

Local Minima in Nonmetric Multidimensional Scaling

Michael A.A. Cox,
School of Business,
University of Newcastle Upon Tyne,
Newcastle Upon Tyne, NE1 7RU.

Trevor F.Cox,
Department of Mathematics,
University of Newcastle,
Newcastle Upon Tyne, NE1 7RU.

Abstract

The problem of locating the true minimum of the stress function in nonmetric multidimensional scaling is highlighted. Particular reference is made to software packages and their documentation that frequently fail to “stress” the possible pit falls.

Introduction

It is well known that solutions of the unidimensional scaling problem (see for example Guttman, 1968, de Leeuw and Heiser, 1977, Pliner, 1984 Hubert and Arabie, 1988 and Hubert at al., 1987) are plagued by the occurrence of numerous local minima. This is also true to an extent in the case of multidimensional scaling, where the practitioner is advised to try a number of trial starting values. However available software does not necessarily lend itself to this approach, and associated manuals do not stress the problem. In fact examples used in the documentation do not always display the true solution. It is not uncommon for published solutions to display properties, which suggest that the procedure has not converged to the assumed (desired) solution.

This paper investigates the occurrence of local (false) minima when employing available software to analyze the standard multidimensional scaling problem. For illustrative purposes, and ease of computation, a subset of 12 individuals is selected from an example included in one of the software

manuals. It is not our intention here to pillory manuals and authors who publish erroneous solutions, but rather to highlight the pitfalls, which might arise.

The Experiment

The dissimilarity data used are displayed in Table 1

Table 1 The Dissimilarities

0.099										
0.033	0.022									
0.183	0.114	0.042								
0.148	0.224	0.059	0.068							
0.198	0.039	0.053	0.085	0.051						
0.462	0.266	0.322	0.435	0.268	0.025					
0.628	0.442	0.444	0.406	0.240	0.129	0.014				
0.113	0.070	0.046	0.047	0.034	0.002	0.106	0.129			
0.173	0.119	0.162	0.331	0.177	0.039	0.089	0.237	0.071		
0.434	0.419	0.339	0.505	0.469	0.390	0.315	0.349	0.151	0.430	
0.762	0.633	0.781	0.700	0.758	0.625	0.469	0.618	0.440	0.538	0.607

These were a subset of the data employed to illustrate the use of one of the pieces of software (NAG, 2000).

For the purposes of comparison the statistical software package SAS (2000) was contrasted to the subroutines in the NAG library (2000) and a specific program (Cox and Cox, 2000).

In each case 1000 independent random restarts were adopted and the algorithms were run until they appeared to converge. None of the SAS solutions indicated any problems. Three of the NAG solutions were rejected. None of the solutions obtained from the Cox and Cox (2000) software were rejected.

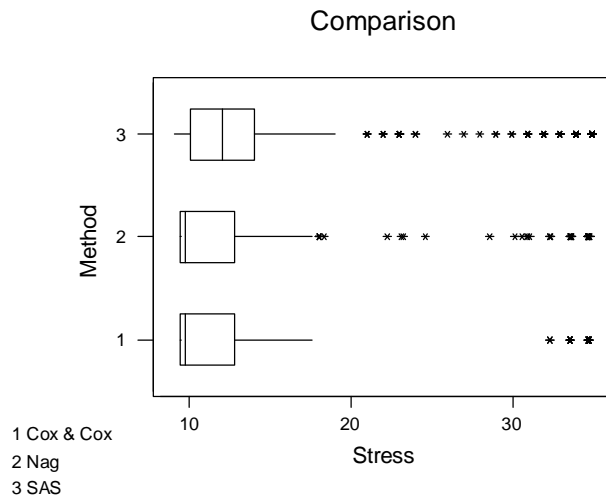
Do the Packages Differ?

The algorithm for nonmetric multidimensional scaling minimises a loss function called STRESS, where

$$STRESS = \sqrt{\frac{\sum_{r,s} (d_{rs} - \hat{d}_{rs})^2}{\sum_{r,s} d_{rs}^2}}$$

$\{d_{rs}\}$ are the distances between points in the multidimensional scaling configuration, and $\{\hat{d}_{rs}\}$ is the least squares isotonic regression of $\{d_{rs}\}$ on the original dissimilarities $\{\delta_{rs}\}$, see Cox and Cox (2000) for further details. A stacked MQMQM (Figure 1) plot illustrates the similarity of the results from the procedures.

Figure 1 A Comparison MQMQM Plot



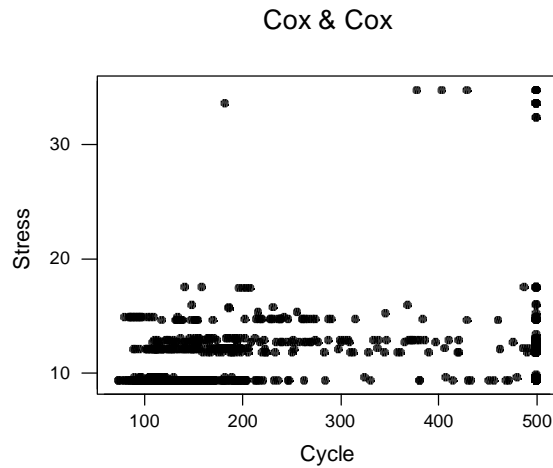
The solution vectors, which corresponded to the observed minimum stress value, were effectively identical, as illustrated by the stress value reported and employing a Procrustes rotation to compare the vectors.

For ease of use attention will now concentrate on the Cox and Cox (2000) software since these solutions were readily accessible.

Does the Number of Cycles Affect the Solutions?

The user may select various parameters. Those, which monitor convergence, will have an obvious effect on the solution. However it is unclear that pre-selecting the number of cycles of the algorithm is important (above a sensible minimum level). On displaying the resulting stress against the number of cycles on convergence (Figure 2), no obvious relationship is revealed.

Figure 2 How Many Cycles?

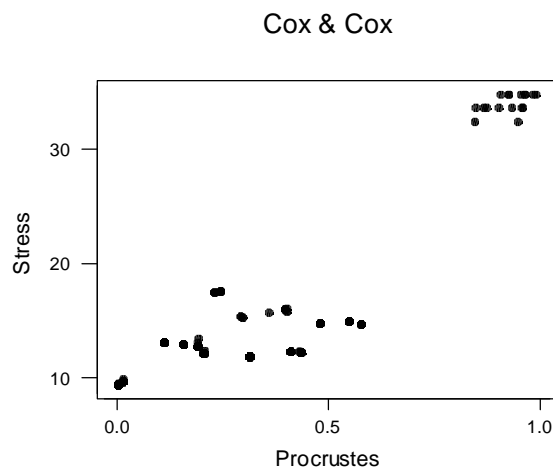


It would appear that the best advice is to adopt as many cycles as the user may reasonably employ, given the constraints of computer time.

Is there a significant difference between the solutions?

The 455 solution vectors from the 1000 overall, which produced the minimum stress of 9.39% were matched using Procrustes rotation. They were effectively identical as revealed by a zero Procrustes statistic. Of the other 545 solutions, 155 have a Procrustes value in excess of 0.4, and 13 a value exceeding 0.9. In fact they are totally dissimilar to the desired solution. These data are summarised in Figure 3.

Figure 3 Relationship between Stress and Procrustes Statistics



In summary around a half of the supposed solutions are effectively wrong!

For users who wish to perform comparative calculations, the minimum stress is 9.39% and the corresponding solution vector is presented in Table 2.

Table 2 The Solution Vector

1.1461	0.0762
0.7073	0.0891
0.6744	0.4548
0.3689	0.8821
0.0538	0.8886
0.0178	0.2989
-0.6866	-0.0026
-1.0906	0.1348
0.0938	0.1511
-0.1920	-0.2261
0.2819	-1.0785
-1.3748	-1.6683

Conclusion

The analysis presented here has revealed that there are numerous pitfalls for the unwary when employing multidimensional scaling. The spread of the assumed “solution” vectors as illustrated by Figure 3 highlights the problem. However, all the cases which resulted in minimum stress values produced identical solutions.

Hence to avoid selecting solutions which correspond to extreme, unwanted, tail values, numerous random starting values must be considered.

References

- Cox, T.F. and Cox, M.A.A., *Multidimensional Scaling*, Chapman & Hall/CRC Press: Boca Rotan, ISBN 1-58488-094-5, 2000.
- De Leeuw, J. and Heiser, W., Convergence of correction matrix algorithms for multidimensional scaling. In Lingoes, J.C. (ed.) *Geometrical Representations of Relational Data*, Ann Arbor, Michigan: Mathesis Press, 1977
- Guttman, L., A general nonmetric technique for finding the smallest co-ordinate space for a configuration of points. *Psychometrika*, 3, 469-506, 1968.

Hubert, L. and Arabie, P., Relying on necessary conditions for optimisation: unidimensional scaling and some extensions. In Bock, H.H. (ed.) Classification and Related Methods of Data Analysis, north Holland, 1988.

Hubert, L., Arabie, P. and Meulman, J., Linear and circular unidimensional scaling for symmetric proximity matrices. Brit. J. Math. And Stat. Psychol., 50, 253-284, 1987.

NAG, <http://www.nag.com>, 2000.

Pliner, V., Metric unidimensional scaling and global optimisation. Journal of Classification, 13, 3-18, 1984.

SAS, <http://www.sas.com>, 2000.

Key phrases

1. Multidimensional Scaling
2. Stress
3. NAG
4. SAS

Mailing address

Dr. M.A.A. Cox,
School of Business,
University of Newcastle Upon Tyne,
Newcastle Upon Tyne, NE1 7RU,
U.K.

Telephone: (011 44) 0191 222 6234

Fax: (011 44) 0191 261 1182

e-mail: mike.cox@ncl.ac.uk

Michael A.A. Cox,
School of Business,
University of Newcastle Upon Tyne,
Newcastle Upon Tyne, NE1 7RU,
U.K.
18 March 2005

Ms. Michele Alexander,
Statistics and Computing,
Editorial Office,
Kluwer Academic Publishers,
101, Philip Drive,
Norwell,
MA 02061,
USA.
Malexander@wkap.com

Ref: STCO24 - 00

Dear Ms. Alexander,

I would like to submit the enclosed manuscript "Local Minima in Nonmetric Multidimensional Scaling", of which I include 5 copies, to the Statistics and Computing Journal.

Yours sincerely,

Mike Cox.

Executing from file: c:\aew3210da\test\cox&cox.mtb
 1000 rows read.
 * ERROR * Duplicate name not permitted.

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C1	1000	11.511	9.720	11.031	3.571	0.113

Variable	Minimum	Maximum	Q1	Q3
C1	9.390	34.800	9.390	12.750

Descriptive Statistics

Variable	N	N*	Mean	Median	TrMean	StDev
C1	706	294	11.095	9.390	10.816	2.674

Variable	SE Mean	Minimum	Maximum	Q1	Q3
C1	0.101	9.390	34.800	9.390	12.260

Histogram

* NOTE * N missing = 294

Boxplot

* NOTE * N missing = 294

Plot

* NOTE * N missing = 294

Plot

* NOTE * N missing = 294

Executing from file: c:\aew3210da\test\g03fcf.mtb
 1000 rows read.

* ERROR * Duplicate name not permitted.

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C1	1000	11.908	9.720	11.169	4.429	0.140

Variable	Minimum	Maximum	Q1	Q3
C1	9.387	38.040	9.389	12.750

Descriptive Statistics

Variable	N	N*	Mean	Median	TrMean	StDev
C1	997	3	11.864	9.720	11.148	4.337

Variable	SE Mean	Minimum	Maximum	Q1	Q3
C1	0.137	9.387	34.820	9.389	12.750

Histogram

* NOTE * N missing = 3

Boxplot

* NOTE * N missing = 3

Executing from file: c:\aew3210da\test\sas.mtb
 1000 rows read.

* ERROR * Duplicate name not permitted.

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C5	1000	13.837	12.000	12.947	6.453	0.204

Variable	Minimum	Maximum	Q1	Q3
C5	9.000	35.000	10.000	14.000

Boxplot

* NOTE * N missing = 297

Histogram

Histogram of Cox&Cox N = 706 N* = 294
 Each * represents 10 observation(s)

Midpoint	Count	
10	408	*****
12	203	*****
14	76	*****
16	8	*
18	7	*
20	0	
22	0	
24	0	
26	0	
28	0	
30	0	
32	0	
34	4	*

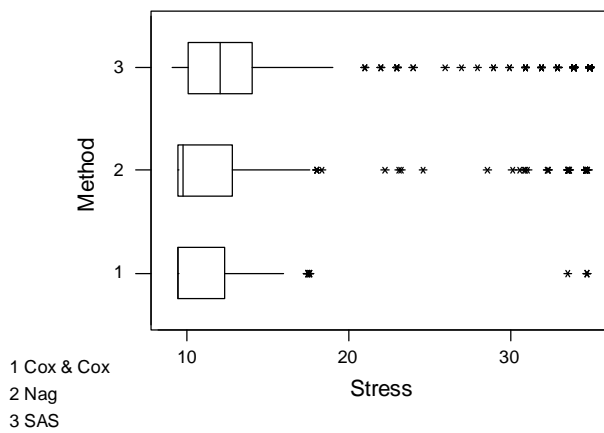
Histogram of Nag N = 997 N* = 3
Each * represents 15 observation(s)

Midpoint	Count	
10	512	*****
12	281	*****
14	136	*****
16	15	*
18	20	**
20	0	
22	1	*
24	3	*
26	0	
28	1	*
30	4	*
32	4	*
34	20	**

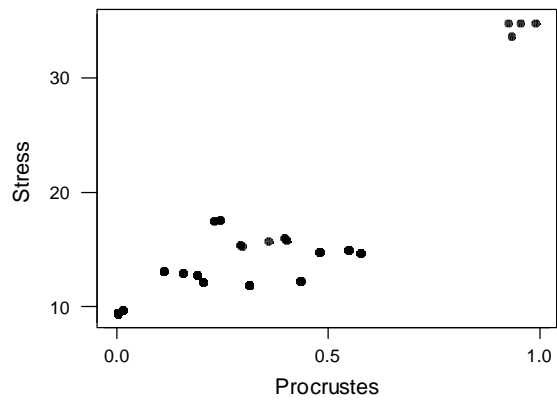
Histogram of SAS N = 1000
Each * represents 10 observation(s)

Midpoint	Count	
10	291	*****
12	280	*****
14	187	*****
16	116	*****
18	28	***
20	1	*
22	6	*
24	8	*
26	1	*
28	2	*
30	4	*
32	14	**
34	27	***
36	35	****

Comparison



Cox & Cox - Selected



Executing from file: c:\aew3210da\test\cox&cox1.mtb
1000 rows read.
* ERROR * Duplicate name not permitted.

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C1	1000	11.511	9.720	11.031	3.571	0.113

Variable	Minimum	Maximum	Q1	Q3
C1	9.390	34.800	9.390	12.750

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C1	1000	11.511	9.720	11.031	3.571	0.113

Variable	Minimum	Maximum	Q1	Q3
C1	9.390	34.800	9.390	12.750

Executing from file: c:\aew3210da\test\g03fcf.mtb
1000 rows read.

* ERROR * Duplicate name not permitted.

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C1	1000	11.908	9.720	11.169	4.429	0.140

Variable	Minimum	Maximum	Q1	Q3
C1	9.387	38.040	9.389	12.750

Descriptive Statistics

Variable	N	N*	Mean	Median	TrMean	StDev
C1	997	3	11.864	9.720	11.148	4.337

Variable	SE Mean	Minimum	Maximum	Q1	Q3
C1	0.137	9.387	34.820	9.389	12.750

Histogram

* NOTE * N missing = 3

Boxplot

* NOTE * N missing = 3

Executing from file: c:\aew3210da\test\sas.mtb
1000 rows read.

* ERROR * Duplicate name not permitted.

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C5	1000	13.837	12.000	12.947	6.453	0.204

Variable	Minimum	Maximum	Q1	Q3
C5	9.000	35.000	10.000	14.000

Boxplot

* NOTE * N missing = 3

Histogram

Histogram of Cox&Cox N = 1000
 Each * represents 15 observation(s)

Midpoint	Count	
10	520	*****
12	335	*****
14	104	*****
16	11	*
18	13	*
20	0	
22	0	
24	0	
26	0	
28	0	
30	0	
32	2	*
34	15	*

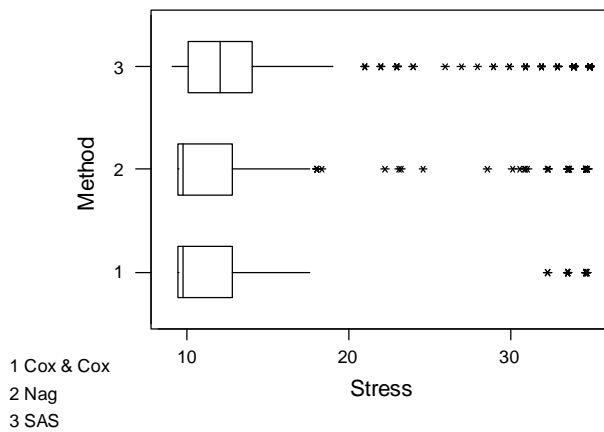
Histogram of Nag N = 997 N* = 3
 Each * represents 15 observation(s)

Midpoint	Count	
10	512	*****
12	281	*****
14	136	*****
16	15	*
18	20	**
20	0	
22	1	*
24	3	*
26	0	
28	1	*
30	4	*
32	4	*
34	20	**

Histogram of SAS N = 1000
 Each * represents 10 observation(s)

Midpoint	Count	
10	291	*****
12	280	*****
14	187	*****
16	116	*****
18	28	***
20	1	*
22	6	*
24	8	*
26	1	*
28	2	*
30	4	*
32	14	**
34	27	***
36	35	****

Comparison



Cox & Cox - All

