A NEW INDEX OF DIMENSIONALITY — DETECT

HAE RIM KIM

ABSTRACT. A data-driven index of dimensionality for an educational or psychological test — DETECT, short for Dimensionality Evaluation To Enumerate Contributing Traits, is proposed in this paper. It is based on estimated conditional covariances of item pairs, given score on remaining test items. Its purpose is to detect whatever multidimensionality structure exists, especially in the case of approximate simple structure. It does so by assigning items to relatively dimensionally homogeneous clusters via attempted maximization of the DETECT over all possible item cluster partitions. The performance of DETECT is studied through real and simulated data analyses.

1. Introduction

Given an educational or psychological test, it is very important to identify the number of latent dimensions, to estimate the amount of multidimensionality, and to assign items to simple structure clusters (i.e., each cluster consisting of dimensionally homogeneous items) when approximate simple structure holds, as can be the case for many kinds of tests. A data-driven index of dimensionality — DE-TECT, short for Dimensionality Evaluation To Enumerate Contributing Traits, is introduced in this paper. Its purpose is to detect whatever multidimensionality structure exists, especially in the case of approximate simple structure. Informally, approximate simple structure holds when a test is composed of disjoint item clusters that are dimensionally distinct from each other while the items are relatively dimensionally homogeneous in each cluster. Operating in an exploratory mode, DETECT searches for the best assignment of items to relatively dimensionally ho-

---

mogeneous clusters. Then, *DETECT* estimates the amount of multidimensionality displayed by the chosen partition into item clusters. Intuitively, if a test has an approximate simple structure, then when two items come from the same cluster, the conditional item pair covariance given total score on the remaining items will be positive. By contrast, if two items come from different clusters, the conditional item pair covariance will be negative. For example, given a test of 50% math items and 50% verbal items, one expects the conditional covariance of a math pair or of a verbal pair (conditional on the score on the remaining items) to be positive. By contrast, the conditional covariance of a math item, verbal item pair should be negative.

Based on this, for a given partition of items into disjoint clusters, *DETECT* combines the estimated conditional item pair covariances by adding conditional item pair covariances when two items come from the same cluster and subtracting conditional item pair covariances when two items come from different clusters. Therefore, intuitively, the maximum value of *DETECT* occurs when the "correct" dimensionality-based cluster formation is utilized. Further, the number of clusters for the cluster formation that maximizes *DETECT* is judged to be the number of dimensions present in the test, and the cluster that an item is located in corresponds to the dominant dimension the item is measuring. In addition, the magnitude of the maximum *DETECT* value is informative in indicating the degree of multidimensionality the test displays.

A Genetic Algorithm is used to calculate the maximum *DETECT* value over all possible item cluster partitions and to identify the maximizing cluster partition. In general, it is computationally prohibitive to search over all possible cluster partitions; instead, an informed choice of cluster formations is usually used in searching for the maximum *DETECT* value. One method is to use a Hierarchical Agglomerative Cluster (HAC) analysis Computer Program, the HAC algorithm customized by the use of a proximity measure sensitive to the dimensional homogeneity of an item pair and programmed by Roussos (Roussos, 1993), to propose various cluster partitions from which one can find the maximum *DETECT* value. However, the

DETECT value gotten from HAC is not always the global maximum. By using a Genetic Algorithm, we have a greater chance to obtain the global maximum *DETECT* value, or at least get very close to it.

The remainder of the paper is organized as follows. The definition of *DETECT* is given in Section 2. Section 3 uses *DETECT* to analyze real data of the GRE. Section 4 contains some simulation results from using *DETECT*. Section 5 summarizes results and gives some discussion.

2. Definition of *DETECT*

2.1 *DETECT*

The index *DETECT* was developed by Kim (1994) as an outgrowth of Junker and Stout's (1994) $\hat{\varepsilon}$ of multidimensionality assessment index. Suppose $\{X_i; 1 \leq i \leq n\}$ is a test of n items. Suppose $A_1, A_2, \ldots, A_r$ are all non-empty subsets of the test $\{X_i\}$, $A_i \cap A_j = \emptyset$ for all $1 \leq i < j \leq r$, and $\bigcup_{i=1}^{r} A_i = \{X_i\}$. Then, $\mathcal{P} = \{A_1, A_2, \ldots, A_r\}$ is called a *r-subset (r-cluster) partition* of the test.

For an item pair (X_i, X_j), define a weighted sum of conditional covariance estimates of X_i and X_j as

$$\widehat{Cov}_{ij} = \frac{1}{J} \sum_{k=0}^{n-2} J_k \widehat{Cov}(X_i, X_j | S_{ij} = k). \tag{1}$$

Here S_{ij} is the observed correct score on the $(n-2)$ remaining items except for items i and j, J_k is the number of examinees with score $S_{ij} = k$, and J is the total number of examinees. The estimated covariance $\widehat{Cov}(X_i, X_j | S_{ij} = k)$ for the index triple (i, j, k) may be computed in the usual way:

$$\widehat{Cov}(X_i, X_j | S_{ij} = k) = \frac{1}{J_k} \sum_{l=1}^{J_k} (x_{ikl} - \bar{x}_{ik})(x_{jkl} - \bar{x}_{jk});$$

here x_{ikl} is the score of item i for the l-th examinee with $S_{ij} = k$, and $\bar{x}_{ik} = (1/J_k) \sum_{l=1}^{J_k} x_{ikl}$, $i = 1, 2, \ldots, n$, $k = 0, 1, \ldots, n-2$, $l = 1, 2, \ldots, J_k$.

Let Ω be the set of all pairs of item indices, i.e.,

$$\Omega = \{(i,j), 1 \le i < j \le n\}.$$

Note that Ω has $n(n-1)/2$ elements.

The index $DETECT$ is defined as

$$DETECT(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{(i,j)\in\Omega} \delta_{ij}(\widehat{Cov}_{ij} - \overline{Cov}), \tag{2}$$

where $\mathcal{P}$ is any specified r-subset (r-cluster) partition of the test, $\overline{Cov}$ is the average of $\widehat{Cov}_{ij}$ over all $n(n-1)/2$ item pairs, and

$$\delta_{ij} = \begin{cases} 1 & \text{if items } X_i \text{ and } X_j \text{ are in the same cluster} \\ -1 & \text{otherwise.} \end{cases}$$

The index δ_{ij} manipulates the $(\widehat{Cov}_{ij} - \overline{Cov})$ term in (2), to be added or subtracted according as items X_i and X_j belong to the same cluster or not; when both items belong to the same cluster the *centered* (it is centered at $\overline{Cov}$) conditional covariance estimate $(\widehat{Cov}_{ij} - \overline{Cov})$ is added, while it is subtracted otherwise. The rationale of centering will be discussed later.

As a function of $\mathcal{P}$, $DETECT$ can take on value ranging from negative to positive. One object here is to find one particular partition of the test *maximizing* $DETECT$. Maximum $DETECT$ is expected when test is classified into dimensionally "correct" subsets. To further understanding, close examination of the behavior of conditional covariance $\widehat{Cov}_{ij}$, which is used as basic building block of $DETECT$, is essential in terms of membership of items i and j into subsets.

2.2 Conditional Covariance

For didactic purposes, we first consider two dimensional tests and assume that all items are modeled by the following two dimensional compensatory logistic model (Reckase, 1985):

$$P_i(\theta_1, \theta_2) = c_i + \frac{1 - c_i}{1 + \exp\{1.7(a_{i1}\theta_1 + a_{i2}\theta_2 - d_i)\}},$$

where

(a_{i1}, a_{i2}) is the discrimination parameter vector,

d_i is a scalar parameter that related to the difficulty of item i,

c_i is the lower-asymptote parameter ($0 \leq c_i < 1$),

(θ_1, θ_2) is the complete latent trait vector.

In the two dimensional (θ_1, θ_2) plane, intuitively the total score (or the remaining total score) can be thought as (best) measuring a linear composite θ of the two traits having weights determined primarily by the influence of each trait; this linear composite is called the *test composite*. Especially in simple structure case, the number of items in each trait and the strength of discrimination of the items in each trait are most influential in determining the weights of the test composite. The point is that $\widehat{Cov}_{ij}$ is interpreted as an estimator of $Cov(X_i, X_j \mid \theta)$ for an appropriate level of θ corresponding to the observed score S_{ij} value.

Adapting the graphical method of representing items of Ackerman(1994), when a test is strictly unidimensional, all item discrimination vectors lie on a straight line in the (θ_1, θ_2) plane, and consequently the test composite considered also lies on the same line. In this case, $\widehat{Cov}_{ij} = 0$ because local independence holds for unidimensional θ approximately estimated by S_{ij}, except for statistical error caused by score unreliability and by estimation noise. Thus, maximum *DETECT* value will also be expected to be zero except for statistical error.

When a test is essentially unidimensional (see Stout, 1990), one possible case is that all item discrimination vectors form one narrow fan-shaped sector in the two dimensional plane, as shown in Figure 1(a). In this case the dominate unidimensional trait is the test composite, and the $\widehat{Cov}_{ij}$ can be expected to be relative small. Hence, the maximum *DETECT* value should be relatively small, indicating approximate unidimensionality or weak multidimensionality.

Assume in the cases considered, as illustrated in Figure 1, the score we condition on has equal weights for both θ_1 and θ_2; that is, the test composite direction has

a 45° angle (the dotted line in Figure 1) in the plane. Then, conditioning on the total score, corresponding to the 45° line trait, it is intuitive that positive covariances should be produced between items on the same side of the 45° line, and negative covariances between items from different sides except for statistical error. Recall from (2) that δ makes the negative conditional covariances positive in calculating $DETECT$ when they are from between-cluster items. Therefore, except for statistical error, $DETECT$ can be expected to be maximized at by a 2-subset partition formed by dividing items by the 45° line, even for the cases as shown in Figure 1(c) and (d) where the degree of simple structure is much less. For more details, consult Kim (1994). Supporting simulation results will be provided in Section 6.
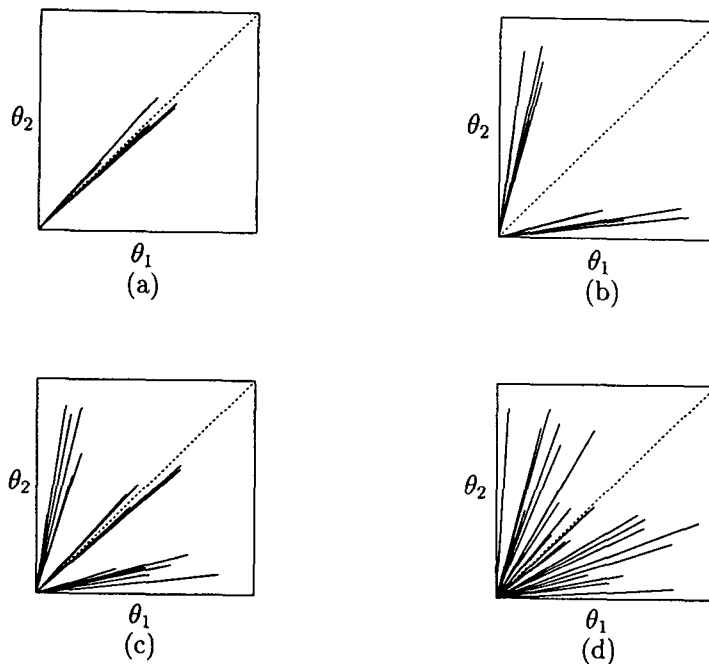

Figure 1

Four Hypothetical Test Structures in the Two Dimensional Plane

(the dotted line represents the test composite direction)

2.3 $DETECT_{max}$

Denote $DETECT_{max}$ to be the maximum $DETECT$ value calculated over all possible partitions of a test. According to the theoretical results in the next section, it becomes clear that the main objective is to find a partition that maximizes, or approximately maximizes, $DETECT$, because then one suspects that this partition, except for statistical error, correctly indicates the underlying multidimensional structure. The number of sizable clusters (that contain at least a certain number of items, say 4) in this partition that maximizes $DETECT$ is judged to be the number of dimensions present in the test, and the average direction of the cluster that an item is located in corresponds to the dominant dimension the item is (best) measuring. The minimal cluster size restriction helps prevent the identification of dimensions having only a minor influence as well as helping reduce the possibility of statistical noise being opportunistically yet incorrectly judged by $DETECT$ as contributing a (minor) dimension.

Since each estimated conditional covariance $\widehat{Cov}_{ij}$ contributes to a measure of the lack of unidimensionality resulting from violation of local independence (LI), the size of $DETECT_{max}$ can be viewed as an indicator that quantifies the *amount* of departure from unidimensionality. This amount of departure from unidimensionality is interpreted as the magnitude of departure from the *unidimensional composite direction* determined by a weighted average of all the underlying latent dimensions, these dimensions represented by item clusters in the approximate simple structure case. This composite direction can be thought of intuitively as the single dimension best measured by the test, somewhat like the psychologist's g on an intelligent test. $DETECT_{max}$ is expected to be close to zero for unidimensional data, while it reaches a substantially larger value for heavily multidimensional data.

Unfortunately, for a finite-length unidimensional test, there exists statistical bias in the index $DETECT$ due to the lack of reliability of the conditioning scores S_{ij}, as recognized by Rosenbaum (1984), Holland & Rosenbaum (1986), Douglas, Kim & Stout (1994) and Kim et al. (in press) among others. That is it can be

proposed under appropriate assumptions that

$$Cov(X_i, X_j | S_{ij}) > 0, \quad \text{for all } 1 \leq i < j \leq n.$$

Therefore,

$$E[Cov(X_i, X_j | S_{ij})] > 0, \quad \text{for all } 1 \leq i < j \leq n.$$

Notice that

$$E[Cov(X_i, X_j | S_{ij})] = \sum_{k=0}^{n-2} \text{Prob} \, (S_{ij} = k) \, Cov(X_i, X_j | S_{ij} = k). \qquad (3)$$

By comparing (1) and (3), we see that $\widehat{Cov}_{ij}$ is a reasonable estimate of $E[Cov(X_i, X_j | S_{ij})]$. Hence, the claimed statistical bias of the $\widehat{Cov}_{ij}$ and hence of $DETECT$ will occur. In order to correct this bias, the average $\overline{Cov}$ is subtracted from each $\widehat{Cov}_{ij}$ before it is combined into $DETECT$. This is why the $(\widehat{Cov}_{ij} - \overline{Cov})$ term is used in (2) rather than $\widehat{Cov}_{ij}$. Indeed, after this bias correction, as will be seen later, $DETECT_{max}$ remains small for unidimensional data as desired. Simulation studies show that this correction, designed for the correction of the positive bias in the unidimensional case, has no visible deleterious impact in the multidimensional case. That is, as desired, $DETECT_{max}$ remains large for strongly multidimensional data while staying near zero for the unidimensional data.

Table 1 below roughly categorizes a suggested quantitative interpretation of the amount of departure from unidimensionality, or the amount of multidimensionality, which is indicated by the maximum $DETECT$ value. It should be stressed that the "amount" of multidimensionality is distinct from the number of dimensions; a two-dimensional data set could display a large amount of multidimensionality if the two dimensions are each well measured and are weakly correlated while an eight-dimensional data set could display very weak multidimensionality if there is only one dominant dimension and/or the multiple dimensions are highly correlated. In Table 1, the $DETECT$ value has been multiplied by 100 for convenience.

Table 1

A Categorization of $DETECT_{max}$ as an Index
of Amount of Multidimensionality

$DETECT_{max}$	Multidimensionality
0.0 – 0.19	unidimensional
0.2 – 0.39	weak
0.4 – 0.79	moderate
0.8 –	strong

Theoretical justification of $DETECT$ is developed by Zhang (1996) and it supports well the use of $DETECT$. Also it is essential to search for the meaningful cluster formation which maximizes the $DETECT$, in fact requiring enormous computation. Recently the Genetic Algorithm is adapted by Zhang (1996). See Zhang (1996) and Zhang, Stout & Kim (1995) for details.

3. Real Data Analysis

Two Analytical Reasoning sections of an administration of the GRE have eight passages with 38 items. We chose four passages for which the data was complete and which seemed likely to be dimensionally distinct. These four passages have a total of 19 items with numbers of items/passage being 5, 4, 4, and 6. The number of examinees we used was 2477. $DETECT$ was maximized at four clusters with the maximum $DETECT$ value being 8.34×10^{-3}. Strikingly, the four clusters found by $DETECT$ corresponded **exactly** to the items associated with each of the four passages.

4. Simulated Data Analysis

The three parameter logistic (3PL) model is used in the generation of dichoto-

mously scored data, 1 for a correct answer and 0 for an incorrect answer. The probability of getting an item X right at ability θ is given by the equation

$$P(X = 1 \mid \theta) = c + \frac{1 - c}{1 + \exp\{-1.7a(\theta - b)\}}$$

where a, b, and c are parameters characterizing the item assumed to be independent each other. Parameter c is called the *guessing parameter* corresponding to the probability that a person completely lacking ability ($\theta = -\infty$) will answer the item correctly. Often it is called the lower asymptote. Parameter a represents the discriminating power of the item along the probability curve $P(X = 1 \mid \theta)$ at its inflexion point, while parameter b determines the position of the curve along the ability scale. It is called the item difficulty. See Lord (1980) for details.

A 40 item test is split into several dimensionally distinct clusters. Each cluster is *unidimensional* to form a simple structure test. That is, all the items within a cluster load on *one* ability trait and the unidimensional ability varies over separate clusters. From unidimensional (1D) to four dimensional (4D) cases are simulated. Table 2 gives the number of items in each dimension.

Table 2

Number of Items in the Test and
in Each Dimensionally Distinct Cluster

	Number of Items in the Test	Number of Items in Each Dimensionally Distinct Cluster
1D	40	40
2D	40	20/20
3D	40	13/13/14
4D	40	10/10/10/10

Item parameters are generated independently of items and of respective parameters within an item from the normal distribution with mean and variance specified by Table 3 below.

Table 3
Summary of Item Parameter Distribution

	Discrimination	Difficulty
Mean	1.0	0.0
Standard Deviation	0.5	0.8
Maximum	2.5	1.5
Minimum	0.35	-1.5

The correlation coefficient between ability traits is one of the important factors to determine the extent of multidimensionality. In this simulation study, six different values, 0.3, 0.5, 0.7, 0.8, 0.85, and 0.9, are employed as the correlation coefficients among ability traits generated from the multivariate normal distribution. In each simulation model all possible pairs of ability traits have identical correlation coefficients. 6000 response vectors are generated per model, and then the data are cross validated with 3000 examinee responses used for constructing item clusters and the other 3000 examinee responses used for calculating $DETECT$. Note that all the values of $DETECT$ presented in this paper are multiplied by 100 for ease of presentation.

The values of $DETECT$ are displayed in Table 4 for the unidimensional simulated data with the increasing number of clusters up to 5. As expected, all 5 $DETECT$ values remain fairly small.

Table 4
$DETECT$ in the Unidimensional Case

	Number of Clusters				
	1	2	3	4	5
1D	0.0000	0.0340	0.0421	0.0524	0.0342

Table 5 shows $DETECT_{max}$ values for the two, three, and four dimensional cases at the different correlation coefficients. Notice that $DETECT$ is maximized at the

correct dimensionally-based cluster partitions in all these cases. It is interesting to observe that the size of $DETECT_{max}$ is a function of correlation coefficient. For example, the smaller the correlation coefficient, the larger $DETECT_{max}$, implying larger amounts of lack of unidimensionality. In all cases when the traits are highly correlated, there exists less multidimensionality revealing smaller $DETECT_{max}$. Also it is noteworthy that the size of $DETECT_{max}$ in Table 5 roughly explains the strength of multidimensionality of the data.

$DETECT$'s behavior in the mixed structure tests are also investigated, even though providing very promising results, but we do not deal with those here. For details see Kim (1994).

<div align="center">

Table 5

$DETECT_{max}$ in the Two, Three, Four Dimensional Cases

</div>

	Correlation Coefficient					
	0.3	0.5	0.7	0.8	0.85	0.9
2D	2.8446	1.7669	0.9105	0.8155	0.5224	0.4401
3D	2.0013	1.5984	1.0471	0.6417	0.4472	0.3706
4D	1.5148	1.1825	0.7834	0.4424	0.3077	0.2748

5. Discussion

The estimated conditional covariance based index $DETECT$ for assessing the dimensionality structure of educational/psychological test data is defined and investigated extensively in order to discern its properties. Through analyses of simulated data, $DETECT$ has been shown to display effective performance in identifying the number of dimensions present in test data as well as in identifying the items contributing to each dimension in the case of approximate simple structure and both the mixed and approximate simple structure cases for two dimensional data. $DETECT$ has been shown to function effectively on identifying the paragraph-based items if a verbal test as producing separate dimensions. Also it quantifies the lack of unidimensionality of the data.

Recently, a theoretical justification for *DETECT* is made by defining its theoretical analogue, called theoretical *DETECT* (see, Zhang and Stout, 1995). We can see that under certain reasonable conditions, the theoretical *DETECT* will be maximized at the correct simple structure cluster partition of the test items with the number of clusters in this partition corresponding to the number of dimensions of the test, for example, the clusters corresponding to items associated with the distinct paragraphs of a reading comprehension test. The properties of this theoretical *DETECT* are under further investigation. More investigation on the asymptotic behavior regarding *DETECT* is also planned for a future study as well as additional simulations to study the performance of *DETECT* when the dimensionality is at least three and approximate simple structure does not hold.

REFERENCES

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*, 255-278.

Douglas, J., Kim, H. R., & Stout, W. F. (1994). Exploring and explaining the lack of local independence through conditional covariance functions. Paper presented at the 1994 annual meeting of the American Educational Research Association, New Orleans, LA.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*, 1523-1543.

Junker, B. W., & Stout, W. F. (1994). Robustness of ability estimation when multiple traits are present with one trait dominant. In D. Laveault, B. Zumbo, M. Gessaroli, & M. Boss (Eds.), *Modern Theories of Measurement: Problems and Issues (pp. 31-61)*. Ottawa, Canada, University of Ottawa Press.

Kim, H. R., et al. (in press). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement.* Invited Paper.

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data. Unpublished doctoral dissertation, Department of Statistics, University of Illinois at Urbana-Champaign.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems.* Lawrence Erlbuam Associates.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Rosenbaum, P. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.

Roussos, L. A. (1993). Hierarchical agglomerative clustering computer program users manual. Unpublished manuscript. Department of Statistics, University of Illinois at Urbana-Champaign.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 55*, 293-326.

Zhang, J. (1996). *Some Fundamental Issues in Item Response Theory with Applications.* Unpublished doctoral dissertation, Department of Statistics, University of Illinois at Urbana-Champaign.

Zhang, J., & Stout, W. F. (1995). Theoretical results concerning *DETECT.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1995.

Zhang, J., Stout, W. F., & Kim, H.R. (1995). A theoretical justification for the use of *DETECT* in assessing latent multidimensionality. To be submitted.

DEPARTMENT OF APPLIED STATISTICS, SANG JI UNIVERSITY, WONJU, KANG-WON - DO 220-702, KOREA.

E-mail: hrkim@soback.kornet.nm.kr