously scored items (Kim, 2000a) has shown that DETECT worked well to measure test multidimensionality and find dimensionally distinct item clusters in a small scale simulation only using 4-category items. It, however, used item score from 0 to 3 directly, resulting in larger value of DETECT from 0 to 17 than in binary item cases. Yu and Nandakumar (2001) suggested poly-Detect, replacing correlations of covariances. Originally DETECT has been designed to collect non-zero covariances resulted from violation of unidimensionality. Hence each amount of multidimensionality, realized in the conditional covariance, between item pair has vital importance for DETECT. That is the original magnitude of covariance is very important for DETECT calculation. However, due to standardization of covariance into correlation it is possible poly-Detect measures less exactly degrees of lack of unidimensionality than when covariances are used. It occurs often in multivariate data.

This study, instead, rescales scores into [0, 1] continuum first, i.e., for example, if $m = 4$ then scores 0, 1, 2, and 3 are rescored into 0, 1/3, 2/3, and 1, respectively. It is obvious there are no changes for binary items. By standardizing scores first, resulted covariances are considered wholly as evidences of lack of unidimensionality. The following section presents simulation results.

3. Simulation Study

3.1 Method of Simulation

This study deals with cases of 1, 2, 3, or 4 dimension(s). Items in a test is divided evenly into two subtests in two dimensional case, into three subtests in three dimensional case, and into four subtests in four dimensional case. For instance, when we use the 40 item test, two subtests with 20 items each are generated for a two dimensional case. We note this case as 20/20. In doing the same work, we get 13/13/14 and 10/10/10/10 situation respectively for three and four dimensional cases. Each subtest consists of items measuring only one dimension of trait. That is, one subtest is involved in only one dimension. Simulation is carried out for three situations; short test and small sample size(20 items with 500 examinees), medium length test and large sample size(40 items with 2000 examinees), and long test and small sample size(100 items with 500 examinees) to discover the behaviour of DETECT in varying conditions.

In an every single dimensional case polytomous items are generated as well as binary items. For binary cases the guessing parameter is fixed to 0.2. For polytomous cases items of 3 or 4 categories are employed and in a test all items are designed to have the same number of categories. Item parameters are simulated independently from the distributions shown in <Table 1> for each dimension when 2, 3, or 4 dimensions are involved. For the cases of 3 categories once difficulty-category parameters are generated independently from the assigned distributions, then the second parameter is added to the first difficulty-category parameter in order to make it more realistic. Situation remains similar for the cases of 4 categories; there needs only one more addition of independently generated third difficulty-category parameter to the second one. <Table 1> below summarizes item parameters.

The cases mentioned beforehand are so-called simple structured ones where each item is associated by only one trait. Not taking account of cases with mixed items, where more than one traits are involved in one item, we vary amounts of correlation between traits to make a test dimensionally heterogeneous. Correlation coefficient of .3, .5, .7, .8, or .9 is used between two traits. When more than two traits are employed we consider only balanced structure, where every pair of traits has the same correlation coefficient.

The abilities of examinees are generated from the standard normal distribution. Cross validation is employed in DETECT caculation; a half of examinees is used for finding dimensionally distinctive item clusters and the other half of examinees is used for calculating DETECT. From now on, the name DETECT indicates the cross-validated DETECT value using item clusters found by the first half of examinees. Each run of simulation follows three steps below;

Step 1. Generates item parameters.

Step 2. Generates item responses.
Step 3. Finds item clusters which maximizes DETECT using the first half examin-ees and calculates DETECT using the second half.

Each case is replicated 1000 times for the cases of medium length test and large sample size and long test and small sample size, respectively. For the short test and small sample size case 2000 replications are carried out. DETECT is multiplied by 100 for ease of representation.

<Table 1> Distribution Summary of Item Parameters

item parameter		number of categories 2	3	4
a_i	mean	1.0	1.0	1.0
	std. dev.	0.5	0.5	0.5
	max	2.5	2.5	2.5
	min	0.35	0.35	0.35
b_{i1}	mean	0.0	-0.3	-0.4
	std. dev.	0.8	0.7	0.8
	max	1.5	1.0	1.0
	min	-2.5	-2.0	-2.5
b_{i2}	mean		0.0	0.5
	std. dev.		0.7	0.5
	max		1.5	1.5
	min		-0.5	-0.5
b_{i3}	mean			0.5
	std. dev.			0.4
	max			1.2
	min			-0.5

3.2 Results of Simulation

As shown in <Table 2> one dimensional cases have almost none multidimensionality across all numbers of categories; the mean DETECTs over 1000/2000 runs are smaller than or equal to 0.050 except two 0.114 and 0.064 in short test and small sample size. It means DETECT works quite well in the unidimensional data. It is almost sure that cases of DETECT under 0.150 are considered as unidimensional without referring to item clustering formation. In addition, the cases of medium length test and large sample size result in smaller deviations as expected to be.

<Tables 3-5> show the means and standard deviations of DETECT for 2, 3, and 4 dimensional cases using 5 different values of correlation coefficient in each

test situation. First of all it is obvious that the less traits are correlated, i.e., as correlation coefficient decreases, the bigger DETECT is obtained, showing stronger multidimensionality. Secondly, it is also noteworthy that as another trait (dimension) is added, without changing total number of items in the test, DETECT value decreases when both correlation coefficient and number of categories are fixed. Thirdly, as number of categories increases DETECT decreases. It is clearly due to the fact that there is more values, which are usable for responses, in [0, 1] continuum. That means as the number of categories increases absolute difference between possible responses becomes smaller, resulting in smaller covariances in magnitude. Finally, in the cases using 2000 examinees we acquire stable DETECT, having smaller standard deviations than in the cases using 500 examinees.

<Table 2> DETECT for One Dimensional Cases

	short test and small sample size			medium length test and large sample size			long test and small sample size		
	number of categories			number of categories			number of categories		
	2	3	4	2	3	4	2	3	4
mean	0.114	0.064	0.032	0.050	0.031	0.018	0.019	0.009	0.004
std. dev.	0.086	0.054	0.036	0.021	0.014	0.009	0.044	0.034	0.036
max	0.434	0.263	0.161	0.130	0.070	0.060	0.264	0.208	0.256
min	-0.146	-0.125	-0.076	-0.019	-0.016	-0.007	-0.108	-0.090	-0.146

<Table 6> shows the percentages of recovering item clusters correctly by DETECT. They are obtained in 100 replication only for the 40 item and 2000 examinee cases. It is quite promising that in cases of correlation upto .5 dimensionally distinctive item clusters are all restored exactly. In cases of correlation .7 only several cases are missed. Also in the cases of correlation .9 at least 71% of item dimension structures are found correctly by DETECT. Accompanied by the magnitude of DETECT, DETECT's solution of item cluster formation supplies very useful information on test dimensional structure. It is noted that the information on both magnitude of DETECT and the clustering should go together in analysis.

It is, therefore, suggested that only when DETECT value is great enough, saying greater than 0.2, cluster formation found by DETECT can be useful as an evidence of test multidimensionality. Overall DETECT enumerates degrees of lack of unidimensionality and finds dimensionally different item clusters for polytomously scored items very effectively as well as for binary items.

<Table 3> DETECT for Two Dimensional Cases

corr. coeff.		short test and small sample size			medium length test and large sample size			long test and small sample size		
		number of categories			number of categories			number of categories		
		2	3	4	2	3	4	2	3	4
.9	mean	0.629	0.368	0.318	0.773	0.492	0.422	0.130	0.076	0.068
	std. dev.	0.230	0.143	0.111	0.093	0.058	0.047	0.110	0.093	0.085
.8	mean	1.360	0.841	0.703	1.474	0.971	0.836	0.483	0.318	0.300
	std. dev.	0.308	0.207	0.182	0.154	0.106	0.083	0.255	0.230	0.253
.7	mean	2.022	1.288	1.086	2.112	1.437	1.255	0.983	0.696	0.617
	std. dev.	0.383	0.272	0.257	0.198	0.141	0.125	0.418	0.361	0.419
.5	mean	3.206	2.196	1.877	3.331	2.333	2.100	2.350	1.522	1.376
	std. dev.	0.506	0.411	0.404	0.283	0.203	0.192	0.783	0.621	0.744
.3	mean	4.332	3.119	2.708	4.500	3.210	2.943	4.136	2.515	2.167
	std. dev.	0.608	0.521	0.516	0.337	0.265	0.249	1.036	0.906	1.030

<Table 4> DETECT for Three Dimensional Cases

corr. coeff.		short test and small sample size			medium length test and large sample size			long test and small sample size		
		number of categories			number of categories			number of categories		
		2	3	4	2	3	4	2	3	4
.9	mean	0.487	0.280	0.247	0.668	0.432	0.364	0.119	0.064	0.058
	std. dev.	0.178	0.110	0.090	0.672	0.045	0.038	0.100	0.079	0.074
.8	mean	1.104	0.719	0.598	1.280	0.864	0.732	0.487	0.273	0.209
	std. dev.	0.232	0.163	0.141	0.118	0.078	0.067	0.228	0.192	0.164
.7	mean	1.662	1.143	0.933	1.851	1.276	1.099	1.007	0.629	0.492
	std. dev.	0.281	0.201	0.180	0.151	0.106	0.094	0.332	0.330	0.282
.5	mean	2.649	1.950	1.630	2.919	2.073	1.826	2.224	1.581	1.156
	std. dev.	0.353	0.279	0.266	0.207	0.152	0.143	0.501	0.493	0.447
.3	mean	3.603	2.756	2.304	3.902	2.875	2.562	3.435	2.534	1.859
	std. dev.	0.430	0.347	0.348	0.239	0.199	0.184	0.550	0.564	0.528

<Table 5> DETECT for Four Dimensional Cases

corr. coeff.		short test and small sample size			medium length test and large sample size			long test and small sample size		
		number of categories			number of categories			number of categories		
		2	3	4	2	3	4	2	3	4
.9	mean	0.280	0.213	0.181	0.550	0.354	0.298	0.080	0.031	0.046
	std. dev.	0.124	0.092	0.074	0.068	0.038	0.029	0.080	0.073	0.072
.8	mean	0.616	0.539	0.457	0.997	0.705	0.603	0.230	0.168	0.159
	std. dev.	0.167	0.139	0.116	0.082	0.061	0.052	0.132	0.100	0.136
.7	mean	1.128	0.864	0.745	1.559	1.043	0.906	0.635	0.421	0.340
	std. dev.	0.206	0.167	0.145	0.112	0.081	0.073	0.195	0.193	0.203
.5	mean	1.605	1.522	1.320	2.094	1.711	1.519	1.499	0.917	0.870
	std. dev.	0.232	0.226	0.210	0.131	0.121	0.113	0.246	0.256	0.355
.3	mean	2.543	2.132	1.869	3.121	2.366	2.117	2.001	1.554	1.445
	std. dev.	0.249	0.271	0.253	0.202	0.158	0.153	0.237	0.419	0.398

<Table 6> Percentage of Finding the Right Item Clustering by DETECT

corr. coeff.	Two Dimension			Three Dimension			Four Dimension		
	number of categories			number of categories			number of categories		
	2	3	4	2	3	4	2	3	4
.9	88	81	100	85	90	98	71	75	98
.8	100	99	100	99	100	100	100	100	100
.7	100	100	100	100	100	100	100	100	100
.5	100	100	100	100	100	100	100	100	100
.3	100	100	100	100	100	100	100	100	100

4. A Real Data Study

An evaluation data for a lecture in a university is examined to reveal its dimensional structure using DETECT. The data was obtained at the end of fall semester in 2004. The test is composed of 18 items and each item is scored from 0 to 2 according to the degrees of satisfaction. Score 0 represents poor satisfaction, score 1 moderate satisfaction, and score 2 good satisfaction, respectively. I.e. they are 3 category items. The number of students evaluating the lecture is 176. DETECT value of 0.349 is resulted in with three clusters; the first cluster has 3 items of item numbers 1, 2, and 3, the second cluster has 8 items of 4, 5, 8, 9, 10, 11, 12, and 17, and the last cluster has 7 items of 6, 7, 13, 14, 15, 16, and 18. It implies this data shows lack of unidimensionality, but is weakly multidimensional. Referring to values in <Table 4> for the short test with small sample size cases, it is understood as three dimensional with correlation coefficient between dimensions are a little less than .9.

On top of finding about existence of multidimensionality in the test as above, it is very interesting and meaningful to look at cluster formation found by DETECT; all 3 items evaluating student's own attitude toward the lecture are grouped together in one cluster, 8 items evaluating lecturer's preparation and faithfulness to the lecture are grouped in another, and the rest 7 items evaluating lecturer's teaching skills and student's overall satisfaction of the lecture are grouped in the other. It is obvious that the first 3 item cluster is measuring different trait from those in the other two clusters. In order to investigate this feature further a study below is performed. For convenience of presentation, let's denote the first 3 item cluster student attitude cluster, the second 8 item cluster lecturer faithfulness cluster, and the last 7 item cluster lecturer skill cluster. When we employ student attitude cluster and lecturer faithfulness cluster together DETECT value is 0.458, showing weak but more multidimensionality than the whole test. Another study using student attitude cluster and lecturer skill cluster results in DETECT value 0.511. However, when we use two clusters concerning lecturer the value of DETECT 0.307 is obtained, meaning weakly multidimensional and weaker than the other two cases as we expected.

5. Closing Remarks

Newly refined DETECT procedure rescales scores into [0, 1] continuum when items are scored as $0, 1, \ldots, m-1 (m \geq 3)$ in order to obtain standardized score range. By doing so, DETECT can extend its role to polytomously scored items. Extended version of DETECT for polytomous items works strikingly excellently in revealing test dimensionality structure by both calculating the amount of multidimensionality and discovering item clusters. Of course the same results are obtained for binary items as well. This study focuses on finding feasibility of DETECT for polytomous items using simulated data and real data. There is a hope that these results give some insights for analysts who try to reveal the dimensional structure of the data. It might be possibile to apply DETECT to test writing/developing procedure when one seeks a dimensionally homogeneous test. More studies on this topic follow as a future work.

References

[1] Kim, H.R. (1994). New Techniques for the Dimensionality Assessment of Standardized Test Data. A Doctoral Dissertation, University of Illinois at Urbana-Champaign.

[2] Kim, H.R. (1997). Dimensionality Structure Analysis in Latent Traits Estimation. *Proceedings of Mathematical Education*, Vol. 5, 524-532.

[3] Kim, H.R. (2000a). A Dimensionality Assessment for Polytomously Scored Items Using DETECT. *The Korean Communications in Statistics*, Vol. 7, 597-603.

[4] Kim, H.R. (2000b). Some Asymptotic Properties of Conditional Covariance in the Item Response Theory. *The Korean Communications in Statistics*, Vol. 7, 959-966.

[5] Masters, G.N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, Vol. 47, 149-174.

[6] Masters, G.N. and Wright, B.D. (1996). The Partial Credit Model. In W. Linden and R. Hambelton (Eds.). *Handbook of Modern Item Response Theory*. New York, NY; Springer.

[7] Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, Vol. 16, 159-176.

[8] Yu, F. and Nandakumar, R. (2001). Poly-Detect for Quantifying the Degree of Multidimensionality of Item Response Data. *Journal of Educational Measurement*, Vol. 38, 99-120.

[9] Zhang, J. and Stout, W.F. (1999). The Theoretical DETECT Index of Dimensionality and Its Application To Approximate Simple Structure. *Psychometrika*, Vol. 64, 213-249.

[Received August 2005, Accepted August 2006]