Review of: "A New Index for Measuring the Difference Between Two Probability Distributions"

Karoly Heberger[1]

1 Research Centre for Natural Sciences

Potential competing interests: No potential competing interests to declare.

Journal: Qeios

link: https://doi.org/10.32388/ABGI6D

Submission date: Apr 15, 2024

Qeios ID: ABGI6D

Title: A New Index for Measuring the Difference Between Two Probability Distributions

Author: Hening Huang

The scope of Qeios (providing a platform for scientific contributions and encouraging discussions) is honored and even admired, but some preliminary filtering would be essential. I am well aware of the fact that the authors may require rigorous peer review before submission, and it cannot be made compulsory. However, the present contribution will enhance the scientific noise without adding much to our previous knowledge.

Further on, I will concentrate on the deficiencies and possibilities for improvement.

Confusion about distances of two distributions

One cannot avoid the feeling that known concepts and terms are renamed simply. The usage of the square of the probability function (PDF), known as a second (central) moment, *i.e.*, the variance, is straightforward. The squared PDF is straightforwardly used in maximum likelihood estimations (also in Bayesian inference). Another example of the square of the wave function has physical meaning in quantum chemistry: It shows the location probability of a particle at a specific place.

If one uses a well-known quantity, as *e.g.*, the Pearson correlation coefficient (*r*), s/he cannot claim to introduce a novel quantity as its square (*r*^2), though *r* and *r*^2 have different ranges
[-1 to +1 and 0 to +1, respectively] and also different meanings [“closeness” to linear relationships (direction included) and explained variance by the model, respectively]. A simple algebraic transformation cannot form the basis of a novelty claim.

It is especially problematic here because the distribution similarity index (DSI), denoted by φ(X1,X2), (X1 and X2 are

random variables), has already been defined earlier [Huang 2023]. Putting the same quantity on a reverse scale does not dedicate to defining a new measure; the distribution discrepancy index (DDI) = 1 – φ(X1,X2).

So much the more as i) the preprint of [Huang, 2023] has not been subjected to rigorous peer review; ii) it is unclear what the justification is for a technical note; iii) similarity and dissimilarity are not to be distinguished in drug design [Willet, 1998], multicriteria decision making (MCDM) [Ipkovich *et al.* 2024], and other disciplines. “Minimum Violation is a measure to check the ordinal consistency of an MCDM method. It penalizes order reversals ....” [Rezaei 2015]. Glover has introduced one (or more) tabu lists to prevent variables from being reversed in artificial intelligence as early as 1986. “The function of such lists is not to prevent a move from being repeated, but to prevent it from being reversed, and the prohibition against reversal is conditional rather than absolute …” [Glover 1986]. The problem is known as Simpson’s Paradox in the machine learning field. “Simpson’s paradox is a type of aggregation bias … arises when an association observed in aggregated data disappears or reverses …” [Mehrabi *et al.* 2021.]

The above short survey proves clearly that direction change should not cause any confusion unless non-experts, students do not pay attention to the different direction of distance (dissimilarity) and similarity measures.

Indeed, there are some confusions about distance measures of distributions (statistical distributions). The author commendably traced the first occurrence of the Bhattacharyya coefficient, but the subsequent interpretation of indices leads nowhere. According to Lopatecki, “Population Stability Index (PSI) … measures the *relative entropy*, or difference in information represented by two distributions.” [Lopatecki 2023a]. However, *relative entropy* and *Kullback–Leibler* (KL) *divergence* (Kullback and Leibler 1951) have not been discriminated recently [Abonyi*et al.* 2023]. “KL divergence is a non-symmetric metric that measures the *relative entropy* or difference in information represented by two distributions.” [Lopatecki 2023b]. Whether the *symmetric PSI* or the *asymmetric KL divergence* should be preferred is at least debatable. Todeschini *et al.* has evaluated 51 similarity coefficients and classified them into four different classes: symmetric, asymmetric, intermediate, and correlation-based. [Todeschini *et al.* 2012]. A variance analysis (ANOVA) revealed significant differences; just the symmetric and intermediate indices were indistinguishable [Rácz et al. 2018]. “The superiority of symmetric and intermediate coefficients contrasts with cheminformatics, where usually asymmetric measures are preferred,” as *e.g.*, the Tanimoto coefficient [Rácz *et al.* 2018].

Internet and social media participants lack rigorous peer review; hence, they add to the confusion and fail to resolve the contradictions despite their helpfulness and well-meant attitude. [Dhinakaran 2020, Lopatecki 2023].

Basic issues about the square of PDF should be resolved: is it an expectation value or variance? Are “informity” and DDI metrics?

Some statements are not substantiated in the discussion and repeated in the conclusion:

“The proposed distribution discrepancy index (DDI) … provides an appropriate measure of the difference between two probability distributions.” – Where are the proofs? When can one claim “appropriate” in this context?

“Since the distribution discrepancy index (DDI) ranges between 0 and 1, its interpretation is intuitive, simple, and

meaningful.” Why? *Ex catedra* statements lead nowhere. As most similarity indices are scaled between 0 and 1 [Todeschini *et al.* 2012], this feature cannot be considered as *differentia specifica.*

“DDI values 0.25, 0.5, and 0.75 can be interpreted as representing small, moderate, and high levels of the difference between the two probability distributions” – Oh no! By no means! These values are arbitrary without any statistical support. Moreover, the thresholds depend heavily on the degree of freedom.

Suggestions for improvements

I was thinking a lot about how to transform this technical note and the preprint [Huang 2023] into an acceptable one. Perhaps the first action should be to submit the preprint to a journal of high reputation. Involving some experts as coauthors would certainly be helpful.

It should be evaluated by independent experts:

i. Whether the “informity concept” is new.
ii. Is the “informity” a better measure of “informativeness” (whatever it is) than its competitors?
iii. Whether the above terms and “cross-informity” and “joint informity” are useful magnitudes and whether they are not linear combinations of earlier, well-known measures.
iv. The difference between entropy, hypervolume, *etc.*, and the suggested concepts should be determined, eventually with empirical examinations and comparisons.

Some more activity of the author seems to be advisable:

i. Practical, real-life examples are necessary to show the superiority of the “novel” similarity and discrepancy indices in some instances; otherwise, their creation is unjustified.
ii. The examples should be carefully validated, and an applicability domain should be assigned (defined).
iii. The author applied a Gaussian distribution, which has well-known characteristics. It would be interesting to know whether the “new” index has some unexpected features for very different distributions.
iv. Significance tests should be elaborated to support arbitrary numbers for “small, moderate, and high levels of the difference between the two probability distributions.”
v. Fair comparison is advised, *i.e.*, the usage of more performance parameters and optimization of their outcome (Post Pareto analysis).

Minor errors

“Bhattacharyya coefficient and overlapping index (Pastore and Calcagni 2019, Mulekar and Mishra 1994)” – It is somewhat odd to cite the “Bhattacharyya coefficient” without using Bhattacharyya's name. The first person should always be cited, who first raised the idea in question. The same is true for the “modified Morisita index of Horn (1966).”

“To Knowa*rize*” – ???

The equations are not numerated but referred to by numbers.

The line between the numerator and denominator should be leveled to the equation sign.

Space usage: “communities*Memoirs*”, *etc*.

Singular and plural agreement: *e.g.*, “high levels of the difference between the two probability distributions”.*Etc*.

Naturally, not all activity is necessary to elaborate a useful paper from these notes and preprints, but a simple correction of minor errors is not sufficient.

References

Abonyi J, Ipkovich J, Dörgő G, Héberger K 2023 Matrix factorization-based multi-objective ranking–What makes a good university? *PLoS ONE* **18**(4) e0284078.

DOI 10.1371/journal.pone.0284078

Dhinakaran A 2020 Using Statistical Distances for Machine Learning Observability Towards Data Science https://towardsdatascience.com/using-statistical-distance-metrics-for-machine-learning-observability-4c874cded78 (access date: May 01 / 2024)

Glover F 1986 Future Paths for Integer Programming and Links to Artificial Intelligence*Computers and Operations Research* **13**(5) 533-549 DOI 10.1016/0305-0548(86)90048-1

Huang H 2023 The theory of informity (preprint) DOI: 10.13140/RG.2.2.28832.97287

Ipkovich A, Héberger K, Sebestyén V, Abonyi J 2024 Utility function-based generalization of sum of ranking differences–country-wise analysis of greenhouse gas emissions *Ecological Indicators* **160** 111734 DOI 10.1016/j.ecolind.2024.111734

Kullback S and Leibler RA 1951 On Information and Sufficiency. *The Annals of Mathematical Statistics* **22**(1) 79-86 DOI: 10.1214/aoms/1177729694

Lopatecki J 2023a Population Stability Index (PSI): What You Need to Knowhttps://arize.com/blog-course/population-stability-index-psi/ (access date: May 01 / 2024)

Lopatecki J 2023b KL Divergence: When to Use Kullback-Leibler divergencehttps://arize.com/blog-course/kl-divergence/ (access date: May 01 / 2024)

Mehrabi N, Morstatter F, Saxena N, Lerman K, and Galstyan A 2021 A Survey on Bias and Fairness in Machine Learning *ACM Computing Surveys* **54**(6) 115 DOI 10.1145/3457607

Rácz A, Andrić F, Bajusz D, Héberger K 2018 Binary similarity measures for fingerprint analysis of qualitative metabolomic profiles *Metabolomics* **14** 29 DOI 10.1007/s11306-018-1327-y

Rezaei J 2015 Best-worst multi-criteria decision-making method *Omega* **53** 49-57 DOI 10.1016/j.omega.2014.11.009

Todeschini R, Consonni V, Xiang H, Holliday J, Buscema M, Willett P 2012 Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling* **52** 2884-2901 DOI 10.1021/ci300261r

Willett P 1998 Chemical Similarity Searching *Journal of Chemical Information and Computer Sciences* **38** 983-996 DOI 10.1021/ci9800211

May 02 / 2024

Karoly Heberger