Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

My Early Interactions with Jan and Some of His Lost Papers

Yoshio Takane
Department of Psychology, University of Victoria

Abstract

It has been over 40 years since I got to know Jan. This period almost entirely overlaps my whole career as a psychometrician. During these years, I have had many contacts with him. This paper reviews some of my early interactions with him, focussing on the following topics: (1) An episode surrounding the inception of the ALSOS project, and (2) Jan's unpublished (and some lost) notes and papers that I cherished and quoted in my work, including (2a) the ELEGANT algorithm for squared distance scaling, (2b) the INDISCAL method for nonmetric multidimensional scaling (MDS), and (2c) notes on DEDICOM.

Keywords: The weighted additive model (WAM), alternating least squares (ALS) algorithms, optimal scaling (OP), The ELEGANT algorithm, The indicator method for nonmetric MDS (the INDISCAL method), DEDICOM (DEcomposition into DIrectional COMponents).

1. The inception of the ALSOS project

I met Jan for the first time in the spring of 1974. I was a first year graduate student (Ph.D. level) at the Thurstone Psychometric Labs at the University of North Carolina at Chapel Hill under the supervision of Forrest Young. Jan was at Bell Labs working with Joe Kruskal and Doug Carroll. He was on leave of absence from Department of Data Theory at the University of Leiden. One day Jan came down to Chapel Hill to give a talk in a colloquium organized by Forrest. I do not remember exactly what he talked about in his talk, but it must have been something about nonmetric multidimensional scaling (MDS), which was then one of the hottest topics in psychometrics. It has been more than 40 years since then, and I have had many interactions with Jan, the most memorable one being a joint project, later known as the "ALSOS" project (43; 45). Between 1976 and 1980, we (Young, de Leeuw, and Takane in some order) published seven papers, all having an "Alternating Least Squares method with

Optimal Scaling features" as their subtitles (8; 39; 40; 41; 44; 45; 46). Here is my personal account of how it started.

During the spring semester of 1974, I took a course in Developmental Psychology. I was not particularly interested in this course, but we were required to take three courses offered outside the Psychometric Labs as one of the requirements for a Ph.D degree. The course used a new textbook by Liebert, Poulos, and Strauss (24), which contained a report on Kempler's (16) study on (the failure of) conservation of quantity in young children. In his study, Kempler constructed a set of 100 rectangles by systematically varying their height and width over ten levels each (ranging from 10 inches to 14.5 inches in half an inch interval), and asked four groups of children of varying age (1-st, 3-rd, 5-th, and 7-th graders) to judge if the rectangles looked large or small. He then calculated the mean height and width of the rectangles judged large for the four groups of children separately. He found that the mean height of the rectangles judged large was disproportionately large for younger children, while just the opposite was true for the mean width. As the children got older, the mean height of the rectangles judged large decreased rather consistently, while the mean width increased, until the two became almost equal in the end. Kempler thought this was because younger children tended to put more emphasis on height than width when they made area judgments of rectangles, but as they got older, they became able to take into account both the height and width of rectangles more equitably.

When I saw this report, I immediately thought that this would be an ideal example for the weighted additive model (WAM). It was mid 1970's, when the weighted distance model like INDSCAL (1) was very popular among psychometricians to account for a certain type of individual differences in (dis)similarity judgments. The INDSCAL model postulates a common spatial representation of stimuli across all individuals, while also assuming differential weights attached to different dimensions by different individuals that give rise to individual differences. I was thinking something similar should hold in more fundamental additive models. I was very excited about my idea, and immediately went to talk about it with Forrest, who was then developing computer programs called POLYCON (polynomial conjoint scaling) and ADCON (additive conjoint scaling). Forrest was very happy about my idea of the weighted additive model, and we immediately started thinking about expanding his projects. We thought it would make more sense to begin with simple additive models (ADDALS; (8)) because the ADCON project was already in progress, and the WAM for all groups or individuals combined presupposed a simple additive model for each group or each individual. It was so decided that my first duty as a research assistant was to rewrite the ADCON program to make it more general in terms of the number of additive factors, accommodating possibly non-orthogonal factors, handling of missing data, etc. I recall POLYCON used the more conventional steepest descent method (a la Kruskal (21)) for optimization, while ADCON used an algorithm, which was later named an alternating least squares (ALS) algorithm. We kept using the latter algorithm in ADDALS as well as in all subsequent programs developed under the ALSOS project. At this stage, an ALS algorithm was used rather heuristically. We knew from our experience that it was monotonically convergent in the sense that it consistently minimized the least squares criterion, but we were unable to prove it theoretically. As a first year Ph.D. student in psychometrics, I had little knowledge in numerical optimization algorithms at the time.

It was under this sort of circumstances Jan arrived in Chapel Hill. We had a hidden agendum for his visit. We asked him about our algorithm used in ADCON, and Jan could

immediately tell us that it was monotonically convergent, and that it was straightforward to show it theoretically. He also suggested that the algorithm might be called alternating least squares (ALS), and that the extended ADCON program might be renamed ADDALS. We talked about other possible models that might be fitted similarly. It was clear by this time that our algorithmic scheme would work not only for simple and weighted additive models (8; 41), but also for other linear models (e.g.,regression analysis models (44)), bilinear models (e.g., principal component analysis models (46)), quadratic models (e.g., weighted and unweighted squared Euclidian models (39)), and common factor analysis models (40). The three of us (Young, de Leeuw, and Takane) agreed to set up a joint project later known as the ALSOS project (43; 45).

ALSOS has two major ingredients, optimal scaling (OS) and alternating least squares (ALS). Psychometrics has a long tradition of "scaling up" the data measured on lower scale levels into higher ones through models of the data (31), called OS. Its proliferation in the 1970's and 80's was, however, mainly inspired by the development of nonmetric multidimensional scaling (20; 21; 26), where observed (dis)similarity data measured on an ordinal scale were monotonically transformed to fit a distance model. This was done by minimizing a least squares (LS) criterion that measures the overall discrepancy between monotonically transformed data and distances between stimuli represented as points in a multidimensional space.

More specifically, let us denote the least squares (LS) criterion by $\phi(\mathbf{X}, \hat{\mathbf{d}})$, where $\mathbf{X}$ indicates the matrix of stimulus coordinates, and $\hat{\mathbf{d}}$ is the vector of monotonically transformed data. We minimize this criterion with respect to both $\hat{\mathbf{d}}$ and $\mathbf{X}$. Kruskal (21) approached this problem by first minimizing the criterion with respect to $\hat{\mathbf{d}}$ conditionally on $\mathbf{X}$, and then unconditionally with respect to $\mathbf{X}$. This process may formally be expressed as

$$\min_{X,\hat{d}} \phi(\mathbf{X}, \hat{\mathbf{d}}) = \min_{X} \min_{\hat{d}|X} \phi(\mathbf{X}, \hat{\mathbf{d}}) = \min_{X} \hat{\phi}(\mathbf{X}), \tag{1}$$

where $\hat{\phi}(\mathbf{X}) = \min_{\hat{d}|X} \phi(\mathbf{X}, \hat{\mathbf{d}})$ is the conditional minimum of $\phi(\mathbf{X}, \hat{\mathbf{d}})$ with respect to $\hat{\mathbf{d}}$ given $\mathbf{X}$, which can only be obtained algorithmically for a specific $\mathbf{X}$ given. The unconditional minimum of $\hat{\phi}(\mathbf{X})$ with respect to $\mathbf{X}$ can in turn be obtained by an iterative procedure such as the steepest descent method. This minimization strategy works well for the simple Euclidean distance model used in nonmetric MDS, where the number of parameters to be estimated is usually not exceedingly large. However, it may not work so well when other models are fitted. For example, the model in principal component analysis (PCA) tends to have a large number of parameters, since there are both component loading and score matrices to be estimated. This may well be the reason why Kruskal and Shepard (22) drew a rather negative conclusion about their nonmetric PCA procedure.

In alternating least squares (ALS) algorithm, on the other hand, a parameter updating procedure is split into two distinct phases:

$$\min_{X|\hat{d}} \phi(\mathbf{X}, \hat{\mathbf{d}})$$

and

$$\min_{\hat{d}|X} \phi(\mathbf{X}, \hat{\mathbf{d}}),$$

each of which is a relatively simple conditional minimization step. These two steps are iterated until no significant improvement in the value of ϕ occurs from one iteration to the next. Since

the model estimation phase (the first phase) can further be broken down into smaller sub-steps, a large number of parameters in the fitted model is no longer a big issue. It is monotonically convergent because each step obtains a conditional minimum of one subset of parameters while fixing all others. An additional benefit comes from the fact that a normalised LS criterion can often be minimized by minimizing an unnormalized LS criterion combined with an actual normalization of model parameters (4). Consequently, ALS quickly became one of the most popular optimization strategies in psychometrics. Long after the ALSOS project was over, I am still benefitting from the handiness of ALS algorithms. I have kept using the algorithms in various contexts, e.g., in DCDD (Different Constraints on Different Dimensions (38)), ERA (Extended Redundancy Analysis (35)), and GSCA (Generalized Structured Component Analysis (13; 14)) and its variants (15; 48). Interestingly, it has been shown that WAM is a special case of all of these models (DCDD, ERA, and GSCA). Under certain circumstances, it is also a special case of CPCA (Constrained Principal Component Analysis (32)).

There still is a lingering problem with the ALS algorithms (30) in my mind, however. The convergence of ALS algorithms is based on the notion that a monotonically decreasing bounded sequence must converge. This is a well known fact in mathematics. However, this condition alone does not guarantee that a convergence point is a local minimum of an objective function. It is possible that the convergence point is some accumulation point that is not a local minimum. This is because the amount of improvement in the value of the objective function can get smaller and smaller as the iteration proceeds, and the algorithm may never get to the desired point. I often use the following anecdote to illustrate my point: Suppose we would like to go from Seattle to Los Angels. Our plan is to go on each day a half way between the current location and San Francisco. This way we are getting closer to Los Angels every day, but we know that we can never get to Los Angels (or anywhere beyond San Francisco). We say that there is a point of discontinuity at San Francisco in our travel plan. Obviously, there should be no such points of discontinuity to be able to reach our desired destination. Essentially the same holds for numerical algorithms. Consequently, Zangwill (47) requires that there are no such points in our algorithm, but he does not show explicitly how we can verify it in practice. In numerical optimization literature (e.g., (9)), the notion of continuity in the algorithm (This should not be confused with the continuity of the objective function.) is replaced by a condition of a sufficient decrease in the value of the objective function in each step of the algorithm (such as Wolfe's condition). The sequential nature of the ALS algorithms, however, makes it difficult to verify this condition directly.

2. Unpublished (and some lost) papers

Toward the end of October 2014, I received an email from Jan asking me if I still had copies of a couple of his unpublished papers. One was de Leeuw (4) entitled "A normalized cone regression approach to alternating least squares algorithms" (this paper happened to be quoted above), and the other was de Leeuw (3) entitled "An alternating least squares algorithm for squared distance scaling." While I could immediately locate the former, I could not find the latter, which regrettably seemed to have gotten lost forever. Those were the days before personal computers, and even published papers were not likely to be on an electronic medium. Let alone unpublished ones. Somewhat fortunately, the content of (3) has been described in some detail in Takane's MDS book in Japanese (28). The algorithm used in this paper was called the ELEGANT algorithm for a reason to be noted below.

When I first met Jan in 1974, he struck me as already a very mature scientist. His mathematical expertise was awesome. I thought I had a lot to learn from him. After he went back to Bell Labs, he generously sent me several working papers of his, and I appreciated very much the opportunities to study his unpublished papers closely. When he was young, Jan was saying that he would not publish a paper unless it was accepted essentially as was in the first round of review. It was kind of rare, however, that papers were accepted without any revisions. So quite a few of his papers remained unpublished, some of which were subsequently lost permanently. In addition to (3), two more papers (notes) were subsequently identified as missing. They are the INDISCAL paper and notes on DEDICOM. The former deals with an indicator method for nonmetric MDS, and the latter concerns some important properties of the algorithm I developed earlier for DEDICOM. The former was quoted as de Leeuw, Takane, and Young (in preparation) in (27), but for some reason, it is not listed in the reference list of the paper. So the exact title of the paper is unknown. (There is a slight chance that it has never been written formally, but a description of the procedure has been given in Takane (28), which may be based on the brief description of the idea in (2).) The latter was referenced as de Leeuw (1983) (6) in Takane (1985) (29). In what follows, I will try my best to reproduce the contents of Jan's three lost papers as faithfully as possible as a token of my appreciation to him.

2.1. The ELEGANT algorithm for squared distance scaling

This is an algorithm for fitting squared Euclidean distances. Let $\mathbf{X}$ denote the n-stimuli by r-dimensions matrix of stimulus coordinates. Let o_{ij} represent the observed dissimilarity between stimuli i and j, and let d_{ij} represent the corresponding Euclidean distance calculated from $\mathbf{X}$. We assume that o_{ij} is symmetric, so that it is sufficient to consider only $j > i$. Let $\mathbf{D}_{o^2}$ denote the diagonal matrix of order $n^* = n(n-1)/2$ with o_{ij}^2's arranged in certain order as its diagonal elements, and let $\mathbf{D}_{d^2}$ denote the diagonal matrix of order n^* with d_{ij}^2 arranged in the same order as the diagonal elements of $\mathbf{D}_{o^2}$. Let $\tau(\mathbf{X})$ represent the least squares criterion defined on squared distances. This may be written as

$$\tau(\mathbf{X}) = \sum_{i,j>i} (o_{ij}^2 - d_{ij}^2)^2 = \mathrm{tr}(\mathbf{D}_{o^2} - \mathbf{D}_{d^2})^2. \tag{2}$$

Let $\mathbf{A}$ denote the n^* by n design matrix for pair comparisons. A row of $\mathbf{A}$ represents a stimulus pair and a column of $\mathbf{A}$ represents a stimulus. If the k-th diagonal element of $\mathbf{D}_{o^2}$ (and $\mathbf{D}_{d^2}$) represents the (squared) dissimilarity (resp. distance) between stimuli i and j, the k-th row of $\mathbf{A}$ has 1 in the i-th column and -1 in the j-th column with all other elements equal to zero. Then $\tau(\mathbf{X})$ above can be rewritten as

$$\tau(\mathbf{X}) = \mathrm{tr}(\mathbf{D}_{o^2} - \mathrm{diag}(\mathbf{A}\mathbf{X}\mathbf{X}'\mathbf{A}'))^2. \tag{3}$$

Define $\mathbf{Q}$ as

$$\mathbf{Q} = \mathbf{A}\mathbf{X}\mathbf{X}'\mathbf{A}' - \mathrm{diag}(\mathbf{A}\mathbf{X}\mathbf{X}'\mathbf{A}'), \tag{4}$$

and $\mathbf{Q}^*$ as

$$\mathbf{Q}^* = \mathbf{D}_{o^2} + \mathbf{Q}. \tag{5}$$

Then $\tau(\mathbf{X})$ can be further rewritten as:

$$\tau(\mathbf{X}) = \mathrm{tr}(\mathbf{Q}^* - \mathbf{A}\mathbf{X}\mathbf{X}'\mathbf{A}')^2. \tag{6}$$

By differentiating the above τ and setting the result equal to zero, we obtain

$$-\frac{1}{2}\frac{\partial\tau(\mathbf{X})}{\partial\mathbf{X}} = \mathbf{A}'(\mathbf{Q}^* - \mathbf{A}\mathbf{X}\mathbf{X}'\mathbf{A}')\mathbf{A}\mathbf{X} = \mathbf{O}, \tag{7}$$

assuming temporarily that $\mathbf{Q}^*$ is constant (i.e., not a function of $\mathbf{X}$). Noticing that

$$\mathbf{A}'\mathbf{A} = n\mathbf{J}_n = n(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'), \tag{8}$$

where $\mathbf{I}_n$ is the identity matrix of order n, and $\mathbf{1}_n$ is the n-component vector of ones, and that

$$\mathbf{J}_n\mathbf{X} = \mathbf{X} \tag{9}$$

(i.e., the origin of the space is set at the centroid of the stimulus configuration), we may further rewrite Equation (7) as

$$(\mathbf{A}'\mathbf{Q}^*\mathbf{A}/n^2)\mathbf{X} = \mathbf{X}\boldsymbol{\Delta}, \tag{10}$$

where it is assumed that

$$\boldsymbol{\Delta} = \mathbf{X}'\mathbf{X} \tag{11}$$

is a diagonal matrix. Let

$$\mathbf{B}(\mathbf{X}) = \mathbf{A}'\mathbf{Q}^*\mathbf{A}/n^2. \tag{12}$$

Then, for fixed $\mathbf{B}(\mathbf{X})$,

$$\mathbf{B}(\mathbf{X})\mathbf{X} = \mathbf{X}\boldsymbol{\Delta} \tag{13}$$

is an eigen-equation for $\mathbf{B}(\mathbf{X})$. Since

$$\tau(\mathbf{X}) = \text{tr}(\mathbf{Q}^*) - n^2\text{tr}(\boldsymbol{\Delta}^2), \tag{14}$$

minimizing τ for fixed $\mathbf{Q}^*$ (and so for fixed $\mathbf{B}(\mathbf{X})$) amounts to obtaining the eigenvectors of $\mathbf{B}(\mathbf{X})$ corresponding to the r largest eigenvalues. The matrix of (normalized) eigenvectors $\mathbf{X}$ is then re-scaled to satisfy Equation (11).

The above procedure has to be applied iteratively, since $\mathbf{B}(\mathbf{X})$ is a function of $\mathbf{X}$ and has to be updated for a new estimate of $\mathbf{X}$. This is repeated until no significant change occurs in the value of τ. This process is similar to the refactoring method in common factor analysis, where communalities are estimated by iterative refactoring. That is, the matrix of factor loadings is estimated with tentative estimates of communalities by solving an eigen-equation. This yields new estimates of communalities, which are then used for an improved estimate of the matrix of factor loadings, and so on. A modification of the above algorithm to ordinal data is straightforward using the ALS algorithmic framework. Note that $\mathbf{B}(\mathbf{X})$ can be further rewritten as

$$\mathbf{B}(\mathbf{X}) = \frac{1}{n^2}\mathbf{A}'(\mathbf{D}_{o^2} - \text{diag}(\mathbf{A}\mathbf{X}\mathbf{X}'\mathbf{A}')\mathbf{A} + \mathbf{X}\mathbf{X}', \tag{15}$$

Let $s_{ij} = (o_{ij}^2 - d_{ij}^2)/n^2$. Then, the ij-th element of the first term in $\mathbf{B}(\mathbf{X})$, denoted as b_{ij}^*, is obtained by

$$b_{ij}^*(\mathbf{X}) = \begin{cases} -s_{ij} & \text{if } i \neq j, \\ \sum_{i\neq j} s_{ij} & \text{if } i = j. \end{cases} \tag{16}$$

It is more convenient and efficient to use this expression for calculating $\mathbf{B}(\mathbf{X})$.

It turned out that the above algorithm was painfully slow, like the refactoring method in common factor analysis. So much so that I mentioned to Jan that although it looked elegant on surface, it was too slow to be useful in practice. My remark was based on the fact that the algorithm was more elegant than the method used in ALSCAL, which also fitted squared distances, but updated each coordinate of a stimulus point at a time with all other coordinates fixed (which amounted to solving a cubic equation in each step). The ELEGANT algorithm, on the other hand, could update the entire set of stimulus coordinates simultaneously. Jan responded to my remark and suggested, half jokingly, to call it ELEGANT. The slow convergence of the algorithm may be seen from the expression of $\mathbf{B}(\mathbf{X})$ in Equation (15), a major portion of which is $\mathbf{X}\mathbf{X}'$, so that the new update of $\mathbf{X}$ is greatly affected by the previous $\mathbf{X}$. The ELEGANT algorithm has never been used in practice because of its slow convergence. However, after having worked on so many acceleration techniques for iterative procedures (25; 36; 37), I now think that its slow convergence can be overcome to the extent that it is practically viable. Squared distance scaling is attractive because (a) the squared Euclidean distance function is a simple quadratic function of its parameters (stimulus coordinates), and (b) an extension to ordinal dissimilarity data is straightforward (Kruskal's LS monotonic transformation algorithm can be used on squared distances just as with unsquared distances).

Takane (27) compared some formal characteristics of the ELEGANT algorithm with classical MDS, the C-matrix method (5; 10), and the fourth type of quantification method (12), which was also used as an initialization method in the original C-matrix method. This paper also describes the ELEGANT algorithm in some detail, and it is written in English.

2.2. INDISCAL: An indicator method for nonmetric multidimensional scaling (MDS)

This is an algorithm for scalar product optimization in nonmetric MDS. Let $o_1, \cdots, o_K$ represent K distinct observed dissimilarities in ascending order (i.e., $o_k < o_{k'}$ for $k < k'$), where K indicates the total number of distinct values in observed dissimilarities. Define $\mathbf{B}_k = [b_{ijk}]$ as

$$b_{ijk} = \begin{cases} 1 & \text{if } o_{ij} \geq o_k, \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

where o_{ij} $(1 \leq i, j \leq n)$ denotes the observed dissimilarity between stimuli i and j measured on an ordinal scale. The matrix of optimally transformed squared dissimilarity data can then be expressed as

$$\hat{\mathbf{D}}^{(2)}(\boldsymbol{\theta}) \equiv \sum_{k=1}^{K} \theta_k \mathbf{B}_k, \tag{18}$$

where θ_k, the k-th element of $\boldsymbol{\theta}$, is assumed non-negative. The scalar product matrix derived from $\hat{\mathbf{D}}^{(2)}(\boldsymbol{\theta})$ can then be written as

$$\mathbf{A}(\boldsymbol{\theta}) \equiv -\frac{1}{2}\mathbf{J}\hat{\mathbf{D}}^{(2)}(\boldsymbol{\theta})\mathbf{J} = \sum_{k=1}^{K} \theta_k(-\frac{1}{2}\mathbf{J}\mathbf{B}_k\mathbf{J}) = \sum_{k=1}^{K} \theta_k \mathbf{C}_k, \tag{19}$$

where $\mathbf{J}$ is, as before, the centering matrix of order n (i.e., $\mathbf{J} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n/n$, where $\mathbf{I}_n$ is the identity matrix of order n and $\mathbf{1}_n$ is the n-element vector of ones), and $\mathbf{C}_k = (-1/2)\mathbf{J}\mathbf{B}_k\mathbf{J}$.

Now the least squares criterion based on the scalar product can be written as

$$\omega(\boldsymbol{\theta}, \mathbf{X}) = \text{tr}(\mathbf{A}(\boldsymbol{\theta}) - \mathbf{X}\mathbf{X}')^2. \tag{20}$$

This criterion may be minimized by ALS, i.e., by alternately minimizing it with respect to $\mathbf{X}$ for fixed $\boldsymbol{\theta}$, and with respect to $\boldsymbol{\theta}$ for fixed $\mathbf{X}$ subject to the non-negativity restrictions on θ_k. The former is particularly simple because it reduces to the eigen-decomposition of the matrix $\mathbf{A}(\boldsymbol{\theta})$. The latter is a nonnegative least squares problem (e.g., (23)), which is a kind of quadratic programming problem involving a slightly more complicated procedure than Kruskal's least squares monotonic regression algorithm, but which can be solved, for example, by an active constraint method in a finite number of steps. A conditional minimization scheme is also possible in which the criterion function is minimized first with respect to $\mathbf{X}$ conditionally upon $\boldsymbol{\theta}$, which is then unconditionally minimized with respect to $\boldsymbol{\theta}$ (2). However, the ALS minimization scheme is simple enough, and is expected to work efficiently.

The basic motivation for fitting squared Euclidean distances in the previous section was because monotonicity of unsquared distances and that of squared distances are completely equivalent, and Kruskal's (20; 21) least squares monotonic transformation algorithm for the former can be applied to the latter without any modification. Minimizing the least squares criterion defined in terms of scalar products (the strain optimization), on the other hand, is attractive because finding the stimulus configuration is straightforward (can be done in closed form), once observed or derived scalar products are given. Optimal monotonic transformations, on the other hand, are more complicated with scalar products. Perhaps for this reason, no fully nonmetric MDS procedure had been developed until late 1990's (42) based on the scalar product optimization. However, it is remarkable that a long time before 1998, Jan developed such a procedure. It is called INDISCAL, an indicator method of nonmetric MDS. Jan called it INDISCAL, again half jokingly, as a "parodic" name for INDSCAL. As alluded to earlier, the method was briefly mentioned in Section 3.3 of (2), but might have never been written up as such. It was cited in (27) without any substantive description, but is described more fully in (28). Since (28) is written in Japanese, INDISCAL has never been widely known in the psychometric community in the world.

2.3. Notes on DEDICOM

DEDICOM (DEcomposition into DIrectional COMponents) is a model for square asymmetric tables, proposed by Harshman (11). The model can be written as

$$\mathbf{A}_0 = \mathbf{X}\mathbf{B}\mathbf{X}', \tag{21}$$

where $\mathbf{A}_0$ is an n by n model matrix describing asymmetric relationships among n stimuli, $\mathbf{X}$ is an n by r-component ($r \ll n$) loading matrix, and $\mathbf{B}$ is an r by r matrix describing asymmetric relationships among the r components. The matrix $\mathbf{X}$ captures the relationships between the r components and the n stimuli. This model attempts to explain asymmetric relationships among n stimuli ($\mathbf{A}_0$) by postulating a smaller number of components which assume asymmetric relationships among themselves ($\mathbf{B}$), and the relationships between the components and the stimuli ($\mathbf{X}$). The model given above is not unique. To remove the model indeterminacy, we impose the following restrictions:

$$\mathbf{X}'\mathbf{X} = \mathbf{I}_r, \tag{22}$$

and

$$\mathbf{B}'\mathbf{B} + \mathbf{B}\mathbf{B}' = \mathbf{\Delta} \quad \text{(diagonal)}. \tag{23}$$

In practice, it is rare to find an asymmetric table that can be exactly represented by Model (21) with prescribed rank r, due to error perturbations. It is more common that the above $\mathbf{A}_0$ is only approximately true, and it is necessary to find the model matrix $\mathbf{A}_0$ that best fits to the observed data matrix $\mathbf{A}$. That is, we seek to find $\mathbf{X}$ and $\mathbf{B}$ of prescribed rank r that minimizes

$$\psi(\mathbf{X}, \mathbf{B}) = \text{tr}(\mathbf{A} - \mathbf{X}\mathbf{B}\mathbf{X}')'(\mathbf{A} - \mathbf{X}\mathbf{B}\mathbf{X}'). \tag{24}$$

Takane (29) developed an iterative algorithm to find an $\mathbf{X}$ and $\mathbf{B}$ that jointly minimize this criterion. The $\mathbf{B}$ that minimizes ψ for given $\mathbf{X}$ can be easily obtained by

$$\mathbf{B} = \mathbf{X}'\mathbf{A}\mathbf{X}. \tag{25}$$

(This corresponds to the conditional minimum of ψ with respect to $\mathbf{B}$ given $\mathbf{X}$.) The minimum of ψ with respect to $\mathbf{X}$ is then obtained by iterating the following steps until convergence:

Step I. Calculate $\mathbf{X}^* = \mathbf{A}'\mathbf{X}\mathbf{X}'\mathbf{A}\mathbf{X} + \mathbf{A}\mathbf{X}\mathbf{X}'\mathbf{A}'\mathbf{X} = \mathbf{A}'\mathbf{X}\mathbf{B} + \mathbf{A}\mathbf{X}\mathbf{B}'$.

Step II. Apply the Schmidt orthogonalization method to $\mathbf{X}^*$ to obtain $\mathbf{X}$, and go back to Step I.

The Schmidt orthogonalization process can be formally written as

$$\mathbf{X}^* = \mathbf{X}\mathbf{R}', \tag{26}$$

where $\mathbf{X}$ is a columnwise orthogonal matrix, and $\mathbf{R}'$ is an upper triangular matrix. At convergence, $\mathbf{R}'$ becomes a diagonal matrix. This may be seen by noting that at convergence $\mathbf{X}'\mathbf{X}^* = \mathbf{B}'\mathbf{B} + \mathbf{B}\mathbf{B}' = \mathbf{R}'$. Since $\mathbf{B}'\mathbf{B} + \mathbf{B}\mathbf{B}'$ is symmetric, and $\mathbf{R}'$ is upper triangular, it must be diagonal.

I observed that the above algorithm was always monotonically convergent, but could not show theoretically that was indeed the case. After a few months of struggles, I decided to ask several people for help, including Jan and Henk (Kiers). Several months later, I received a response from Jan (6). Unfortunately, this note seems to have been lost permanently. However, as far as I can remember, it contained two important messages:

i) Let the rank of the matrix $\mathbf{A}\mathbf{A}' + \mathbf{A}'\mathbf{A}$ be called the Harshman rank of a square asymmetric matrix $\mathbf{A}$, and when it holds (i.e., $\text{rank}(\mathbf{A}\mathbf{A}' + \mathbf{A}'\mathbf{A}) = r$), there is an exact solution to Model (21), where $\mathbf{A}_0 = \mathbf{A}$. Let

$$\mathbf{A}\mathbf{A}' + \mathbf{A}'\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}' \tag{27}$$

represent the compact singular value decomposition (SVD) or the eigen-decomposition of the symmetric matrix $\mathbf{A}\mathbf{A}' + \mathbf{A}'\mathbf{A}$, where $\mathbf{P}$ is the n by r matrix of columnwise orthogonal matrix of singular vectors (or eigenvectors), and $\mathbf{D}$ is the r by r positive definite diagonal matrix of singular values (or eigenvalues). Then $\mathbf{X} = \mathbf{P}$ and $\mathbf{B} = \mathbf{P}'\mathbf{A}\mathbf{P}$ will exactly solve Equation (21) with the restrictions given in Equations (22) and (23). Furthermore, if the diagonal elements of $\mathbf{D}$ are all distinct, $\mathbf{X}$ is unique up to reflections and permutations of its columns. Also, note that Model (21) with the restrictions (22) and (23) implies that

$$\mathbf{A}'\mathbf{A} + \mathbf{A}\mathbf{A}' = \mathbf{X}(\mathbf{B}'\mathbf{B} + \mathbf{B}\mathbf{B}')\mathbf{X}' = \mathbf{X}\mathbf{\Delta}\mathbf{X}', \tag{28}$$

so that $\mathbf{X} = \mathbf{P}$, and $\mathbf{\Delta} = \mathbf{D}$. (Although the exact infallible case as above rarely occurs in practice, this result is useful to obtain a good initial estimate of $\mathbf{X}$ for iterative solutions.)

ii) A sufficient condition for Takane's algorithm to be monotonically convergent is that the matrix $(\mathbf{A} \otimes \mathbf{B}) + (\mathbf{A} \otimes \mathbf{B})'$ is positive definite. This implies that Takane's algorithm is generally not monotonically convergent. Note that this condition depends on iterations, since $\mathbf{B}$ is a function of the current update of $\mathbf{X}$.

A few weeks later, I received another letter, this time from Henk (Kiers), in response to my inquiry, reporting that he solved my problem completely. He suggested that we wrote a joint paper, which culminated in (19). His letter indicated, consistent with Jan's result, that my algorithm was not always monotonically convergent. His letter also offered a prescription necessary to make my algorithm always monotonically convergent. In essence, it was suggested to modify Step I of Takane's algorithms as follows:

Step I'. Calculate $\mathbf{X}^* = \mathbf{AXX'A'X} + \mathbf{A'XX'AX} + 2\alpha\mathbf{X}$, where α is a scalar not smaller than the largest eigenvalue of the symmetric part of $-(\mathbf{A} \otimes \mathbf{B})$.

The above modification makes Takane's algorithm consistent with majorization algorithms (5) (See also (7; 17; 18).), another great invention and contribution of Jan to psychometrics/statistics. Note that α introduced above also depends on iterations through $\mathbf{B}$ (just as in the sufficient condition for monotonic convergence of Takane's original algorithm). A choice of alpha that does not depend on iterations is given by the largest eigenvalue the symmetric part of $-(\mathbf{A} \otimes \mathbf{A})$.

The above modification guarantees monotonic convergence of Takane's algorithm, while it may slow down the convergence (when α is positive) of the algorithm. Takane and Zhang (36) went the opposite direction. They incorporated the minimal polynomial extrapolation method to speed up the convergence at the sacrifice of monotonic convergence.

3. Concluding remarks

I have had many other memorable contacts with Jan, although I must close this essay soon. I published two more papers with him (33; 38) in late 1980's and early 1990's. By then, Jan became one of my valuable resource persons, that is, those I solicited some help whenever I encountered some problems that I could not solve myself. (One instance of this was already mentioned above in the case of Takane's algorithm for DEDICOM.) In these papers, I had to get some advice from Jan and Henk (Kiers) about existing literature on the topics, and about the convergence and uniqueness properties of the algorithms I came up with. My joint papers with Jan were predominantly first-authored by me, but this merely reflected the fact that I sometimes needed his help, but not vice versa.

There was one problem I raised to Jan, to which I never received his response. He has probably forgotten it by now, but this seems to be a great opportunity to remind him about it. I am still hopeful to get some hint from him. Let me describe the problem briefly. My former graduate student, Heungsun Hwang, and I were developing an MDS model for Burt tables (34) around year 2000. Let f_{ij} denote the observed joint frequency of the i-th category of item I and the j-th category of item J, and let p_{ij} represent the corresponding probability,

for which we postulated the following model:

$$p_{ij} = \frac{w_i w_j \exp(-d_{ij}^2)}{\sum_k \sum_\ell w_k w_\ell \exp(-d_{k\ell}^2)}, \tag{29}$$

where w_i (≥ 0 and $\sum_{i\subset I} w_i = 1$) is the bias parameter for category i, d_{ij} is the Euclidian distance between categories i and j represented as points in a space of dimensionality r, and the summations in the denominator extend over all categories in the two items. In line with the tradition of MDS, the squared Euclidean distance is parameterized by $d_{ij}^2 = \sum_a^r (x_{ia} - x_{ja})^2$, where x_{ia} is the coordinate of stimulus i on dimension a. Being consistent with the implicit constraints in multiple correspondence analysis, we required that $\sum_{i\subset I} f_i x_{ia} = 0$ for all items and dimensions ($1 \leq a \leq r$), where f_i is the marginal frequency of category i. We derived estimates of covariances among observed f_{ij}'s in off-diagonal blocks of Burt tables, and constructed a generalized LS criterion using the inverse of the estimated covariance matrix as weights, to estimate the bias parameters and the coordinates of the category points. We thought the model was quite plausible, and the estimation procedure quite attractive, yielding efficient estimates. We, however, experienced some anomaly with our Gauss-Newton algorithm. We frequently observed that solutions degenerated with the category points for one item drifting away from the centroid of the category configuration. Although we never observed this phenomenon in the case of only two items (a single two-way contingency table), the frequency of degenerate solutions seemed to pick up as the number of items increased. How could this happen? I am still puzzled by this phenomenon.

References

[1] Carroll, J. D., and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, **35**, 282-319.

[2] de Leeuw, J. (1974). Finding a positive semidefinite matrix of prescribed rank r in a nonlinear differetiable manifold. Technical Report, Bell Labs, Murry Hill, NJ.

[3] de Leeuw, J. (1975). An alternating least squares approach to squared distance scaling. Unpublished manuscript. Department of Data Theory, The University of Leiden, Leiden, The Netherlands.

[4] de Leeuw, J. (1977a). A normalized cone regression approach to alternting least squares algorithms. Unpublished manuscript. Department of Data Theory, The University of Leiden, Leiden, The Netherlands.

[5] de Leeuw, J. (1977b). Application of convex analysis to multidimensional scsling. In J. R. Barra, F. Brodeau, G. Romier, and B. van Cutem (Eds.), *Recent Developments in Statistics*, (pp. 133-145). Amsterdum: North Holland Publishing Co.

[6] de Leeuw, J. (1983). Notes on DEDICOM. Department of Data Theory, The University of Leiden, Leiden, The Netherlands.

[7] de Leeuw. J. (1988). Convergence of the majorization method for multidimensional scaling. *Journal of Classification*, **5**, 163-180.

[8] de Leeuw, J., Young, F. W., and Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, **41**, 471-504.

[9] Fletcher, R. (1987). *Practical Methods of Optimization*. New York: Wiley.

[10] Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, **33**, 469-506.

[11] Harshman, R. A. (1978). Models for analysis of asymmetrical relationships among N objects or stmuli. Paper presented at the First Joint Meeting of the Psychometric Society and the Society of Mathematical Psychology, Hamilton, Ontario.

[12] Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, **3**, 69-98.

[13] Hwang, H., and Takane, Y. (2004). Generalized structured component analysis. *Psychometrika*, **69**, 81-99.

[14] Hwang, H., and Takane, Y. (2014). *Generalized structured component analysis: A component-based approach to structural equation modeling*. Boca Raton: Chapman and Hall/CRC Press.

[15] Jung, K., Takane, Y., Hwang, H., and Woodward, T. S. (2012). Dynamic GSCA (Generalized Structured Component Analysis) with applications to the analysis of effective connectivity in functional neuroimaging data. *Psychometrika*, **77**, 827-848.

[16] Kempler, B. (1971). Stimulus correlates of area judgments: A psychophysical developmental study. *Developmental Psychology*, **4**, 158-163.

[17] Kiers, H. A. L. (1990). Majorization as a tool for optimizing a class of matrix functions. *Psychometrika*, **55**, 417-428.

[18] Kiers, H. A. L., and ten Berge, J. M. F. (1992). Minimization of a class of matrix trace functions by means of refined majorization. *Psychometrika*, **57**, 371-382.

[19] Kiers, H. A. L., ten Berge, J. M. F., Takane, Y., and de Leeuw, J. (1990). A generalization of Takane's algorithm for DEDICOM. *Psychometrika*, **55**, 151-158.

[20] Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1-29.

[21] Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, **29**, 115-129.

[22] Kruskal, J. B., and Shepard, R. N. (1974). A nonmetric variety of linear factor analysis. *Psychometrika*, **39**, 123-157.

[23] Lawson, C. L., and Hanson, R. J. (1974). *Solving least squares problems*. Englewood Cliffs, NJ: Prentice Hall.

[24] Liebert, R. N., Poulos, R. W., and Strauss, G. (1974). *Developmental Psychology.* Englewood Cliffs, NJ: Prentice Hall.

[25] Loisel, S., and Takane, Y. (2011). Generalized GIPSCAL Re-revisited: A fast convergent algorithm with acceleration by the minimum polynomial extrapolation. *Advances in Data Analysis and Classification*, **5**, 57-75.

[26] Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function, I and II. *Psychometrika*, **27**, 125-140 and 219-246.

[27] Takane, Y. (1977). On the relations among four methods of multidimensional scaling. *Behaviormetrika*, **9**, 29-43.

[28] Takane, Y. (1980). *Multidimensional Scaling.* Tokyo: University of Tokyo Press, (in Japanese).

[29] Takane, Y. (1985). Diagonal estimation in DEDICOM. In the *Proceedings of the Behviormetric Society Meeting* (pp. 100-101), Hokkaido University, Sapporo, Japan.

[30] Takane, Y. (1992). "Gifi, A. (1990). *Nonlinear Multivariate Analysis.* Chichester: Wiley." reviewed for the *Journal of the American Statistical Association*, **87**, 587-588.

[31] Takane, Y. (2005). Optimal scaling. In Everitt, B., and Howell, D. (Eds.), *Encyclopedia of Statistics for Behavioral Sciences*, (pp. 1479-1482). Chichester: Wiley.

[32] Takane Y. (2013). *Constrained principal component analysis and related techniques.* Boca Raton: Chapman-Hall/CRC Press.

[33] Takane, Y., and de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, **52**, 393-408.

[34] Takane, Y., and Hwang, H. (2000). A multidimensional scaling model for analysis of multiple-choice categorical data. Unpublished manuscript, Department of Psychology, McGill University, Montreal.

[35] Takane, Y., and Hwang, H. (2005). An extended redundancy analysis and its applications to two practical examples. *Computational Statistics and Data Analysis*, **49**, 785-808.

[36] Takane, Y., and Zhang, Z. (2009). Algorithms for DEDICOM: Acceleration, deceleration, or neither? *Journal of Chemometrics*, **23**, 364-370.

[37] Takane, Y., Jung, K., and Hwang, H. (2010). An acceleration technique for ten Berge et al.'s algorithm for orthogonal INDSCAL. *Computational Statistics*, **25**, 409-428.

[38] Takane, Y., Kiers, H. A. L., and de Leeuw, J. (1995). Component analysis with different constraints on different dimensions. *Psychometrika*, **60**, 259-280.

[39] Takane, Y., Young, F. W., and de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, **42**, 8-67.

[40] Takane, Y., Young, F. W., and de Leeuw, J. (1979). Nonmetric common factor analysis: An alternating least squares method with optimal scaling features. *Behaviormetrika*, **6**, 45-56.

[41] Takane, Y., Young, F. W., and de Leeuw, J. (1980). An individual differences additive model: An alternating least squares method with optimal scaling features. *Psychometrika*, **45**, 183-209.

[42] Trosset, M. W. (1998). A new formulation of the nonmetric strain problem in multidimensional scaling. *Journal of Classifiction* **15**, 15-35.

[43] Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika*, **46**, 347-388.

[44] Young, F. W., de Leeuw, J., and Takane, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, **41**, 505-529.

[45] Young, F.W., de Leeuw, J., and Takane, Y. (1980). Quantifying qualitative data. In Lantermann, E. D., and Feger, H. (Eds.), *Similarity and Choice*, (pp. 150-179). Bern: Hans Huber.

[46] Young, F. W., Takane, Y., and de Leeuw, J. (1978). Principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features.*Psychometrika*, **43**, 279-281.

[47] Zangwill, W. I. (1969). *Nonlinear Programming: A Unified Approach*. Englewood Cliffs, NJ: Prentice Hall.

[48] Zhou, L., Takane, Y., and Hwang, H. (in press). Dynamic GCANO (Generalized Structured Canonical Correlation Analysis) with applications to the analysis of effective connectivity in functional neuroimaging data. *Computational Statistics and Data Analysis*.

Affiliation:

Yoshio Takane
Department of Psychology
University of Victoria
5173 Del Monte Avenue, Victoria, BC, Canada
E-mail: yoshio.takane@mcgill.ca
URL: http://takane.brinkster.net/Yoshio/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd