Journal of Statistical Software

May 2015, Volume 65, Code Snippet 2. *http://www.jstatsoft.org/*

MF Calculator: A Web-Based Application for Analyzing Similarity

Célia M. D. Sales
Universidade de Évora
Cis/ISCTE-IUL, UNIDEP-ISMAI

Peter P. Wakker
Erasmus University

Paula C. G. Alves
Cis/ISCTE-IUL,
King's College London

Luís Faísca
Universidade do Algarve,
DPCE-FCHS, IBB/CBME

Abstract

This paper presents the metric-frequency calculator (**MF Calculator**), an online application to analyze similarity. The **MF Calculator** implements a metric-frequency similarity algorithm for the quantitative assessment of similarity in ill-structured data sets. It is widely applicable as it can be used with nominal, ordinal, or interval data when there is little prior control over the variables to be observed regarding number or content. The **MF Calculator** generates a proximity matrix in CSV, XML or DOC format that can be used as input to traditional statistical techniques such as hierarchical clustering, additive trees, or multidimensional scaling. The **MF Calculator** also displays a graphical representation of outputs using additive similarity trees. A simulated example illustrates the implementation of the MF calculator. An additional example with real data is presented, in order to illustrate the potential of combining the **MF Calculator** with cluster analysis. The **MF Calculator** is a user-friendly tool available free of charge. It can be accessed from `http://mfcalculator.celiasales.org/Calculator.aspx`, and it can be used by non-experts from a wide range of social sciences.

Keywords: metric-frequency calculator, metric similarity, feature similarity, software implementation, ill-structured data, metric distance, additive similarity trees.

1. Introduction

The measurement of similarity is important in many domains. Being similar can be formal-

ized as being close with respect to some distance measure, where the distance measure, and being close, depends on the context. For example, in case-based decision theory, options are evaluated by the similarity-weighted average utility of their predecessors (Gilboa and Schmeidler 2001). Plagiarism accusations are based on similarity assessments (Prechelt, Malpohl, and Philippsen 2003). Similarity is also central in categorizations (Sneath and Sokal 1973). Its use is so widespread that Ashby and Ennis (2007) wrote "A review, or even a listing of all the uses of similarity is impossible" (Introduction, p. 4116).

Measures of similarity have been studied almost exclusively for well-structured data sets, where the number of variables is specified *a priori*. However, in social sciences, data sets are often ill-structured. This is typically the case when people are given open-ended instructions and thus can come up with items that are impossible to determine in advance. We often compare subjective judgments, preferences, or emotional states where the type and number of alternatives is unknown beforehand, and can be very diverse. In such situations, an approach that limits judgments to the same *a priori* set of items for every subject is not appropriate.

To deal with this issue, Sales and Wakker (2009) introduced a new theoretical measure for quantifying similarity, the metric-frequency measure (MF). They applied the MF to a psychotherapy study, comparing complaints of family members. There, the number and nature of issues raised by the members were unpredictable. More recently, the MF has been integrated into the Individualized Patient Progress System (IPSS; Sales and Alves 2012), a web-based monitoring system for psychological treatments that allows clinicians to compare patients based on their personal complaints. This system is currently being tested with family therapists, psychodramatists and drug addiction group therapists (Alves, Sales, and Ashworth 2013).

To facilitate the application of the MF by other researchers, a user-friendly and effective tool is required. This paper presents such a tool, the **MF Calculator**, which is now freely available as a web application.

2. The MF

One commonly used kind of similarity measurement is metric (or dimensional). Then the variables can take a range of numerical values, and a metric distance measure is used to specify similarity. Other similarity measurements are qualitative (or featural). Then variables are qualitative, usually dichotomous (present or absent; Tversky 1977). Carroll (1976) recommended combining metric and qualitative aspects, because the perception of similarity is a complex mental phenomenon that usually involves both inputs. Navarro and Lee (2003) first proposed a model for such measurements. They coded qualitative items as 1 (present) or 0 (absent), and could then apply metric distances, subsequently used in maximum likelihood fittings of data. The metric frequency measure (MF), defined below, combines metric and qualitative aspects in a different manner, designed to incorporate other generalizations.

Example 1. *This example illustrates the MF. Assume that two people, A and B, are independently presented with a picture and are asked to list the emotions (items) associated with this picture and to rate the intensity of each emotion. As each person is free to raise any item according to their subjective evaluation, the number and nature of items in the data set are unpredictable: There can be many emotions raised by one person and not by the other. The number of items raised by A and B can also differ. For instance, A may associate only*

one emotion with the picture and B may experience seven emotions. A measure of similarity between A and B can be constructed in the following way: First, the difference of scores is calculated for items that both persons raise. Second, when both persons mention the same items, then this in itself is a signal of their similarity, while items raised by one person and not by the other in themselves reflect a distance (difference) between the two people (Tversky 1977). Third, the number of items raised also provides information about similarity.

In the example above, person B raises seven items, and person A raises only one item. This difference in itself also generates distance (difference) between the two people. For an efficient application of such frequency similarities in linguistic studies, see Maki, Krimsky, and Munoz (2006). Such information is not accounted for in traditional approaches. For another efficient method to measure similarity, based on two-dimensional sorting on a computer screen, see Goldstone (1994). The MF combines the metric component and the frequency component for the measurement of similarity, and thus is composite.

The MF consists of 1) the metric similarity, based on differences of the scores on joint items (the lower the differences, the more similarity there is), and 2) the frequency similarity (the more similar the number of items raised by both people, the more similarity there is). The MF formula is shown below (3).

$$\frac{\sum\left(1-|diff|\right)}{j+f+m} \tag{1}$$

$$\frac{\sqrt{j/N}+1-\left|\sqrt{f/N}-\sqrt{m/N}\right|}{2} \tag{2}$$

$$\frac{1}{2}\frac{\sum\left(1-|diff|\right)}{j+f+m}+\frac{1}{4}+\frac{1}{4}\sqrt{j/N}-\frac{1}{4}\left|\sqrt{f/N}-\sqrt{m/N}\right| \tag{3}$$

These equations represent the following quantities: Summation Σ in (1): over all items raised by either person A or person B[1]; $|diff|$: absolute value of the difference in $0-1$ normalized scores that the two persons assign to the item under consideration, with $1-|diff|$ the resulting similarity; j: the number of (*joint*) items raised by both person A and person B; f: the number of items raised by person A and not by person B; m: the number of items raised by person B and not by person A; and N: an upper bound for the number of items that can be raised by one person (explained later).

The score similarity (or metric similarity) proceeds as in most metric approaches to similarity measurement. Scores should be at least at an interval scale level, so that differences are meaningful. Some items may be raised by one person and not by the other. Accordingly, when an item is not raised, it is entered as 0 in the difference and this is taken as the minimum score. The MF thus allows no negative scores, and can only be applied to uni-directional items. For incorporating negative scores, see Sales and Wakker (2009).

The scaling of absence as 0, and its difference with the minimal positive score of items if present, should obviously agree with the other score levels and their differences. Those levels should therefore be chosen deliberately by the researchers when scoring the data (Sales and

[1]Or, equivalently, over all conceivable items because those not raised will all contribute zero to the summation.

Wakker 2009). Given the meaningful score 0, our scales are in fact ratio scales. To avoid arbitrary, unequal weightings of different items in what follows, they are normalized to a [0,1] scale by dividing the scores of each item by the maximum of their range.

Then for each item we compute the difference between the scores that the two people give to this item, *diff*. The similarity is $(1 - |diff|)$ and the score similarity is the sum, $\Sigma(1 - |diff|)$ over the total number of items raised, $j + f + m$.

For the frequency similarity, the frequencies are normalized by dividing by the upper bound N. We usually take N equal to the maximum number of items raised by any person in our sample. Then the frequencies are normalized to a [0,1] scale, and the frequency similarity is weighted likewise as the score similarity. Sometimes N can be chosen larger than the mentioned maximum. Then the frequency score is more compressed around 0.25, and is bounded away from the minimum 0 and the maximum 0.5. The overall effect is that the frequency similarity then has less influence on the MF. Thus N is an extra free parameter that a researcher can increase if the information contained in the observed frequencies is less reliable than that contained in the scores, and should have less weight.

A variation could be developed where N could be taken below the maximum frequency (number of items) observed, in which case the frequency similarity impacts the MF more than the score similarity. The MF can then take negative values, which can be rescaled to a [0,1] scale. In most situations, the frequency similarity will be less reliable than the score similarity, because a score assigned to some specified item is a deliberate and contemplated act, whereas the remembering or forgetting of some item is more coincidental. Hence we focus on cases where either, as default, N is equal to the maximum frequency as an objective default, or N is larger if the researcher has reasons to give less weight to the frequency similarity. In general, the relative importance of the frequency information concerning the score information should be determined by researchers who know the context of the application.

Because all frequencies are divided by N, the results are normalized, which ensures that j/N, f/N, and m/N never exceed 1, and we properly weight the score similarity viz-a-viz the frequency similarity. In the above example, B was the person raising most items, i.e., seven emotional states. Therefore, $N = 7$.

Instead of the number j/N, the frequency similarity model uses its square root, $\sqrt{j/N}$. This transformation is curved downwards (concave), meaning that similarity increases less for high values of j (and j/N) than for low values. For instance, if persons A and B list completely different emotional states ($j = 0$), and if they then both raise one identical emotion, then this increase of one item (from $j = 0$ to $j = 1$) has more impact than if there are already many items in common (concerning an increase from, say, $j = 18$ to $j = 19$). Such an evaluation is plausible. Following the same rationale, the square-root transformation is also applied to f/N and m/N. Other concave transformations could obviously also be considered. The square root transformation, steep at the minimum 0, and with moderate derivatives in between that do not vanish at the maximum 1, fits our application well. The steepness at 0 captures the categorical difference between no and some overlap, and gives satisfactory results in applications.

The frequency similarity is the average of the similarity due to the number of items that both people raise, and the similarity due to the difference in the number of items raised by person A and person B. Finally, the MF results as the average of the score similarity and the

Person A		Person B		Person C	
Item	Score	Item	Score	Item	Score
Green	7	Brown	6	Green	7
Blue	6	Orange	6		
Red	4	Yellow	2		
Pink	3				

Table 1: Color preferences of persons A, B and C, rated by intensity of preference.

	Green	Blue	Red	Pink	Brown	Orange	Yellow
Person A	7	6	4	3	0	0	0
Person B	0	0	0	0	6	6	2
Person C	7	0	0	0	0	0	0

Table 2: Input database for the **MF Calculator**.

frequency similarity:

$$\frac{1}{2}\frac{\sum(1-|diff|)}{j+f+m}+\frac{1}{2}\frac{\sqrt{j/N}+1-\left|\sqrt{f/N}-\sqrt{m/N}\right|}{2}$$

which can be re-written as (3).

3. Handling the MF Calculator

We now turn to the explanation of the **MF Calculator**. The MF is computed online at http://mfcalculator.celiasales.org/Calculator.aspx. In brief, the user prepares the input database in the CSV format, uploads to this website, chooses between the available analyses (i.e., score similarity, frequency similarity, or overall similarity), and obtains the results on screen. Output files are available in three formats (CSV, XML and DOC), and can be used in subsequent analyses. To ensure confidentiality and data protection, all data and respective outputs generated by this website will not be stored automatically on any server. When the user closes the internet browser, everything is lost and cannot be retrieved.

3.1. Preparing the input database for MF Calculator

For illustration, we use a hypothetical example involving three persons who were asked to identify their favorite colors and rate how much they liked them on a 7-point scale, ranging from 1 to 7 (Table 1). The question is how similar the persons are regarding their color preferences.

The **MF Calculator** supports input databases in the CSV format with the following structure: a) Rows represent cases to be compared and b) columns represent items or variables. Cells display scores. Table 2 displays our data. Whenever an item was not raised by an individual, the score is 0. The difference of 2 between score 0 and the minimum positive score, 2, reflects that this difference is taken as twice as significant as the other minimal differences of scores.

The **MF Calculator** handles databases containing up to 500 cases. CSV files are generated by saving your database as a CSV file (*comma-separated values*), an option available in com-

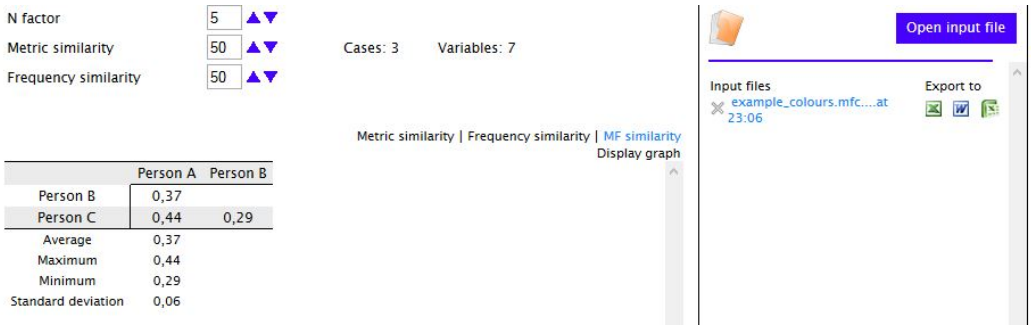

Figure 1: MF similarity between persons A, B and C in terms of color preferences (rated by preference).

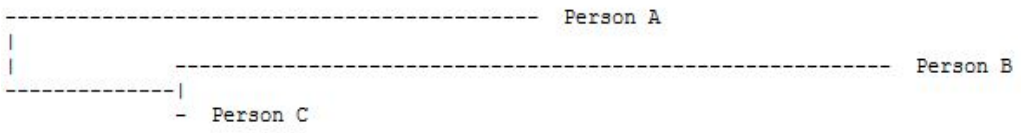

Figure 2: Similarity tree for persons A, B and C in terms of color preferences (rated by preference).

mercial software such as MS Excel, SPSS (IBM Corporation 2013) or SAS (SAS Institute Inc. 2011). The **MF Calculator** assumes empty cells (missings) as zero values. Therefore, an input database can have empty cells instead of zero values.

3.2. Calculating the MF similarity

The MF similarity matrix is computed in the Calculator section of the website. To do so, the input database is uploaded and the N factor should be specified. In the example, person A raised the highest number of colors (four colors), and $N \geq 4$ is natural. For illustrative purposes let us assume that the mere presence or absence of colors is considered to be somewhat more coincidental and less informative than the scores given to the colors. We hence choose $N = 5$. With the database imported and the N factor chosen, the MF is ready to be calculated by clicking on the input database file.

The **MF Calculator** generates several outputs: 1) similarity matrices (overall similarity, score similarity and frequency similarity) in DOC, XML and CSV formats; 2) descriptive statistics of the similarities (mean, maximum, minimum, and standard deviation); and 3) the graphical representation of results, using additive similarity trees (Sattath and Tversky 1977); see Figures 1 and 2.

Nodes in the tree represent the cases, and the length of the path joining them represents their proximity. These similarity trees produced by the **MF Calculator** are recommended for visualization of samples of up to 50 cases. For larger samples, we recommend downloading the MF output file and using statistical software for graphical representation (explained later).

The MF output files are available for download and must be saved to a local computer or to an external driver. Otherwise all data will be lost once the web page is closed.

Person A	Person B	Person C
Item	Item	Item
Green	Brown	Green
Blue	Orange	
Red	Yellow	
Pink		

Table 3: Color preferences of persons A, B and C, rated nominally (purely qualitative data).

	Green	Blue	Red	Pink	Brown	Orange	Yellow
Person A	1	1	1	1	0	0	0
Person B	0	0	0	0	1	1	1
Person C	1	0	0	0	0	0	0

Table 4: Input database for the **MF Calculator** (1 = present; 0 = absent).

3.3. Using the MF Calculator with nominal data

In a qualitative setting where only the presence or absence of items can be observed, the **MF Calculator** can also be used. Suppose that in the preceding example individuals were only asked to identify their favorite colors, without rating how much they liked them (Table 3). The database would then assign scores 1 (present) and 0 (absent) to each item (Table 4).

4. Contexts in which the MF Calculator can be applied

The **MF Calculator** can be used with nominal, ordinal or interval data. Cases compared can either be people (as in our hypothetical example), or any other entity. It allows the comparison of user-generated items, such as patient-generated measures in clinical settings (Ashworth, Evans, and Clement 2009), content analysis judgments with no *a priori* categorization system, or virtually any kind of answers resulting from open-ended instructions, with no limits to the content or number of resulting outputs. The **MF Calculator** can also serve to analyze multiple items questionnaires where participants are free to choose among a large number of options and then rate only the selected items (for an example, see the real data illustration below).

The similarity matrix resulting from the **MF Calculator** can serve as an input data matrix for further statistical analyses, including visualization, clustering and scaling. The CSV format allows its direct use with most statistical software packages.

4.1. Illustration of the MF Calculator with a real example

We present a real life example illustrating the combined use of the MF similarity and cluster analysis. In order to study the incidence of early negative life experiences, 100 Portuguese university students (n = 79 females) filled in the Negative Life Events Inventory (Brás and Cruz 2008). Participants indicated which of the 25 negative life events (NLE) listed in the Inventory had happened in their lives up to 12 years of age. The perceived impact of each selected event was then evaluated on a 5-point scale, ranging from 1 = no impact to 5 = extremely negative impact. In the input database for the **MF Calculator**, rows display the students to be compared, and columns refer to the NLE. Each cell gives the impact score

of the NLE. When the event has not been selected, its value is 0. The database was built in Excel, saved in CSV format, and uploaded to the **MF Calculator**. $N = 25$ was used (the maximum number of NLE any participant could select). The resulting MF similarity matrix was downloaded in CSV format and used in SPSS for a hierarchical cluster analysis using the average linkage within groups agglomerative method (`WAVERAGE` algorithm).

The resulting dendrogram suggests a two-clusters solution. In order to interpret the meaning of these two distinct patterns of response, we compare both clusters based on their mean values. We found a first cluster of students with a history of more severe early life experiences ($n = 51$) that is characterized by a high number of NLE (mean of 9.9 NLE per student), and more negative consequences due to perceived dysfunctional family environment, health problems in their families, and experiences of psychological neglect and abuse. The second cluster ($n = 49$) includes students who typically report a reduced number of NLE (mean of 3.9 NLE), particularly own health problems and adverse living conditions. Participants from the first cluster evaluated their experiences of living in a dysfunctional family environment and of being psychologically abused as having had a much stronger negative impact in their lives than participants from the second cluster (Cohen's $d > 1.25$).

5. Conclusion

We have implemented a composite measure of similarity, the MF measure. It combines metric information, based on differences in scores on given items, with featural information that is derived from the co-occurrence and difference in numbers of items raised. The MF measure is especially useful in complex ill-structured settings where the number and content of variables are unpredictable. We introduce the **MF Calculator**, which is the first implementable software for the computation of similarity in ill-structured settings, and illustrate it in examples. The **MF Calculator** is flexible and can be applied to both metric and qualitative variables. The proximity matrix generated by the **MF Calculator** can be used as input of traditional statistical techniques such as hierarchical clustering, additive trees, multidimensional scaling, and so on. It, thus, can be widely applied in behavioral and social sciences. It greatly enhances our possibilities to measure and analyze similarity, which is central in numerous fields.

Acknowledgments

The development of the software, as well as the preparation of this manuscript, was undertaken with the support of funds allocated by the Portuguese Foundation for Science and Technology (Grant nr. PTDC/PSI-PCL/098952/2008).

The authors are grateful to Prof. Cláudia Carmo (Universidade do Algarve) for providing the Negative Life Events Inventory data analysed as an illustrative example in this article, to Mr. Kenneth Gunn for his assistance in editing the English text, and to Professor Bruno Maia for his support when preparing the manuscript in LaTeX format.

References

Alves PCG, Sales CMD, Ashworth M (2013). "Enhancing the Patient Involvement in Out-

comes: A Study Protocol of Personalised Outcome Measurement in the Treatment of Substance Misuse." *BMC Psychiatry*, **13**(337).

Ashby FG, Ennis DM (2007). *Similarity Measures*. Scholarpedia.

Ashworth M, Evans C, Clement S (2009). "Measuring Psychological Outcomes After Cognitive Behaviour Therapy in Primary Care: A Comparison Between a New Patient-Generated Measure PSYCHLOPS (Psychological Outcome Profiles) and HADS (Hospital Anxiety and Depression Scale)." *Journal of Mental Health*, **18**(2), 169–177.

Brás M, Cruz JP (2008). "Inventário de Acontecimentos de Vida Negativos, Construção e Validação numa População Adulta." In AP Noronha, C Machado, L Almeida, M Gonçalves, S Martins, V Ramalho (eds.), *Actas da XIII Conferência Internacional de Avaliação Psicológica: Formas e Contextos*. Psiquilíbrios Edições, Braga, Portugal.

Carroll JD (1976). "Spatial, Nonspatial and Hybrid Models for Scaling." *Psychometrika*, **41**(4), 439–463.

Gilboa I, Schmeidler D (2001). *A Theory of Case-Based Decisions*. Cambridge University Press, Cambridge, United Kingdom.

Goldstone R (1994). "An Efficient Method for Obtaining Similarity Data." *Behavior Research Methods*, **26**(4), 381–386.

IBM Corporation (2013). *IBM SPSS Statistics 22*. IBM Corporation, Armonk. URL http://www.ibm.com/software/analytics/spss/.

Maki WS, Krimsky M, Munoz SOL (2006). "An Efficient Method for Estimating Semantic Similarity Based on Feature Overlap: Reliability and Validity of Semantic Feature Ratings." *Behavior Research Methods*, **38**(1), 153–157.

Navarro DJ, Lee MD (2003). *Advances in Neural Information Processing Systems*. Massachusetts Institute of Technology, Cambridge, MA.

Prechelt L, Malpohl G, Philippsen M (2003). "Finding Plagiarisms Among a Set of Programs With JPlag." *Journal of Universal Computer Science*, **8**(11), 1016–1038.

Sales CMD, Alves PCG (2012). "Individualized Patient-Progress Systems: Why We Need to Move Towards a Personalized Evaluation of Psychological Treatments." *Canadian Psychology*, **53**(2), 115–121.

Sales CMD, Wakker PP (2009). "The Metric-Frequency Measure of Similarity for Ill-Structured Data Sets, with an Application to Family Therapy." *The British Journal of Mathematical and Statistical Psychology*, **62**(3), 663–682.

SAS Institute Inc (2011). *The SAS System, Version 9.3*. SAS Institute Inc., Cary. URL http://www.sas.com/.

Sattath S, Tversky A (1977). "Additive Similarity Trees." *Psychometrika*, **42**(3), 319–345.

Sneath PH, Sokal RR (1973). *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. W. H. Freeman, San Francisco, CA.

Tversky A (1977). "Features of Similarity." *Psychological Review*, **84**(4), 327–352.

Affiliation:

Célia Maria Dias Sales
Instituto Universitário de Lisboa (ISCTE-IUL), Cis-IUL
and
UNIDEP-ISMAI
and
Universidade de Évora, Departamento de Psicologia
Colégio Pedro da Fonseca, Rua da Barba Rala, 1 PITE
7005-345 Évora, Portugal
E-mail: celiasales@uevora.pt

Peter P. Wakker
Econometric Institute
Erasmus University Rotterdam
E-mail: wakker@ese.eur.nl

Paula Cristina Gomes Alves
Instituto Universitário de Lisboa (ISCTE-IUL), Cis-IUL
and
King's College London
E-mail: paulagomesalves@hotmail.com

Luís Faísca
Universidade do Algarve, FCHS
Departamento de Psicologia e Ciências da Educação
and
IBB/CBME
Email: lfaisca@ualg.pt

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 65, Code Snippet 2 *Submitted:* 2011-10-31
May 2015 *Accepted:* 2014-06-18