On The Recovery Of The Consecutive Ones Property By Generalized Reciprocal Averaging Algorithms

Willem J. Heiser and Matthijs J. Warrens
Psychometrics and Research Methodology Group, Leiden University Institute for Psychological Research, Leiden University, Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands

Abstract

In this note we present a general proof of a phenomenon demonstrated in Heiser (1981): if the columns of a (0,1)-table can be permuted such that the 1s in each row form a consecutive interval, then the correct order of the columns can be found using the reciprocal averaging algorithm. Several variants of the reciprocal averaging algorithm for which the same property can be proved are considered.

1. Introduction

Let $\mathbf{A} = \{a_{ij}\}$ be a (0,1)-table of order $n \times m$. Matrix $\mathbf{A}$ has the consecutive 1s property if it is possible to order the columns so that, in every row, the 1s form a consecutive interval (are bunched together). If $\mathbf{A}$ is a re-ordered subject by attribute matrix with consecutive 1s in each row, all subjects have single-peaked preference functions, that is, they always check contiguous stimuli. If all runs of ones have the same length, the table has a parallelogram structure as defined by Coombs (1964, chapter 4). A (0,1)-table with consecutive 1s may also be interpreted as an intuitively meaningful and simple archaeological model: an artifact comes into use at a certain point in time, it remains in use for a certain period, and after some time it goes out of use.

Table 1. A (0,1)-table of order 15×5 with consecutive 1s.

					Table cont.				
1	0	0	0	0					
1	1	0	0	0	0	0	1	0	0
0	1	0	0	0	0	1	1	1	1
1	1	1	0	0	0	0	1	1	0
0	1	1	0	0	0	0	1	1	1
1	1	1	1	0	0	0	0	1	0
0	1	1	1	0	0	0	0	1	1
1	1	1	1	1	0	0	0	0	1

An example of a (0,1)-table with consecutive 1s is presented in Table 1. The same table can be found in Heiser (1981, p. 73). Table 1 is used in Heiser to demonstrate in great detail (pp. 72–79) that the order of the columns of Table 1, a table with consecutive 1s in the rows, is reflected in the final column scores obtained with the reciprocal averaging algorithm. In the next section we present a general proof of this property. It should be noted that the essential ingredients of the proof are already in the example presented in Heiser (1981). We also consider several variants of the reciprocal averaging algorithm for which the same property may be proved.

2. Reciprocal averaging

Let $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_m)$ be a n-tuple and m-tuple of real numbers. Let x and y be two sets of scores for the rows and columns of $\mathbf{A}$ respectively. Furthermore, let

$$a_{i+} = \sum_{j=1}^{m} a_{ij} \quad \text{and} \quad a_{+j} = \sum_{i=1}^{n} a_{ij}$$

denote, respectively, the row and column totals of table $\mathbf{A}$. We define the reciprocal averaging algorithm first considered in Horst (1935):

1. assign arbitrary values to y
2. calculate arithmetic mean $x_i = a_{i+}^{-1} \sum_j a_{ij} y_j$
3. calculate arithmetic mean $y_j^* = a_{+j}^{-1} \sum_i a_{ij} x_i$
4. replace the old y by new values y^*.

In this algorithm, the column scores y are the averages of the row scores x and, reciprocally, the row scores are the averages of the column scores. With some form of normalization the reciprocal averaging algorithm will almost always converge to a unique solution.

Correspondence analysis is an exploratory technique for analyzing the interaction of the rows and columns of a table with nonnegative entries (cf. Greenacre, 1984). The technique finds multi-dimensional scores for the rows and columns that redefine the dimensions of the space so that the principal dimensions capture a maximal amount of inertia (variance). One way of obtaining the scores of the first correspondence analysis dimension is using the reciprocal averaging algorithm defined above.

Next, we present a general proof of the phenomenon first demonstrated in Heiser (1981).

Theorem. *If (0,1)-table* $\mathbf{A}$ *has the consecutive 1s property, then the correct order of the columns is reflected in the final scores* y *of the reciprocal averaging algorithm.*

Proof: Because the reciprocal averaging algorithm converges to a unique solution, it is sufficient to show that the basic combined iteration step for obtaining y^* does not change the order of the columns if $\mathbf{A}$ has consecutive 1s in the rows and the elements of y are correctly ordered. We consider two adjacent columns of $\mathbf{A}$, j and $j+1$. Because $\mathbf{A}$ is (0,1)-table, only elements $a_{ij} = 1$ and $a_{ij+1} = 1$ for some i are relevant for obtaining values y_j^* and y_{j+1}^*. We therefore have the three possibilities presented in Table 2, where # denotes the number of row profiles.

Suppose that there are respectively p, q and r row profiles were the three possibilities in Table 2 occur. Let $u = (u_1, u_2, ..., u_p)$, $v = (v_1, v_2, ..., v_q)$ and $w = (w_1, w_2, ..., w_r)$ be respectively a p-tuple, q-tuple and r-tuple containing the arithmetic means of the $(p + q + r)$ rows were the three possibilities in Table 2 occur. Let $f(u, v)$ denote the arithmetic mean of the elements of u and v. If $y_j \leq y_{j+1}$ then all elements of u are equal or smaller than the elements in w. Therefore $f(u, v) \leq f(v, w)$ and $y_j^* \leq y_{j+1}^*$.

The proof of uniqueness for the column scores is more tedious. It is a matter of verifying that two columns can only have the same score if (i) they are identical or (ii) there are degeneracies, that is, groups of rows and columns are totally unconnected. With respect to (i), note that if $y_j < y_{j+1}$, then $y_j^* < y_{j+1}^*$ if either $p > 0$ or $r > 0$. □

Table 2. Three types of row profiles for two columns j and $j+1$ when the rows contain consecutive 1s.

a_{ij}	a_{ij+1}	#
1	0	p
1	1	q
0	1	r

Heiser (1981, p. 77) already discovered that the order of columns j and $j+1$ does not depend on the elements they have in common, but on the elements that are specific for j and $j+1$. The essential ingredient in the general proof is that all elements of u, which are specific to j, are smaller than the elements of w, which are specific to $j+1$.

3. Alternative algorithms

Several variants of the reciprocal averaging algorithm have been proposed in the literature. Some of these methods were proposed as techniques that are more robust to outliers compared to ordinary reciprocal averaging. Nishisato (1984) and Heiser (1987) considered the method of reciprocal (generalized) medians. Nishisato (1984) also discussed the method of trimmed reciprocal averages, whereas Sachs (1994) studied the method of reciprocal biweighted means using Tukey's biweight. The consecutive 1s property derived in the previous section can also be derived for other algorithms, that is, algorithms that do not use the weighted arithmetic mean, but some other measure of central tendency.

Median. The measure of central tendency known as the median is the number separating the higher half of a list of numbers from the lower half. The median can be found by arranging the numbers from lowest value to highest value and picking the middle one. If there is an even number of observations, the median is not unique. The median then equals the arithmetic mean of the two middle values.

Let us consider the proof of the theorem for the median. Let $g()$ denote the median and let u, v and w contain the medians of the $(p+q+r)$ rows were the three possibilities in Table 2 occur. If $y_j \leq y_{j+1}$ then all elements of u are equal or smaller than the elements in w. Therefore $g(u,v) \leq g(v,w)$ and $y_j^* \leq y_{j+1}^*$. Thus, the method of reciprocal medians recovers the correct order in the weak sense. However, the median cannot ensure uniqueness of the scores. After one combined iteration step we may only show that if $y_j < y_{j+1}$, then $y_j^* \leq y_{j+1}^*$. This clustering phenomenon is already reported in Heiser (1987), and is discussed in more detail in Michailidis and De Leeuw (2004). □

Trimmed mean. The trimmed mean involves the calculation of the arithmetic mean after discarding a specific percentage of the ordered numbers at the high and low end. It is usual to discard an equal amount at both ends. With respect to the theorem, the proof for the trimmed mean is similar to that of the ordinary arithmetic mean. However, if increasingly more numbers at the ends are trimmed, it becomes less likely that $y_j^* < y_{j+1}^*$ if $y_j < y_{j+1}$, and we may run in the same difficulties with respect to uniqueness of the column scores, i.e. a clustered solution, as encountered with the median. □

4. Discussion

The consecutive ones property is important in several fields, as diverse as DNA sequencing, archaeological seriation, and psychological choice analysis. There is an extensive literature on its graph-theoretical characterization and algorithms to identify it (cf. McConnell, 2004; Hsu, 2002). If $\mathbf{A}$ has consecutive 1s in each row, then the matrix $\mathbf{A}'\mathbf{A}$ has the Robinson property (Wilkinson, 1971, p. 279). Therefore, it is of both theoretical and practical interest to see that not only arithmetic averaging, but also a more general notion of averaging is sufficient to identify these properties with a straight-forward algorithm.

Other generalized reciprocal averaging algorithms could be considered. For instance, the arithmetic mean together with the geometric and harmonic means, are sometimes known as the Pythagorean means. Let $y = (y_1, y_2, ..., y_m)$ be a m-tuple of positive real numbers. The weighted geometric and harmonic means are given by, respectively

$$x_i = \exp\left(\frac{\sum_{j=1}^{m} a_{ij} \ln y_j}{\sum_{j=1}^{m} a_{ij}}\right) \quad \text{and} \quad x_i = \frac{\sum_{j=1}^{m} a_{ij}}{\sum_{j=1}^{m} y_j^{-1} a_{ij}}.$$

The geometric and harmonic means are attractive alternatives to the arithmetic mean, because they both tend to mitigate the impact of large outliers. However, the geometric and harmonic means may only be used in a reciprocal averaging algorithm if we adopt a normalization procedure that ensures that the scores in tuples x and y are positive real numbers. At present it is unknown if such an algorithm converges to a unique solution.

References

Coombs, C. H. (1964). *A Theory of Data.* New York: Wiley.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis.* New York: Academic Press.

Heiser, W. J. (1981). *Unfolding Analysis of Proximity Data.* Unpublished Ph.D. thesis, Leiden University.

Heiser, W. J. (1987). Correspondence analysis with least absolute residuals. *Computational Statistics & Data Analysis, 5*, 337–356.

Horst, P. (1935). Measuring complex attitudes. *Journal of Social Psychology, 6*, 369–374.

Hsu, W.-L. (2002). A simple test for the consecutive ones property. *Journal of Algorithms, 43*, 1–16.

McConnell, R. M. (2004). A certifying algorithm for the consecutive ones property. *Proceedings of the fifteenth annual ACM-SIAM symposium on discrete algorithms.* Philadelphia, P.A.: SIAM, pp. 768–777.

Michailidis, G. & De Leeuw, J. (1987). Homogeneity analysis using least absolute deviations. *Computational Statistics & Data Analysis, 48*, 587–603.

Nishisato, S. (1984). Dual scaling of reciprocal medians. *Proceedings of the 32nd Scientific Conference of the Italian Statistical Society* (Estratto Dagli Atti della XXXII Riunione Scientifica), Sorrento, Italy. Rome, Italy: Societa Italiana di Statistica, pp. 141–147.

Sachs, J. (1994). Robust dual scaling with Tukey's biweight. *Applied Psychological Measurement, 18*, 301–309.

Wilkinson, E. M. (1971). Archaeological seriation and the traveling salesman problem. In F. R. Hodson, D. G. Kendall & P. Tautu (Eds.), *Mathematics in the Archaeological and Historical Sciences.* Edinburgh: University Press, pp. 276–283.