Qeios

Two New Indices for Measuring the Difference Between Two Probability Distributions

Hening Huang
Teledyne RD Instruments (retired)
San Diego, California, USA

Abstract This paper proposes two new indices for measuring the difference between two probability distributions: one is named "distribution similarity index (DSI)" and the other is named "distribution discrepancy index (DDI)". These two indices are derived based on the concepts of informity and cross-informity in the recently proposed informity theory. Both indices range between 0 and 1. A low DSI value or a high DDI value indicates a large difference between two probability distributions. A high DSI value or a low DDI value indicates a small difference. Three examples are provided to compare the proposed indices with existing similarity and discrepancy indices.

Keywords Cross-informity; distribution similarity; informity; probability distribution

1. Introduction

Measuring the difference between two probability distributions is an important task in many fields of science and engineering, including statistics, data science, machine learning, and imaging processing. There are two types of indices for measuring the difference: similarity index and discrepancy (or divergence) index. A commonly used similarity index is the Bhattacharyya coefficient (Bhattacharyya 1943). The Bhattacharyya coefficient was originally proposed by Bhattacharyya (1943), but was reinvented by Matusita (1955), so some authors referred to it as the Matusita measure (e.g. Dhaker et al. 2019). Another commonly used similarity index is the overlapping index (Weitzman 1970). Both the Bhattacharyya coefficient and the overlapping index tell us the degree of overlap between two distributions. Both range between 0 and 1, where 0 indicates that two distributions have no overlap and 1 indicates that the two distributions completely overlap. Table 1 shows the calculation formulas of these two similarity indices

Table 1. Two existing similarity indices for comparing two distributions

	Two discrete distributions X_1 and X_2	Two continuous distributions Y_1 and Y_2
Bhattacharyya coefficient	$B(X_1, X_2) = \sum_{i=1}^{N} \sqrt{P_1(x_i)P_2(x_i)}$.	$B(Y_1, Y_2) = \int_{-\infty}^{\infty} \sqrt{p_1(y)p_2(y)}dy$
Overlapping index	$\Omega(X_1, X_2) = \sum_{i=1}^{N} \min[P_1(x_i)P_2(x_i)]$	$\Omega(Y_1, Y_2) = \int_{-\infty}^{\infty} \min[p_1(y), p_2(y)]\, dy$

A well-known discrepancy index is the Kullback–Leibler (KL) divergence (Kullback and Leibler 1951) and its extension called population stability index (PSI) (e.g. Yurdakul 2018). According to Lopatecki (2023), "The advantage of PSI over KL divergence is that it is a symmetric metric. PSI can be thought of as the round trip loss of entropy – the KL Divergence going from one distribution to another, plus the reverse of that." The PSI is 0 if two distributions are identical. However, the

https://doi.org/10.32388/ABGI6D.2

PSI is not upper bounded, and its sensitivity can arise numerical issues in applications (Nielsen 2019). Two other discrepancy indices are the Hellinger distance and the total variation distance. The Hellinger distance is related to the Bhattacharyya coefficient. The total variation distance is the complement of the overlapping index. Both the Hellinger distance and the total variation distance range between 0 and 1. Table 2 shows the calculation formulas of these two discrepancy indices (Hellinger distance and total variation distance). The Kullback–Leibler (KL) divergence and the PSI are not shown in the table because we do not consider them in this study (remark: they are calculated in the log-transformed probability space and therefore cannot be compared with the other two indices).

Table 2. Two existing discrepancy indices for comparing two distributions

	Two discrete distributions X_1 and X_2	Two continuous distributions Y_1 and Y_2
Hellinger distance	$H(X_1,X_2)=\sqrt{\frac{1}{2}\sum_{i=1}^{N}\left(\sqrt{P_1(x_i)}-\sqrt{P_2(x_i)}\right)^2}$ $=\sqrt{1-B(X_1,X_2)}$	$H(Y_1,Y_2)=\sqrt{\frac{1}{2}\int_{-\infty}^{\infty}\left(\sqrt{p_1(y)}-\sqrt{p_2(y)}\right)^2 dy}$ $=\sqrt{1-B(Y_1,Y_2)}$
Total variation distance	$\Delta(X_1,X_2)=\frac{1}{2}\sum_{i=1}^{N}\lvert P_1(x_i)-P_2(x_i)\rvert$ $=1-\Omega(X_1,X_2)$	$\Delta(Y_1,Y_2)=\frac{1}{2}\int_{-\infty}^{\infty}\lvert p_1(y)-p_2(y)\rvert dy$ $=1-\Omega(Y_1,Y_2)$

In this paper, we derive two new indices for measuring the difference between two probability distributions: one is named "distribution similarity index (DSI)" and the other is named "distribution discrepancy index (DDI)", based on the concepts of informity and cross-informity in the recently proposed informity theory. In the following sections, section 2 reviews the concepts of informity and cross-informity. Sections 3 defines a quantity called informity divergence. Section 4 presents the distribution similarity index (DSI) and the distribution discrepancy index (DDI). Section 5 presents discussion. Section 6 gives three application examples and compares the proposed DSI and DDI with the existing indices shown in Tables 1 and 2. Section 7 presents conclusion.

2. The concepts of informity and cross-informity

The concepts of informity and cross-informity were introduced in the recently proposed informity theory (Huang 2023). For a discrete random variable X with its probability mass function (PMF) $P(x)$, each outcome x_i has a probability $P(x_i)$. The discrete informity of X, denoted by $\beta(X)$, is defined as (Huang 2023)

$$\beta(X)=\sum_{i=1}^{N}[P(x_i)]^2=\mathrm{E}[P(x)], \tag{1}$$

where N is the number of all passible outcomes.

For a continuous random variable Y with the probability density function (PDF) $p(y)$, the continuous informity of Y, denoted by $\beta(Y)$, is defined as (Huang 2023)

$$\beta(Y) = \int [p(y)]^2 \, dy = \mathrm{E}[p(y)]. \tag{2}$$

Informity is a measure of the overall informativeness of an information-probability system represented by the PMF or PDF of a random variable. It has the opposite meaning of the information entropy (Huang 2023).

The cross-informity between two discrete random variables X_1 and X_2 is defined as (Huang 2023)

$$\beta(X_1 \cap X_2) = \sum_{i=1}^{N} P_1(x_i) P_2(x_i) = \mathrm{E}_{P_1}[P_2(x)] = \mathrm{E}_{p_2}[P_1(x)]. \tag{3}$$

The discrete cross-informity is symmetric, i.e. $\beta(X_1 \cap X_2) = \beta(X_2 \cap X_1)$.

On the other hand, the cross-informity between two continuous random variables Y_1 and Y_2 is defined as (Huang 2023)

$$\beta(Y_1 \cap Y_2) = \int p_1(y) p_2(y) dy = \mathrm{E}_{p_1}[p_2(y)] = \mathrm{E}_{p_2}[p_1(y)]. \tag{4}$$

The continuous cross-informity is also symmetric, i.e. $\beta(Y_1 \cap Y_2) = \beta(Y_2 \cap Y_1)$.

The cross-informity measures the similarity of two distributions.

3. Informity divergence

We define "informity divergence" as a measure of the difference between two probability distributions. For two discrete random variables X_1 and X_2, informity divergence is denoted by $D(X_1, X_2)$ and written as

$$D(X_1, X_2) = \sum [P_1(x) - P_2(x)]^2 = \beta(X_1) - 2\beta(X_1 \cap X_2) + \beta(X_2). \tag{5}$$

For two continuous random variables Y_1 and Y_2, informity divergence is denoted by $D(Y_1, Y_2)$ and written as

$$D(Y_1, Y_2) = \int [p_1(y) - p_2(y)]^2 dy = \beta(Y_1) - 2\beta(Y_1 \cap Y_2) + \beta(Y_2). \tag{6}$$

Note that Eg. (5) can be rewritten as

$$\frac{1}{2} D(X_1, X_2) = \frac{1}{2}[\beta(X_1) + \beta(X_2)] - \beta(X_1 \cap X_2), \tag{7}$$

and Eg. (6) can be rewritten as

$$\frac{1}{2} D(Y_1, Y_2) = \frac{1}{2}[\beta(Y_1) + \beta(Y_2)] - \beta(Y_1 \cap Y_2). \tag{8}$$

This shows that one-half of the informity divergence is the average of the two informities after removing the cross-informity.

4. Distribution similarity index (DSI) and distribution discrepancy index (DDI)

We define the "distribution similarity index (DSI)" as the square root of the ratio between the cross-informity of two distributions and the average informity of the two distributions. For comparing two discrete distributions of X_1 and X_2, the DSI is denoted by $\varphi(X_1, X_2)$ and written as

$$\varphi(X_1, X_2) = \sqrt{\frac{\beta(X_1 \cap X_2)}{\frac{1}{2}[\beta(X_1) + \beta(X_2)]}}. \tag{9}$$

For comparing two continuous distributions of Y_1 and Y_2, the DSI is denoted by $\varphi(Y_1, Y_2)$ and written as

$$\varphi(Y_1, Y_2) = \sqrt{\frac{\beta(Y_1 \cap Y_2)}{\frac{1}{2}[\beta(Y_1) + \beta(Y_2)]}}. \tag{10}$$

Furthermore, we define "distribution discrepancy index (DDI)" as the square root of the ratio between the informity divergence of two distributions and the sum of the two informities. For comparing two discrete random variables X_1 and X_2, the DDI is denoted by $\phi(X_1, X_2)$ and written as

$$\phi(X_1, X_2) = \sqrt{\frac{D(X_1, X_2)}{[\beta(X_1) + \beta(X_2)]}} = \sqrt{\frac{\beta(X_1) - 2\beta(X_1 \cap X_2) + \beta(X_2)}{[\beta(X_1) + \beta(X_2)]}}. \tag{11}$$

For comparing two continuous random variables Y_1 and Y_2, the DDI is denoted by $\phi(Y_1, Y_2)$ and written as

$$\phi(Y_1, Y_2) = \sqrt{\frac{D(Y_1, Y_2)}{[\beta(Y_1) + \beta(Y_2)]}} = \sqrt{\frac{\beta(Y_1) - 2\beta(Y_1 \cap Y_2) + \beta(Y_2)}{[\beta(Y_1) + \beta(Y_2)]}}. \tag{12}$$

The author noticed through internet search that the squared DSI for discrete distributions is the same as the Morisita (1959) index for measuring the similarity between communities in comparative ecological studies, and the squared DSI for continuous distributions is the same as the modified Morisita index of Horn (1966). Therefore, the proposed DSI can be considered as a modification of the Morisita index. The modification using the square root is necessary to make the DSI comparable to other two similarity indices: the Bhattacharyya coefficient and the overlapping index. This is also to be consistent with the DDI.

The DSI is related to the DDI. It is readily to show that $\mathrm{DSI} = \sqrt{1 - (\mathrm{DDI})^2}$ and $\mathrm{DDI} = \sqrt{1 - (\mathrm{DSI})^2}$.

5. Discussion

Note that both the DSI and DDI are standardized quantities, ranging between 0 and 1. They have opposite meanings. A low DSI value or a high DDI value is interpreted to mean that the two distributions are highly dissimilar or their discrepancy is large. On the other hand, a high DSI value or a low DDI value is interpreted to mean that the similarity between the two probability distributions is high or their discrepancy is small. When DSI=0 and DDI=1, the two distributions in question are widely separated without overlap. When DSI=1 and DDI=0, the two distributions are identical and completely overlap. However, there is no universal rule to define small, moderate, and high levels of the similarity or discrepancy. Intuitively and tentatively, we suggest that DSI values 0.25, 0.5, and 0.75 be interpreted as indicating small, moderate and high levels of similarity between the two distributions, whereas DDI values 0.25, 0.5, and 0.75 be interpreted as indicating small, moderate and high levels of discrepancy between the two distributions. These proposed benchmarks are similar to the I^2 values of 25%, 50%, and 75%, which are interpreted as indicating small, moderate, and high levels of heterogeneity in meta-analysis (where I^2 is the heterogeneity index, Borenstein et al. (2017)). These benchmarks are thought to provide a convenient context for discussing the results of the similarity or discrepancy analysis.

The proposed DDI can be thought of as a standardized Euclidean distance (L_2 norm). It satisfies four requirements (conditions) of a distance metric. Consider the DDI for two discrete distributions: $\phi(X_1, X_2)$. It is readily to show that $\phi(X_1, X_2)$ satisfies (a) $\phi(X_1, X_2) \geq 0$ (non-negativity), (b) $\phi(X_1, X_2) = 0$ if and only if $X_1 = X_2$ (identity), (c) $\phi(X_1, X_2) = \phi(X_2, X_1)$ (symmetry), and (d) $\phi(X_1, X_2) \leq \phi(X_1, X_3) + \phi(X_3, X_2)$ (triangle inequality). These conditions are also satisfied by the DDI for two continuous distributions: $\phi(Y, Y_2)$.

Note that the total variation distance is a standardized Manhattan distance (L_1 norm). However, the Hellinger distance is neither a L_1 nor L_2 norm.

6. Comparison examples

In this section, we consider three examples and compare the proposed DSI with two existing similarity indices (the Bhattacharyya coefficient and the overlapping index). We also compare the proposed DDI with two existing discrepancy indices (the Hellinger distance and the total variation distance).

6.1 Example 1: grade distribution of credit scores

Yurdakul (2018) showed an example of the calculation of the population stability index (PSI) for grade distribution of credit scores. His data are shown in Table 3. Our calculation results for this example are shown in Table 4.

Table 3. Data for grade distribution of credit scores (Yurdakul 2018)

Grade	Base	Target
A	0.253	0.177
B	0.302	0.262
C	0.204	0.285
D	0.134	0.158
E	0.072	0.088
F	0.026	0.025
G	0.008	0.006

Table 4. Results for measuring the difference between two distributions (example 1)

Similarity measure	Discrepancy measure
DSI $\varphi(\text{base}, \text{target}) = 0.9829$	DDI $\phi(\text{base}, \text{target}) = 0.1842$

Bhattacharyya coefficient $B(\text{base}, \text{target}) = 0.9916$	Hellinger distance: $H(\text{base}, \text{target}) = 0.0919$
Overlapping index $\Omega(\text{base}, \text{target}) = 0.8800$	Total variation distance: $\Delta(\text{base}, \text{target}) = 0.1200$

6.2 Example 2: comparison of two non-unimodal distributions

Kang and Wildes (2015) presented an example of comparing two non-unimodal distributions $f(x)$ and $g(x)$ using the Bhattacharyya coefficient. Their data are shown in Table 5. They obtained the Bhattacharyya coefficient $B(f, g) = 0.8124$. Our calculation results for this example are shown in Table 6.

Table 5. Data for comparison of two non-unimodal distributions (Kang and Wildes 2015)

x	$f(x)$	$g(x)$
0	0	0
1	0.075	0.05
2	0.35	0.23
3	0.075	0.05
4	0	0.085
5	0	0.17
6	0	0.085
7	0.075	0.05
8	0.35	0.23
9	0.075	0.05
10	0	0

Table 6. Results for measuring the difference between two distributions (example 2)

Similarity measure	Discrepancy measure
DSI $\varphi(\text{base}, \text{target}) = 0.9083$	DDI $\phi(\text{base}, \text{target}) = 0.4183$
Bhattacharyya coefficient $B(\text{base}, \text{target}) = 0.8124$	Hellinger distance: $H(\text{base}, \text{target}) = 0.4331$
Overlapping index $\Omega(\text{base}, \text{target}) = 0.6600$	Total variation distance: $\Delta(\text{base}, \text{target}) = 0.3400$

6.3 Example 3: comparison of two normal distributions

Consider Y_1 and Y_2 are normally distributed: $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. The informity of Y_1 is

$$\beta(Y_1) = \frac{1}{2\sigma_1\sqrt{\pi}}. \tag{13}$$

The informity of Y_2 is

$$\beta(Y_2) = \frac{1}{2\sigma_2\sqrt{\pi}}. \tag{14}$$

The cross-informity of Y_1 and Y_2 is

$$\beta(Y_1 \cap Y_2) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\left[-\frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]. \tag{15}$$

Note that in the special case where $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2 = \sigma$, the cross-informity reduces to the informity. That is, $\beta(Y_1 \cap Y_2) = \frac{1}{2\sigma\sqrt{\pi}} = \beta(Y_1)$ (or $\beta(Y_2)$).

It is readily to show that the DSI for comparing two normal distributions is

$$\varphi(Y_1, Y_2) = \frac{\beta(Y_1 \cap Y_2)}{\frac{1}{2}[\beta(Y_1)+\beta(Y_2)]} = \frac{2\sqrt{2}\sigma_1\sigma_2}{(\sigma_1+\sigma_2)\sqrt{(\sigma_1^2+\sigma_2^2)}} \exp\left[-\frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]. \tag{16}$$

The Bhattacharyya coefficient (*BC*) for comparing two normal distributions is (Mathoverflow 2022)

$$B(Y_1, Y_2) = \exp\left\{-\frac{1}{4}\ln\left[\frac{1}{4}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} + 2\right)\right] + \frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}\right\}. \tag{17}$$

We consider two scenarios, A: $\sigma_1 = \sigma_2 = 1$, $|\mu_1 - \mu_2|$ ranges from 0 to 6, and B: $\mu_1 = \mu_2$, $\sigma_1 = 1$, σ_2 ranges from 0.1 to 6.

When $\sigma_1 = \sigma_2$, the overlapping index $\Omega(Y_1, Y_2) = 2\Phi\left(-\frac{|\mu_1-\mu_2|}{2}\right)$. When $\mu_1 = \mu_2$, the overlapping index is (Mulekar and Mishra 1994)

$$\Omega(Y_1, Y_2) = \begin{cases} 1 - 2\Phi(b) + 2\Phi(Cb) & if\ 0 < C < 1, \\ 1 + 2\Phi(b) - 2\Phi(Cb) & if\ C \geq 1, \end{cases} \tag{18}$$

where $b = \sqrt{-\ln\left(C^2/(1 - C^2\right)}$ and $C = \sigma_1/\sigma_2$.

Figure 1 shows the comparison of the three similarity indices for Scenario A. It can be seen from Figure 1 that, in this special case that $\sigma_1 = \sigma_2 = 1$, the proposed DSI is the same as the Bhattacharyya coefficient and both are greater than the overlapping index. As expected, when the difference between the two means is zero ($\mu_1 - \mu_2 = 0$), the three similarity indices are equal to 1 because the two distributions completely overlap. When the difference between the two means is large (say $|\mu_1 - \mu_2| = 6$), the three similarity indices are close to 0 because the two distributions are widely separated.

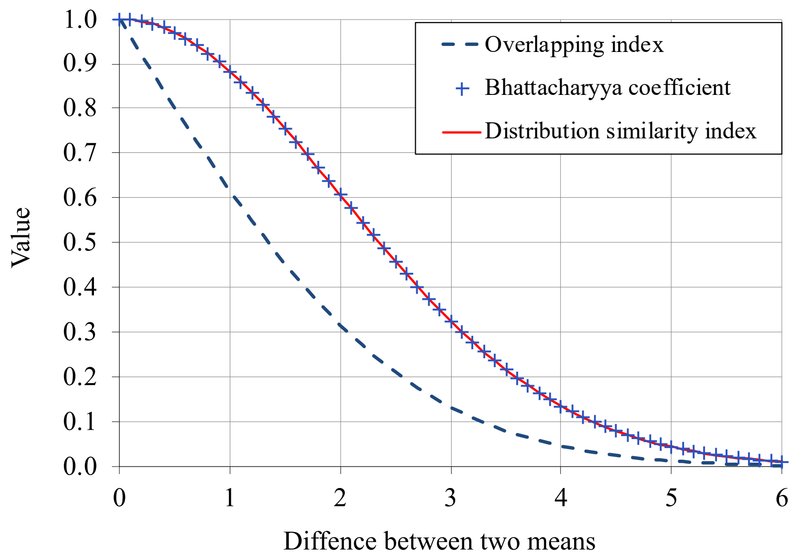

Figure 1. Comparison of the of the three similarity indices as a function of the difference between the means of two normal distributions with $\sigma_1 = \sigma_2 = 1$ (Scenario A)

Figure 2 shows the comparison of the three similarity indices for Scenario B. Note that, as expected, when $\sigma_2 = \sigma_1 = 1$, the three similarity indices are equal to 1 because the two distributions completely overlap. The proposed DSI is slightly greater than the Bhattacharyya coefficient. The overlapping index is the smallest among the three similarity indices.

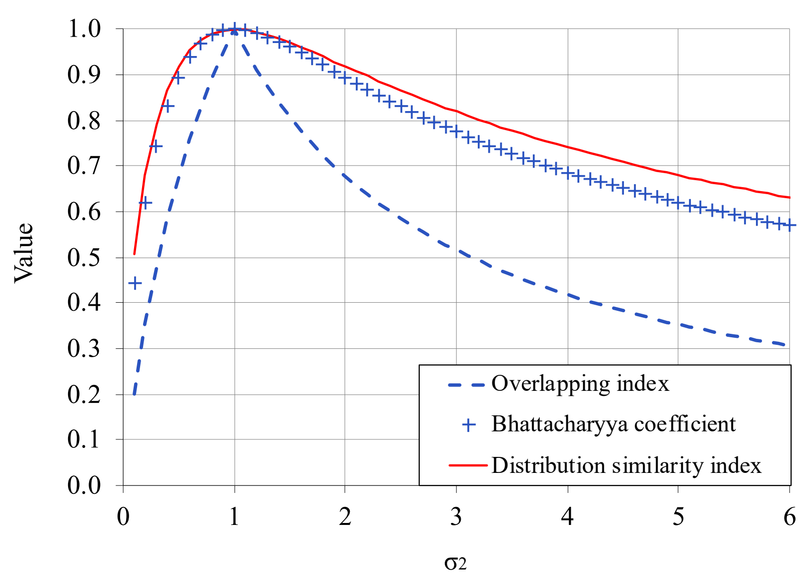

Figure 2. Comparison of the three similarity indices for the two normal distributions with $\mu_1 = \mu_2$, $\sigma_1 = 1$, and σ_2 ranges from 0.1 to 6 (Scenario B)

Figure 3 shows the comparison of the three discrepancy indices for Scenario A. Note that, as expected, when the difference between the two means is zero ($|\mu_1 - \mu_2| = 0$), the three discrepancy indices are equal to 0 because the two distributions completely overlap. When the difference between the two means is large (say $|\mu_1 - \mu_2| = 6$), the three discrepancy indices are close to 1 because the two distributions are widely separated.

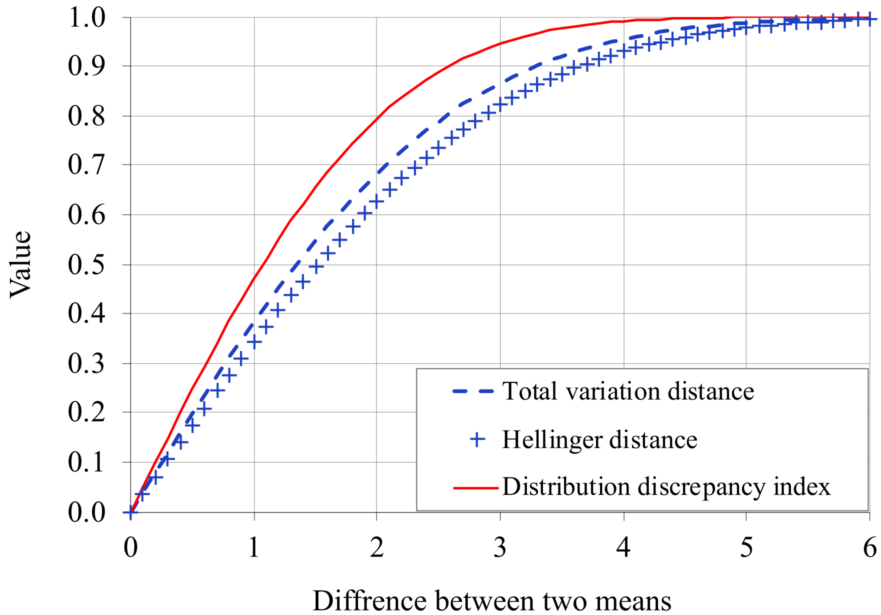

Figure 3. Comparison of the three discrepancy indices as a function of the difference between the means of two normal distributions with $\sigma_1 = \sigma_2 = 1$ (Scenario A)

Figure 4 shows the comparison of the three discrepancy indices for Scenario B. Note that, as expected, when $\sigma_2 = \sigma_1 = 1$, the three discrepancy indices are zero because the two distributions completely overlap. The proposed DDI is the greatest among the three discrepancy indices.

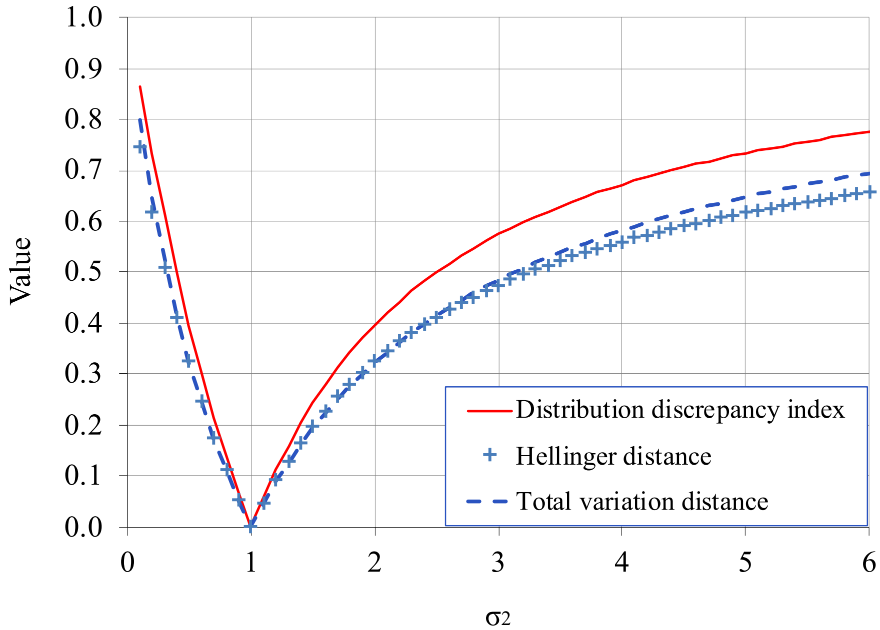

Figure 4. Comparison of the three discrepancy indices for two normal distributions with equal means ($\mu_1 = \mu_2$), $\sigma_1 = 1$, and σ_2=0.1–6 (Scenario B)

7. Conclusion

The proposed distribution similarity index (DSI) and the distribution discrepancy index (DDI) quantify the difference between two probability distributions from different perspectives. The DSI and the DDI have opposite meanings. The three examples show that among the three similarity indices, the DSI and the Bhattacharyya coefficient are consistent; the overlapping index is not consistent with the DSI or the Bhattacharyya coefficient. On the other hand, among the three discrepancy indices, the proposed DDI seems to be almost always larger than the Hellinger distance or the total variation distance. Further research is needed to examine the performance of the proposed DSI and DDI using more application examples or different distributions.

Acknowledgements An earlier version of this paper was posted as a preprint on Qeios and received comments from 14 reviewers. The author would like to thank these reviewers for their time and comments, which helped to improve the quality of this paper. The author would also like to thank Qeios for providing an open peer review planform.

References

Bhattacharyya A 1943 On a measure of divergence between two statistical populations defined by their probability distributions *Bulletin of the Calcutta Mathematical Society* **35** 99–109 MR 0010358

Borenstein M, Higgins J P T, Larry V, Hedges L V, and Rothstein H R 2017 Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity *Research Synthesis Methods* https://doi.org/10.1002/jrsm.1230

Dhaker H, Ngom P, Ibrahimouf B, and Mbodj M 2019 Overlap coefficients based on Kullback-Leibler of two normal densities: equal means case *Journal of Mathematics Research, Canadian Center of Science and Education* **11**(2) 114-124

Hellinger, E (1909) Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen *Journal für die reine und angewandte Mathematik (in German), **1909** (136): 210–271, doi:10.1515/crll.1909.136.210, JFM 40.0393.01, S2CID 121150138*

Horn H S 1966 Measurement of "Overlap" in Comparative Ecological Studies *The American Naturalist* **100**(914) 419-424

Huang H 2023 The theory of informity (preprint) *ResearchGate* https://www.researchgate.net/publication/376206296_The_theory_of_informity DOI: 10.13140/RG.2.2.28832.97287

Kang S and Wildes R P 2015 *The n-Distribution Bhattacharyya Coefficient* Technical Report EECS-2015-02, York University, Toronto, Ontario, Canada https://www.eecs.yorku.ca/research/techreports/2015/EECS-2015-02.pdf

Kullback S and Leibler R A 1951 On Information and Sufficiency *Annals of Mathematical Statistics* **22** 79-86

Lopatecki J 2023 Population Stability Index (PSI): What You Need To Know *arize* https://arize.com/blog-course/population-stability-index-psi/ accessed March 2024

Mathoverflow 2022 How close are two Gaussian random variables? URL (version: 2022-04-30): https://mathoverflow.net/q/421397 accessed November 2023

Matusita K 1955 Decision rules, based on the distance, for problems of fit, two samples, and estimation *The Annals of Mathematical Statistics* **26**(4) 631-640 https://doi.org/10.1214/aoms/1177728422

Morisita M 1959 Measuring of interspecific association and similarity between communities *Memoirs of the Faculty of Science, Kyushu Univ., Series E (Biology)* **3** 65-80

Mulekar M S and Mishra S N 1994 Overlap coefficients of two normal densities: equal means case *J. Japan Statistican Soc.* **24**(2) 169-180

Nielsen F 2019 On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means *Entropy* 21(5) 485. https://doi.org/10.3390/e21050485

Weitzman, M. S. (1970). Measures of overlap of income distributions of white and Negro families in the United States (Vol. 22). US Bureau of the Census.

Yurdakul B 2018 *Statistical Properties of Population Stability Index* Dissertations 3208 Western Michigan University https://scholarworks.wmich.edu/dissertations/3208