Les Cahiers du GERAD ISSN: 0711–2440

Identification of the Minkowski Parameter for Multidimensional Scaling

G. Caporossi

G–2008–61

September 2008

La publication de ces rapports de recherche bénéficie d'une subvention du Fonds québécois de la recherche sur la nature et les technologies.

Identification of the Minkowski Parameter for Multidimensional Scaling

Gilles Caporossi

GERAD and HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada, H3T 2A7

gilles.caporossi@gerad.ca

September 2008

Les Cahiers du GERAD

G–2008–61

Abstract

Based upon a robust optimization technique, Variable Neighborhood Search (VNS), we use simulation to find rules for identifying the correct Minkowski parameter r to use in Multidimensional Scaling (MDS). Claims from the paper of Shepard (1974) are confirmed, and others are nuanced. We confirm the value of r cannot be defined precisely, that the proper r value cannot be found in dimension 2 or when the data is too noised, but it is possible to find good estimates in the other cases.

Key Words: Multidimensional Scaling, Minkowski parameter, Monte Carlo Simulation, Optimization.

Résumé

En s'appuyant sur une technique d'optimisation robuste, la Recherche à Voisinages Variables, nous utilisons la simulation pour identifier le paramètre de Minkowski r approprié à utiliser en *Multidimensional Scaling* (MDS). Certaines affirmations de l'article classique de Shepard (1974) sont confirmées, et d'autres sont nuancées. Nous confirmons que la valeur de r ne peut être identifiée avec précision, que la valeur de r ne peut pas être trouvée en dimension 2 ou quand les données sont trop bruitées, mais il est possible de trouver de bonnes estimations dans les autres cas.

1 Introduction

Multidimensional scaling (MDS) consists in determining coordinates to stimuli in a low dimension space from proximity/dissimilarity information. It was proposed about fifty years ago by Shepard (1957, 1962,a). Few years later, Kruskal proposed a measure to evaluate the quality of a model (Kruskal, 1964) and a first algorithm to solve the problem (Kruskal, 1964a). MDS, its extensions and some related researches are described in (Cox & Cox, 2001) or (Borg & Groenen, 2005). To shortly describe the problem, suppose n stimuli and note δ_{ij} the dissimilarities between stimuli i and j. A dissimilarity indicates a notion of distance between stimuli (for example in marketing, a degree of difference between goods); the larger it is, the more different are the corresponding goods. The purpose of MDS is to find coordinates x_{ik} of these stimuli in a D dimensions space in such a way that distance d_{ij} between the representation of stimuli i and j best reflects the original dissimilarity information δ_{ij}.

From this low dimensional representation of the data, analysis could be made.

The MDS parameters
As stressed by Sherman (1972) and Shepard (1974), the choice of the appropriate parameters is an important step toward a good MDS analysis.

- The first parameter to define is the error function, a value indicating the quality of the solution, the smaller the better. The raw stress and normalized stress (Kruskal, 1964) are most commonly used. The raw stress could be viewed as the sum of squared difference between the inter-stimuli distances d_{ij} and the corresponding dissimilarities δ_{ij} as described by Eq. (1)

$$raw\ stress\ =\ S^* = \sum_{i<j}(d_{ij} - \delta_{ij})^2. \tag{1}$$

  If we note $T^* = \sum_{i<j} d_{ij}^2$ the scaling factor, the normalized stress is described by Eq. (2)

$$stress\ =\ S = \sqrt{\frac{S^*}{T^*}} = \sqrt{\frac{\sum_{i<j}(d_{ij} - \delta_{ij})^2}{\sum_{i<j} d_{ij}^2}}. \tag{2}$$

  Kruskal's normalized stress has the property to provide a smoother function, which makes the number of local optima smaller ; however, the use of the normalized stress does not completely avoid this problem.
- The next parameter is the dimension of the space in which the stimuli are to be represented. The dimension should correspond to the nature of the stimuli under study. Choosing a smaller dimension would cause some errors as all the pertinent information cannot be represented (Lee, 2001). On the other hand, choosing a higher dimension increases the importance of noise in the representation. A classical way to identify the correct dimension is by looking at the stress/dimension curve and find a break in the curve (known as the *scree test* for identifying the number of relevant factors in principal component analysis or the correct number of clusters in clustering).

  Finding the correct dimension by the scree test is sometimes difficult as the stress/dimension curve may be smooth, in which case the breaking point is difficult to identify. To overcome this problem, Lee (2001) proposes a BIC based test to validate the choice of the dimension in MDS.

- Another parameter is the choice of the distance measure to use. The distance computation is usually achieved using the Minkowski formula (3).

$$d_{ij} = (\sum_{k=1}^{D} |x_{ik} - x_{jk}|^r)^{1/r} \tag{3}$$

Where d_{ij} is the distance between stimuli i and j, x_{ik} is the k^{th} coordinate of stimuli i, D the dimension of the space and r the Minkowski parameter. If $r = 2$, the formula represents the Euclidian distance while it corresponds to the city-block (or Manhattan) distance if $r = 1$. As stressed by Arabie (1991), the Euclidian distance is probably not always the most appropriate. On the other hand, the Euclidian distance is more robust than other distance measures in term of the optimization in MDS. Indeed, as the Euclidian distance is not sensitive to the rotation, any initial solution that could be used prior to a gradient descent could be expected to be closer to a global minima in this case.

The optimization issue
The number of papers on that topic indicates that optimization problem underlying MDS is far from trivial. Except for the unidimensional scaling case (Simantiraki, 1996; Brsuco & Stahl, 2005), to our knowledge, there exists no efficient exact algorithm for MDS.

One of the major difficulties in MDS is the combinatorial issue which is particularly important in unidimensional scaling and was widely studied (Pliner, 1996; Simantiraki, 1996; Hubert et al., 1997; Lau et al., 1998; Brusco & Stahl, 2000; Brsuco & Stahl, 2005; Hubert et al., 2002; Brusco, 2002, 2006), but also in multidimensional scaling where it causes a large number of local optima. The problem of the local optima is known since the beginning of studies on MDS (Kruskal, 1964) and it has particularly important consequences in the case of non-Euclidian distances. Important researches were achieved to improve the optimization in MDS, more specifically using the city-block distance. Some authors proposed the use of majorization technique (deLeeuw, 1988; Groenen et al., 1995) to help finding a good solution. However, the local optima issue was not completely avoided and tunneling was suggested (Groenen & Heiser, 1996) to improve the solution. Other researches proposes the use a combinatorial approach (Hubert & Arabie, 1986; Hubert et al., 1992), nonlinear programming (Lau et al., 1998), distance smoothing (Groenen et al., 1999) or simulated annealing (Brusco, 2001; Murillo et al., 2005; Vera et al., 2007) to handle the problem. The algorithmical issue is closely related to the reliability of the analysis and has an important impact on the choice of the parameters of the model.

In the present paper, we propose the use of the Variable Neighborhood Search (VNS) (Mladenović & Hansen, 1997; Hansen & Mladenović, 1997; Hansen & Mladenović, 2001), a metaheuristic developed to solve combinatorial optimization problems, but that is also well suited for continuous optimization problems. From the beginning of its use, VNS has proved to be an efficient metaheuristic and usually provides good results.

The VNS algorithm and the monte carlo simulation procedure are described in the second and third section. The fourth section describes the results of the study and the fifth section explains a suggested procedure toward the correct parameters choice. Some real data applications are then described and the last section concludes.

2 The optimization method, VNS

Probably the most intuitive way of improving a solution after a local optimum is found is multistart, which consists in applying a large number of successive local searches (for example a gradient descent in the case of MDS) from various initial random configurations. Unfortunately, such an approach is likely to fail if the number of local optima becomes very large.

One important feature of VNS is to keep information acquired from previous searches and from the best known solution. The idea underlying this method is that local optima often share a large amount of characteristics and it would be a loss of time to explore solutions that are very far away from good known solutions (as it is done with multistart). For example, in the case of MDS, it would be surprising that two stimuli with important dissimilarity would appear close to each other in a good solution.

The VNS algorithm could be described as follows: first apply a local search to a random initial solution. Keep in memory the obtained solution as *best_solution*. Then, apply a small perturbation (shaking) to *best_solution* before doing a local search again. The solution obtained could either be better than *best_solution* or not. If so, keep the obtained solution as *best_solution* in memory. If not, increase the magnitude of the perturbation and repeat the process. In order to avoid searches that are too far away from *best_solution*, when the magnitude of the perturbation becomes too large, it is again reduced to its smallest size. Each time *best_solution* is improved, the magnitude of the perturbation is reduced to its smallest size to better explore the vicinity of this new *best_solution*.

The stopping criterion of the algorithm could either be the total CPU time (which is most often used), the CPU time elapsed from the last improvement, a given number of iterations or a combination of those.

Apart from the maximum magnitude for a perturbation, only a local search method and a perturbation scheme are required to apply VNS to a problem, which makes it rather easy to implement. In the case of MDS, the local search was a simple gradient search. As applying small perturbations to each of the stimuli would not have enough impact, the basic perturbation scheme used consists in completely moving a stimuli at random. This approach takes the combinatorial underlying structure of MDS into account. The description of the VNS implementation used here is described in (Caporossi & Taboubi, 2005).

The results obtained suggest that this optimization technique is rather robust for any Minkowski distance and for different dimensions. It is thus well adapted for a numerical study on ways to determine the best Minkowski parameter to use.

3 Monte Carlo analysis

To study the impact of the Minkowski parameter upon the result, a Monte Carlo simulation was conducted.

In a similar way to Lee (2001), 20 stimuli were generated in a hypercube with dimension 2, 3 or 4. For each dimension, 5 initial configurations were generated using different random seeds. The dissimilarities between stimuli were then computed with Minkowski parameter $l = 1, 1.5, 2, 2.5$ and 3, the letter l is used here to denote the Minkowski parameter used for the generation in order to avoid confusion with the the parameter used when attempting to recover the configuration (r). A random uniform noise with magnitude e: ±0% (noiseless data), ±5% and ±10%, was then added to each original dissimilarity. The VNS algorithm was used to find configurations minimizing the raw stress for each of these 225 problems with Minkowski parameter $r = 1.0 \ldots 5.0$ by increments of 0.1; which consists in solving 9225 problems by VNS (each run was limited to 120 seconds maximum CPU time on a SUN with Dual Core AMD 2200 MHz processor and 4 Go memory running Linux operating system).

4 Results

The curves indicating the raw stress value as a function of the Minkowski parameter are similar for the different problems, which tend to indicate that the conclusions we can make from these simulations are not due to random effects. Except some rare cases for extremal r values in higher dimensions, the curves are very smooth and regular, which indicates a rather reliable performance of the optimization. Figure 1 represents one of the 225 curves created to study the problem. Here, the dimension was 2, the Minkowski parameter used for the generation of the problem was $l = 2$ and 10% noise was added.

Two remarks could be made after a quick look at the curve from Figure 1:

- even if the data was generated with Minkowski parameter 2, the minimum stress is not achieved for $r = 2$. Furthermore, $r = 2$ seems to correspond to a local maxima. This experiment confirms a remark by Shepard (1974) stating that *"while finding that the lowest stress is attainable for $r = 2$ may be evidence that the underlying metric is Euclidian, the finding that the lowest stress is attainable for a value of r that is much smaller or larger may be artifactual"* (this phenomena will be referred to as the *Euclidian artefact*).

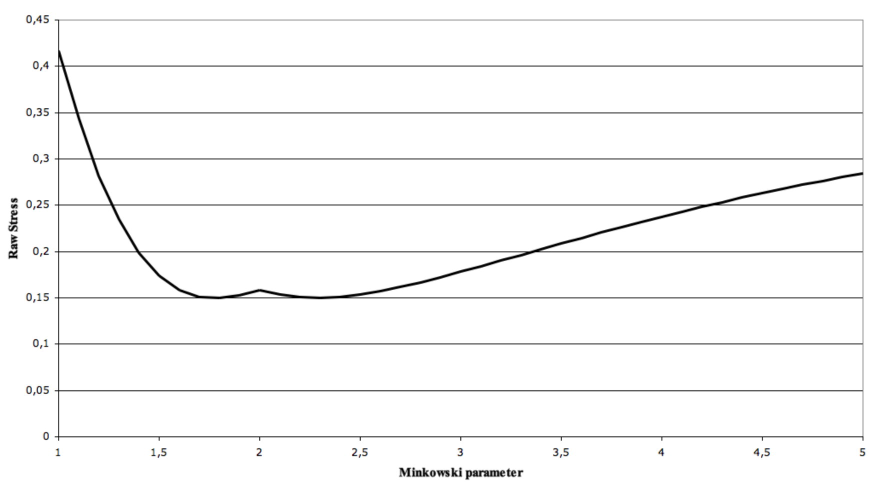

Figure 1: Minimal stress for (±10%) noised data generated with $r = 2$ in the plane according to the Minkowski parameter used for the MDS algorithm.

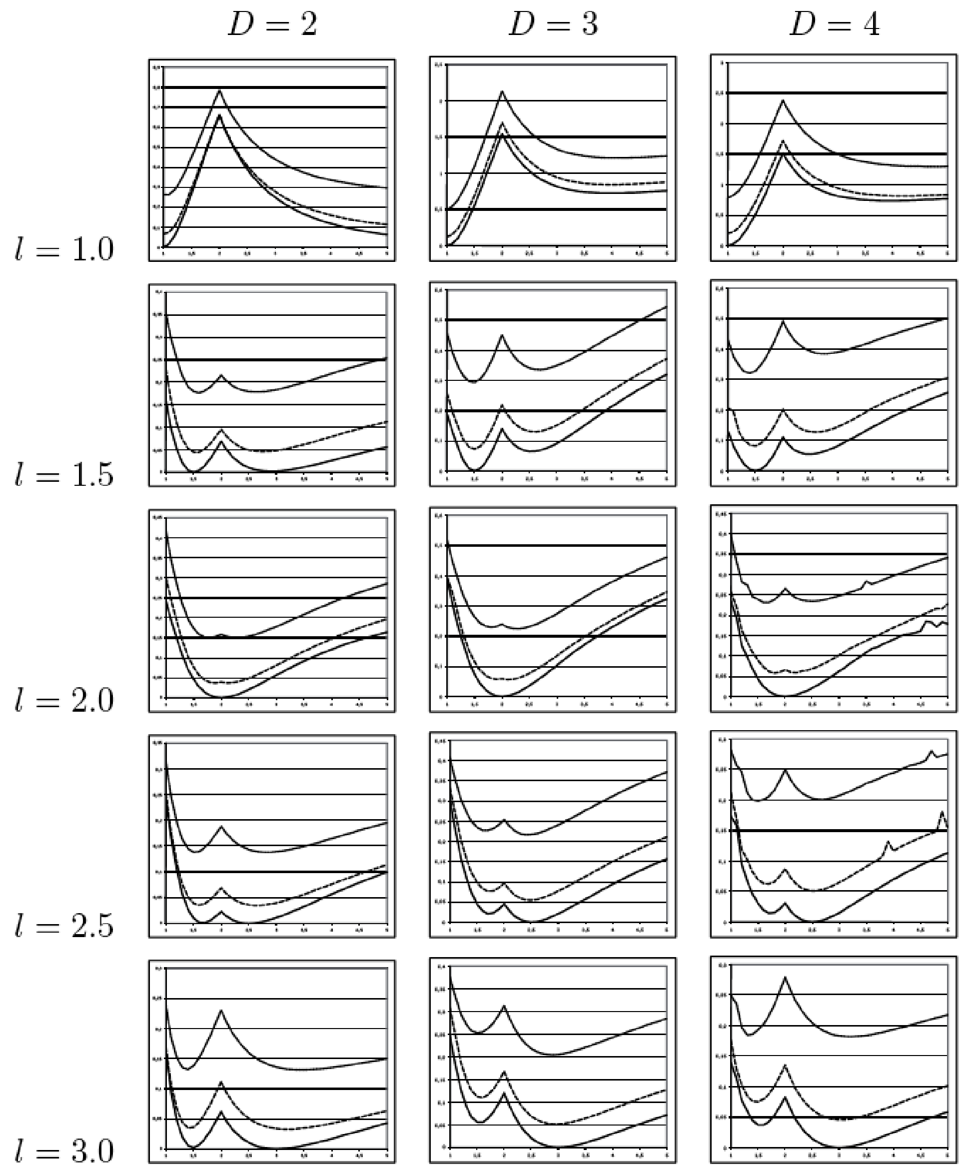

Figure 2: Curves representing the raw stress as a function of the parameter r. Dots are associated to $e = \pm 10\%$ (above), dashed lines to $e = \pm 5\%$ (middle) and solid lines to $e = 0\%$ (below). Each figure corresponds to a value of the Minkowski parameter for the generation l (rows) and a dimensionality D (columns).

- If we note r and r^* two Minkowski parameters, r and r^* correspond to dual norms if

$$\frac{1}{r} + \frac{1}{r^*} = 1 \tag{4}$$

  is respected. Shepard (1974) stated that the minimal stress value attained for a given r is close to that obtained for r^* if r and r^* correspond to dual norms. This statement was confirmed if the dimension $D = 2$.

For moderately noised data, except when the dimension is 2, the value r for which the raw stress S^* is minimized is close to l. However, if $l = 2$, due to the *euclidian artifact*, we find two distinct values r and r^* minimizing S^*. Note that in any case, the two local minima for r correspond to dual norms, even if one of them is clearly better (this property could be used to better identify the best value for r).

When analysing more deeply the curves obtained for various dimensions, Minkowski parameter for the generation and noise level, the first of the above mentioned always holds, except in the case data is noiseless and was generated with $l = 2$. However, the second was only observed in dimension $D = 2$.

When noise increases, the *euclidian artifact* increases its effect and the curve changes from a "U" shape (for noiseless data) to the "W" shape curve. To see what happens for very important noise, we generated random numbers instead of dissimilarities in a 10 stimuli dataset. The stress curve obtained on Figure 3 has a reversed "V" shape with maximum at $r = 2$, which could be expected.

As no Minkowski parameter is adapted to such data, it is not surprising to notice that the *Euclidian artifact* plays a dominant role. This remark does not mean that noised data could not analyzed with MDS, but rather that the impact of the Minkowski parameter is then small compared to the noise and the choice of the r parameter will not have an important impact on the MDS analysis. As a consequence, using $r = 1$ is certainly a reasonable choice.

5 Suggested procedure to identify the value r to use

The current study was not intended to identify the proper dimensionality of the data, but to concentrate on the Minkowski parameter. Test was therefore achieved for the same dimensionality as the generation. However, as this parameter has an impact on the potential results, it must be considered when the exact dimension is not known.

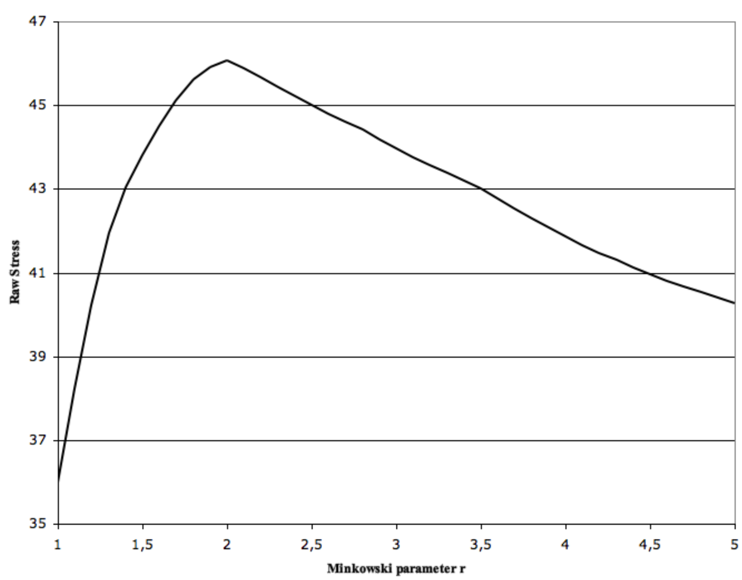

Figure 3: Minimal stress for random dissimilarities as a function of the Minkowski parameter.

From some isolated tests, we found some cases of data generated in dimension 4 with $l = 2$ for which the stress curve on dimension 2 suggested $l = 1$ (or that the data was very noisy). Even for these cases, however, the simple scree test for the dimension indicates the correct value regardless to the value r used. A good way to proceed, in case both the dimension D and the Minkowski parameter r are unknown, is to identify the dimension first.

If the dimension is 2, it will be difficult to identify wether $1 < r < 2$ or $r > 2$ as the stress will be close for r and r^* if $\frac{1}{r} + \frac{1}{r^*} = 1$. If the noise is very important, or if the best Minkowski parameter is 1, the curve will look like a reversed "V". In the case of very noisy data, the effect of the Minkowski parameter will be overwhelmed by the *Euclidian artefact*, and the difference between distances (according to r) is smaller than the noise. The choice of the Minkowski parameter will then have small impact upon the MDS analysis. If the noise is moderate and the Minkowski parameter is different than 1, the curve will have a "W" shape, the two local minima corresponding to r_1 and r_2. If the noise is small enough, the best solution indicates the best Minkowski parameter. If the stress for r_1 and r_2 are close, is may either be because the best r to choose is 2 or because of the noise. If there is a difference between the stress for r_1 and r_2, choose the one with best stress.

6 Applications to real data

To evaluate the procedure and help the researcher better understand the data, the proposed methodology was applied to two real datasets: the "Morse code data" (Rothkopf, 1957) and the "Nations" data (Wish, 1971).

6.1 The Morse code data

The original dataset (Rothkopf, 1957) is a confusion matrix. Letters were quickly transmitted to a morse operator. The matrix indicates the number of times each letter was recorded by the operator given the letter that was sent. This information is not symmetric. The transformation used here was the same as used by Hubert et al. (1997) and Brusco (2001), *i.e.,* from the asymmetric confusion matrix $C = \{c_{ij}\}$, compute the dissimilarity matrix $\Delta = \{\delta_{ij} = 2 - (c_{ij} + c_{ji})\}$.

The problem was first solved for $D = 1 \dots 6$ and for $r = 1.0$ to $r = 6.0$ by increments of 0.1. A quick look at the curve from Figure 4 indicates that the dimension for this data is most likely 2 or 3.

Considering dimension is 2, we draw the curve on Figure 5. This figure could typically be associated to $r = 1$ if the dimension is (close to) appropriate and the noise is moderated, which is a reasonable hypothesis in this case.

6.2 The nations similarity data

Wish (1971), asked 18 student to rate the similarity between 12 nations on a 1-9 scale (1=very different, 9=very similar). Average ratings were subtracted from 10 in order to get dissimilarity values. Due to the small number of stimuli, using a dimension higher than 2 would not be reasonable. The stress curve made with this data is represented on Figure 6.

Again, the most appropriate Minkowski parameter seems to be 1.

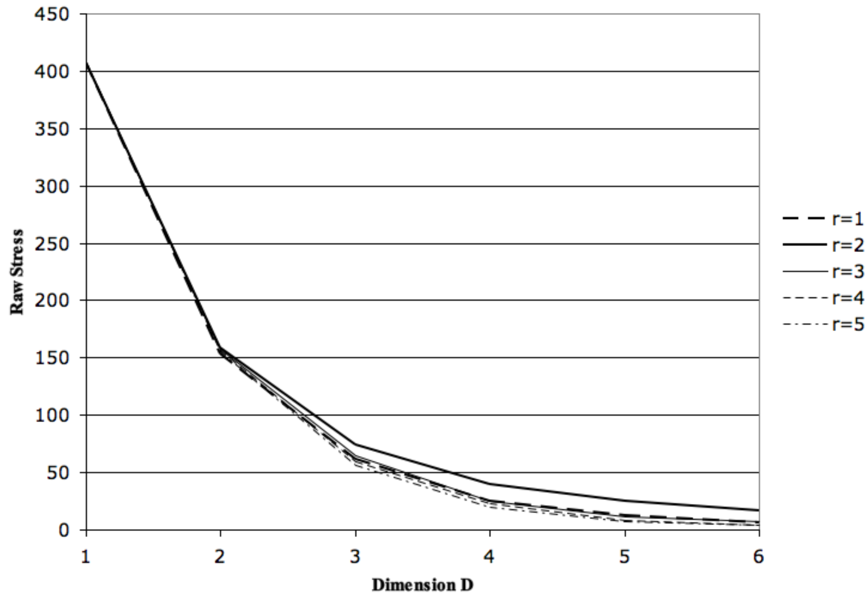

Figure 4: Minimum raw stress as a function of the dimension for various Minkowski parameters (r).

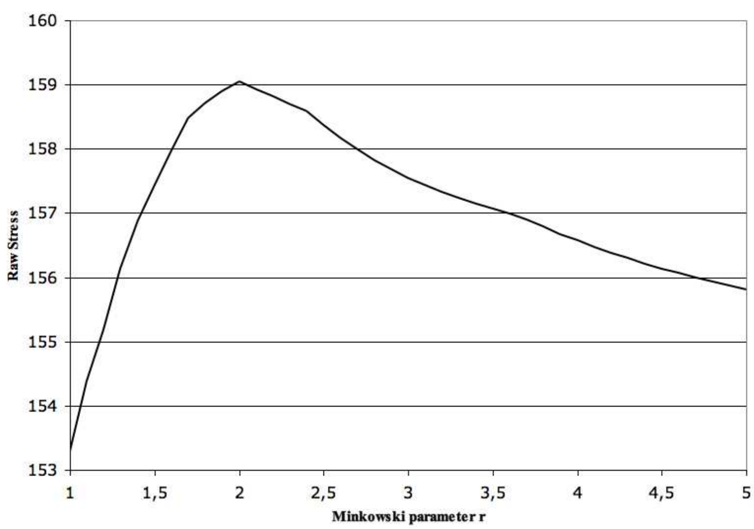

Figure 5: Minimal stress for the Morse code data as a function of the Minkowski parameter.

7 Conclusions

From a practical point of view, determining the value l from the final raw stress obtained is not always possible. However, there are several cases where it is possible: if $l = 1.0$, the curve has a reversed "V" shape with a maximum for $r = 2$. In the other cases, as soon as the data has some noise, the curve has a "W" shape with a local minima at $1 < r_1 < 2$ and $r_2 > 2$. Our result tend to show that the relation $\frac{1}{r_1} + \frac{1}{r_2} \approx 1$ holds.

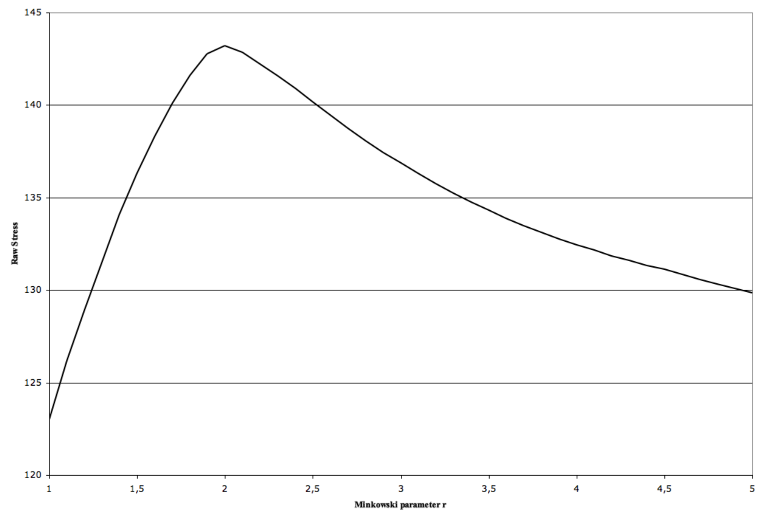

Figure 6: Minimal stress for the "nations" data as a function of the Minkowski parameter.

When the dimension $D = 2$, it is difficult to identify the original l value due to the relation (4) by which two configurations with different value r have approximatively the same stress value. When the dimension increases, the two values r_1 and r_2, which still approximatively respect the relation (4), corresponds to different stress values, which indicates which of r_1 and r_2 is close to the original l value. Unfortunately, when noise becomes important, it is difficult to identify l from r_1 and r_2. However, this is not surprising as different values l could correspond to equivalent dissimilarities after noise is added.

A complete analysis with a large number of real datasets would certainly be interesting but already, looking at few real data, gives the hint that choosing $r = 2$ is probably not the best choice in many circumstances.

References

Arabie, P. (1991). Was Euclid an Unnecessarily Sophisticated Psychologist? *Psychometrika, 56(4)*, 567–587.

Borg, I., & Groenen, P.J.F. (2005). *Modern Multidimensional Scaling: Theory and Applications* (2nd ed.). New York: Springer.

Brusco, M.J., & Stahl, S. (2000). Using Quadratic Assignment Methods to Generate Initial Permutations for Least-Squares Unidimentional Scaling of Symmetric Proximity Matrices. *Journal of Classification, 17*, 197–223.

Brusco, M.J. (2001). A Simulated Annealing Heuristic for Unidimensional and Multidimensional (City-Block) Scaling of Symmetric Proximity Matrices. *Journal of Classification, 18*, 3–33.

Brusco, M.J. (2002). Integer Programming Methods for Seriation and Unidimensional Scaling of Proximity Matrices: A Review and Some Extensions. *Journal of Classification, 19*, 45–67.

Brusco, M.J., & Stahl, S (2005). Optimal Least-Squares Unidimentional Scaling: Improved Branch-and-Bound Procedures and Comparison to Dynamic Programming. *Psychometrika, 70(2)*, 253–270.

Brusco, M.J. (2006). On the Performance of Simulated Annealing for Large-Scale L2 Unidimensional Scaling. *Journal of Classification, 23*, 255–268.

Caporossi, G., & Taboubi, S. (2005). A Variable Neighborhood Search Algorithm for Multidimensional Scaling in Arbitrary Norm. *Les Cahiers du GERAD, G–2005–60.*

Cox, T.F., & Cox, M.A.A. (2001). *Multidimensional Scaling* (2nd ed.). Chapman and Hall/CRC.

de Leeuw, J. (1988). Convergence of the Majorization Method for Multidimensional Scaling. *Journal of Classification, 5*, 163–180.

Groenen, P.J.F., Heiser, W.J., & Mathar, R. (1995). The Majorization Approach to Multidimensional Scaling for Minkowski Distances. *Journal of Classification, 12*, 3–19.

Groenen, P.J.F., & Heiser, W.J. (1996). The Tunneling Method for Global Optimization in Multidimensional Scaling. *Psychometrika, 61(3)*, 529–550.

Groenen, P.J.F., Heiser, W.J., & Meulman, J.J. (1999). Global Optimization in Least-Squares Multidimensional Scaling by Distance Smoothing. *Journal of Classification, 16*, 225–254.

Hansen, P., & Mladenović, N. (1997). Variable Neighborhood Search for the p-Median. *Location Science, 5(4)*, 207–226.

Hansen, P., & Mladenović, N. (2001). Variable Neighborhood Search: Principles and Applications. *European Journal of Operations Research, 130(3)*, 449–467.

Hubert, L.J., & Arabie, P. (1986). Unidimensional Scaling and Combinatorial Optimization. In J. de Leeuw, W. Heiser, J. Meulman, & F. Critchley (Eds.), *Multidimensional Data Analysis*, Leiden, The Netherlands: DSWO Press, 181–196.

Hubert, L.J., Arabie, P., & Hesson-Mcinnis, M. (1992). Multidimensional Scaling in the City-Block Metric: A Combinatorial Aproach. *Journal of Classification, 9*, 211–236.

Hubert, L., Arabie, P., & Meulman, J.J. (1997). Linear and Circular Unidimentional Scaling for Symmetric Proximity Matrices. *British Journal of Mathematical and Statistical Psychology, 50*, 253–284.

Hubert, L.J., Arabie, P., & Meulman, J.J. (2002). Linear Unidimensional Scaling in the L2-Norm: Basic Optimization Methods Using MATLAB. *Journal of Classification, 19*, 303–328.

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

Kruskal, J.B.(1964a). Nonmetric Multidimensional Scaling: a Numerical Method. *Psychometrika, 29(2)*, 115–129.

Lau, K.-N., Leung, P.L., & Tse, K.-K.(1998). A Nonlinear Programming Approach to Metric Unidimensional Scaling. *Journal of Classification, 15*, 3–14.

Lee, M.D.(1997). The Connectionnist Construction of Psychological Spaces. *Connection Science, 9(4)*, 323–351.

Lee, M.D. (2001). Determining the Dimensionality of Multidimensional Scaling Representations for Cognitive Modeling. *Journal of Mathematical Psychology, 45*, 119–166.

Mladenović, N., & Hansen, P. (1997). Variable Neighbourhood Search. *Computers and Operations Research, 24(11)*, 1097–1100.

Murillo, A., Vera, J.F., & Heiser, W.J. (2005). A Permutation-Translation Simulated Annealing Algorithm for L1 and L2 Unidimensional Scaling. *Journal of Classification, 22*, 119–138.

Pliner, V. (1996). Metric Unidimensional Scaling and Global Optimization. *Journal of Classification, 13*, 3–18.

Rothkopf, E.Z. (1957). A Measure of Stimulus Similarity and Errors in Some Paired-Associate Learning Tasks. *Journal of Experimental Psychology, 53*, 94–101.

Shepard, R.N. (1957). Stimulus and Response Generalization: a Stochastic Model Relating Generalization to Distance in Psychological Space. *Psychometrika, 22(4)*, 325–345.

Shepard, R.N. (1962). The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. I. *Psychometrika, 27(2)*, 125–140.

Shepard, R.N. (1962a). The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. II. *Psychometrika, 27(3)*, 219–246.

Shepard, R.N. (1974). Representation of Structure in Similarity Data-Problems and Prospects. *Psychometrika, 39(4)*, 373–421.

Sherman, C.R. (1972). Nonmetric Multidimensional Scaling: A Monte Carlo Study of the Basic Parameters. *Psychometrika, 37(3)*, 323–355.

Simantiraki, E. (1996). Unidimensional Scaling: A Linear Programming Approach Minimizing Absolute Deviation. *Journal of Classification, 13*, 19–25.

Vera, J.F., Heiser, W.J., & Murillo, A.(2007). Global Optimization in Any Minkowski Metric: A Permutation-Translation Simulated Annealing Algorithm for Multidimensional Scaling. *Journal of Classification, 24*, 277–301.

Wish, M. (1971). Individual Differences in Perceptions and Preferences among Nations. In C.W. King & D.J. Tigert (Eds.), *Attitude Research Reaches New Heights*, Chicago: American Marketing Association.