ON DEGENERATE NONMETRIC UNFOLDING SOLUTIONS

JAN DE LEEUW

ABSTRACT. Loss functions proposed for nonmetric unfolding are discussed critically. Practical experience suggests that they do not work, mathematical reasons are sought why this is. Although the loss functions are constructed in such a way that they are ill-behaved at trivial solutions, it is shown that they are rather well-behaved along differentiable paths to trivial solutions. It is shown that in the neighborhood of trivial solutions we can find infinitely many nonmetric unfolding solutions which cannot be distinguished from stationary points, and which can differ very considerably in loss function value. The conclusion is that nonmetric unfolding, as currently formalized, is an inherently ill-posed problem and that a different approach is called for.

This paper originally appeared as a technical report of the Department of Data Theory, University of Leiden, in May 1983. It was never published.

1. INTRODUCTION

The nonmetric unfolding problem is defined as follows. The data are n rankings of m objects. We want to represent both rankings and objects in a low-dimensional space. The rankings are represented by vectors $x_1, \cdots, x_n$ and the objects by vectors $y_1, \cdots, y_m$. The representation is chosen in such a way that if object j precedes object ℓ in ranking i then we must have $d(x_i, y_j) \leq d(x_i, y_\ell)$,

Date: March 21, 2007.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Unfolding, nonmetric scaling, loss functions, trivial solutions.

where $d(.,.)$ is Euclidean distance. To use a common social science example: if subject i prefers j to ℓ, then y_j must be not farther from x_i than y_ℓ. The unfolding model was introduced in this generality by Coombs. Compare for example Coombs [1964].

The unfolding model defines a system of nonlinear inequalities, which generally does not have an exact solution, for example because there is some sort of random disturbance in the preference judgments. Thus we want an approximate solution. We need a loss function to tell us how good a particular solution is. The general idea of using loss functions in unfolding and related problems is due to Kruskal and Guttman. Important early references are Roskam [1968] and Kruskal and Carroll [1969]. The current state of affairs is discussed very competently in the books of Hartmann [1979] and Borg [1981]. The general conclusion seems to be that the Kruskal-Guttman approach to multidimensional scaling has been very successfull in many areas in which it has been applied. It has been very unsuccessfull in nonmetric unfolding. We shall briefly discuss the reasons for this unfortunate diagnosis, a more thoroughgoing analysis and some possible remedies are discussed in Heiser [1981].

2. A SHORT HISTORY OF NONMETRIC UNFOLDING LOSS FUNCTIONS

The major trouble is the fact that on many, if not most, occasions nonmetric unfolding algorithms find *perfect* but *trivial* solutions to the inequalities of the unfolding model. In order to make this clear we introduce the loss function used by Kruskal and Roskam for nonmetric unfolding. It is

$$\sigma_1(X,Y) = \frac{1}{n}\sum_{i=1}^n \min_{\delta_{ij}} \frac{\sum_{j=1}^m (d_{ij}-\delta_{ij})^2}{\sum_{j=1}^m (d_{ij}-d_{i\star})^2}.$$

Here d_{ij} is short for $d(x_i,y_j)$, $d_{i\star}$ is the average d_{ij} for ranking i, and the δ_{ij} are the disparities. They are restricted to be monotone with the rankings. Thus if i prefers j to ℓ then we must have

$\delta_{ij} \leq \delta_{i\ell}$. Nonmetric unfolding programs minimize $\sigma_1(X,Y)$ over X and Y, where X contains the n vectors x_i and Y contains the m vectors y_j. Of course this implies that we have to choose a *dimensionality* p of the Euclidean space we work in, in almost all applications of unfolding p is either one or two. Why did Kruskal and Roskam choose $\sigma_1(X,Y)$?

Loss function $\sigma_1(X,Y)$ differs in some importent respects from the loss function introduced by Kruskal [1964a,b] for nonmetric scaling. This is

$$\sigma_2(X,Y) = \min_{\delta_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^m (d_{ij} - \delta_{ij})^2}{\sum_{i=1}^n \sum_{j=1}^m d_{ij}^2}.$$

Loss function $\sigma_2(X,Y)$ works very well for the scaling problems for which it was designed, but it does not work for unfolding. The reason for this is clear: if we set $x_i = x$ for all i and $y_j = y$ for all j, then all d_{ij} are equal and $\sigma_2(X,Y) = 0$. Thus this *two-point configuration* always gives a perfect, but trivial, solution to the nonmetric unfolding problem.

In a related context Kruskal [1965] introduced the loss function

$$\sigma_3(X,Y) = \min_{\delta_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^m (d_{ij} - \delta_{ij})^2}{\sum_{i=1}^n \sum_{j=1}^m (d_{ij} - d_{\star\star})^2},$$

where $d_{\star\star}$ is the average of all nm "between-set" distances d_{ij}. In the appropriate context this works nicely. It excludes the two-point configuration from consideration, because $\sigma_3(X,Y)$ is undefined if all distances are equal. In fact it excludes a more subtle trivial solution as well: if $x_i = 0$ for all i and all y_j have equal length, then all distances are also equal. The same thing is true if all y_j are zero and all x_i have equal length. These configurations are called the *objects-sphere* and the *rankings-sphere*, where we have to remember that if the rankings are on the sphere then the objects are in the center and if the objects are on the sphere then the rankings are in the center.

Another more general trivial configuration, of which the rankings-sphere is a special case, merely has $y_j = 0$ for all j. The x_i are quite arbitrary. Thus d_{ij} is equal to the length of x_i, which means that the d_{ij} are constant within rankings, although not necessarily between rankings. Thus if the x_i have different lengths, then $\sigma_3(X, Y)$, and we have a perfect trivial solution, which we call the *object-point configuration*. Observe that all trivial solutions we discuss satisfy the nonlinear inequalities defining the unfolding problem. This has nothing to do with the definition of loss functions. Loss functions are merely designed to make the trivial solutions *inadmissible* in the minimization problem, basically by making loss equal to $0/0$ at the trivial configurations.

How must we eliminate the object-point configuration ? Roskam and Kruskal have experimented with

$$\sigma_4(X, Y) = \min_{\delta_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^m (d_{ij} - \delta_{ij})^2}{\sum_{i=1}^n \sum_{j=1}^m (d_{ij} - d_{i\star})^2},$$

This is undefined at the object-point configuration, and in this sense it does what it was supposed to do. But again another trivial solution exists for which $\sigma_4(X, Y) = 0$. It is discussed by Kruskal and Carroll [1969]. We call it the *two-plus-two configuration*. Suppose the rankings and objects are numbered in such a way that object 1 comes last in ranking 1 (and should thus correspond to the largest distance in the first row of d_{ij}). Consider the one-dimensional configuration with $y_1 = 0, y_2 = \cdots = y_m = 1, x_2 = \cdots = x_n = 2, x_1 = 3$. Then $d_{ij} = 1$ for all i, j except for d_{11}, which is equal to three. A little reflection shows that for this configuration indeed $\sigma_4(X, Y) = 0$. Thus loss function $\sigma_4(X, Y)$ also fails. Finally, Kruskal and Roskam arrived at $\sigma_1(X, Y)$. It has not been possible to find additional trivial solutions, not depending on the data in any essential way, for which $\sigma_1(X, Y)$ is always equal to zero.

Does it follow that the search for an appropriate nonmetric unfolding loss function has been successfull ? Some people are rather optimistic (Roskam [1977, p. 329] is an example). Others are more reserved. Borg [1981, ch. 17] calls the unfolding model "ziemlich tuckisch" from the technical point of view, and points out that the degeneracy problems are not yet solved. Heiser [1981, chapters 7 and 8] is even more pessimistic. He argues that there is not enough information in the data to fit the unfolding model nonmetrically in a reliable way. The only possible remedy is to either strengthen the model or to restrict the class of admissible transformations (i.e. minimize over smaller sets of δ_{ij}). In their recent survey De Leeuw and Heiser [1980] conclude that it is already very difficult to construct a program for metric multidimensional unfolding that works reliably (in metric unfolding the δ_{ij} are given numbers). Nonmetric unfolding is even more difficult.

We have reviewed the history of loss function construction in nonmetric unfolding in some detail, because it makes the strategy of the people developing the techniques perfectly clear. Degenerate or trivial solutions kept popping up. They tried to get rid of them by adjusting the normalization factors (the denominators) of the loss functions in such a way that they were not defined (equal to $0/0$ at trivial solutions. It was hoped, obviously, that this also kept the iterative process away from these trivial solutions. In practice it turns out that the iterative process still tends to degenerate the solution in the great majority of the cases. In fact, Heiser (personal communication, 1982) conjectures that published nontrivial solutions are probably nontrivial only because the iterations were stopped before the process had properly converged. In this paper we show clearly that the whole idea of hoping that a clever choice of the denominator solves all problems is basically unsound. There is no reason at all why the iterative process should keep away from $0/0$.

3. Perturbation of trivial solutions

In this section we try to find out what happens in the neighborhood of a trivial solution (X_0, Y_0) for which $\sigma_1(X, Y)$ is $0/0$. Thus we use perturbations $X(\epsilon)$ and $Y(\epsilon)$, depending on a positive parameter ϵ, such that $(X(\epsilon), Y(\epsilon))$ converges to (X_0, Y_0) if ϵ decreases to zero. In fact we use the stronger assumption that

$$\lim_{\epsilon\downarrow 0}\frac{X(\epsilon)-X_0}{\epsilon}=X_1,$$
$$\lim_{\epsilon\downarrow 0}\frac{Y(\epsilon)-Y_0}{\epsilon}=Y_1,$$

where X_1 and Y_1 are matrices, not both equal to the zero matrix. Triviality of (X_0, Y_0) implies that the unperturbed distances are equal within rows. We use λ_i for the common value of the distances in row i. Thus $\lambda_i = d(x_i(0), y_j(0))$. The λ_i are supposed to be nonzero. Define

$$c_{ij}=\frac{(x_i^0-y_j^0)'(x_i^1-y_j^1)}{\lambda_i}$$

and suppose that for each i there are at least two different c_{ij}.

Theorem 1. *Under the stated assumptions* $\lim_{\epsilon\downarrow 0}\sigma_1(X(\epsilon), Y(\epsilon))$ *exists, and is equal to*

$$\sigma_5(X_0, Y_0, X_1, Y_1)=\frac{1}{n}\sum_{i=1}^n\min_{y_{ij}}\frac{\sum_{j=1}^m(c_{ij}-y_{ij})^2}{\sum_{j=1}^m(c_{ij}-c_{i\star})^2}.$$

Here the y_{ij} *must satisfy the same ordinal constraints as the* δ_{ij} *earlier.*

Proof. In the first place we compute perturbed distances in row i. They are $d_i(\epsilon) = \lambda_i u + \epsilon c_i + o(\epsilon)$, where u is an m-vector with all elements equal to $+1$. Thus if $\epsilon \downarrow 0$ we find that $\frac{d_i(\epsilon)-\lambda_i u}{\epsilon}$ converges to c_i, row i of C. If Π_i is the metric projection on the cone of vectors which are in the correct order for row i, then $\Pi_i(\frac{d_i(\epsilon)-\lambda_i u}{\epsilon})$ converges to $\Pi_i(c_i) = y_i$, because the metric projection is continuous. Thus if $\delta_i(\epsilon) = \Pi_i(d_i(\epsilon))$ then $\delta_i(\epsilon) = \lambda_i u + \epsilon y_i + o(\epsilon)$. Also

$d_{ij}(\epsilon) - d_{i\star}(\epsilon) = \epsilon(c_{ij} - c_{i\star}) + o(\epsilon)$. Combining all this gives the stated result. □

Before we proceed we discuss the precise contents of the theorem. If ϵ decreases to zero the configuration $(X(\epsilon), Y(\epsilon))$ converges to the trivial solution (X_0, Y_0), but the corresponding loss function value $\sigma_1(\epsilon)$ converges to a finite value, depending on X_0, Y_0, X_1, Y_1. A nonmetric unfolding *algorithm* is not concerned with the fact that loss is undefined at a trivial solution. In the neighborhood of such a trivial solution it merely searches for a solution with a low value of $\sigma_5(X_0, Y_0, X_1, Y_1)$. If we have a (X_0, Y_0, X_1, Y_1) such that $\sigma_5(X_0, Y_0, X_1, Y_1)$ is low and (X_0, Y_0) is trivial, then we can construct a solution to the unfolding problem, which is almost trivial and which has loss almost equal to $\sigma_5(X_0, Y_0, X_1, Y_1)$.

In the case of the specific trivial configurations we have discussed earlier we can specialize our theorem somewhat. For the object-point configuration we have $Y_0 = 0$, and we can choose without loss of generality $X_1 = 0$. Thus $c_{ij} = -\frac{(x_i^0)'y_j^1}{\lambda_i}$. In the neighborhood of the object-point configurations nonmetric unfolding algorithms fit the *vector model* (also known as the nonlinear factor analysis, nonlinear component analysis, or the compensatory model). If there exists a perfect solution to the nonlinear inequalities defining the vector model, then there also is an unfolding solution that cannot be distinguished from a trivial object-point solution and which has loss that cannot be distinguished from zero.

For the objects-sphere configuration we have $X_0 = 0$ and $\mathbf{diag}\, Y_0 Y_0' = I$. Thus $c_{ij} = -(y_j^0)'(x_i^1 - y_j^1)$, which can be considered as a sort of signed version of the compensatory distance model discussed by Coombs [1964] and Roskam [1968]. For the two-point configuration all rows of X_0 are the same and all rows of Y_0 are the same. Thus $c_{ij} = \zeta'(x_i^1 - y_j^1)$, with $\zeta = \frac{x_0 - y_0}{\lambda}$. Because $\zeta' x_i^1$ is constant in a row this actually means that in the neighborhood of the two-point

configuration nonmetric unfolding fits $c_{ij} = -\zeta' y_j^1$, or $c_{ij} = \theta_j$, a row-conditional version of the additive model.

It is of some interest to observe that Theorem 1 remains true if there are ties in the data. It does not matter whether we use the primary or the secondary approach to ties discussed by Kruskal [1964a,b]. Theorem 1 also remains true if we use Guttman's *rank image principle* to compute the δ_{ij}, instead of the *monotone regression method* used by Kruskal and Roskam.

4. Stationary values

Using more complicated, but essentially identical, computations we can prove Theorem 2, below. In this theorem the following notation is used, If f is a function of s vector variables, then $\mathcal{D}_t f(\underline{x}_1, \cdots, \underline{x}_s)$ is the vector of partials of f with respect to x_t, evaluated in $\underline{x}_1, \cdots, \underline{x}_s$.

Theorem 2. *Under the same assumptions as in Theorem 1*

$$\lim_{\epsilon \downarrow 0} \epsilon \mathcal{D}_1 \sigma_1(X(\epsilon), Y(\epsilon)) = \mathcal{D}_3 \sigma_5(X_0, Y_0, X_1, Y_1),$$
$$\lim_{\epsilon \downarrow 0} \epsilon \mathcal{D}_2 \sigma_1(X(\epsilon), Y(\epsilon)) = \mathcal{D}_4 \sigma_5(X_0, Y_0, X_1, Y_1).$$

Proof. As Theorem 1. □

Again a verbal description of the contents of Theorem 2 is useful. If the partials of $\sigma_5(X_0, Y_0, X_1, Y_1)$ with respect to X_1 and Y_1 are zero, then we can find (X, Y) arbitrary close to (X_0, Y_0) with arbitrary small derivatives. For a given (X_0, Y_0) we can find the corresponding (X_1, Y_1) that minimizes $\sigma_5(X_0, Y_0, X_1, Y_1)$. For small ϵ we have, by Theorem 1,

$$\sigma_1(X(\epsilon), Y(\epsilon)) \approx \sigma_5(X_0, Y_0, X_1, Y_1),$$

and by Theorem 2

$$\epsilon\mathcal{D}_1\sigma_1(X(\epsilon),Y(\epsilon))\approx 0,$$
$$\epsilon\mathcal{D}_2\sigma_1(X(\epsilon),Y(\epsilon))\approx 0.$$

Observe that for fixed(X_0,Y_0) the function c_{ij} is linear in (X_1,Y_1). Finding the optimum (X_1,Y_1) is consequently a linear nonmetric problem.

Alternatively it becomes interesting to minimize $\sigma_5(X_0,Y_0,X_1,Y_1)$ over all four arguments, with the restriction that (X_0,Y_0) is trivial for the nonmetric unfolding problem. We have seen that this triviality constraint can be decomposed into the union of $Y_0=0$ (object-point configuration) and $X_0=0$ with **diag** $Y_0'Y_0=I$ (objects-circle configuration). In the first case minimizing $\sigma_5(X_0,Y_0,X_1,Y_1)$ amounts to fitting the vector model to the data, in the second case the signed compensatory distance model. Although each trivial solution to the unfolding inequalities can be made to correspond with a flat point of $\sigma_1(X,Y)$, it seems sensible to compute these "optimum trivial soIutions" first by fitting a vector or compensatory distance model.

5. A SMALL EXAMPLE

We generated 10 random rankings of 5 objects as our data, and 10 random points of the circle as our X_0. We study our loss functions in the neighborhood of the object-point configuration. Thus $Y_0=0$ and $X(\epsilon)\equiv X_0$. We first minimized $\sigma_5(X_0,Y_0,X_1,Y_1)=\sigma_5(X_0,0,0,Y_1)$ over Y_1 for fixed X_0. This was done in APL, very precisely, by an ad hoc program which found a point where largest partial derivative was less than 1E-10 (this is APL-notation for floating point numbers). The value of $\sigma_5(X_0,Y_0,X_1,Y_1)$ at this point was .4023140247. In Table 1 we have collected information on $\sigma_1*X_9,\epsilon Y_1)$ for various values of ϵ. In the two last columns of the table we give upper bounds on the size of the largest partials.

Thus, for example, for ϵ =1E-7 the largest element in $Y(\epsilon) = \epsilon Y_1$ is 3E-7, and the configuration is for all practical purposes an object-point. The partials of σ_1 with respect to X are all less than 1E-8 and the partials with respect to Y are all less than 1E-7, which means that for all practical purposes we have a stationary point here. Certainly all the existing nonrnetric unfolding programs will consider it as a stationary point. For ϵ which is too small the round-off makes the approximation of $\sigma_1(X_0, \epsilon Y_1)$ deteriorate, while Theorem 2 tells us that the partials become very large again.

In Table 2 we have done essentially the same thing, except for the fact that $\sigma_5(X_0, 0, 0, Y_1)$ was minimized over X_0 and Y_1 first by the same ad hoc APL program. Thus X_0 and Y_1 define a stationary point for the vector model. The value of $\sigma_5(X_0, 0, 0, Y_1)$ at this point was .0956953055. Table 2 shows that around 1E-7 we have an object-point configuration with almost exactly this value for unfolding loss σ_1 which is almost exactly stationary. Any program would except it as a local minimum.

6. Discussion

The problem discussed in the introduction was that nonmetric unfolding often gives solutions which are wholly or partly degenerate. Designers of nonmetric unfolding programs have tried to avoid the obvious trivial solutions, which are not even data-dependent, by a clever choice of the normalization factors in the loss functions. This has produced the Kruskal-Roskam loss function σ_1 which is "partitioned by rows". It is clear, from practical experience with unfolding reported for example in Heiser [1981] and Borg [1981], that use of this loss function still leads to partial degeneracies. We show that use of this loss function cannot even guarantee that the completely trivial solutions do not occur.

Technically speaking it is clear that $\sigma_1(X, Y)$ does not exist at the points (X, Y) which define trivial solutions. It certainly does not

make sense to investigate if $\sigma_1(X, Y)$ is continuous or differentiable at these points. We have investigated the loss function along paths of the form $(X(\epsilon), Y(\epsilon))$ which tend to a trivial solution if the parameter ϵ decreases to zero and which are differentiable functions of the parameter. We found that $\sigma_1(X(\epsilon), Y(\epsilon))$ converges if $\epsilon \downarrow 0$ to a value which could be related to fit of the vector model or compensatory distance model along the path. Because it is the job of an algorithm to find a path along which the function decreases, it is clear that in the neighborhood of a trivial point nonmetric unfolding algorithms will choose their paths according to the fit of the vector or compensatory distance model. It is not true that the partials of $\sigma_1(X(\epsilon), Y(\epsilon))$ converge if $\epsilon \downarrow 0$, but if the perturbations X_1 and Y_1 are close to stationary points of the vector or compensatory distance models we can find points arbitrary close to a trivial solution with arbitrary small derivatives.

Our discussion can be extended in principle to partial degeneracy. If the objects are on a circle, and not all rankings are in the center of the circle, then we have deqeneracy only in some rows. In other rows the additional freedom can be used to improve the fit. There are obviously very many possibilities of partial degeneracy, although clearly the objects-circle is far more flexible in this respect than the object-point. It seems to us that by using this partial deqereracy clever nonmetric unfolding programs will almost always find very good solutions, which tell us something about the data, but not very much. We repeat the statement made by De Leeuw and Heiser in their discussion of metric unfolding. “As a matter of fact, even the best metric unfolding methods do not work very well. Nonmetric unfolding methods do not work at all. " [De Leeuw and Heiser, 1980, p. 305]

References

I. Borg. *Anwendungsorientierte Multidimensionale Skalierung*, volume 1 of *Lehr- und Forschungstexte Psychologie*. Springer, 1981.

C. H. Coombs. *A Theory of Data*. Wiley, 1964.

J. De Leeuw and W. J. Heiser. Theory of Multidimensional Scaling. In P.R. Krishnaiah, editor, *Handbook of Statistics, Volume II*. North Holland Publishing Company, Amsterdam, The Netherlands, 1980.

W. Hartmann. *Geometrische Modelle zur Analyse empirischer Data*. Akademie Verlag, 1979.

W.J. Heiser. *Unfolding Analysis of Proximity Data*. PhD thesis, University of Leiden, 1981.

J. B. Kruskal. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika*, 29:1–27, 1964a.

J. B. Kruskal. Analysis of Factorial Experiments by Estimating Monotone Transformations of the Data. *Journal of the Royal Statistical Society*, B27:251–263, 1965.

J.B. Kruskal. Nonmetric Multidimensional Scaling: a Numerical Method. *Psychometrika*, 29:115–129, 1964b.

J.B. Kruskal and J.D. Carroll. Geometrical Models and Badness of Fit Functions. In P.R. Krishnaiah, editor, *Multivariate Analysis, Volume II*, pages 639–671. North Holland Publishing Company, 1969.

E.E. Roskam. *A General System for Nonmetric Data Analysis*, chapter 13. Mathesis Press, 1977.

E.E. Roskam. *Metric Analysis of Ordinal Data in Psychology*. PhD thesis, University of Leiden, 1968.

TABLE 1. approximation of loss function and gradient at arbitrary object-point

eps	loss	$\mathcal{D}_X$	$\mathcal{D}_Y$
1E-3	.4024913570	1E-4	1E-3
1E-5	.4028107802	1E-6	1E-5
1E-7	.4028139922	1E-8	1E-7
1E-9	.4028140340	1E-7	1E-8
1E-11	.4026087144	1E-6	1E-2

TABLE 2. approximation of loss function and gradient at optimal object-point

eps	loss	$\mathcal{D}_X$	$\mathcal{D}_Y$
1E-3	.0956136560	1E-5	1E-3
1E-5	.0956944676	1E-9	1E-5
1E-7	.0956953971	1E-9	1E-7
1E-9	.0956953046	1E-7	1E-7
1E-11	.0947034739	1E-5	1E-2

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://www.cuddyvalley.org