I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12355

Comparison Dimensions and Similarity: Addressing Individual Heterogeneity

Pavel Jelnov

MAY 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12355

Comparison Dimensions and Similarity: Addressing Individual Heterogeneity

Pavel Jelnov
Leibniz University Hannover and IZA

MAY 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Comparison Dimensions and Similarity: Addressing Individual Heterogeneity

How many comparison dimensions individuals consider when they are asked to judge how similar two different objects are? I address individual heterogeneity in the number of comparison dimensions with data from a laboratory experiment. I estimate the smallest number of dimensions such that objects may be represented in space where distance corresponds to similarity. I find that the mean smallest number of dimensions in real data is one standard deviation smaller than in randomly simulated data. Furthermore, I find that individuals who find the objects relatively similar to each other are also the ones who implicitly consider fewer dimensions.

JEL Classification: D12

Keywords: multidimensional scaling, similarity, latent dimensions, dimensionality

Corresponding author:
Pavel Jelnov
Wirtschaftswissenschaftliche Fakultät
Institut für Arbeitsökonomik
Leibniz Universität Hannover
Königsworther Platz 1
30167 Hannover
Germany

E-mail: jelnov@aoek.uni-hannover.de

1 Introduction

This paper asks how many comparison dimensions individuals consider when they are required to judge how similar two different objects are. The approach is based on multidimensional scaling (MDS), a concept where objects are represented as points in space such that similarity between objects corresponds to distance between points. I go beyond the existing MDS literature in allowing individual heterogeneity in the number of comparison dimensions.

This approach contributes to two bodies of research. First, the existing MDS studies consider individual heterogeneity in terms of preferences (a method known as *unfolding*) and in terms of weighting of pre-determined dimensions (a method known as *dimensional weighting model*). However, the existing literature does not address, to the best of my knowledge, the possibility that individuals may consider a different number of comparison dimensions. The second related literature is search theory. Search strategies essentially depend on the level of similarity between the alternatives. Recently, a few studies advanced this literature by addressing search in a multidimensional environment (Chiappori et al. (2012), Coles and Francesconi (2018)). The present paper is related to both contexts, as it links similarity with dimensionality.

In particular, I show that dimensionality is negatively correlated with perceived similarity. Why should we care? For example, this finding suggests that a small number of attributes that the consumer considers or knows about is associated with perception of the products as similar to each other. My results are also relevant to the above-mentioned search theory. A fundamental result in this theory is that a higher level of similarity between alternatives is associated with a shorter duration of search. With respect to the present paper's findings, dimensionality can be used to account for the level of perceived similarity in search models.

Interpretation of similarity, one of the fundamental issues in the research of perception, can be traced back through the seminal works of Guttman (1968) and Tversky (1977). The question in the present paper is whether similarity is related to the number of latent dimensions an individual considers when she or he is asked to judge how similar two different objects are. Intuition, inspired by triangle inequality, implies that a larger number of dimensions should be associated with a longer "distance" between objects. I provide evidence for this intuition in data. I collect a data set that consists of rankings-by-similarity and direct similarity judgments for a small set of objects. The participants of the study do not report which and how many comparison dimensions they consider. However, I estimate the implicit number of considered dimensions on individual level. That is, for each participant, I find the smallest number of dimensions n needed to represent the objects as points in $\mathbb{R}^n$, such that the distance between the points decreases in the similarity between the corresponding objects. The two following findings pose the results of this study. First, the smallest number of dimensions is not a spurious statistic, mechanically derived from rankings of pairs of objects by similarity. By contrast, I find that the average smallest number of dimensions in real data is one standard deviation smaller than in randomly simulated data. Second, individuals who find the objects more similar to each other by average are also the ones who implicitly consider fewer dimensions.

In particular, data was collected in a laboratory where each participant performed two tasks with the same set of seven objects. In the first task, the participants had to order pairs of objects by similarity. In the second task, the participants had to evaluate similarity of each pair on a scale from “not similar at all” to “very similar.” The collected data is utilized in three stages of analysis. First, the smallest number of dimensions is derived from

the individual's ranking of the pairs of objects by similarity. Second, the mean estimated number of dimensions is compared with the mean number of dimensions from randomly simulated data. This comparison leads to a conclusion that individuals consider significantly fewer dimensions than if data are randomly simulated. Third, the estimated number of dimensions is regressed on the average similarity between the objects, separately evaluated by the same individuals in the second task. The result of this regression is that an increase of one unit in the average similarity between the objects is associated with a 10 to 30 percentage points lower probability to consider a high number of dimensions.

In the remainder of the paper, I first briefly introduce the MDS concept and my innovation in its implementation. I proceed with a description of the study, the method of data analysis, and results. Finally, I provide a brief discussion of the main finding before concluding the paper.

2 The concept of the smallest number of dimensions

I estimate the number of comparison dimensions from the individual ranking of pairs of objects by similarity. For many decades, the analysis of similarity has been dominated by geometric models (Tversky (1977)). One family of such methods, called multidimensional scaling (MDS), represents similarity between objects as distances between points in a low-dimensional space. Objects are placed in space, such that distance between any two objects corresponds to the level of similarity between them. An important difference of MDS from other methods of dimensionality reduction is that the dimensions are not reduced from observed variables.

The classical ordinal MDS considers a set $\mathcal{L}$ of objects and the input is the rank order r of similarity between objects, such that $r = 1$ for the most similar pair and $r = \binom{|\mathcal{L}|}{2}$ for the least similar one. The output is a set of points $\{x_i\} \subset \mathbb{R}^n$ (non-real spaces and non-Euclidean distance metrics may also be considered) that monotonically translates rank of similarity into increasing distance between points. The solution minimizes a loss function called "stress" given the number of dimensions.

In the classical MDS, the number of dimensions is predetermined. For example, in a frequently cited review of pattern recognition methods, Jain et al. (2000) define MDS as a method "to represent a multidimensional data set in two or three dimensions." A similar definition is given in Borg et al. (2017). The main motivation for this restriction to a low number of dimensions is the use of MDS for visualization of similarity. As a result of the restriction of the number of dimensions to two or three, the MDS configuration is not perfect and generates a positive stress.

Moreover, in the basic MDS the similarities are averaged over individuals and no individual heterogeneity in terms of the space where the objects should be located is allowed. More complex models assume that individuals may be heterogeneous in terms of their preferences over the objects (the unfolding method) or in terms of weights they put on the pre-determined dimensions (dimensional weighting model). For a detailed discussion of the different MDS and unfolding models, see Carroll and Green (1997), Borg and Groenen (2005) and Borg et al. (2017).[1]

[1]To the best of my knowledge, the most comprehensive reviews of MDS in marketing research date back to Cooper (1983) and Carroll and Green (1997). None of the reviewed studies estimate the number of dimensions on individual level. By contrast, they seek interpretation of pre-specified dimensions (Carroll and Green (1997)). Later literature that links similarity and marketing in general and is related to MDS is particular includes Leftkoff-Hagius and Mason (1993), Bijmolt et al. (1998), and Fuchs and Diamantopoulos (2012). This literature is quite rare but, probably, not because of lack of interest. The problems in using MDS

In the present study, I relax the assumption that all individuals consider the same space when they compare the objects. This relaxation generates two differences from the existing MDS models. First, MDS is implemented on individual rather than on aggregate level. Second, the number of dimensions is not set in advance.

What is the smallest number of dimensions needed to perfectly (with zero stress) represent the ordinal similarity data such that $||x_i - x_j|| < ||x_k - x_l||$ iff $r_{ij} < r_{kl}$, where $i \neq j$ and $k \neq l$? Three objects may always be represented in a single dimension. Generally, the smallest number of dimensions for k objects is at most $k - 2$ (Guttman (1968)). In the present study, I consider a set of seven objects. Thus, the smallest number of dimensions needed to represent data perfectly is between one and five. I find evidence for variation in this number on individual level. In my data set of 136 observations, the mean smallest number of dimensions is 3.45 and the standard deviation is 0.58. In the core of the analysis, I seek a behavioral interpretation of this variation.

3 The study

Data for this article was collected through a study conducted at the department of economics of one of German universities.[2] The focus was on similarity between fruits. Participants were exposed to pictures of seven fruits: apple, strawberry, banana, apricot, orange, kiwi, and cherry.[3] Participants were instructed to consider the level of similarity between the fruits

in marketing research include the limited ability of MDS procedures to fully portray the structure in such volumes of data (DeSarbo et al. (1994)) and "the problem of controllability/manipulability of the dimension" (Carroll and Green (1997)). For an up-to-date review of MDS applications in different disciplines, see Lin and Fong (2019).

[2]The study took place between December 10 and December 15, 2015.

[3]All pictures had a white background.

and were not restricted to any definition of similarity.

Setup

The participants were all students who were each given 20 Euros for their participation. The study consisted of ten sessions, held at the same room, with up to 19 volunteers participating in each session. The total number of participants was 138; data of two of the participants were lost for technical reasons. Thus, the sample consists of 136 observations. The sample with control variables of age and gender consists of 132 observations, because four participants did not submit the personal questionnaire or did not fill in gender and/or age. The duration of each session was one-and-a-half hours, of which about fifteen minutes (6 minutes for the fastest participant and 33 minutes for the slowest one) were devoted to working on the present study. The rest of the time was devoted to listening to instructions (these included a presentation and oral explanations), participating in a study for another research project, and completing a personal questionnaire. Participants worked on computers, separated by opaque barriers, and did not communicate with each other during the study. The study was conducted using a website hosted free of charge on the somee.com server. The language of the website and of the explanations was German.

First task

In the first part of the study, participants had to order the pairs of fruits from the most similar pair to the least similar one. To this end, they were exposed to a sequence of screens, each of which presented two pairs of fruits. When exposed to each screen, a participant had to choose the more similar pair. For example, one had to decide which two fruits are more

similar to each other - grapefruit and apple or grapefruit and banana (see a screenshot in Figure 1). The order of screens followed the bubble sort algorithm (Clapham and Nicholson (2009); see Appendix for the algorithm). This is a classical algorithm for sorting a vector. The algorithm avoids repeated comparisons, does not allow ties, and does not generate loops that would make further analysis impossible. For a short vector of $\binom{7}{2} = 21$ elements, the bubble sort is a relatively fast algorithm (requires a low number of comparisons). The output for each participant is a full and strictly sorted-by-similarity vector of 21 pairs of fruits. The algorithm is adaptive, and participants had to go over a different number of screens. It varied between 39 and 159, with both the median and the mean being 103 screens. After every 10 screens, there was a ten-second break during which an unrelated to the study picture appeared. This was done to break the monotony of the procedure.

Figure 1: A screenshot of part 1 of the study

Second task

In the second part of the study, participants were sequentially exposed to 21 screens (in the same order for all participants), each of which presented pictures of one of the pairs of fruits. For each pair, participants had to choose the level of similarity between the two fruits: not similar at all, not similar, somewhat similar, similar, very similar (see a screenshot in Figure 2).

To summarize, for each participant, the output of the study consists of the 21 pairs of fruits ordered by similarity and the direct evaluation of the similarity of each pair. In addition, for each participant, the website recorded her working time and the number of

screens she was exposed to.

Figure 2: A screenshot of part 2 of the study

4 Deriving the smallest number of dimensions

The use of the collected data is as follows. For each participant, I calculate the smallest number of dimensions n needed to represent in $\mathbb{R}^n$ his ranking of pairs by similarity. Formally, for each of the participants, I find the smallest n, such that one can find $\{x_i\} \subset \mathbb{R}^n$ such that $||x_i - x_j|| < ||x_k - x_l||$ iff $r_{ij} < r_{kl}$ for each $1 \leqslant i, j, k, l \leqslant 7$ where $i \neq j$ and $k \neq l$.

The calculation of the smallest number of dimensions is numeric. The algorithm, implemented in MATLAB, solves a system of 20 strict inequalities corresponding to the 21 monotonically increasing distances between 7 points in $\mathbb{R}^n$. First, the program tries to locate the objects in a single dimension (n=1). If the program fails to find a solution after trying 50 different random initial guesses, it proceeds to two dimensions and starts over with 50 new initial guesses and so on until the system of inequalities is solved. Conditional on finding a

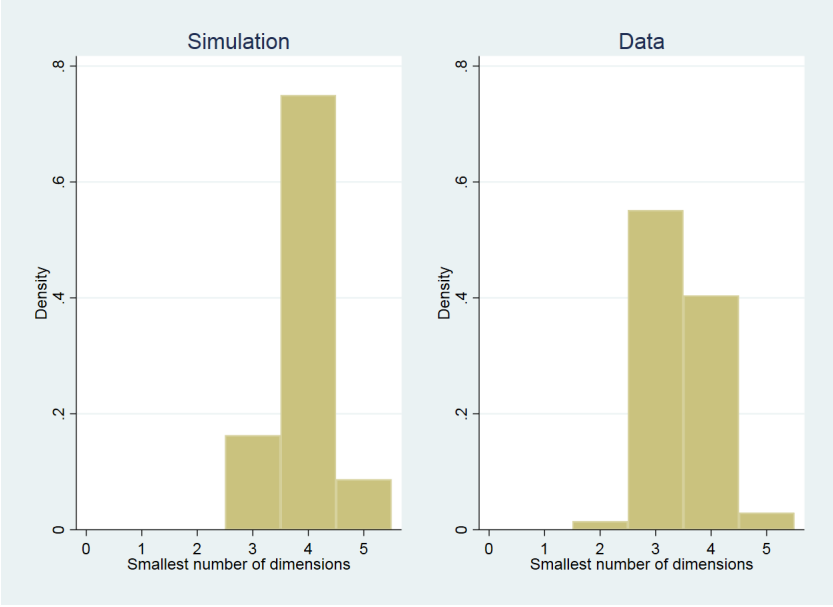
solution in n dimensions, the solution is achieved with probability 0.86 at the first random guess. Moreover, conditional on finding a solution in n dimensions, none of the observations requires more than 5 initial guesses. Thus, giving the algorithm a chance to find a solution with 50 initial guesses is a safe way to avoid a situation when a possible solution is not found.

Table 1 presents the descriptive statistics of the participants, their similarity judgments, and the derived smallest number of dimensions.

Table 1: Summary statistics

Variable	N	Mean	S.D.
Smallest number of dimensions	136	3.45	0.58
Average similarity	136	2.63	0.49
Male	133	0.54	0.50
Age	132	23.19	3.14
Seconds per click	136	6.23	2.36

Figure 3: The smallest number of dimensions: data versus simulated spurious rankings



5 Results

The smallest number of dimensions

The above-described procedure that derives the smallest number of dimensions is fully agnostic with regard to any underlying behavioral structure. Hence, the first question is whether the smallest number of dimension, derived on the individual level, is not a spurious statistic. A spurious smallest number of dimensions should not be statistically different from the one derived from a random ranking of pairs of objects. I perform this test by comparing the smallest number of dimensions derived from data and the smallest number of dimensions derived from simulated random rankings. I simulate 1000 random permutations of numbers from 1 to 21. For each permutation, I derive the smallest number of dimensions similarly to how it is done with real data. Figure 3 shows the histograms of the smallest number of dimensions derived from data and of the smallest number of dimensions derived from the simulated random rankings. The modes of the two distributions are 3 and 4, respectfully. The mean smallest number of dimensions in data is 3.45 (0.58) versus 3.92 (0.49) in the simulated random rankings (the figures in parentheses are the standard deviations). It means that the smallest number of dimensions in data is almost one standard deviation smaller than in spurious rankings. The t-statistic of the difference is as high as 9. Thus, the null hypothesis that data and spurious permutations generate the same mean smallest number of dimensions is rejected at any level of statistical significance. Moreover, two out of the 136 observations require only two dimensions, whereas none of the 1000 simulated rankings do. This result justifies the inquiry for a behavioral interpretation of the smallest number of dimensions.

The smallest number of dimensions and similarity

In order to interpret the smallest number of dimensions, I explore the correlation between this number and the average perceived similarity between the objects. Because for 95% of the observations the smallest number of dimensions is either 3 or 4, I estimate a Probit regression where the dependent variable is one for 4 or 5 dimensions and zero otherwise (i.e., for 2 or 3 dimensions).[4] I name this binary dependent variable "a high smallest number of dimensions." The main explanatory variable is the average similarity between the objects on individual level (the similarity between each pair of objects is evaluated on a scale from 1 [not similar at all] to 5 [very similar]). Thus, the dependent variable is derived from data collected in the first task, while the main explanatory variable is derived from data collected in the second task.

Table 2 presents the average marginal effects of Probit regressions with robust standard errors. Column 1 considers the whole sample, column 2 considers participants with time per click above the median, and column 3 considers the same sample as column 2 but with age and gender controls added to the regression. As the estimates show, the average similarity is negatively correlated with the smallest number of dimensions. The results are stronger and statistically significant when only thoughtful participants are considered (columns 2 and 3). The average marginal effect is -0.10 in the full sample without controls and it is -0.25 and -0.3 in the subsample of thoughtful participants, without and with controls, respectfully. Thus, each increase by one in average similarity (scaled from one to five) is associated with a decrease of 10-30 percentage points in the probability to have a high smallest number of

[4] The standard deviation of the smallest number of dimensions decreases only by 14% as a result of grouping the values.

dimensions.

Table 2: Probit regression results

	Average marginal effects from Probit regressions		
	(1)	(2)	(3)
Variable	All participants	Participants with working time above median	
Average similarity	-0.107	-0.248**	-0.299***
	(0.084)	(0.102)	(0.341)
Male			0.173
			(0.122)
Age			-0.018
			(0.015)
N	136	68	65

Dependent variable: Dummy for more than 3 dimensions. Robust standard errors are reported in parentheses.
Statistical significance: * for 0.1, ** for 0.05, *** for 0.01.

6 Discussion

Regressions in Table 2 show a clear relationship between the smallest number of dimensions and similarity, but they cannot show the direction of the causal link. Does an individual find the objects more similar to each other because she considers fewer attributes or is it another way around?

It is easy to explain the causal effect of the number of considered attributes on the average subjective similarity. Let us assume that similarity can be indeed approximated by Euclidean distance between points that represent objects. When one considers fewer comparison dimensions, by triangle inequality the average distance between every two points is shorter.

However, also the causal effect in the opposite direction may be explained. Let us assume that an individual finds all objects quite similar to each other. In the context of the present study, she finds all fruits similar. For instance, she does not eat fruits at all and is indifferent about them. Because all fruits seem to her similar (and/or irrelevant), she considers only few attributes when she compares them. For example, she takes into account only color and size.

It may be the case that both directions exist in data. Given the results of the present study, which is the first one to estimate the number of dimensions on individual level and to document its association with similarity, the natural next step is to design a study that can illuminate the direction of the causal link.

7 Conclusions

This paper presents new evidence contributing to the literature on the geometric representation of similarity data. The novelty is in deriving the smallest number of dimensions on individual level, comparing the mean smallest number of dimensions in data to the one derived from a random simulation, and in regressing the smallest number of dimensions on the average similarity between the objects, separately evaluated by the same individuals. I find that the smallest number of dimensions is one standard deviation smaller in data than when similarity rankings are randomly simulated. For a set of seven objects, the mode of the smallest number of dimensions in real data is 3, while for randomly simulated data it is 4. Furthermore, when subjective similarity is scaled between one and five, each unit of average subjective similarity is associated with a 10-30 percentage points lower probability

to consider a high number of dimensions. These results support a behavioral interpretation of the smallest number of dimensions, even though this number is implicit and is not reduced from observed variables. However, the direction of the causal link between the number of considered attributes and similarity is let to be identified in further research.

References

Bijmolt, Tammo HA, T. H., Wedel, M., Pieters, R. G., and DeSarbo, W. S.. "Judgments of brand similarity". *International Journal of Research in Marketing*, 15(3), (1998): 249-268.

Borg, Ingwer, and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, (2005).

Borg, Ingwer, Patrick JF Groenen, and Patrick Mair. *Applied multidimensional scaling and unfolding*. New York, NY: Springer, 2017.

Carroll, J. Douglas, and Paul E. Green. "Psychometric methods in marketing research: Part II, multidimensional scaling," *Journal of Marketing research* 34.2 (1997): 193-204.

Chiappori, Pierre-André, Sonia Oreffice, and Climent Quintana-Domeque. "Fatter attraction: anthropometric and socioeconomic matching on the marriage market." *Journal of Political Economy* 120.4 (2012): 659-695.

Clapham, Christopher, and James Nicholson. "Bubble sort algorithm." In *The Concise Oxford Dictionary of Mathematics*. : Oxford University Press, (2009).

Coles, Melvyn G., and Marco Francesconi. "Equilibrium Search with Multiple Attributes and the Impact of Equal Opportunities for Women." (2018).

Cooper, Lee G. "A review of multidimensional scaling in marketing research." *Applied Psychological Measurement* 7.4 (1983): 427-450.

DeSarbo, W. S., Manrai, A. K., and Manrai, L. A.. Latent class multidimensional scaling: A review of recent developments in the marketing and psychometric literature. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 190–222). Cambridge, MA: Blackwell (1994).

Fuchs, Christoph, and Adamantios Diamantopoulos. "Positioning Bases' Influence on Product Similarity Perceptions." In *Quantitative Marketing and Marketing Management*. Gabler Verlag, Wiesbaden. (2012): 325-351.

Guttman, Louis. "A general non metric technique for finding the smallest coordinate space for a configuration of points." *Psychometrika* 33.4 (1968): 469-506.

Jain, Anil K., Robert P. W. Duin, and Jianchang Mao. "Statistical pattern recognition: A review." *IEEE Transactions on pattern analysis and machine intelligence* 22.1 (2000): 4-37.

Lefkoff-Hagius, Roxanne, and Charlotte H. Mason. "Characteristic, beneficial, and image attributes in consumer judgments of similarity and preference." *Journal of Consumer Research* 20.1 (1993): 100-110.

Lin, L., and D. K. H. Fong. "Bayesian multidimensional scaling procedure with variable selection." *Computational Statistics & Data Analysis* 129 (2019): 1-13.

Tversky, Amos. "Features of similarity." *Psychological review*, 84.4 (1977).

Appendix: The Bubble Sort Algorithm

In the first part of the study, participants had to order pairs of fruits by similarity. At each click, they were exposed to a screen with two pairs of pictures of fruits and they must choose the more similar pair (see a screenshot in Figure 1). The output of the procedure is a strictly ordered-by-similarity list of all pairs of fruits. The order of the screens follows the bubble sort algorithm:

Given vector X of length n, we sort X in ascending (without loss of generality) order.

1. Set i to be n.
2. Is $i = 1$?
   - 2.1 If yes, proceed to 3.
   - 2.2 Set j to be 1.
   - 2.3 Is $j = i$?
     - 2.3.1 If yes, proceed to 2.4.
     - 2.3.2 Is $X[j] < X[j+1]$?
       - 2.3.2.1 If yes, proceed to 2.3.3.
       - 2.3.2.2 Replace $X[j]$ and $X[j+1]$.
     - 2.3.3 Increase j by 1 and proceed to 2.3.
   - 2.4 Decrease i by 1 and proceed to 2.
3. Finish.