A New Index for Measuring the Difference Between Two Probability Distributions

Hening Huang

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

This technical note proposes a new index for measuring the difference between two probability distributions, named distribution discrepancy index (DDI). The proposed distribution discrepancy index is derived based on the concepts of informity, cross-informity, and informity divergence in the recently proposed informity theory. It is defined as the ratio between the informity divergence of two probability distributions and the sum of the two informities. The proposed distribution discrepancy index ranges between 0 and 1, which makes its interpretation intuitive, simple, and meaningful. A low DDI value (e.g. close to 0) indicates that the difference between the two probability distributions is small. A high DDI value (e.g. close to 1) indicates that the difference is large. Two examples are presented to demonstrate the application of the proposed distribution discrepancy index.

Hening Huang

Teledyne RD Instruments (retired)

San Diego, California, USA

Keywords: Cross-informity; distribution discrepancy; informity; probability distribution.

1. Introduction

There are two types of indices for comparing two probability distributions: similarity index and divergence index. Bhattacharyya coefficient and overlapping index (Pastore and Calcagni 2019, Mulekar and Mishra 1994) are two well-known similarity indices. The Bhattacharyya coefficient was originally proposed by Bhattacharyya (1943), but Matusita (1955) seemed to have reinvented it, so some authors referred to it as the Matusita measure (e.g. Dhaker et al. 2019). Huang (2023) proposed an informity-based index for measuring the similarity between two distributions, named distribution similarity index (DSI). The DSI for discrete distributions is the same as the Morisita (1959) index for measuring the similarity between communities in ecological studies, and the DSI for continuous distributions is the same as the

modified Morisita index of Horn (1966).

A well-known divergence index is 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.”

In this technical note, we propose a new index for measuring the difference between two distributions, named distribution discrepancy index (DDI). In the following, section 2 reviews the concepts of informity and cross-informity. Sections 3 defines informity divergence and distribution discrepancy index. Section 4 presents discussion. Section 5 presents two examples. Section 6 presents conclusion.

2. The concepts of informity and cross-informity

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

$$\beta(X) = \sum_{i=1}^{N} [P\left(x_i\right)]^2 = E[P(x)].$$

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)

$$\beta(Y) = \int [p(y)]^2 dy = E[p(y)].$$

Informity is a measure of the overall informativeness of an information-probability system represented by a 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 with PMFs $P_1(x)$ and $P_2(x)$ is defined as (Huang 2023)

$$\beta\left(X_1 \cap X_2\right) = \sum_{i=1}^{N} P_1\left(x_i\right) P_2\left(x_i\right) = E_{P_1}\left[P_2(x)\right] = E_{p_2}\left[P_1(x)\right]$$

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

On the other hand, the cross-informity between two continuous random variables Y_1 and Y_2 with PDFs $p_1(y)$ and $p_2(y)$ is defined as (Huang 2023)

$$\beta\left(Y_1 \cap Y_2\right) = \int p_1(y)p_2(y)dy = E_{p_1}\left[p_2(y)\right] = E_{p_2}\left[p_1(y)\right].$$

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

The cross-informity measures the similarity of two distributions.

3. Informity divergence and distribution discrepancy index (DDI)

3.1. *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

$$D(X_1, X_2) = \sum [P_1(x) - P_2(x)]^2 = \beta\left(X_1\right) - 2\beta\left(X_1 \cap X_2\right) + \beta\left(X_2\right).$$

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

$$D(Y_1, Y_2) = \int [p_1(y) - p_2(y)]^2 dy = \beta\left(Y_1\right) - 2\beta\left(Y_1 \cap Y_2\right) + \beta\left(Y_2\right).$$

Note that Eg. (5) can be rewritten as

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

and Eg. (6) can be rewritten as

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

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

3.2. *Distribution discrepancy index (DDI)*

Furthermore, we define “distribution discrepancy index (DDI)” as 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 distribution discrepancy index (DDI), denoted by $\phi(X_1, X_2)$, is written as

$$\phi\left(X_1, X_2\right) = \frac{D\left(X_1, X_2\right)}{\left[\beta\left(X_1\right) + \beta\left(X_2\right)\right]} = 1 - \varphi\left(X_1, X_2\right),$$

where $\varphi(X_1, X_2)$ is the distribution similarity index (DSI) for comparing two discrete distributions ofX_1 and X_2 (Huang 2023)

$$\varphi(X_1, X_2) = \frac{\beta\left(X_1 \cap X_2\right)}{\frac{1}{2}[\beta\left(X_1\right) + \beta\left(X_2\right)]} = \frac{2\sum_{i=1}^{N} P_1\left(x_i\right)P_2\left(x_i\right)}{\sum_{i=1}^{N}[P_1\left(x_i\right)]^2 + \sum_{i=1}^{N}[P_2\left(x_i\right)]^2}.$$

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

$$\phi\left(Y_1, Y_2\right) = \frac{D\left(Y_1, Y_2\right)}{\left[\beta\left(Y_1\right) + \beta\left(Y_2\right)\right]} = 1 - \varphi\left(Y_1, Y_2\right),$$

where $\varphi(Y_1, Y_2)$ is the distribution similarity index (DSI) for comparing two continuous distributions ofY_1 and Y_2 (Huang 2023)

$$\varphi(Y_1, Y_2) = \frac{\beta\left(Y_1 \cap Y_2\right)}{\frac{1}{2}[\beta\left(Y_1\right) + \beta\left(Y_2\right)]} = \frac{2\int_{-\infty}^{\infty} p_1(y)p_2(y)dy}{\int_{-\infty}^{\infty}[p_1(y)]^2dy + \int_{-\infty}^{\infty}[p_2(y)]^2dy}.$$

4. Discussion

It is important to note that the proposed distribution discrepancy index (DDI) ranges between 0 and 1. This makes the interpretation of the DDI is intuitive, simple, and meaningful. A low DDI value is interpreted to mean that the difference between the two probability distributions is small. A high DDI value is interpreted to mean that the difference is large. For example, 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.

Moreover, the distribution discrepancy index (DDI) is related to the distribution similarity index (DSI): DDI=1-DSI. This relationship is a good property of both indices. In the case that DDI=0, the two distributions in question are identical and DSI=1. On the other hand, in the case that DDI=1, the two distributions in question are widely separated and DSI=0.

5. Application examples

5.1. *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 1.

Table 1. 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

Yurdakul (2018) calculated the population stability index $PSI(base, target)$=0.068. We calculated informity divergence $D(base, target)$ = 0.015 and the distribution discrepancy index $\phi(base, target)$ = 0.034. Thus, the difference between the target and the base is small.

5.2. *Comparison of two normal distributions*

Consider Y_1 and Y_2 are normally distributed: $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with PDFs $p_1(y)$ and $p_2(y)$ respectively. The informity of Y_1 is

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

The informity of Y_2 is

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

The cross-informity of Y_1 and Y_2 is

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

The distribution discrepancy index (DDI) for comparing the two normal distributions is

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

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.

Figure 1 shows the distribution discrepancy index (DDI) for scenario (a). Note that, as expected, when the difference between the two means is zero ($|\mu_1 - \mu_2|$=0), DDI=0 because the two distributions are identical. When the difference is large (say $|\mu_1 - \mu_2|$=6), the DDI value is close to 1 because the two distributions are widely separated.

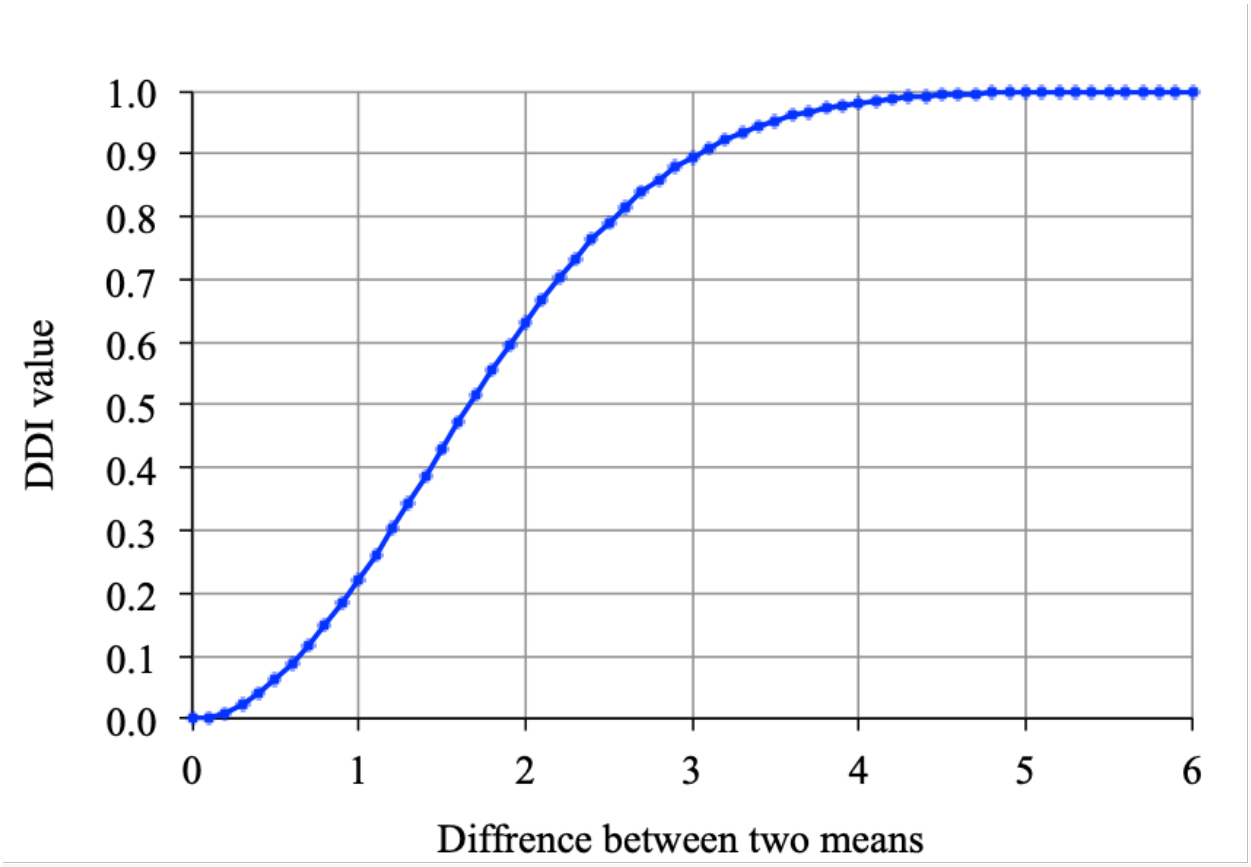

Figure 1. Distribution discrepancy index (DDI) as a function of the difference between the means of two normal distributions with $\sigma_1 = \sigma_2 = 1$

Figure 2 shows the distribution discrepancy index (DDI) for scenario (b). Note that, as expected, when $\sigma_2 = \sigma_1 = 1$, the DDI value is zero because the two distributions are identical.

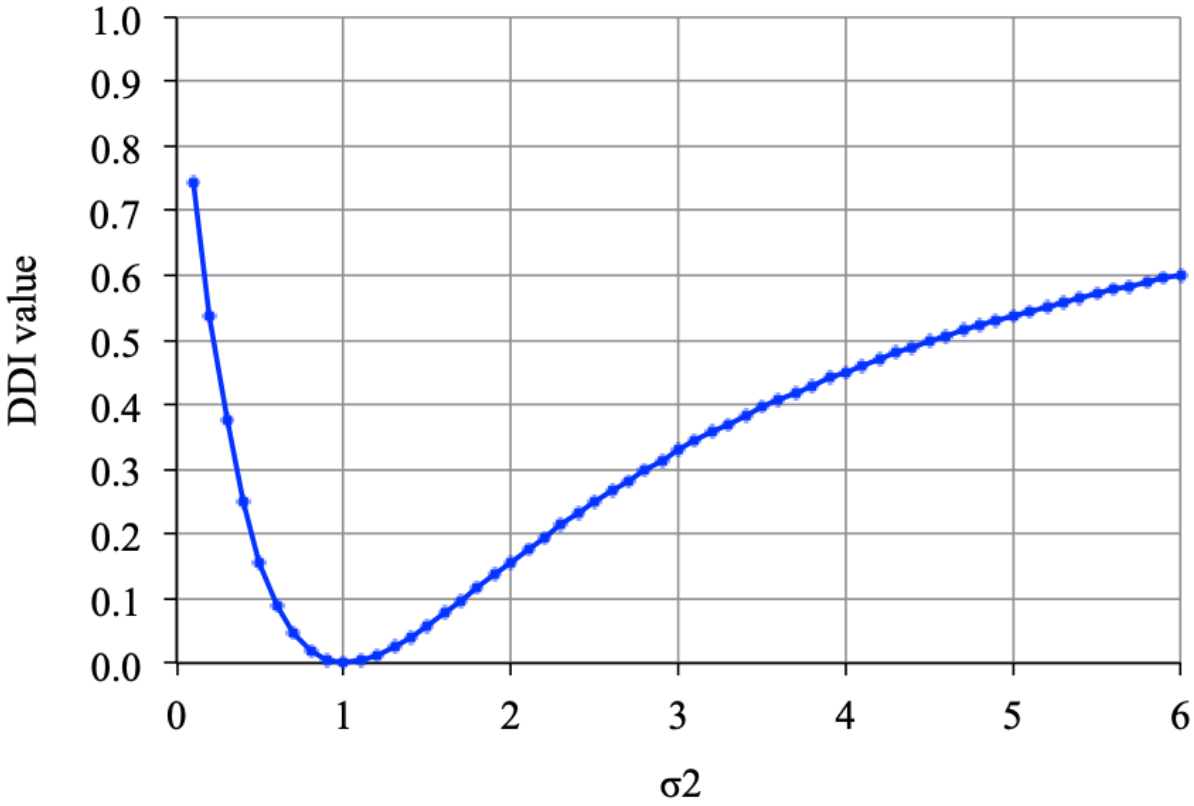

Figure 2. Distribution discrepancy index (DDI) for two normal distributions with equal means ($\mu_1 = \mu_2$), $\sigma_1 = 1$, and σ_2=0.1–6

6. Conclusion

The proposed distribution discrepancy index (DDI) is the ratio between the informity divergence of two distributions and the sum of the two informities. It provides an appropriate measure of the difference between two probability distributions. Since the distribution discrepancy index (DDI) ranges between 0 and 1, its interpretation is intuitive, simple, and meaningful. 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. Moreover, the distribution discrepancy index (DDI) is related to the distribution similarity index (DSI): DDI=1-DSI. This relationship is a good property of both indices.

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
- 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
- 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
- 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
- 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
- Pastore M and Calcagnì A 2019 Measuring distribution similarities between samples: A distribution-free overlapping index *Front Psychol.* **21**(10) 1089 doi: 10.3389/fpsyg.2019.01089
- Yurdakul B 2018 *Statistical Properties of Population Stability Index* Dissertations 3208 Western Michigan University https://scholarworks.wmich.edu/dissertations/3208