A Note-So-Comprehensive List of Dissimilarity Measures for Probability Distributions

Kisung You

kyoustat@gmail.com

April 27, 2021

Contents

1 Introduction

2 The List

We begin this section by introducing notations. Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space with sample space $\mathcal{X}$, its σ-algebra $\mathcal{F}$, and some Lebesgue or counting measure μ. We mostly consider two probability measures P and Q, both of which are dominated by μ with respect to Radon-Nikodym densities $p = dP/d\mu$ and $q = dQ/d\mu$.

Abbr.	Full Name	Definition
BD	Bhattacharyya Distance	$D_{BD}[P:Q] = -\log\left(\int_{\mathcal{X}} \sqrt{p(x)q(x)} d\mu(x)\right)$
CSD	Cauchy-Schwarz Divergence	$D_{CSD}[P,Q] = -\log\left(\frac{\int_{\mathcal{X}} p(x)q(x)d\mu(x)}{\sqrt{\int_{\mathcal{X}} p(x)^2 d\mu(x) \int_{\mathcal{X}} q(x)^2 d\mu(x)}}\right)$
HD	Hellinger Distance	$D_{HD}[P:Q] = \sqrt{\frac{1}{2}\int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 d\mu(x)}$
KLD	Kullback-Leibler Divergence	$D_{KLD}[P:Q] = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x)$
JD	Jeffreys Divergence	$D_{JD}[P:Q] = D_{KL}[P:Q] + D_{KL}[Q:P]$
RD	Rényi Divergence	$D_{RD,\alpha}[P:Q] = \frac{1}{\alpha-1} \log\left(\int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha} d\mu(x)\right)$

Table 1: Summary table of dissimilarity measures.

2.1 Bhattacharyya Distance

Bhattacharyya (1946) proposed a dissimilarity measure between two multinomial populations, which was later generalised for arbitrary measures. Bhattacharyya distance (BD) is closely related to the Bhattacharyya coefficient $\rho(P,Q)$ which measures the amount of overlap between two populations

$$\rho(P,Q) = \int_{\mathcal{X}} \sqrt{p(x)q(x)} d\mu(x)$$

in that the BD is defined using the coefficient

$$D_{BD}[P:Q] = -\log(\rho(P,Q)) = -\log\left(\int_{\mathcal{X}} \sqrt{p(x)q(x)} d\mu(x)\right) \tag{1}$$

Property 1. $D_{BD}[P:Q]$ is non-negative and symmetric.

Property 2. $0 \leq \rho(P,Q) \leq 1$ so that $D_{BD} \in [0,\infty)$.

Proof of Property 2. The Cauchy-Schwarz inequality for two densities gives that

$$\begin{aligned}\rho(P,Q) = \int_{\mathcal{X}} \sqrt{p(x)q(x)} d\mu(x) &= |\langle \sqrt{p(x)}, \sqrt{q(x)} \rangle| \leq \|\sqrt{p(x)}\| \|\sqrt{q(x)}\| \\ &= \sqrt{\int_{\mathcal{X}} p(x) d\mu(x)} = \sqrt{\int_{\mathcal{X}} q(x) d\mu(x)} = 1 \cdot 1 = 1\end{aligned}$$

and taking the negative of the log of the Bhattacharyya coefficient gives the range. □

2.2 Cauchy-Schwarz Divergence

The Cauchy-Schwarz inequality states that for vectors u and v of an inner product space,

$$|\langle u, v\rangle|^2 \leq \langle u, u\rangle \cdot \langle v, v\rangle$$

which motivates a dissimilarity measure as follows,

$$|\langle u, v\rangle| \leq \sqrt{\langle u, u\rangle} \cdot \sqrt{\langle v, v\rangle} \Rightarrow \frac{|\langle u, v\rangle|}{\|u\| \cdot \|v\|} \leq 1 \Rightarrow -\log\left(\frac{|\langle u, v\rangle|}{\|u\| \cdot \|v\|}\right) \geq 0.$$

From the observation above, Kampa et al. (2011) proposed Cauchy-Schwarz Divergence (CSD)

$$D_{CSD}[P_1 : P_2] = -\log\left(\frac{\int p_1(x)p_2(x)d\mu(x)}{\sqrt{\int p_1(x)^2 d\mu(x) \int p_2(x)^2 d\mu(x)}}\right) \tag{2}$$

2.3 Hellinger Distance

Hellinger distance (HD) is a metric for probability distributions (Hellinger; 1909). It is closely related to the Bhattacharyya distance since HD can be defined using Bhattacharyya coefficient shown in Equation (1).

$$D_{HD}[P : Q] = \sqrt{1 - \rho(P, Q)} = \left(1 - \int_{\mathcal{X}} \sqrt{p(x)q(x)} d\mu(x)\right)^{1/2} \tag{3}$$

where the Equation (3) is a bit different from Table 1 but two are equivalent expressions, which can be derived from simple algebra.

$$\begin{aligned} D_{HD}^2 &= \frac{1}{2}\int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 d\mu(x) = \frac{1}{2}\int_{\mathcal{X}} \{p(x) + q(x)\} d\mu(x) + -\frac{2}{2}\int_{\mathcal{X}} \sqrt{p(x)q(x)} d\mu(x) \\ &= \frac{1}{2} + \frac{1}{2} - \int_{\mathcal{X}} \sqrt{p(x)q(x)} d\mu(x) = 1 - \int_{\mathcal{X}} \sqrt{p(x)q(x)} d\mu(x). \end{aligned}$$

Property 1. $D_{HD}[P : Q]$ is a metric.

2.4 Jeffreys Divergence

Jeffreys (1946) proposed a divergence measure that symmetrizes the Kullback-Leibler divergence

$$D_{JD}[P : Q] = D_{KL}[P : Q] + D_{KL}[Q : P] \tag{4}$$

by summing two KL divergences of opposite directions.

2.5 Kullback-Leibler Divergence

Kullback-Leibler divergence (KLD), also called relative entropy, is one of the most fundamental measure of discrepancy between two probability measures with long history since its inception by Kullback and Leibler (1951). For two measures P and Q, KLD is defined as

$$D_{KLD}[P:Q] = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x) \tag{5}$$

Property 1. $D_{KLD}[P:Q]$ is non-negative and asymmetric.

2.6 Rényi Divergence

Rényi divergence (RD) can be considered as a generalization of several dissimilarities (Rényi; 1961). RD involves a single parameter $\alpha \in (0,\infty),\ \alpha \neq 1$ known as an order that controls balance between two distributions is defined as

$$D_{RD,\alpha}[P:Q] = \frac{1}{\alpha-1} \log \left(\int_{\mathcal{X}} p(x)^{\alpha} q(x)^{1-\alpha} d\mu(x) \right). \tag{6}$$

Some special cases for Rényi divergence of order α in the limit sense include (1) twice the Bhattacharyya distance ($\alpha = 1/2$) and (2) the Kullback-Leibler divergence ($\alpha = 1$).

2.7 Wasserstein Distance

Wasserstein Distance (WD) uses the language from the theory of optimal transport **NEEDREF** d, which is a distance over the set of measures with the finite moment of order p. Usually noted as W_p, the p-Wasserstein distance for two measures P and Q on a metric space $(\mathbb{X}, d)$ is defined as

$$D_{WD,p}[P:Q] = \left(\inf_{\pi \in \Pi(P,Q)} d(x,y)^p d\pi(x,y) \right)^{\frac{1}{p}} \tag{7}$$

where ..

3 Case Study : Gaussian Distributions

For a d-dimensional random variable X, we say it is normally distributed $X \sim \mathcal{N}(\mu, \Sigma)$ with two parameters $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ for mean and variance respectively. The density function is written as

$$f(x|\mu, \Sigma) = (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right) \tag{8}$$

where $\det(\Sigma)$ is the determinant of a given square matrix Σ. In this section, we derive some closed-form formula of dissimilarities introduced in Section 2 for a pair of Gaussian distributions $P_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $P_2 = \mathcal{N}(\mu_2, \Sigma_2)$ whose densities are denoted as $p_1(x)$ and $p_2(x)$ respectively.

Bhattacharyya Distance

$$D_{BD}[P_1 : P_2] = \frac{1}{8}(\mu_1 - \mu_2)^\top \Sigma_*^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\log\left(\frac{|\Sigma_*|}{\sqrt{|\Sigma_1||\Sigma_2|}}\right) \tag{9}$$

where $\Sigma_* = (\Sigma_1 + \Sigma_2)/2$.

Proof. We utilize the fact that Rényi divergence of order $\alpha = 1/2$ is twice the Bhattacharyya distance.

$$\begin{aligned} D_{RD,1/2}[P_1 : P_2] &= -2\log\left(\int_{\mathcal{X}} \sqrt{p_1(x)}\sqrt{p_2(x)}d\mu(x)\right) \\ &= 2\left\{-\log\left(\int_{\mathcal{X}} \sqrt{p_1(x)p_2(x)}d\mu(x)\right)\right\} = 2D_{BD}[P_1 : P_2]. \end{aligned}$$

By plugging $\alpha = 1/2$ in the Equation (12), we get

$$D_{RD,1/2}[P_1 : P_2] = \frac{1}{4}\Delta\mu^\top \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} \Delta\mu + \log\left(\frac{|(\Sigma_1 + \Sigma_2)/2|}{|\Sigma_1|^{1/2}|\Sigma_2|^{1/2}}\right)$$

where $\Delta\mu = \mu_1 - \mu_2$. Dividing the above by 2, we acquire the result as shown in Equation (9). □

Cauchy-Schwarz Divergence

$$D_{CSD}[P_1 : P_2] = -\log\left(\frac{\mathcal{N}(\mu_1|\mu_2, \Sigma_1 + \Sigma_2)}{\sqrt{\mathcal{N}(\mu_1|\mu_1, 2\Sigma_1)\cdot\mathcal{N}(\mu_2|\mu_2, 2\Sigma_2)}}\right) \tag{10}$$

Proof. We use **Fact 1** repeatedly where

$$\int p_1(x)p_2(x)d\mu(x) = \mathcal{N}(\mu_1|\mu_2, \Sigma_1 + \Sigma_2)\int_{\mathcal{X}} \mathcal{N}(x|\mu_{12}, \Sigma_{12})d\mu(x) = \mathcal{N}(\mu_1|\mu_2, \Sigma_1 + \Sigma_2)$$

$$\int p_i(x)^2 d\mu(x) = \mathcal{N}(\mu_i|\mu_i, \Sigma_i + \Sigma_i)\int_{\mathcal{X}} \mathcal{N}(x|\mu_{ii}, \Sigma_{ii})d\mu(x) = \mathcal{N}(\mu_i|\mu_i, 2\Sigma_i) \text{ for } i = 1, 2$$

and plugging the above in the definition gives the closed-form expression in Equation (10). □

Hellinger Distance

$$D_{HD}[P_1 : P_2] = \left[1 - \frac{(|\Sigma_1||\Sigma_2|)^{1/4}}{|\Sigma_*|^{1/2}} \cdot \exp\left(-\frac{1}{8}(\mu_1 - \mu_2)^\top \Sigma_*^{-1}(\mu_1 - \mu_2)\right)\right]^{1/2} \tag{11}$$

where $\Sigma_* = (\Sigma_1 + \Sigma_2)/2$.

Proof. We use the following relation with respect to the Bhattacharyya coefficient ,

$$\begin{aligned} D_{BD}[P_1 : P_2] &= -\log \rho(P_1, P_2) \quad &\leftrightarrow \quad \rho(P_1, P_2) &= \exp(-D_{BD}[P_1 : P_2]) \\ D_{HD}[P_1 : P_2] &= \sqrt{1 - \rho(P_1, P_2)} \quad &\leftrightarrow \quad \rho(P_1, P_2) &= 1 - D_{HD}[P_1 : P_2]^2 \end{aligned}$$

so that we have

$$\exp(-D_{BD}[P_1 : P_2]) = 1 - D_{HD}[P_1 : P_2]^2 \ \rightarrow \ D_{HD}[P_1 : P_2] = \sqrt{1 - \exp(-D_{BD}[P_1 : P_2])}.$$

Since we are given close-form formulae of the Bhattacharyya distance in Equation (9), re-arranging the terms with respect to the above relation gives the result. □

Rényi Divergence

We assume the order $\alpha \neq 1$ since the Equation (6) is not properly defined and the equivalence to KL divergence makes sense only in the limiting sense.

$$\begin{aligned} D_{RD,\alpha}[P_1 : P_2] =& \frac{\alpha}{2}(\mu_1 - \mu_2)^\top [\alpha\Sigma_2 + (1-\alpha)\Sigma_1]^{-1}(\mu_1 - \mu_2) \\ &- \frac{1}{2(\alpha - 1)} \log\left(\frac{|\alpha\Sigma_2 + (1-\alpha)\Sigma_1|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^{\alpha}}\right) \end{aligned} \tag{12}$$

Proof. The Rényi divergence of order α for two densities p_1 and p_2 is defined as

$$D_{RD,\alpha}[P_1 : P_2] = \frac{1}{\alpha - 1} \log \int p_1(x)^\alpha p_2(x)^{1-\alpha} d\mu(x)$$

so we focus on the integral term,

$$
\begin{aligned}
\int p_1(x)^\alpha p_2(x)^{1-\alpha} d\mu(x) &= \int \left[(2\pi)^{-d/2}|\Sigma_1|^{-1/2}\exp\left(-\frac{(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1)}{2}\right)\right]^\alpha \\
&\times \left[(2\pi)^{-d/2}|\Sigma_2|^{-1/2}\exp\left(-\frac{(x-\mu_2)^\top\Sigma_2^{-1}(x-\mu_2)}{2}\right)\right]^{1-\alpha} \\
&= (2\pi)^{-d/2}|\Sigma_1|^{-\frac{\alpha}{2}}|\Sigma_2|^{-\frac{1-\alpha}{2}} \\
&\times \int \exp\left(-\frac{(x-\mu_1)^\top\alpha\Sigma_1^{-1}(x-\mu_1)+(x-\mu_2)^\top(1-\alpha)\Sigma_2^{-1}(x-\mu_2)}{2}\right) d\mu(x).
\end{aligned}
$$

We can integrate out the second term by completing the square in a multivariate manner

$$
(x-\mu_1)^\top\alpha\Sigma_1^{-1}(x-\mu_1)+(x-\mu_2)^\top(1-\alpha)\Sigma_2^{-1}(x-\mu_2) = (x-\tilde{\mu})^\top S^{-1}(x-\tilde{\mu}) + C
$$

where

$$
\begin{aligned}
S &= \left[\alpha\Sigma_1^{-1}+(1-\alpha)\Sigma_2^{-1}\right]^{-1} \\
C &= \alpha(1-\alpha)(\mu_1-\mu_2)^\top\left[\alpha\Sigma_2+(1-\alpha)\Sigma_1\right]^{-1}(\mu_1-\mu_2) \\
\tilde{\mu} &= S\left(\alpha\Sigma_1^{-1}\mu_1+(1-\alpha)\Sigma_2^{-1}\mu_2\right).
\end{aligned}
$$

We denote $\Delta\mu = \mu_1-\mu_2$ and the above simplification gives

$$
\begin{aligned}
\int p_1(x)^\alpha p_2(x)^{1-\alpha} d\mu(x) &= (2\pi)^{-d/2}|\Sigma_1|^{-\frac{\alpha}{2}}|\Sigma_2|^{-\frac{1-\alpha}{2}}\exp\left(-\frac{1}{2}C\right)(2\pi)^{d/2}|S|^{1/2} \\
&= \frac{|\alpha\Sigma_1^{-1}+(1-\alpha)\Sigma_2^{-1}|^{-1/2}}{|\Sigma_1|^{\alpha/2}|\Sigma_2|^{(1-\alpha)/2}}\exp\left(-\frac{\alpha(1-\alpha)\Delta\mu^\top\left[\alpha\Sigma_2+(1-\alpha)\Sigma_1\right]^{-1}\Delta\mu}{2}\right)
\end{aligned}
$$

and multiply $(|\Sigma_1||\Sigma_2|)^{-1/2}$ in the denominator and numerator of the first term using the fact that $|AB| = |A||B|$ naturally implies $|AB|^{-1/2} = (|A||B|)^{-1/2}$ so that

$$
= \frac{|\alpha\Sigma_2+(1-\alpha)\Sigma_1|^{-1/2}}{|\Sigma_1|^{\frac{\alpha-1}{2}}|\Sigma_2|^{-\frac{\alpha}{2}}}\exp\left(\frac{\alpha(\alpha-1)\Delta\mu^\top\left[\alpha\Sigma_2+(1-\alpha)\Sigma_1\right]^{-1}\Delta\mu}{2}\right)
$$

so that finally we reach the following;

$$
\begin{aligned}
D_{RD,\alpha}[P_1:P_2] &= \frac{\alpha(\alpha-1)\Delta\mu^\top\left[\alpha\Sigma_2+(1-\alpha)\Sigma_1\right]^{-1}\Delta\mu}{2(\alpha-1)} + \frac{1}{\alpha-1}\log\left(\frac{|\alpha\Sigma_2+(1-\alpha)\Sigma_1|^{-1/2}}{|\Sigma_1|^{\frac{\alpha-1}{2}}|\Sigma_2|^{-\frac{\alpha}{2}}}\right) \\
&= \frac{\alpha}{2}\Delta\mu^\top[\alpha\Sigma_2+(1-\alpha)\Sigma_1]^{-1}\Delta\mu - \frac{1}{2(\alpha-1)}\log\left(\frac{|\alpha\Sigma_2+(1-\alpha)\Sigma_1|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha}\right)
\end{aligned}
$$

which completes the derivation. □

4 Miscellaneous Facts

In this section, we introduce some miscellaneous facts. We use $P_i = \mathcal{N}(\mu_i, \Sigma_i)$ for $i \in \mathcal{I}$ for Gaussian distributions in $\mathbb{R}^d$.

Fact 1 (product of two Gaussians)

For P_1 and P_2, product of two normal densities is a scaled normal density,

$$p_1(x) \cdot p_2(x) = \mathcal{N}(x|\mu_1, \Sigma_1) \cdot \mathcal{N}(x|\mu_2, \Sigma_2) = c_{12} \cdot \mathcal{N}(x|\mu_{12}, \Sigma_{12})$$

where

$$\begin{aligned} c_{12} &= \mathcal{N}(\mu_1|\mu_2, (\Sigma_1 + \Sigma_2)) \\ \Sigma_{12} &= (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} \\ \mu_{12} &= \Sigma_{12} \cdot (\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2) \end{aligned}$$

according to Section 8.1.8 of Petersen and Pedersen (2012).

Fact 2 (entropy of Gaussian distribution)

The differential entropy is of a random variable X with density $p(x)$ is defined as

$$H(X) = -\int p(x) \log p(x) dx = -\mathbb{E}_p[\log p(x)]$$

and for $X \sim \mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^d$,

$$H(X) = \frac{d}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma) + \frac{1}{2}d$$

which can be derived as follows;

$$\begin{aligned} H(X) &= -\mathbb{E}\left[\log\left\{(2\pi)^{-d/2}\det(\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\right\}\right] \\ &= -\mathbb{E}\left[-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det(\Sigma) - \frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right] \end{aligned}$$

and we can pull the constant terms out of the expectation

$$= \frac{d}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma) + \frac{1}{2}\mathbb{E}\left[(x-\mu)^\top\Sigma^{-1}(x-\mu)\right].$$

The last term is reduced to $d/2$ by the *trace trick*;

$$\begin{aligned} \mathbb{E}\left[(x-\mu)^\top\Sigma^{-1}(x-\mu)\right] &= \mathbb{E}\left[\mathrm{tr}\left((x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\right] = \mathbb{E}\left[\mathrm{tr}\left((x-\mu)(x-\mu)^\top\Sigma^{-1}\right)\right] \\ &= \mathrm{tr}\left[\mathbb{E}\left((x-\mu)(x-\mu)^\top\right)\Sigma^{-1}\right] = \mathrm{tr}\left(\Sigma\Sigma^{-1}\right) = \mathrm{tr}(I_d) = d \end{aligned}$$

in that we can acquire the following form,

$$H(X) = \frac{d}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma) + \frac{1}{2}d.$$

Fact 3 (integral of square root density)

The Bhattacharyya coefficient involves evaluation for the integral of square root density. We derive the result with respect to a single Gaussian distribution $P = \mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^d$, then

$$\int_{\mathbb{R}^d} \sqrt{p(x)} dx = (8\pi)^{d/4} |\Sigma|^{1/4}$$

which can be derived as follows;

$$\begin{aligned}
\int_{\mathbb{R}^d} \sqrt{p(x)} dx &= \int (2\pi)^{-d/4} |\Sigma|^{-1/4} \exp\left(-\frac{1}{2}(x-\mu)^\top (2\Sigma)^{-1}(x-\mu) \right) \\
&= (2\pi)^{-d/4} \cdot |\Sigma|^{-1/4} \cdot (2\pi)^{d/2} \cdot |2\Sigma|^{1/2} \int \cdots dx \\
&= (2\pi)^{d/4} \cdot |\Sigma|^{-1/4} \cdot 2^{d/2} \cdot |\Sigma|^{1/2} \\
&= \pi^{d/4} \cdot 2^{3d/4} \cdot |\Sigma|^{1/4} = (8\pi)^{d/4} |\Sigma|^{1/4}
\end{aligned}$$

where the integral with $\cdots$ is a Gaussian distribution with scaled variance parameter which integrates to 1.

References

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations, *Sankhyā: The Indian Journal of Statistics (1933-1960)* **7**(4): 401–406.

Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen., *Journal für die reine und angewandte Mathematik* **1909**(136): 210–271.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems, *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **186**(1007): 453–461.

Kampa, K., Hasanbelliu, E. and Principe, J. C. (2011). Closed-form cauchy-schwarz PDF divergence for mixture of Gaussians, *The 2011 International Joint Conference on Neural Networks*, IEEE, San Jose, CA, USA, pp. 2578–2585.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency, *The Annals of Mathematical Statistics* **22**(1): 79–86.

Petersen, K. B. and Pedersen, M. S. (2012). *The Matrix Cookbook*, Technical University of Denmark.

Rényi, A. (1961). On Measures of Entropy and Information, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, University of California Press, Berkeley, Calif., pp. 547–561.