ISSN 2007-9737

Analysis of Relationships between Co-Symmetric Dissimilarity Measures of Probability Distributions with Involutive Negations

Maria Elena Ensastegui-Ortega[1], Ildar Batyrshin[1,*], Alexander Gelbukh[1], Nailya Kubysheva[2]

[1] Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico

[2] Kazan (Volga Region) Federal University, Republic of Tatarstan, Russia

{batyr1, elena.ensastegui}@gmail.com

Abstract. Learning from data in almost any human activity is a very important task, usually using similarity or dissimilarity between data. Recently, it was shown the importance of considering the involution operation defined on the data domain which reflects a symmetry of data structures. This symmetry should be taken into account in data analysis. Co-symmetric similarity and dissimilarity measures defined over a set with involution play an important role in data analysis. In this paper, four dissimilarity functions over the set of probability distributions are created that meet the property of co-symmetry with respect to the involutive negation of distributions. Scatter graphs are generated from their respective dissimilarity matrices to compare the similarity between them. Additionally, the Pearson, Kendall, and Spearman correlation coefficients are calculated to numerically assess the relationship that exists. Subsequently, four dissimilarity functions are considered due to their higher correlation with those studied in this paper. They are divided into two groups, and an analysis is conducted to determine which are more correlated.

Keywords: Co-symmetry, correlation, dissimilarity, involution, probability distribution.

1 Introduction

Many similarity and dissimilarity measures are proposed for probability distributions [1,2]. Recently, an involutive negation of probability distributions [3] and measure of correlation between distributions were introduced [4, 5]. This correlation measure used co-symmetric distance between probability distributions based on involutive negation of probability distributions. Co-symmetric similarity and dissimilarity measures are important for applications because they take into account the symmetry of data related to involution operation [6]. In this paper, four new co-symmetric dissimilarity functions for probability distributions are created and compared with the other co-symmetric distances between probability distributions considered in [7].

In Section 1, a small outline of the theory used to support the results is given. In Section 2, four distances are used to create four dissimilarity functions that comply with the co-symmetry property. In section 3, they are compared with four other co-symmetric distances introduced in [7].

2 Preliminary Definitions

2.1 Negator and Negation of Probability Distributions

Let $X = (x_1, \dots, x_d)$ be a set of alternatives ordered in some way. A probability distribution over X is a sequence of non-negative numbers $P = (P_1, \dots, P_d)$ such that $\sum_{i=1}^{d} P_i = 1$. Here, for all $i = 1, \dots, d$, P_i is considered as a probability of x_i.

The first example of negation of probability distributions was introduced in [8]. In [9], the

Table 1. Original distances that were considered for this analysis.

Name	Distance
Soergel	$d_{sg} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} \max(P_i, Q_i)}$
Sørensen	$d_{sor} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} (P_i + Q_i)}$
Jaccard	$d_{Jac} = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} {P_i}^2 + \sum_{i=1}^{d} {Q_i}^2 - \sum_{i=1}^{d} P_i Q_i}$
Dice	$d_{Jac} = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} {P_i}^2 + \sum_{i=1}^{d} {Q_i}^2}$

Table 2. New distances created from the original distances and equation (1)

Distance	$\boldsymbol{d_{Name+\,Co-Pro}(P,Q)}$
Soergel	$d_{sg-Co-Pro} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert \ \sum_{i=1}^{d} \lvert N(P_i) - N(Q_i) \rvert}{\sum_{i=1}^{d} \max(P_i, Q_i) \sum_{i=1}^{d} \max(N(P_i), N(Q_i))}$
Sørensen	$d_{sor-Co-Pro} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert \sum_{i=1}^{d} \lvert N(P_i) - N(Q_i) \rvert}{\sum_{i=1}^{d} (P_i + Q_i) \sum_{i=1}^{d} (N(P_i) + N(Q_i))}$
Jaccard	$d_{Jac-Co-Pro} = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2 \sum_{i=1}^{d} (N(P_i) - N(Q_i))^2}{\left(\sum_{i=1}^{d} {P_i}^2 + \sum_{i=1}^{d} {Q_i}^2 - \sum_{i=1}^{d} P_i Q_i\right)\left(\sum_{i=1}^{d} N(P_i)^2 + \sum_{i=1}^{d} N(Q_i)^2 - \sum_{i=1}^{d} N(P_i) N(Q_i)\right)}$
Dice	$d_{Dice-Co-Pro} = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2 \ \sum_{i=1}^{d} \left(N(P_i) - N(Q_i)\right)^2}{\sum_{i=1}^{d} \left({P_i}^2 + {Q_i}^2\right) \sum_{i=1}^{d} (N(P_i)^2 + N(Q_i)^2)}$

general properties of negations of probability distributions and the class on linear negations of probability distributions are considered. In [2], it was introduced an involutive negation of probability distributions.

Relationships of negation with entropy of probability distributions are studied in [10]. Interpretation of probability distributions as fuzzy distribution sets and extension on probability distributions parametric negations of fuzzy sets is considered in [11, 12].

A negator N is a function that transforms point to point one probability distribution $P = (\mathrm{P}_1, \dots, \mathrm{P}_\mathrm{d})$ into another probability distribution $neg(P) = N(P_1), \dots N(P_n))$ called negation of P [9], such that for all $i, j = 1, \dots, n$, from $P_i \le P_j$ it follows $N(P_i) \ge \left(P_j\right)$.

A negation is called an involutive if $neg\big(neg(P)\big) = P$. In [3], Batyrshin introduced a negator:

$$N_B(P_i) = \frac{\max(P) + \min(P) - P_i}{n(\max(P) + \min(P)) - 1} = \frac{MP - P_i}{nMP - 1},$$

where $\max(P) = \max_{i=1,\dots,n}\{P_i\}$, $\min(P) = \min_{i=1,\dots,n}\{P_i\}$, $MP = \max(P) + \min(P)$. This negator generates an involutive negation of probability distributions: $neg_B(P) = \big(N_B(p_1), \dots, N_B(p_n)\big)$ such that $neg_B\big(neg_B(P)\big) = P$.

2.2 Co-symmetric Dissimilarity Functions

Suppose $P = (P_1, \dots, P_d)$ and $Q = (Q_1, \dots, Q_d)$ are two probability distributions. A dissimilarity function $D(P, Q)$ takes values in the interval [0,1] and satisfy the following properties:

symmetry:

$D(P, Q) = D(Q, P)$,

irreflexivity:

$D(P, P) = 0$.

Table 3. Pearson, Kendall and Spearman coefficients for distances creating from equation (1)

Distance	Pearson	Kendall	Spearman
Sorensen Co-Pro Vs Soergel Co-Pro	0,9919	0,9613	0,9976
Sorensen Co-Pro Vs Jaccard Co-Pro	0,9398	0,7828	0,9346
Sorensen Co-Pro Vs Dice Co-Pro	0,9310	0,7705	0,9268
Soergel Co-Pro Vs Jaccard Co-Pro	0,9372	0,7750	0,9299
Soergel Co-Pro Vs Dice Co-Pro	0,9137	0,7580	0,9187
Jaccard Co-Pro Vs Dice Co-Pro	0,9903	0,9634	0,9978

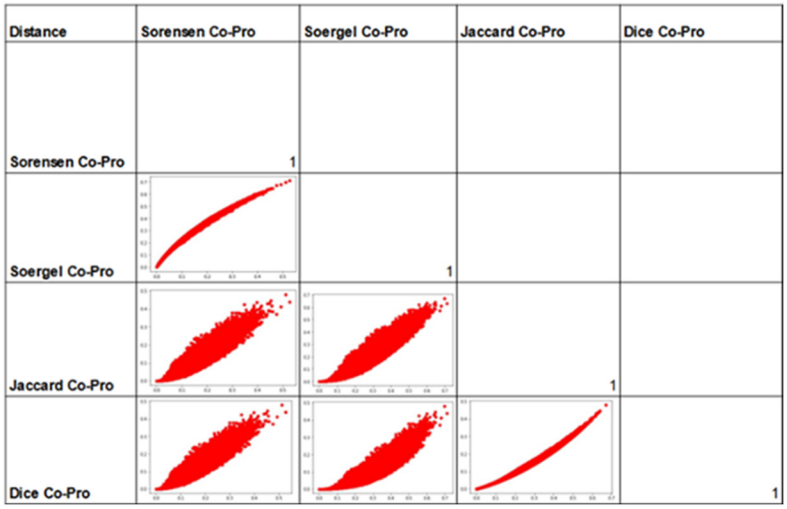


Fig. 1. Scatter graphs comparing distances created from equation (1)

Dissimilarity function is co-symmetric if for all probability distributions P and Q of the length n, it is fulfilled:

$$D\left(neg_B(P), neg_B(Q)\right) = D(P, Q).$$

2.3 Correlation Coefficients

Pearson's correlation coefficient, commonly used in statistical analyses, allows the evaluation of the presence and strength of a linear relationship between two quantitative variables. It varies between -1 and 1.

A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 suggests no linear correlation. On the other hand, Spearman and Kendall correlations are useful tools to investigate monotonic relationships between variables.

While Spearman's is based on the ranges of observations, Kendall's focuses on the agreement of data pairs. Both correlations can vary between -1 and 1 and are designed to be robust to outlier data and not assume specific distributions.

3 New Co-Symmetric Dissimilarity Functions

In [1], different similarity and dissimilarity measures that are usually used to compare distributions of probability functions are considered. They are not

Table 4. Pearson, Kendall and Spearman coefficients of comparing distances from one group.

Distance	Pearson	Kendall	Spearman
Soergel Co-Avg Vs. Soergel Co-Pro	0.9758	0.9137	0.9867
Soergel Co-Avg Vs. Sorensen Co-Avg	0.9906	0.9446	0.9933
Soergel Co-Pro Vs. Sorensen Co-Pro	0.9919	0.9613	0.9976
Soergel Co-Pro Vs. Sorensen Co-Avg	0.9656	0.8494	0.9628
Soergel Co-Avg Vs. Sorensen Co-Pro	0.9587	0.9411	0.9934
Sorensen Co Avg Vs. Sorensen Co-Pro	0.9629	0.8825	0.9766

Table 5. Pearson Kendall and Spearman coefficients of comparing distances from two group.

Distance	Pearson	Kendall	Spearman
Jaccard Co-Avg Vs. Jaccard Co-Pro	0.9566	0.8788	0.9781
Jaccard Co-Avg Vs. Dice Co-Avg	0.988	0.9375	0.9939
Jaccard Co-Pro Vs. Dice Co-Pro	0.9903	0.9634	0.9978
Jaccard Co-Pro Vs. Dice Co-Avg	0.9437	0.8164	0.9501
Jaccard Co-Avg Vs. Dice Co-Pro	0.9318	0.9153	0.9894
Dice Co-Avg Vs. Dice Co-Pro	0.9365	0.8529	0.9688

co-symmetric. We apply the method of co-symmetrization of similarity and dissimilarity functions proposed in [13] to create new dissimilarity measures of probability distributions that comply with the co-symmetry property:

$$D_{Co-Pro}(P,Q) = D(P,Q) * D\big(neg_B(P), neg_B(Q)\big), \quad (1)$$

where * is the product of real numbers. It is easy to show that the distances obtained from (1) are co-symmetric dissimilarity functions. Table 2 shows co-symmetric dissimilarity functions obtained from the four known [1] dissimilarity functions presented in Table 1.

4 Comparative Analysis of New Co-Symmetric Dissimilarity Functions

For comparative analysis of new dissimilarity measures, we used one thousand probability distributions created randomly, each with 10 elements. For the first analysis, the dissimilarity matrices were constructed for the four new co-symmetric dissimilarity measures created by equation (1).

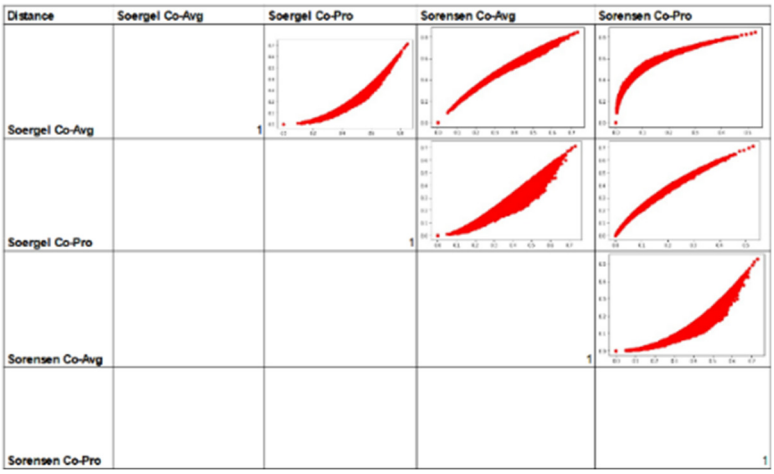
Subsequently, each dissimilarity matrix is transformed into a vector, and the correlation is calculated between two vectors corresponding to two dissimilarity matrixes obtained for two different methods.

The scatter graphs for each pair of vectors are created to graphically observe the correlation that exists between dissimilarity functions, see Fig. 1. In the same way, the correlation between the dissimilarity functions is calculated using Pearson, Kendall and Spearman correlation coefficients.

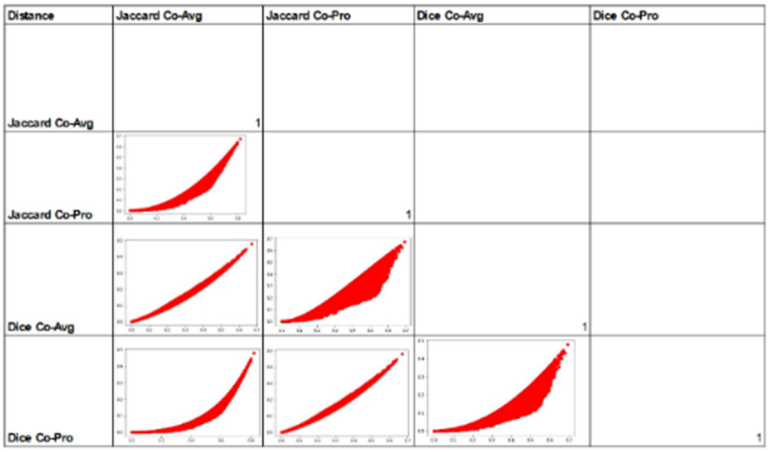
5 Comparing Similarity Functions with Higher Similarity

In [7], new co-symmetric dissimilarity measures based on average operation were obtained. It can be seen from the analysis carried out in the paper that the dissimilarity measures with the greatest correlations were two distances Sorensen Co-Avg and Soergel Co-Avg, and two distances Jaccard Co-Avg and Dice Co-Avg.

In this paper, we obtained the same result for the product-based co-symmetrization (1) of these pair of distances, see Table 3 and Fig. 1, where scatter graphs demonstrate the almost strict monotone dependence between Sorensen Co-Pro and Soergel Co-Pro distances, and between Jaccard Co-Pro and Dice Co-Pro distances.



a)



b)

Fig. 2. a) Scatter plots comparing group one and b) Scatter plots comparing group two

For further analysis of the correlation-based similarity of obtained co-symmetric distances, we divided them into two groups of most correlated distances. The first group contains co-symmetric dissimilarity functions (distances) obtained from Sorensen and Soergel distances, and the second group contains co-symmetric dissimilarity functions obtained from Jaccard and Dice distances.

The results are presented in Tables 4 and 5 and on Figures 2a) and 2b). As was expected, for most co-symmetric dissimilarity functions, different co-symmetrization of the same distance usually gives co-symmetric distances without the correlation less than 0.99. We have paid more attention to the results of the Spearman correlation, which is a measure of monotonic relationship. Only one unexpected result was obtained for Soergel Co-Avg and Sorensen Co-Pro co-symmetric dissimilarity functions, see Table 4.

6 Results

We applied the procedure of co-symmetrization based on product aggregation to the four most popular distances between probability distributions [1]. The correlation analysis of similarity between

these distances show high similarity between them with highest correlation between Soergel and Sorensen based co-symmetric distances, and between Jaccard and Dice based co-symmetric distances. These two pairs of distance are considered as two classes of similar co-symmetric distances with mutual Spearman correlation greater than 0.997 between distances from the same class.

Although we applied three correlation coefficients, Pearson, Spearman, and Kendall correlation, we paid more attention to the Spearman correlation, which is a measure of monotonic relationship. This property is important in the comparison of similarity and dissimilarity measures [14, 15].

Further, we compared co-symmetric distances in each class based on product co-symmetrization with co-symmetric distances obtained in our previous paper [7] based on average co-symmetrization of the same initial distances. The correlation between distances from the same class based on different co-symmetrization of distances is higher than 0.95.

7 Conclusion

We introduced new co-symmetric dissimilarity functions that can serve as distances between probability distributions. These dissimilarity functions take into account the symmetry of the space of finite probability distributions with respect to the uniform distribution $P_U = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$, which is the fixed point of the negation of probability distributions defined over the set with n elements [9], such that $neg(P_U) = P_U$.

Co-symmetrization of four popular distance measures and further correlation analysis of these functions showed highest correlation between Soergel and Sorensen based co-symmetric distances, and between Jaccard and Dice based co-symmetric distances. The same results were obtained for co-symmetric distances obtained previously for another co-symmetrization method.

The obtained results give us a better understanding of known distances and co-symmetric distances obtained from them, which can be used to select suitable distances between probability distributions.

Acknowledgments

This work was partially supported by the Government of Mexico through the grant A1-S-47854 from CONACYT, Mexico, by the projects SIP 20231387 and 20240936 of Instituto Politécnico Nacional, Mexico, and by the program of developing the Scientific-Educational Mathematical Center of Volga Federal District. The authors acknowledge CONACYT for the computing resources provided through the Platform of Deep Learning for Language Technologies of the Supercomputing Laboratory of INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America Ph.D. Award.

References

1. **Cha, S. H. (2007)**. Comprehensive survey on distance/similarity measures between probability density functions. International journal of mathematical models and methods in applied sciences, Vol. 1, No. 2, pp. 1–8.
2. **Sáez, C., Robles, M., García-Gómez, J. M. (2016)**. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. Statistical Methods in Medical Research, Vol. 26, No. 1, pp. 312–336. DOI: 10.1177/0962280214545122.
3. **Batyrshin, I. Z. (2021)**. Contracting and involutive negations of probability distributions. Mathematics, Vol. 9, No. 19, pp. 2389. DOI: 10.3390/math9192389.
4. **Rudas, I. J., Batyrshin, I. Z. (2023)**. Explainable correlation of categorical data and bar charts. Vol. 1, pp. 81–88. DOI: 10.1007/978-3-031-20153-0_7.
5. **Batyrshin, I. Z., Rudas, I. J., Kubysheva, N., Akhtyamova, S. (2022)**. Similarity correlation of frequency distributions of categorical data in analysis of cognitive decline severity in asthmatics. Computación y Sistemas, Vol. 26,

No. 4, pp. 1603–1609. DOI: 10.13053/cys-26-4- 4439.

6. **Batyrshin, I. Z., Tóth-Laufer, E. (2022)**. Bipolar dissimilarity and similarity correlations of numbers. Mathematics, Vol. 10, No. 5, pp. 797. DOI: 10.3390/math 10050797.
7. **Ensastegui-Ortega, M. E., Batyrshin, I., Cárdenas–Perez, M. F., Kubysheva, N., Gelbukh, A. (2024)**. Dissimilarity functions co-symmetry property: a focus on probability distributions with involutive negation. Journal of Intelligent & Fuzzy Systems, pp. 1–10. DOI: 10.3233/jifs-219363.
8. **Yager, R. R. (2015)**. On the maximum entropy negation of a probability distribution. IEEE Transactions on Fuzzy Systems, Vol. 23, No. 5, pp. 1899–1902. DOI:10.1109/tfuzz.2014. 2374211.
9. **Batyrshin, I., Villa-Vargas, L. A., Ramírez-Salinas, M. A., Salinas-Rosales, M., Kubysheva, N. (2021)**. Generating negations of probability distributions. Soft Computing, Vol. 25, No. 12, pp. 7929–7935. DOI: 10.1007/s00500-021-05802-5.
10. **Klein, I. (2022)**. Some technical remarks on negations of discrete probability distributions and their information loss. Mathematics, Vol. 10, No. 20, pp. 3893. DOI:10.3390/math 10203893.
11. **Batyrshin, I. Z. (2022)**. Fuzzy distribution sets. Computación y Sistemas, Vol. 26, No. 3, pp. 1411–1416. DOI: 10.13053/cys-26-3-4360.
12. **Batyrshin, I., Rudas, I., Kubysheva, N. (2023)**. Parametric negations of probability distributions and fuzzy distribution sets. Computación y Sistemas, Vol. 27, No. 3, pp. 619–625. DOI: 10.13053/cys-27-3- 4709.
13. **Batyrshin, I. (2019)**. Towards a general theory of similarity and association measures: similarity, dissimilarity and correlation functions. Journal of Intelligent & Fuzzy Systems, Vol. 36, No. 4, pp. 2977–3004. DOI: 10.3233/jifs-181503.
14. **Batagelj, V., Bren, M. (1995)**. Comparing resemblance measures. Journal of Classification, Vol. 12, No. 1, pp. 73–90. DOI: 10.1007/bf01202268.
15. **Omhover, J., Rifqi, M., Detyniecki, M. (2006).** Ranking invariance based on similarity measures in document retrieval. Adaptive Multimedia Retrieval: User, Context, and Feedback: AMR 2005, Revised selected papers 3, pp. 55–64. DOI: 10.1007/11670834_5.

Article received on 15/03/2024; accepted on 17/05/2024.
**Corresponding author is Ildar Batyrshin.*