Austrian Journal of Statistics
July 2022, Volume 51, 1–9.
http://www.ajs.or.at/
doi:10.17713/ajs.v51i4.1423

A Note on Procrustean Fittings of Noisy Configurations

Ingwer Borg
WWU Münster

Patrick Mair
Harvard University

Abstract

When comparing two or more multidimensional scaling (MDS) configurations, one usually first eliminates meaningless differences by Procrustean transformations. Such fittings lead to a number of unresolved issues such as the typical shrinkage of the fitted configuration relative to the target or how to interpret major similarity measures under various conditions of noise in the data. We here prove that the shrinkage ratio is equivalent to the correlation of the coordinates of the target and the fitted configuration. Thus, in real-life applications, the fitted configuration is always smaller than the target configuration. Both coefficients approach 0 as the noise level goes up. The congruence coefficient of the configurations' distances, in contrast, remains at a high level even in case of pure noise, falsely suggesting that the configurations are somewhat similar. This is important information for the user of Procrustean analyses.

Keywords: Procrustes, MDS, configurational similarity.

1. Introduction

A frequent issue when using multidimensional scaling (MDS) is comparing two or more MDS solutions. Typical examples are studying to what extent an MDS result can be replicated with new data or whether MDS structures are similar in cross-cultural research. Consider a case. Borg and Braun (1996) were interested in work values of East and West German employees shortly after Germany re-united in 1990. They ran a survey, asking the respondents to rate 13 components of their work life (such as "high income" or "good chances for advancement") on a scale from "not important" to "very important" to them personally. Scaling the intercorrelations of the items for two samples leads to 2-dimensional MDS solutions for each data set (Figure 1)[1]. Obviously, the MDS plots generated by the MDS procedure are hard to compare, even though the configurations have only 13 points each. What must be ignored in such comparisons are *meaningless* differences of the MDS plots, such as their orientation or size. It is like comparing two maps of different size, and one is upside down, for example.

When comparing MDS plots, one can eliminate meaningless differences optimally by Procrustean transformations (Borg and Groenen 2005; Borg *et al.* 2018; Cox and Cox 1994;

[1] The configurations were generated using the R-package `smacof` (De Leeuw and Mair 2009; Mair, Groenen, and De Leeuw 2021), under choice of ordinal MDS with a random starting configuration. The Stress values are 0.172 (West) and 0.112 (East), indicating a highly significant model fit in both cases. For further details, see Borg, Groenen, and Mair (2018).

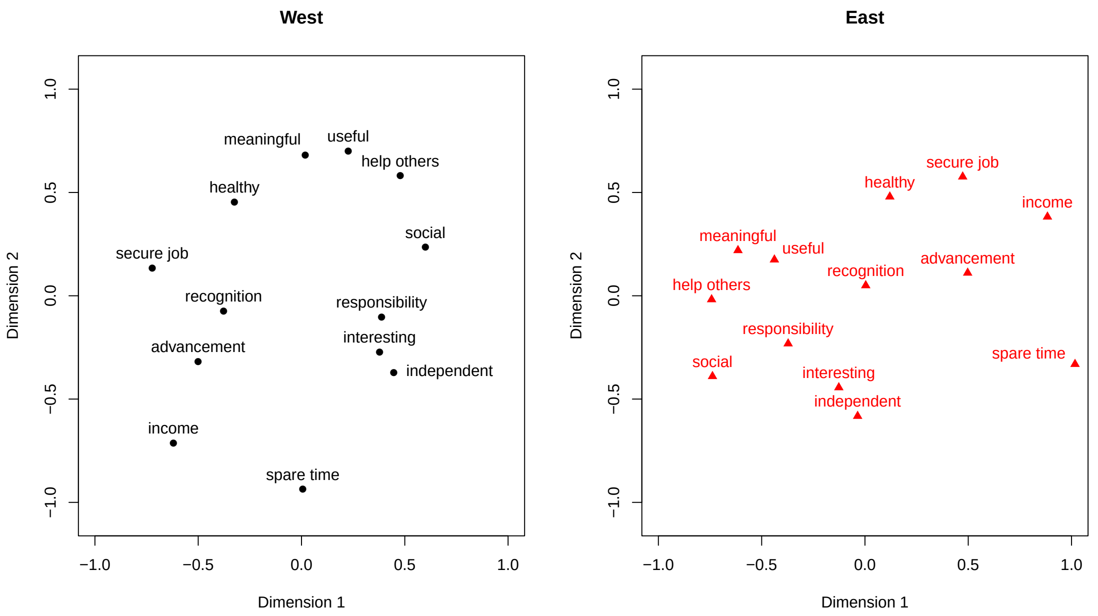

Figure 1: MDS configurations of persons' work values in West (left panel) and East Germany (right panel).

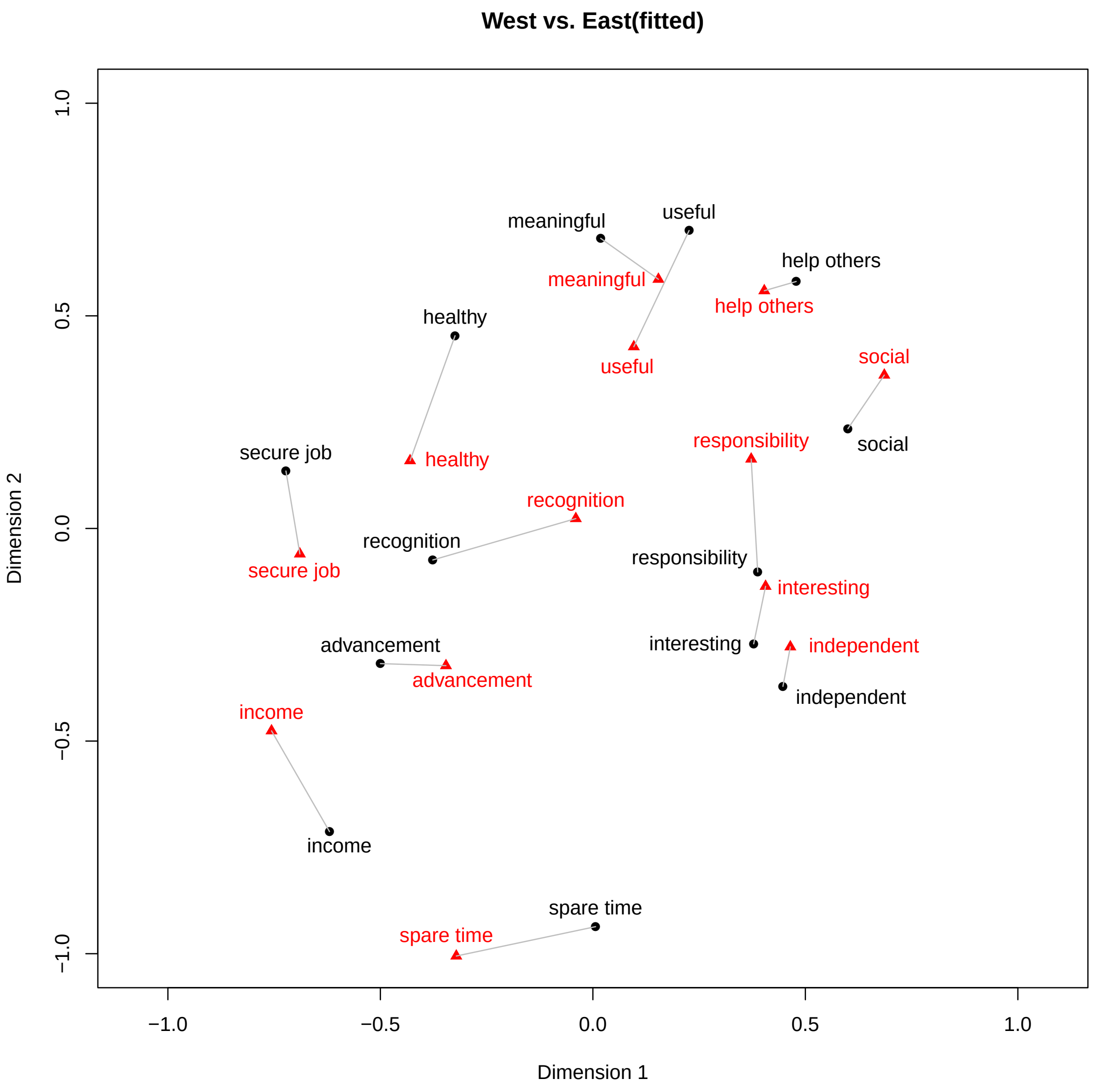

Figure 2: MDS configurations of Figure 1, with the East German configuration fitted Procrustean-wise to the West German configuration; corresponding points connected by line segments.

Gower and Dijksterhuis 2004): If configuration $\mathbf{X}$ is taken as the target, the other configuration, $\mathbf{Y}$, is rotated, reflected, translated, and adjusted in its size to optimally match $\mathbf{X}$. Such similarity transformations do not change the ratios of the distances among the points of the MDS configurations, but bring the configurations to an optimal match. Figure 3 shows that the West and the East German configurations are actually quite similar – after removing meaningless differences.

Formally, given the coordinate matrices $\mathbf{X}$ and $\mathbf{Y}$, both[2] of order $n \times m$, one wants to find the global scaling factor s, the rotation/reflection matrix $\mathbf{T}$, and the translation vector $\mathbf{t}$ such that the following loss function is minimized:

$$L(s, \mathbf{t}, \mathbf{T}) = tr\ [\mathbf{X} - (s\mathbf{YT} + \mathbf{et}^T)]^T[\mathbf{X} - (s\mathbf{YT} + \mathbf{et}^T)], \tag{1}$$

where $\mathbf{e}$ is a vector of ones. The solution $\hat{\mathbf{Y}} = s\mathbf{YT} + \mathbf{et}^T$ is found quickly and robustly as follows (Borg and Groenen 2005; Schönemann and Carroll 1970; Sibson 1978):

1. Compute $\mathbf{C} = \mathbf{X}^T\mathbf{JY}$, where $\mathbf{J}$ is the centering matrix $\mathbf{I} - n^{-1}\mathbf{ee}^T$ and $\mathbf{I}$ the identity matrix.
2. Compute the singular value decomposition $\mathbf{C} = \mathbf{U\Sigma V}^T$.
3. The optimal rotation matrix is $\mathbf{T} = \mathbf{VU}^T$.
4. The optimal dilation factor is $s = trace(\mathbf{X}^T\mathbf{JYT})/trace(\mathbf{Y}^T\mathbf{JY})$.
5. The optimal translation vector is $\mathbf{t} = n^{-1}(\mathbf{X} - s\mathbf{YT})^T\mathbf{e}$.

2. Configurational similarity and shrinkage

Procrustean fittings are used often in practical applications. They are easily realized with most statistical packages. They are also offered as `R` functions, for example in the `psych` package (Revelle 2021) or in `smacof` (De Leeuw and Mair 2009; Mair *et al.* 2021). `smacof` also computes the congruence coefficient (Tucker 1951) as an overall measure of the extent to which the configurations $\mathbf{X}$ and $\mathbf{Y}$ correspond to each other,

$$c(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i<j} d_{ij}(\mathbf{X})d_{ij}(\mathbf{Y})}{[\sum_{i<j} d_{ij}^2(\mathbf{X})]^{1/2}[\sum_{i<j} d_{ij}^2(\mathbf{Y})]^{1/2}} \quad , \tag{2}$$

with $d_{ij}(\mathbf{X})$ the Euclidean distance between points i and j in configuration $\mathbf{X}$. Since distances are always non-negative, c is always positive, with a maximum of 1. Thus, the minimum value of c is not 0, as occasionally stated (see, e.g., Borg and Groenen (2005, p. 400)).

Because it always holds that $c > 0$, the user may interpret this coefficient as a sign that $\mathbf{X}$ and $\mathbf{Y}$ are at least somewhat similar. Technically, perfect similarity means that the configurations are congruent, except for a possible difference in size. There is, however, no definition of minimal similarity or of zero similarity. Statistically, one may consider minimal similarity as the expected distance of corresponding points of $\mathbf{X}$ and $\hat{\mathbf{Y}}$, given that both $\mathbf{X}$ and $\hat{\mathbf{Y}}$ are fully random.

A similarity measure that is compatible with this notion of similarity is the Pearson correlation r of the corresponding coordinates x_{ia} of $\mathbf{X}$ and $\hat{y}_{ia}$ of the Procrustean-fitted $\hat{\mathbf{Y}}$. If both $\mathbf{X}$ and $\mathbf{Y}$ are centered, the optimal translation vector $\mathbf{t}$ is the null vector (Cox and Cox 1994, p. 94) so that $\hat{\mathbf{Y}}$ simplifies to $s\mathbf{YT}$ and the correlation r can be written as

$$r(\mathbf{X}, s\mathbf{YT}) \quad = \quad \frac{tr\ (\mathbf{X}^T\mathbf{YT})}{\sqrt{tr\ (\mathbf{XX}^T) \cdot tr\ (\mathbf{YTT}^T\mathbf{Y}^T)}}, \tag{3}$$

[2] Procrustean fittings can also be used for configurations that differ in the number of points and in their dimensionalities. If the number of points differ, one can compute the transformation operators based on corresponding sub-sets or centroids of points that represent similar content classes. In case of different dimensionalities, one can simply add column vectors with only zeroes to $\mathbf{X}$ or to $\mathbf{Y}$.

where $tr\ (\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$, the trace function of an $n \times n$ matrix $\mathbf{A}$. Because $\mathbf{T}$ is orthonormal, $tr\ (\mathbf{YTT}^T\mathbf{Y}^T) = tr\ (\mathbf{YY}^T)$ and, thus,

$$r(\mathbf{X}, s\mathbf{YT}) = \frac{tr\ (\mathbf{X}^T\mathbf{YT})}{\sqrt{tr\ (\mathbf{XX}^T) \cdot tr\ (\mathbf{YY}^T)}}. \tag{4}$$

Without loss of generality, one can also norm $\mathbf{X}$ to unit length. This yields

$$r(\mathbf{X}, \hat{\mathbf{Y}}) = \frac{tr\ (\mathbf{X}^T\mathbf{YT})}{\sqrt{tr\ (\mathbf{YY}^T)}}. \tag{5}$$

Practitioners are sometimes concerned by the observation that Procrustean-fitted configurations seem to be "too small", in particular when the number of points is large. In other words, the global scaling factor s appears to be incorrect so that the user of Procrustean fittings gets the impression that $\hat{\mathbf{Y}}$ should be dilated to make it as congruent to $\mathbf{X}$ as possible.

To check the validity of this impression, assume again that both $\mathbf{X}$ and $\mathbf{Y}$ are centered, and that $\mathbf{X}$ is also stretched to unit length (i.e., the Frobenius norm of $\mathbf{X}$ becomes $norm(\mathbf{X}) = \sqrt{tr\ (\mathbf{XX}^T)} = \sqrt{tr\ (\mathbf{X}^T\mathbf{X})} = 1$). The shrinkage of $\hat{\mathbf{Y}}$ relative to $\mathbf{X}$ can be measured by the shrinkage ratio

$$sr(\hat{\mathbf{Y}}) = \frac{norm(s\mathbf{YT})}{norm(\mathbf{X})} = norm(s\mathbf{YT}) = \sqrt{tr\ (s\mathbf{YTT}^T\mathbf{Y}^T s)} \tag{6}$$

$$= s \cdot \sqrt{tr\ (\mathbf{YY}^T)} \tag{7}$$

Replacing s with the optimal dilation factor from above, and using the assumption that $\mathbf{X}$ and $\mathbf{Y}$ are column-wise centered so that $\mathbf{X}^T = \mathbf{X}^T\mathbf{J}$ and $\mathbf{Y}^T = \mathbf{Y}^T\mathbf{J}$,

$$sr(\hat{\mathbf{Y}}) = \frac{tr\ (\mathbf{X}^T\mathbf{YT})}{tr\ (\mathbf{Y}^T\mathbf{Y})} \cdot \sqrt{tr\ (\mathbf{YY}^T)} \tag{8}$$

$$= \frac{tr\ (\mathbf{X}^T\mathbf{YT})}{\sqrt{tr\ (\mathbf{Y}^T\mathbf{Y})}}. \tag{9}$$

Because $tr\ (\mathbf{Y}^T\mathbf{Y}) = tr\ (\mathbf{YY}^T)$, we thus find that $sr(\hat{\mathbf{Y}}) = r(\mathbf{X}, \hat{\mathbf{Y}})$. Hence, the shrinkage ratio itself indicates the similarity of $\mathbf{X}$ and $\mathbf{Y}$, and the practitioners' impression that the fitted $\mathbf{Y}$ is "small" – but, of course, not "too small" – is correct in all but trivial applications. In the following, we will show how and why shrinkage occurs, and how it can be interpreted in practice.

3. Configurational similarity under noisy conditions

To check shrinkage in Procrustean fittings of noisy data structures, we run a set of simulation studies. First, we define a fixed $\mathbf{X}$ by randomly picking $n = 100$ points within a unit disk. We then define $\mathbf{Y}$ as $(1-g) \cdot \mathbf{X} + g \cdot \mathbf{E}$, with $0 \leq g \leq 1$. The coordinates of the noise matrix $\mathbf{E}$ are also picked randomly from the unit disk. The weight g is taken from $0, 0.1, ..., 1$ so that we have eleven $\mathbf{Y}$'s ranging from $\mathbf{Y}$ being equal to $\mathbf{X}$ to the case where $\mathbf{Y}$ is completely random and independent of $\mathbf{X}$.

For each combination of $\mathbf{X}$ and $\mathbf{Y}$, we find the optimal Procrustean match of the configurations. The configurations $\mathbf{X}$ and $\hat{\mathbf{Y}}$ are then displayed in one overlay plot, connecting corresponding points (i.e., the points representing the same rows of $\mathbf{X}$ and $\hat{\mathbf{Y}}$). These plots should show how and why $\hat{\mathbf{Y}}$ shrinks relative to its target $\mathbf{X}$. We also compute the fit measures c and r, and the shrinkage factor sr, and study how they are related.

We then repeat these analyses with $n = 15,000$ points. Overlay plots for that many points are uninterpretable, of course, because they are densely filled with points. However, the relation of the fit measures can be assumed to be stable and not dependent on the particular random sampling on which $\mathbf{X}$ and the noise matrix $\mathbf{E}$ are based.

Figure 3 shows the result of the simulation study with $n = 100$ points. The top left panel exhibits an overlay plot of the target configuration $\mathbf{X}$ and of a fitted $\mathbf{Y}$ configuration that contains little noise. Obviously, the target configuration (with points shown as black circles) and the fitted $\mathbf{Y}$ configuration (with points shown as red triangles) are highly similar. The corresponding points are close to each other: The various line segments visualize their distances.

When adding more noise to $\mathbf{Y}$, increasing the noise level to $g = 0.5$, the top right panel of Figure 3 shows considerably larger distances between corresponding points. Moreover, the red triangles that represent $\hat{\mathbf{Y}}$ move more towards the center of the plot. When choosing $\mathbf{Y}$ completely at random and independent of $\mathbf{X}$, one notices in the left bottom plot of Figure 3 that the fitted $\mathbf{Y}$-configuration shrinks strongly towards the center of the plot.

Because sr and r are algebraically equal, the right bottom plot of Figure 1 represents both r and sr on the Y-axis and c on the X-axis. One notes that r (and sr) and c are all equal to 1 in case of no noise or $g = 0$, that is if $\mathbf{X} = \hat{\mathbf{Y}}$. At the other end of the range of noise levels ($g = 1$), r (and sr) approach a value of 0 ($r = sr = 0.084$) while $c = 0.813$.

Figure 4 displays the relation of r, c, and sr for $n = 15,000$ points. One can see that the function curve is similar to the one in Figure 3 (right bottom panel) but due to the large number of points it is very smooth. For high noise levels, both r and sr are very close to zero ($= 0.011$), while $c = 0.818$ remains at a high level far above zero. One also notes that with a linearly increasing level of noise, the misfit of $\mathbf{X}$ and $\hat{\mathbf{Y}}$ accelerates.

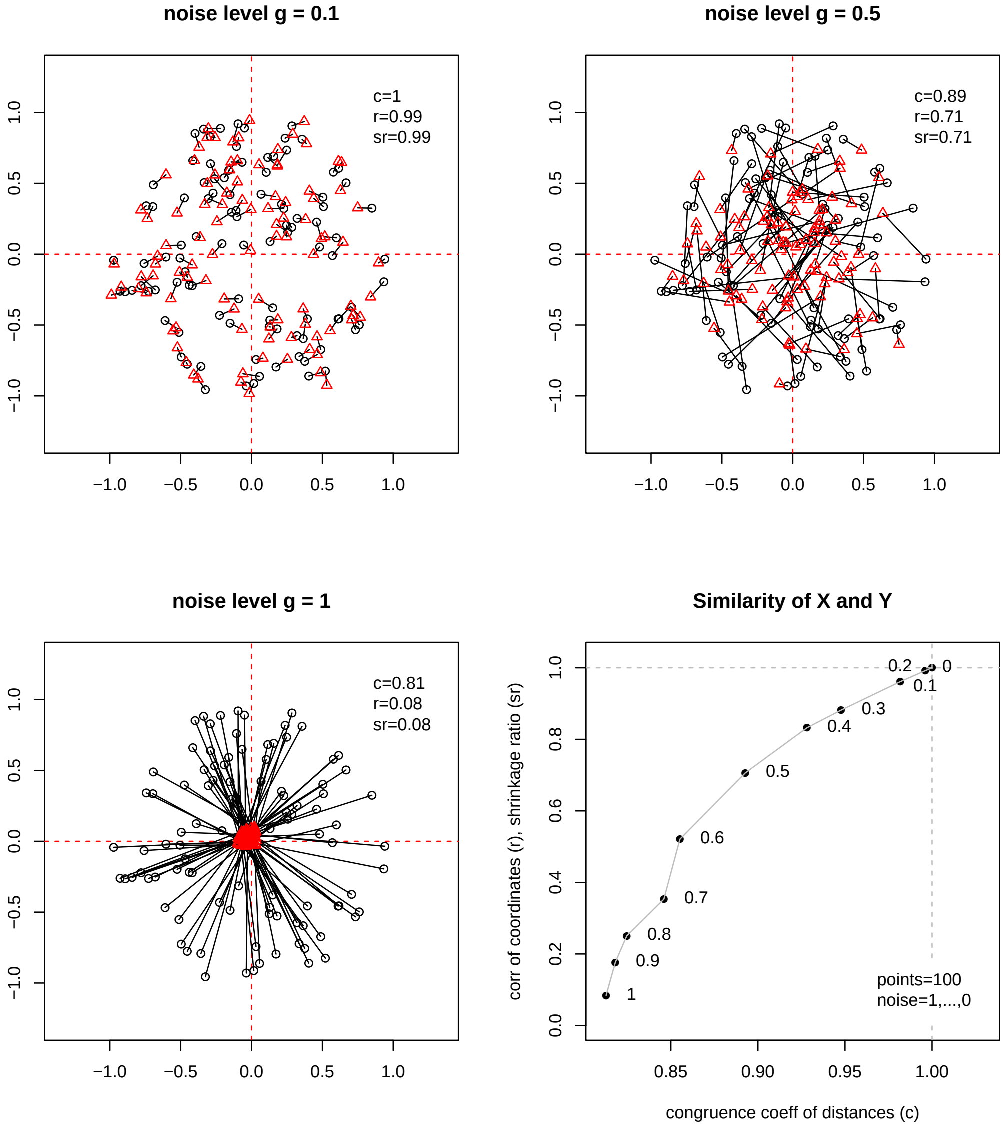

Figure 3: Three cases of fitting a noisy $\mathbf{Y}$ to $\mathbf{X}$; $n = 100$ points; noise level $g = 0.1$, $g = 0.5$, and $g = 1$; bottom right panel: plot of fit measures r and sr, resp., vs. c.)

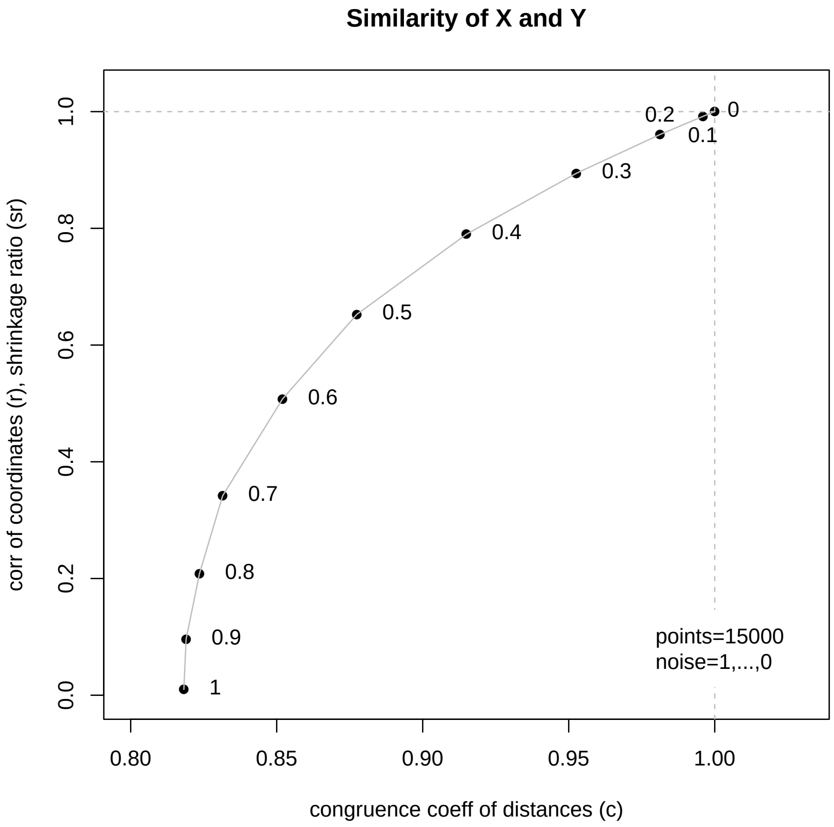

Figure 4: Fitting noisy $\mathbf{Y}$'s to $\mathbf{X}$; $n = 15,000$ points; noise level $g = 0.0, 0.1, ..., 1$; fit measures r and sr, resp., vs. c.)

4. Discussion

Our analyses have shown that the occasional observation that a configuration $\mathbf{Y}$, after it is fitted by Procrustean transformations to another configuration $\mathbf{X}$, tends to be "too small", is actually correct. The fitted $\mathbf{Y}$ is *always* smaller than $\mathbf{X}$ in real-life applications, and this simply expresses that the two configurations are not perfectly similar. Indeed, the larger the fitted $\mathbf{Y}$, $\hat{\mathbf{Y}}$, the more similar is $\mathbf{Y}$ to its target $\mathbf{X}$. For high noise levels and many points, the $\hat{\mathbf{Y}}$ tends to shrink to the centroid of $\mathbf{X}$, i.e. to the point that is closest to all points of $\mathbf{X}$.

The reason for the shrinkage of $\mathbf{Y}$ when fitting it Procrustean-wise to $\mathbf{X}$ is that a central dilation s of $\mathbf{Y}$ is a transformation that radially shifts all points in space relative to the origin. This transformation cannot eliminate random displacements of the points that shift the points independently of each other in any direction of space and, in particular, also across the centroid of $\mathbf{X}$ in opposite directions. With much noise, shrinking $\hat{\mathbf{Y}}$ towards the mean (in case of 1d) or towards a central cluster (in higher-dimensional spaces) is, thus, the best solution to optimally compensate for such displacements.

The shrinkage of $\hat{\mathbf{Y}}$ relative to $\mathbf{X}$ yields a proper measurement of the similarity of $\mathbf{X}$ and $\mathbf{Y}$. We have shown that the shrinkage ratio sr is numerically equivalent to the Pearson correlation $r(\mathbf{X}, \hat{\mathbf{Y}})$, a coefficient often used in practice to measure the similarity of two configurations.

Another finding of practical relevance is that the congruence coefficient c – also often used in applications – is difficult to interpret if the noise level is high. Indeed, c becomes almost useless in case of a large noise proportion, because, as Figure 4 shows, its values become very similar for all noise levels above 0.6. This problem is reduced if one converts c to the alienation coefficient $K(\mathbf{X}, \hat{\mathbf{Y}}) = \sqrt{1 - c^2}$ (Mair *et al.* 2021). However, K does not fully linearize the function in Figure 4 and it also leads to other problems. In particular, it has an upper bound that suggests that the alienation between $\mathbf{X}$ and $\hat{\mathbf{Y}}$ is substantial but not maximal. That this is the wrong interpretation is easy to see. For the noise level $g = 1$, Figure 4 shows a congruence value of $c = 0.813$. This corresponds to $K = 0.582$ and not to $K = 1$ as one might expect when considering the notion of "alienation". Hence, K may lead the user to overestimate the configurational similarity of $\mathbf{X}$ and $\hat{\mathbf{Y}}$. To avoid such misinterpretations we therefore recommend to always compare such coefficients to the similarity measures resulting from fitting a large number of random configurations to each other. Simulations are easily computed using `smacof`'s Procrustean fitting function.

If the user prefers the more familiar linear correlation of corresponding coordinates as a measure of configurational similarity, benchmarking is also recommended, because, in case of few points in $\mathbf{X}$ and $\mathbf{Y}$, Procrustean transformations will always lead to r values that suggest "some" similarity. For example, with $n = 13$ points as in the study by Borg and Braun (1996) discussed in the Introduction, purely random configurations $\mathbf{X}$ and $\mathbf{Y}$ with 13 points each can be expected to lead to an average similarity of $r = 0.366$ ($sd = 0.110$). In typical MDS applications, the number of points is rarely larger than 50 at most, and for 50 points, the expected Procrustean similarity is not $r = 0$ but $r = .190$ ($sd = .064$), as simulation studies show.

Finally, as a technical remark, we note here that picking $\mathbf{X}$ and $\mathbf{E}$ in the above simulations not from a unit disk but from a rectangular random distribution for each dimension (i.e., a unit square) does not change the results. This replicates similar findings by Langeheine (1982).

References

Borg I, Braun M (1996). "Work Values in East and West Germany: Different Weights but Identical Structures." *Journal of Organizational Behavior*, **17**, 541–555. URL `https://doi.org/10.1002/(SICI)1099-1379(199612)17:1+<541::AID-JOB822>3.0.CO;2-M`.

Borg I, Groenen PJF (2005). *Modern Multidimensional Scaling*. 2nd edition. Springer, New York.

Borg I, Groenen PJF, Mair P (2018). *Applied Multidimensional Scaling and Unfolding*. 2nd edition. Springer, New York.

Cox TF, Cox MAA (1994). *Multidimensional Scaling*. Chapman & Hall, London.

De Leeuw J, Mair P (2009). "Multidimensional Scaling Using Majorization: SMACOF in R." *Journal of Statistical Software*, **31**(3), 1–30. URL `http://www.jstatsoft.org/v31/i03/`.

Gower JC, Dijksterhuis GB (2004). *Procrustean Problems*. Oxford.

Langeheine R (1982). "Statistical Evaluation of Measures of Fit in the Lingoes-Borg Procrustean Individual Differences Scaling." *Psychometrika*, **47**, 427–442. URL `https://doi.org/10.1007/BF02293707`.

Mair P, Groenen PJF, De Leeuw J (2021). "More on Multidimensional Scaling in R: SMACOF Version 2." *Journal of Statistical Software*. Forthcoming.

Revelle W (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.1.9, URL `https://CRAN.R-project.org/package=psych`.

Schönemann PH, Carroll RM (1970). "Fitting One Matrix to Another under Choice of a Central Dilation and a Rigid Motion." *Psychometrika*, **35**, 245–256. URL `https://doi.org/10.1007/BF02291266`.

Sibson R (1978). "Studies in the Robustness of Multidimensional Scaling: Procrustes Statistics." *Journal of the Royal Statistical Society*, **B 40**, 234–238. URL https://doi.org/10.1111/j.2517-6161.1978.tb01669.x.

Tucker LR (1951). *A Method for Synthesis of Factor Analysis Studies.* PRS Report 984. ETS.

Affiliation:

Ingwer Borg
Fachrichtung Psychologie
Westfälische-Wilhelms Universität
Fliednerstr. 21
48149 Münster, Germany

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 51 *Submitted:* 2021-09-20
July 2022 *Accepted:* 2022-02-04