PSYCHOMETRIKA—VOL. 49, NO. 1, 111–113
MARCH 1984

DIFFERENTIABILITY OF KRUSKAL'S STRESS AT A LOCAL MINIMUM

JAN DE LEEUW

LEIDEN UNIVERSITY

It is shown that Kruskal's multidimensional scaling loss function is differentiable at a local minimum. Or, to put it differently, that in multidimensional scaling solutions using Kruskal's stress distinct points cannot coincide.

Key words: multidimensional scaling, loss functions

Introduction

In one type of metric multidimensional scaling the loss function

$$\sigma(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(\delta_{ij} - d_{ij}(X))^2, \tag{1}$$

is minimized over all $n \times p$ *configuration matrices* X. Loss function (1) is patterned after Kruskal's famous STRESS (1964a, b). In (1) the w_{ij} are nonnegative *weights*, the δ_{ij} are nonnegative *dissimilarities*. Without loss of generality it can be assumed that both weights and dissimilarities are symmetric and hollow (De Leeuw, 1977, section 3). The $d_{ij}(X)$ are Euclidean *distances* between rows of X, i.e.

$$d_{ij}^2(X) = (x_i - x_j)'(x_i - x_j). \tag{2}$$

While our results deal with $\sigma(\cdot)$, they can be extended easily to the corresponding loss function of nonmetric MDS.

It was pointed out by De Leeuw (1977) that $d_{ij}(\cdot)$ is not differentiable at configurations X with $x_i = x_j$. This may cause trouble in the definition of gradient algorithms, but this trouble turns out to be not very serious. In De Leeuw (1977) and De Leeuw and Heiser (1980) it is shown how the usual gradient algorithms can be modified quite simply to deal with the problem. The resulting algorithm is unambiguously defined, and convergent. A second, more serious, problem is that the algorithm may converge to a configuration X for which the gradient $\nabla\sigma(X)$ does not exist, and for which the gradient is nonzero in any neighborhood of X. While this situation does not affect the convergence of the De Leeuw-Heiser algorithm, it does complicate the additional study of its properties. The proof of linear convergence of the algorithm to a configuration X, for example, uses the assumption that the loss function is twice continuously differentiable at X (De Leeuw, note 2).

In this note we show that if a configuration X is a local minimizer of $\sigma(\cdot)$, then x_i and x_j cannot coincide and $d_{ij}(X) \neq 0$ (unless $w_{ij} = 0$ or $\delta_{ij} = 0$). This result is of some practical importance, because we must bear it in mind if we interpret MDS solutions based on STRESS. It is also theoretically important, because we have seen that it implies that the De Leeuw-Heiser algorithm converges linearly to local minima of $\sigma(\cdot)$. A similar result, for the special case of one-dimensional metric scaling, was given in a nice paper of Defays (1978).

Requests for reprints should be sent to Jan de Leeuw, Department of Data Theory FSW/RUL, Middelstegracht 4, 2312 TW Leiden, Netherlands.

0033-3123/84/0300-5064/0$00.75/0

© 1984 The Psychometric Society

At this point it seems necessary to point out that the problems with differentiability of the loss function do not apply to SSTRESS used in ALSCAL (Takane et al, 1977) or to STRAIN used in INDSCAL (Carroll, 1981). These alternative loss functions are sums of squared deviations between squared dissimilarities and squared distances (SSTRESS) or between inner products and the doubly centered squared dissimilarities (STRAIN). Both squared distances and inner products are differentiable everywhere. We also indicate that the results of Kruskal (1971) are concerned with a different problem. He shows that monotone regression does not destroy differentiability. If the metric loss function is differentiable at X, then the corresponding nonmetric loss function, Kruskal's original STRESS, is also differentiable at X.

Basic Result

We first state and prove our basic result. As indicated below a similar result may be proved for STRESS in nonmetric MDS.

Theorem. If X is a local minimum of $\sigma(\cdot)$, then $d_{ij}(X) = 0$ can occur only if $w_{ij}\delta_{ij} = 0$.

Proof. We use the fact that $\sigma(\cdot)$ is differentiable in all directions. This follows from a direct computation of the directional derivatives, which are defined by

$$D\sigma(X; Y) = \lim_{\varepsilon \downarrow 0} \frac{\sigma(X + \varepsilon Y) - \sigma(X)}{\varepsilon}. \tag{3}$$

If X is a local minimum it follows that X must satisfy $D\sigma(X; Y) \geq 0$ for all $n \times p$ matrices Y.

Firstly

$$Dd_{ij}^2(X; Y) = 2(x_i - x_j)'(y_i - y_j), \tag{4}$$

so

$$Dd_{ij}(X; Y) = d_{ij}^{-1}(X)(x_i - x_j)'(y_i - y_j), \tag{5}$$

if $d_{ij}(X) > 0$. Also, by direct calculation,

$$Dd_{ij}(X; Y) = d_{ij}(Y), \tag{6}$$

if $d_{ij}(X) = 0$. We now set $s_{ij}(X) = d_{ij}^{-1}(X)$ if $d_{ij}(X) > 0$, and $s_{ij}(X) = 0$ if $d_{ij}(X) = 0$. We also set $t_{ij}(X) = 1$ if $d_{ij}(X) = 0$, and $t_{ij}(X) = 0$ if $d_{ij}(X) > 0$. By using these definitions, and by combining (4), (5), (6), we find

$$\begin{aligned} D\sigma(X; Y) &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} Dd_{ij}^2(X; Y) - 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\delta_{ij} Dd_{ij}(X; Y) \\ &= 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)'(y_i - y_j) \\ &\quad + 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\delta_{ij} s_{ij}(X)(x_i - x_j)'(y_i - y_j) \\ &\quad - 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\delta_{ij} t_{ij}(X) d_{ij}(Y). \end{aligned} \tag{7}$$

Because X is a local minimum we must have both $D\sigma(X; Y) \geq 0$ and $D\sigma(X; -Y) \geq 0$, for all possible directions Y. Thus the sum $D\sigma(X; Y) + D\sigma(X; -Y)$ must be nonnega-

tive too. But if we change Y to $-Y$, then the first two terms of (7), which are linear in Y, also change sign. The third term does not change. Thus

$$D\sigma(X; Y) + D\sigma(X; -Y) = -4\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\delta_{ij}t_{ij}(X)\underline{d}_{ij}(Y) \geq 0. \tag{8}$$

Because this is true for all Y, it is also true for all Y with $d_{ij}(Y) > 0$ for all $i \neq j$. But for these Y (8) is true if and only if $w_{ij}\delta_{ij}t_{ij}(X) = 0$ for all i and j. i.e. if and only if $w_{ij}\delta_{ij} = 0$ for all i, j such that $d_{ij}(X) = 0$. □

A similar result for nonmetric scaling can be obtained if we combine our theorem with the results of Kruskal (1971). In nonmetric scaling the δ_{ij} must, of course, be replaced by the disparities $\hat{d}_{ij}$. The nonmetric result is not true for loss based on the rank image principle of Guttman and Lingoes. In fact, more or less the opposite is true: it is easy to see that these loss functions are never differentiable at their stationary points (De Leeuw, note 1, and Kruskal, 1977). Our results also do not apply to the loss functions in Ramsay's MULTISCALE (Ramsay, 1978). Because MULTISCALE measures loss as the sum of squares of the deviations between log-dissimilarities and log-distances, it follows that MULTISCALE loss is not even defined at configurations with coinciding points. Thus for STRAIN and SSTRESS we can have coinciding points at local minima. For STRESS we can only have them at local minima if $w_{ij}\delta_{ij} = 0$. In metric unfolding, for instance, within-set points can coincide but between-set points cannot. In MULTISCALE coinciding points are ruled out even more radically.

REFERENCE NOTES

1. De Leeuw, J. (1974). Smoothness properties of nonmetric loss functions. Unpublished paper, Bell Telephone Labs.
2. De Leeuw, J. (1981). Linear convergence of multidimensional scaling algorithms. Unpublished paper, Department of Data Theory, Leiden University.

REFERENCES

Carroll, J. D. INDSCAL. (1981). In S. S. Schiffman, M. L. Reynolds, & F. W. Young (Eds.), *Introduction to multidimensional scaling.* New York: Academic Press.

Defays, D. (1978). A short note on a method of seriation. *British Journal of Mathematical and Statistical Psychology, 31,* 49–53.

De Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In J. R. Barra, F. Brodeau, G. Romier, & B. Van Cutsem (Eds.), *Recent developments in statistics.* Amsterdam: North Holland Publishing Co.

De Leeuw, J. & Heiser, W. (1980). Multidimensional scaling with restrictions on the configuration. In P. R. Krishnaiah (Ed.), *Multivariate analysis* V. New York: Academic Press.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29,* 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika, 29,* 115–129.

Kruskal, J. B. (1971). Monotone regression: continuity and differentiability properties. *Psychometrika, 36,* 57–62.

Kruskal, J. B. (1977). Multidimensional scaling and other methods for discovering structure. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Mathematical methods for digital computers* III. New York: Wiley.

Ramsay, J. O. (1978). *MULTISCALE: four programs for multidimensional scaling by the method of maximum likelihood.* Chicago: National Educational Resources Inc.

Takane, Y., Young, F. W., & De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika, 42,* 7–67.

Manuscript received 1/25/83
Final version received 10/18/83