analytics

Article

Survey of Distances between the Most Popular Distributions

Mark Kelbert

Department of Statistics and Data Analysis, The National Research University Higher School of Economics, Moscow 101000, Russia; mkelbert@hse.ru

Abstract: We present a number of upper and lower bounds for the total variation distances between the most popular probability distributions. In particular, some estimates of the total variation distances in the cases of multivariate Gaussian distributions, Poisson distributions, binomial distributions, between a binomial and a Poisson distribution, and also in the case of negative binomial distributions are given. Next, the estimations of Lévy–Prohorov distance in terms of Wasserstein metrics are discussed, and Fréchet, Wasserstein and Hellinger distances for multivariate Gaussian distributions are evaluated. Some novel context-sensitive distances are introduced and a number of bounds mimicking the classical results from the information theory are proved.

Keywords: probability distribution; total variation distance; Pinsker's inequality; Le Cam's inequalities; Lévy–Prohorov distance; Wasserstein distance; Jensen–Shannon distance; context-sensitive metrics

1. Introduction

Measuring a distance, whether in the sense of a metric or a divergence, between two probability distributions (PDs) is a fundamental endeavor in machine learning and statistics [1]. We encounter it in clustering, density estimation, generative adversarial networks, image recognition and just about any field that undertakes a statistical approach towards data. The most popular case is measuring the distance between multivariate Gaussian PDs, but other examples such as Poisson, binomial and negative binomial distributions, etc., frequently appear in applications too. Unfortunately, the available textbooks and reference books do not present them in a systematic way. Here, we make an attempt to fill this gap. For this aim, we review the basic facts about the metrics for probability measures, and provide specific formulae and simplified proofs that could not be easily found in the literature. Many of these facts may be considered as a scientific folklore known to experts but not represented in any regular way in the established sources. A tale that becomes folklore is one that is passed down and whispered around. The second half of the word, lore, comes from Old English lār, i.e., 'instruction'. The basic reference for the topic is [2], and, in recent years, the theory has achieved substantial progress. A selection of recent publications on stability problems for stochastic models may be found in [3], but not much attention is devoted to the relationship between different metrics useful in specific applications. Hopefully, this survey helps to make this treasure more accessible and easy to handle.

The rest of the paper proceeds as follows: In Section 2, we define the total variation, Kolmogorov–Smirnov, Jensen–Shannon and geodesic metrics. Section 3 is devoted to the total variation distance for 1D Gaussian PDs. In Section 4, we survey a variety of different cases: Poisson, binomial, negative-binomial, etc. In Section 5, the total variation bounds for multivariate Gaussian PDs are presented, and they are proved in Section 6. In Section 7, the estimations of Lévy–Prohorov distance in terms of Wasserstein metrics are presented. The Gaussian case is thoroughly discussed in Section 8. In Section 9, a relatively new topic of distances between the measures of different dimensions is briefly discussed. Finally, in Section 10, new context-sensitive metrics are introduced and a number of inequalities mimicking the classical bounds from information theory are proved.

check for updates

Citation: Kelbert, M. Survey of Distances between the Most Popular Distributions. *Analytics* **2023**, *2*, 225–245. https://doi.org/10.3390/analytics2010012

Academic Editor: Fabio Postiglione

Received: 6 December 2022
Revised: 19 January 2023
Accepted: 13 February 2023
Published: 1 March 2023

2. The Most Popular Distances

The most interesting metrics on the space of probability distributions are the total variation (TV), Lévy–Prohorov, Wasserstein distances. We will also discuss Fréchet, Kolmogorov–Smirnov and Hellinger distances. Let us remind readers that, for probability measures $\mathbf{P}, \mathbf{Q}$ with densities p, q,

$$\mathrm{TV}(\mathbf{P},\mathbf{Q}) = \sup_{A\subset\mathbf{R}^d} |\mathbf{P}(A) - \mathbf{Q}(A)| = \frac{1}{2}\int_{\mathbf{R}^d} |p(u) - q(u)|\mathrm{d}u \tag{1}$$

We need the coupling characterization of the total variation distance. For two distributions, $\mathbf{P}$ and $\mathbf{Q}$, a pair (X, Y) of random variables (r.v.) defined on the same probability space is called a coupling for $\mathbf{P}$ and $\mathbf{Q}$ if $X \sim \mathbf{P}$ and $Y \sim \mathbf{Q}$. Note the following fact: there exists a coupling (X, Y) such that $\mathbf{P}(X \neq Y) = \mathrm{TV}(\mathbf{P}, \mathbf{Q})$. Therefore, for any measurable function f, we have $\mathbf{P}(f(X) \neq f(Y)) \leq \mathrm{TV}(X, Y)$ with equality iff f is reversible.

In a one-dimensional case, the Kolmogorov–Smirnov distance is useful (only for probability measures on $\mathbf{R}$): $\mathrm{Kolm}(\mathbf{P},\mathbf{Q}) = \sup_{x\in\mathbf{R}} |\mathbf{P}(-\infty, x) - \mathbf{Q}(-\infty, x)| \leq \mathrm{TV}(\mathbf{P},\mathbf{Q})$. Suppose $X \sim \mathbf{P}, Y \sim \mathbf{Q}$ are two r.v.'s, and Y has a density w.r.t. Lebesgue measure bounded by a constant C. Then, $\mathrm{Kolm}(\mathbf{P},\mathbf{Q}) \leq 2\sqrt{C\mathrm{Wass}_1(\mathbf{P},\mathbf{Q})}$. Here, $\mathrm{Wass}_1(\mathbf{P},\mathbf{Q}) = \inf[\mathbf{E}|X - Y| : X \sim \mathbf{P}, Y \sim \mathbf{Q}]$.

Let X_1, X_2 be random variables with the probability density functions p, q, respectively. Define the Kullback–Leibler (KL) divergence

$$\mathrm{KL}(\mathbf{P}_{X_1}||\mathbf{P}_{X_2}) = \int p \log\frac{p}{q}. \tag{2}$$

Example 1. *Consider the scale family* $\{p_s(x) = \frac{1}{s}p\left(\frac{x}{s}\right), s \in (0,\infty)\}$. *Then,*

$$\mathrm{KL}(p_{s_1}||p_{s_2}) = \mathrm{KL}(p_{\frac{s_1}{s_2}}||p_1) = \mathrm{KL}(p_1||p_{\frac{s_2}{s_1}}).$$

The total variance distance and the Kullback–Leibler (KL) divergence appear naturally in statistics. Say, for example, in the testing of binary hypothesis $H_0{:}X \sim \mathbf{P}$ versus $H_1{:}X \sim \mathbf{Q}$, the sum of errors of both types

$$\inf_d[\mathbf{P}(d(X) = H_1) + \mathbf{Q}(d(X) = H_0)] = \int \min[p, q] = 1 - \mathrm{TV}(\mathbf{P},\mathbf{Q}) \tag{3}$$

as the infimum over all reasonable decision rules $d{:}X \to \{H_0, H_1\}$ or the critical domains W is achieved for $W^* = \{p(x) < q(x)\}$. Moreover, when minimizing the probability of type-II error subjected to type-I error constraints, the optimal test guarantees that the probability of type-II error decays exponentially in view of Sanov's theorem

$$\lim_{n\to\infty} -\frac{\ln \mathbf{Q}(d(X) = H_0)}{n} = \mathrm{KL}(\mathbf{P}||\mathbf{Q}). \tag{4}$$

where n is the sample size. In the case of selecting between $M \geq 2$ distributions,

$$\inf_d \max_{1\leq j\leq M} \mathbf{P}_j(d(X) \neq j) \geq 1 - \frac{\frac{1}{M^2}\sum_{j,k=1}^M \mathrm{KL}(\mathbf{P}_j, \mathbf{P}_k) + \log 2}{M-1}. \tag{5}$$

The KL-divergence is not symmetric and does not satisfy the triangle inequality. However, it gives rise to the so-called Jensen–Shannon metric [4]

$$\mathrm{JS}(\mathbf{P},\mathbf{Q}) = \sqrt{D(\mathbf{P}||\mathbf{R}) + D(\mathbf{Q}||\mathbf{R})}$$

with $\mathbf{R} = \frac{1}{2}(\mathbf{P}+\mathbf{Q})$. It is a lower bound for the total variance distance

$$0 \le \mathrm{JS}(\mathbf{P},\mathbf{Q}) \le \mathrm{TV}(\mathbf{P},\mathbf{Q}). \tag{6}$$

The Jensen–Shannon metric is not easy to compute in terms of covariance matrices in the multi-dimensional Gaussian case.

A natural way to develop a computationally effective distance in the Gaussian case is to define first a metric between the positively definite matrices. Let $\lambda_1,\ldots,\lambda_d$ be the generalized eigenvalues, i.e., the solutions of $\det(\Sigma_1 - \lambda\Sigma_2) = 0$. Define the distance between the positively definite matrices by $d(\Sigma_1,\Sigma_2) = \sqrt{\sum_{j=1}^{d}(\ln\lambda_j)^2}$, and a *geodesic* metric between Gaussian PDs $X_1 \sim\mathrm{N}(\mu_1,\Sigma_1)$ and $X_2 \sim\mathrm{N}(\mu_2,\Sigma_2)$:

$$d(X_1,X_2) = \left(\delta^T S^{-1}\delta\right)^{1/2} + \left(\sum_{j=1}^{d}(\ln\lambda_j)^2\right)^{1/2} \tag{7}$$

where $\delta = \mu_1 - \mu_2$ and $S = \frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2$. Equivalently,

$$d^2(\Sigma_1,\Sigma_2) = \mathrm{tr}\left[(\ln(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}))^2\right]. \tag{8}$$

Remark 1. *It may be proved that the set of symmetric positively definite matrices $M^+(d,\mathbf{R})$ is a Riemannian manifold, and (8) is a geodesic distance corresponding to the bilinear form $B(\mathbf{X},\mathbf{Y}) = 4\mathrm{tr}(\mathbf{XY})$ on the tangent space of symmetric matrices $M(d,\mathbf{R})$.*

3. Total Variation Distance between 1D Gaussian PDs

Let Φ and φ be the standard normal distribution and its density. Let $X_i \sim \mathrm{N}(\mu_i,\sigma_i^2)$, $i = 1,2$. Define $\tau = \tau(X_1,X_2) = \mathrm{TV}(\mathrm{N}(\mu_1,\sigma_1^2),\mathrm{N}(\mu_2,\sigma_2^2))$. Note that τ depends on the parameters $\Delta = |\delta|$, with $\delta = \mu_1 - \mu_2$, and σ_1^2,σ_2^2.

Proposition 1. *In the case $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the total variation distance is computed exactly: $\tau(X_1,X_2) = 2\Phi(\frac{|\mu_1-\mu_2|}{2\sigma}) - 1$.*

Proof. By using a shift, we can assume that $\mu_1 = 0$ and $\mu_2 = \Delta > 0$. Then, the set $A = \{x : p_1(x) > p_2(x)\}$ is specified as

$$A = \{e^{-\frac{x^2}{2\sigma^2}} > e^{-\frac{(x-\Delta)^2}{2\sigma^2}}\} = (-\infty,\Delta/2).$$

Hence,

$$\tau(X_1,X_2) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\Delta/2}\left(e^{-\frac{x^2}{2\sigma^2}} - e^{-\frac{(x-\Delta)^2}{2\sigma^2}}\right)\mathrm{d}x = \Phi(b) - \Phi(-b) \tag{9}$$

where $b = \frac{\Delta}{2\sigma}$. Using the property $\Phi(-b) = 1 - \Phi(b)$ leads to the answer. □

Theorem 1.

$$\frac{1}{200}\min\left[1,\max[\frac{|\sigma_1^2-\sigma_2^2|}{\min[\sigma_1^2,\sigma_2^2]},\frac{40\Delta}{\min[\sigma_1,\sigma_2]}]\right] \le \tau \le \frac{3|\sigma_1^2-\sigma_2^2|}{2\max[\sigma_1^2,\sigma_2^2]} + \frac{\Delta}{2\max[\sigma_1,\sigma_2]} \tag{10}$$

The proof is sketched in Section 6. The upper bound is based on the following.

Proposition 2 (Pinsker's inequality). *Let X_1, X_2 be random variables with the probability density functions p, q, and the Kullback–Leibler divergence* $\mathrm{KL}(\mathbf{P}_{X_1}||\mathbf{P}_{X_2})$. *Then, for* $\tau(X_1, X_2) = \mathrm{TV}(X_1, X_2)$,

$$\tau(X_1, X_2) \leq \min[1, \sqrt{\mathrm{KL}(\mathbf{P}_{X_1}||\mathbf{P}_{X_2})/2}]. \tag{11}$$

Proof of Pinsker's inequality. We need the following bound:

$$|x-1| \leq \sqrt{\left(\frac{4}{3} + \frac{2x}{3}\right)\phi(x)}, \phi(x) := x\ln x - x + 1 \tag{12}$$

If $\mathbf{P}$ and $\mathbf{Q}$ are singular, then $\mathrm{KL} = \infty$ and Pinsker's inequality holds true. Assume $\mathbf{P}$ and $\mathbf{Q}$ are absolutely continuous. In view of (7) and Cauchy–Schwarz inequality,

$$\begin{aligned} &\tau(X,Y) = \tfrac{1}{2}\int|p-q| = \tfrac{1}{2}\int q|\tfrac{p}{q} - 1|\mathbf{1}_{\{q>0\}} \\ &\leq \tfrac{1}{2}\Big(\int\Big(\tfrac{4q}{3} + \tfrac{2p}{3}\Big)\Big)^{1/2}\Big(\int q\phi(\tfrac{p}{q})\mathbf{1}_{\{q>0\}}\Big)^{1/2} \\ &= \Big(\tfrac{1}{2}\int p\ln(\tfrac{p}{q})\mathbf{1}_{\{q>0\}}\Big)^{1/2} = (\mathrm{KL}(\mathbf{P}||\mathbf{Q})/2)^{1/2} \end{aligned} \tag{13}$$

To check (12), define

$$g(x) = (x-1)^2 - \left(\frac{4}{3} + \frac{2x}{3}\right)\phi(x)$$

Then, $g(1) = g'(1) = 0, g''(x) = -\frac{4\phi(x)}{3x} < 0$. Hence,

$$g(x) = g(1) + g'(1)(x-1) + \frac{1}{2}g''(\xi)(x-1)^2 = -\frac{4\phi(\xi)}{6\xi}(x-1)^2 \leq 0.$$

□

[Mark S. Pinsker was invited to be the Shannon Lecturer at the 1979 IEEE International Symposium on Information Theory, but could not obtain permission at that time to travel to the symposium. However, he was officially recognized by the IEEE Information Theory Society as the 1979 Shannon Award recipient].

For one-dimensional Gaussian distributions,

$$\mathrm{KL}(\mathbf{P}_{X_1}||\mathbf{P}_{X_2}) = \frac{1}{2}\left(\frac{\sigma_2^2}{\sigma_1^2} - 1 + \frac{\Delta^2}{\sigma_1^2} - \ln\frac{\sigma_2^2}{\sigma_1^2}\right).$$

In the multi-dimensional Gaussian case,

$$\mathrm{KL}(\mathbf{P}_{X_1}||\mathbf{P}_{X_2}) = \frac{1}{2}\Big(\mathrm{tr}\Big(\Sigma_1^{-1}\Sigma_2 - \mathbf{I}\Big) + \delta^T\Sigma_1^{-1}\delta - \ln\det(\Sigma_2\Sigma_1^{-1})\Big) \tag{14}$$

Next, define the Hellinger distance

$$\eta(X,Y) = \frac{1}{\sqrt{2}}\left(\int(\sqrt{p_X(u)} - \sqrt{p_Y(u)})^2 \mathrm{d}u\right)^{1/2} \tag{15}$$

and note that, for one-dimensional Gaussian distributions,

$$\eta(X,Y)^2 = 1 - \frac{\sqrt{2\sigma_1\sigma_2}}{\sqrt{\sigma_1^2 + \sigma_2^2}}e^{-\frac{\Delta^2}{4(\sigma_1^2+\sigma_2^2)}} \tag{16}$$

For multi-dimensional Gaussian PDs with $\delta = \mu_1 - \mu_2$,

$$\eta(X,Y)^2 = 1 - \frac{2^{d/2}\det(\Sigma_1)^{1/4}\det(\Sigma_2)^{1/4}}{\det(\Sigma_1+\Sigma_2)^{1/2}} \exp\left(-\frac{1}{8}\delta^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}\delta\right). \tag{17}$$

In fact, the following inequalities hold:

$$\tau(X,Y) \le \sqrt{2}\eta(X,Y) \le \sqrt{2}\sqrt{\mathrm{KL}(\mathbf{P}_X||\mathbf{P}_Y)} \le \sqrt{2}\sqrt{\chi^2(X,Y)} \tag{18}$$

where $\chi^2(\mathbf{P},\mathbf{Q}) = \int \frac{(p(x)-q(x))^2}{p(x)}dx$. These inequalities are not sharp. For example, the Cauchy–Schwarz inequality immediately implies $\tau(X,Y) \le \frac{1}{2}\sqrt{\chi^2(X,Y)}$. There are also reverse inequalities in some cases.

Proposition 3 (Le Cam's inequalities). *The following inequality holds:*

$$\eta(X,Y)^2 \le \tau(X,Y) \le \eta(X,Y)\left(2-\eta(X,Y)^2\right)^{1/2} \tag{19}$$

Proof of Le Cam's inequalities. From $\tau(X,Y) = \frac{1}{2}\int|p-q| = 1 - \int\min[p,q]$ and $\min[p,q] \le \sqrt{pq}$, it follows that $\tau(X,Y) \ge 1 - \int\sqrt{pq} = \eta^2(X,Y)$. Next, $\int\min[p,q] + \int\max[p,q] = 2$. Therefore, by Cauchy–Schwarz:

$$\begin{aligned}&\left(\int\sqrt{pq}\right)^2 = \left(\int\sqrt{\min[p,q]\max[p,q]}\right)^2 \le \int\min[p,q]\int\max[p,q]\\ &= \int\min[p,q](2-\int\min[p,q])\end{aligned} \tag{20}$$

Hence,

$$\begin{aligned}&(1-\eta(X,Y)^2)^2 \le (1-\tau(X,Y))(1+\tau(X,Y))\\ &\Rightarrow \tau(X,Y) \le \eta(X,Y)(2-\eta(X,Y)^2)^{1/2}.\end{aligned} \tag{21}$$

□

Example 2. *Let* $\mathbf{X} \sim N(0,\Sigma_1)$, $\mathbf{Y} \sim N(0,\Sigma_2)$ *be d-dimensional Gaussian vectors. Suppose that* $\Sigma_2 = (1+\Delta)\Sigma_1$, *where* Δ *is small enough. Let* $r < d$ *and* A *be* $r \times d$ *semi-orthogonal matrix* $AA^T = \mathbf{I}_r$. *Define* $\tau := \tau(A\mathbf{X}, A\mathbf{Y})$. *Then,*

$$\frac{1}{16}\Delta^2 r \le \tau \le \frac{1}{2^{3/2}}\Delta\sqrt{r}.$$

Proof. In view of Le Cam's inequalities, it is enough to evaluate η^2. Note that all r eigenvalues of $\Sigma_1\Sigma_2^{-1}$ equal $(1+\Delta)^{-1}$. Thus,

$$\eta^2 = 1 - \frac{4^{r/4}(1+\Delta)^{r/4}}{(2+\Delta)^{r/2}} = \frac{1}{8}\Delta^2[-r(\frac{r}{4}-1)+\frac{r}{2}(\frac{r}{2}-1)]+o(\Delta^2) = \frac{1}{16}\Delta^2 r + o(\Delta^2). \tag{22}$$

□

[Ernst Hellinger was imprisoned in Dachau but released by the interference of influential friends and emigrated to the US].

4. Bounds on the Total Variation Distance

This section is devoted to the basic examples and partially based on [5]. However, it includes more proofs and additional details (Figure 1).

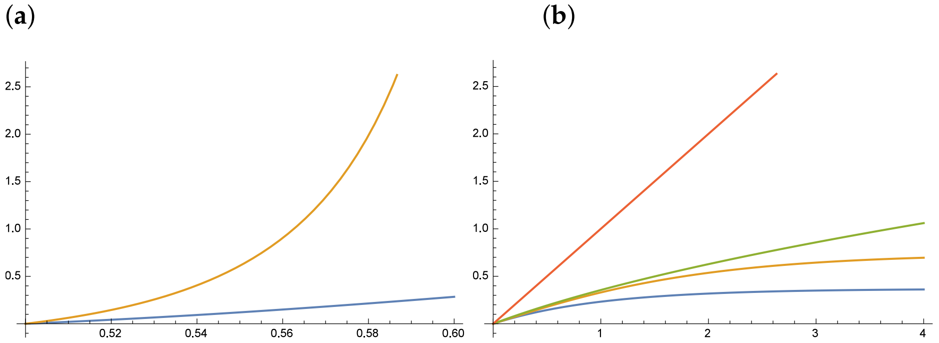

Figure 1. Exact TV distance and the upper bound for (**a**) TV(Bin$(20, \frac{1}{2})$, Bin$(20, \frac{1}{2} + a)$) and (**b**) TV(Pois(1), Pois$(1 + a)$). (**a**) Note that the upper bound becomes useless for $p_2 - p_1 \geq 0.07$; (**b**) blue and orange curves – exact TV distance: the blue curve works for $1 \leq \frac{\lambda_2}{\lambda_1} \leq 2$ and the orange curve for $2 \leq \frac{\lambda_2}{\lambda_1} \leq 4$. Note that the linear upper bound (red curve) is not relevant and the square root upper (green curve) bound becomes useless for $\frac{\lambda_2}{\lambda_1} \geq 4$.

Proposition 4 (Distances between exponential distributions). *(a) Let $X \sim Exp(\lambda)$, $Y \sim Exp(\mu)$, $0 < \lambda \leq \mu < \infty$. Then,*

$$\tau(X, Y) = \left(\frac{\lambda}{\mu}\right)^{\frac{\lambda}{\mu-\lambda}} - \left(\frac{\lambda}{\mu}\right)^{\frac{\mu}{\mu-\lambda}}. \tag{23}$$

(b) Let $\mathbf{X} = (X_1, \ldots, X_d)$, $\mathbf{Y} = (Y_1, \ldots, Y_d)$, each with d i.i.d. components $X_i \sim Exp(\lambda)$, $Y_i \sim Exp(\mu)$. Then,

$$\tau(\mathbf{X}, \mathbf{Y}) = \int_{z^*}^{\infty} (\lambda^d e^{-\lambda y} - \mu^d e^{-\mu y}) \frac{(\sqrt{2}y)^{d-1}}{(d-1)!} \mathrm{d}y \tag{24}$$

where $z^* = \frac{d}{\mu-\lambda} \ln \frac{\mu}{\lambda}$.

Proof. (a) Indeed, the set $A = \{x > 0 : \lambda e^{-\lambda x} > \mu e^{-\mu x}\}$ coincides with the half-axis (y^*, ∞) with $y^* = \frac{1}{\mu-\lambda} \ln \frac{\mu}{\lambda}$. Consequently, $\tau(\mathbf{X}, \mathbf{Y}) = e^{-\lambda y^*} - e^{-\mu y^*}$. (b) In this case, the set $A = \{\mathbf{X} : x_i > 0, \sum\limits_{j=1}^{d} x_j > z^*\}$ with $z^* = \frac{d}{\mu-\lambda} \ln \frac{\mu}{\lambda}$. Given $y > 0$, the area of an $(d-1)$-dimensional simplex $\{\mathbf{x} : x_i > 0, \sum\limits_{j=1}^{d} x_j = y\}$ equals $\frac{(\sqrt{2}y)^{d-1}}{(d-1)!}$. Then, $\tau(\mathbf{X}, \mathbf{Y}) = \int_{\mathbf{x} \in A} [\prod\limits_{j=1}^{d} \lambda e^{-\lambda x_i} - \prod\limits_{j=1}^{d} \mu e^{-\mu x_i}] \mathrm{d}\mathbf{x}$ coincides with (24). □

Proposition 5 (Distances between Poisson distributions). *Let $X_i \sim Po(\lambda_i)$, where $0 < \lambda_1 < \lambda_2$. Then,*

$$\tau(X_1, X_2) = \int_{\lambda_1}^{\lambda_2} \mathbf{P}(N(u) = l - 1) \mathrm{d}u \leq \min\left[\lambda_2 - \lambda_1, \sqrt{\frac{2}{e}} (\sqrt{\lambda_2} - \sqrt{\lambda_1})\right] \tag{25}$$

where $N(u) \sim Po(u)$ and

$$l = l(\lambda_1, \lambda_2) = \lceil (\lambda_2 - \lambda_1)(\ln(\lambda_2/\lambda_1))^{-1} \rceil \tag{26}$$

with $\lceil \lambda_1 \rceil \leq l \leq \lceil \lambda_2 \rceil$.

Proof. Let $N(t) \sim \mathrm{Po}(t)$; then, via iterated integration by part,

$$\mathbf{P}(N(t) \le n) = \sum_{k=0}^{n} e^{-t}\frac{t^k}{k!} = \int_t^\infty e^{-u}\frac{u^n}{n!}\mathrm{d}u = \int_t^\infty \mathbf{P}(N(u) = n)\mathrm{d}u. \tag{27}$$

Hence, $\mathrm{Kolm}(X_1, X_2) = \tau(X_1, X_2) = \mathbf{P}(X_2 \ge l) - \mathbf{P}(X_1 \ge l) =$

$$\mathbf{P}(X_1 \le l-1) - \mathbf{P}(X_2 \le l-1) = \int_{\lambda_1}^{\lambda_2} \mathbf{P}(N(u) = l-1)\mathrm{d}u$$

where

$$l = \min[k \in \mathbf{Z}_+ : f(k) \ge 1] = \lceil(\lambda_2 - \lambda_1)(\ln(\lambda_2/\lambda_1))^{-1}\rceil$$

and $f(k) = \frac{\mathbf{P}(N(\lambda_2)=k)}{\mathbf{P}(N(\lambda_1)=k)}$. □

Proposition 6 (Distances between binomial distributions). *X_i ∼$Bin(n, p_i)$, $0 < p_1 < p_2 < 1$.*

$$\tau(X_1, X_2) = n\int_{p_1}^{p_2} \mathbf{P}(S_{n-1}(u) = l-1)\mathrm{d}u \le \frac{\sqrt{e}}{2}\frac{\psi(p_2 - p_1)}{(1 - \psi(p_2 - p_1))^2} \tag{28}$$

where $S_{n-1}(u)$ ∼$Bin(n-1, u)$ and $\psi(x) = x\sqrt{\frac{n+2}{2p_1(1-p_1)}}$. Finally, define

$$l = \left\lceil \frac{-n\ln(1 - \frac{p_2-p_1}{1-p_1})}{\ln(1 + \frac{p_2-p_1}{p_1}) - \ln(1 - \frac{p_2-p_1}{1-p_1}))} \right\rceil \tag{29}$$

with $\lceil np_1\rceil \le l \le \lceil np_2\rceil$.

Proof. Let us prove the following inequality:

$$np \le \frac{-n\ln(1 - x/q)}{\ln(1 + x/p) - \ln(1 - x/q)} \le n(p + x), 0 < x < q \tag{30}$$

where $p = p_1$, $p + x = p_2$ and $q = 1 - p$. By concavity of the ln, given $p \in (0,1)$ and $q = 1 - p$,

$$f(x) = p\ln(1 + x/p) + q\ln(1 - x/q) \le \ln 1 = 0, 0 < x < q. \tag{31}$$

This gives the bound $\lceil np_1\rceil \le l$ as follows:

$$\begin{aligned}&p\ln(1 + x/p) + q\ln(1 - x/q) \le 0 \Rightarrow np\ln(1 + x/p) - np\ln(1 - x/q) \le -n\ln(1 - x/q)\\ &\Rightarrow np \le \tfrac{-n\ln(1-x/q)}{\ln(1+x/p)-\ln(1-x/q)}.\end{aligned} \tag{32}$$

On the other hand,

$$h(x) = (p + x)\ln(1 + x/p) + (q - x)\ln(1 - x/q) \ge 0, 0 \le x \le q \tag{33}$$

as $h(0) = 0$ and $h'(x) = \ln(1 + x/p - \ln(1 - x/q) \le 0$; this implies the bound $l \le \lceil np_2\rceil$. Indeed:

$$\begin{aligned}&(p + x)\ln(1 + x/p) + (q - x)\ln(1 - x/q) \ge 0 \Rightarrow\\ &n(p + x)\ln(1 + x/p) + n(p + x)\ln(1 - x/q) \ge -n\ln(1 - x/q)\\ &\Rightarrow n(p + x) \ge \tfrac{-n\ln(1-x/q)}{\ln(1+x/p)-\ln(1-x/q)}.\end{aligned} \tag{34}$$

The rest of the solution goes in parallel with that of Proposition 5. Equation (27) is replaced with the following relation: if $S_n(p) \sim \mathrm{Bin}(n, p)$; then,

$$\mathbf{P}(S_n(p) \geq k) = n \int_0^p \mathbf{P}(S_{n-1}(u) = k-1)\mathrm{d}u \tag{35}$$

In fact, iterated integration by parts yields the RHS of (35)

$$\begin{aligned} &= \tfrac{n(n-1)\ldots(n-k+1)}{1\ldots k} p^k (1-p)^{n-k} + \tfrac{n(n-1)\ldots(n-k)}{1\ldots(k+1)} p^{k+1} (1-p)^{n-k+1} \\ &+ \ldots + p^n = \end{aligned} \tag{36}$$

the LHS of (35). □

Proposition 7 (Distance between binomial and Poisson distributions). *$X \sim Bin(n, p)$ and $Y \sim Po(np)$, $0 < np < 2 - \sqrt{2}$*

$$\tau(X, Y) = np[(1-p)^{n-1} - e^{-np}] \tag{37}$$

Alternative bound

$$\mathrm{TV}(\mathrm{Bin}(n, \frac{\lambda}{n}), \mathrm{Pois}(\lambda)) \leq 1 - \left(1 - \frac{\lambda}{n}\right)^{1/2}. \tag{38}$$

For the sum of Bernoulli r.v.'s $S_n = \sum_{j=1}^{n} X_j$ with $\mathbf{P}(X_i = 1) = p_i$,

$$\tau(S_n, Y_n) = \frac{1}{2} \sum_{k=1}^{\infty} |\mathbf{P}(S_n = k) - \frac{\lambda_n^k}{k!} e^{-\lambda_n}| < \sum_{i=1}^{n} p_i^2 \tag{39}$$

where $Y_n \sim Po(\lambda_n)$, $\lambda_n = p_1 + p_2 + \ldots + p_n$ (Le Cam). A stronger result: for $X_i \sim$ Bernoulli (p_i) and $Y_i \sim Po(\lambda_i = p_i)$, there exists a coupling s.t.

$$\tau(X_i, Y_i) = \mathbf{P}(X_i \neq Y_i) = p_i(1 - e^{-p_i}).$$

The stronger form of (39):

$$\frac{1}{32}\left(1 \wedge \lambda_n^{-1}\right) \sum_{j=1}^{n} p_i^2 \leq \tau(X_n, Y_n) \leq \lambda_n^{-1}\left(1 - e^{-\lambda_n}\right) \sum_{j=1}^{n} p_i^2. \tag{40}$$

Proposition 8 (Distance between negative binomial distributions). *Let $X_i \sim NegBin$ (m, p_i), $0 < p_1 < p_2 < 1$*

$$\tau(X_1, X_2) = (m + l - 1) \int_{p_1}^{p_2} \mathbf{P}(S_{m+l-2}(u) = m-1)\mathrm{d}u \tag{41}$$

where $S_n(u) \sim Bin(n, u)$ and

$$l = \lceil -m \frac{\ln(1 + \frac{p_2 - p_1}{p_1})}{\ln(1 - \frac{p_2 - p_1}{1 - p_1})} \rceil \tag{42}$$

with $\lceil m \frac{1-p_2}{p_2} \rceil \leq l \leq \lceil m \frac{1-p_1}{p_1} \rceil$.

5. Total Variance Distance in the Multi-Dimensional Gaussian Case

Theorem 2. *Let $\tau = \mathrm{TV}(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2))$, and Σ_1, Σ_2 be positively definite. Let $\delta = \mu_1 - \mu_2$ and Π be a $d \times (d-1)$ matrix whose columns form a basis for the subspace orthogonal* *to δ. Let $\lambda_1, \ldots, \lambda_{d-1}$ denote the eigenvalues of the matrix $(\Pi^T \Sigma_1 \Pi)^{-1} \Pi^T \Sigma_2 \Pi - \mathbf{I}_{d-1}$ and $\lambda = \sqrt{\sum_{i=1}^{d-1} \lambda_i^2}$. In $\mu_1 \neq \mu_2$, then*

$$\frac{1}{200} \min[1, \varphi(\delta, \Sigma_1, \Sigma_2)] \leq \tau \leq \frac{9}{2} \min[1, \varphi(\delta, \Sigma_1, \Sigma_2)] \tag{43}$$

where

$$\varphi(\delta, \Sigma_1, \Sigma_2) = \max[\frac{\delta^T(\Sigma_1 - \Sigma_2)\delta}{\delta^T \Sigma_1 \delta}, \frac{\sqrt{\delta^T \delta}}{\sqrt{\delta^T \Sigma_1 \delta}}, \lambda] \tag{44}$$

In the case of equal means $\mu_1 = \mu_2$, the bound (43) is simplified:

$$\frac{1}{100} \min[1, \lambda] \leq \tau \leq \frac{3}{2} \min[1, \lambda]. \tag{45}$$

Here, $\lambda = \sqrt{\sum_{j=1}^{d} \lambda_j^2}$, $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\Sigma_1^{-1} \Sigma_2 - \mathbf{I}_d$ for positively definite Σ_1, Σ_2.

Proof is given in Section 6.

Suppose $r \ll d$, and we want to find a low-dimensional projection $A \in \mathbf{R}^{r \times d}, AA^T = \mathbf{I}_r$ of the multidimensional data $\mathbf{X} \sim \mathrm{N}(\mu_1, \Sigma_1)$ and $\mathbf{Y} \sim \mathrm{N}(\mu_2, \Sigma_2)$ such that $\mathrm{TV}(A\mathbf{X}, A\mathbf{Y}) \rightarrow \max$. The problem may be reduced to the case $\mu_1 = \mu_2 = 0$, $\Sigma_1 = \mathbf{I}_n$, $\Sigma_2 = \Sigma$, cf. [6]. In view of (44), it is natural to maximize

$$\min[1, \sum_{i=1}^{r} g(\gamma_i)] \tag{46}$$

where $g(x) = \left(\frac{1}{x} - 1\right)^2$ and γ_i are the eigenvalues of $A\Sigma A^T$. Consider all permutations π of these eigenvalues. Let

$$\pi^* = \mathrm{argmax}_\pi \sum_{i=1}^{r} g(\lambda_{\pi(i)}), \gamma_i = \lambda_{\pi^*(i)}, i = 1, \ldots, r. \tag{47}$$

Then, rows of matrix A should be selected as the normalized eigenvectors of Σ associated with the eigenvalues γ_i.

Remark 2. *For zero-mean Gaussian models, this procedure may be repeated mutatis mutandis for any of the so-called f-divergences $D_f(\mathbf{P}||\mathbf{Q}) := \mathbf{E}_{\mathbf{P}}\left[f\left(\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\mathbf{P}}\right)\right]$, where f is a convex function such that $f(1) = 0$, cf. [6]. The most interesting examples are:*

(1) *KL-divergence: $f(t) = t \log t$ and $g(x) = \frac{1}{2}(x - \log x - 1)$;*
(2) *Symmetric KL-divergence: $f(t) = (t-1)\log t$ and $g(x) = \frac{1}{2}(x + \frac{1}{x} - 2)$;*
(3) *The total variance distance: $f(t) = \frac{1}{2}|t-1|$ and $g(x) = \left(\frac{1}{x} - 1\right)^2$;*
(4) *The square of Hellinger distance: $f(t) = (\sqrt{t} - 1)^2$ and $g(x) = \left(\frac{x+1}{x}\right)^2$;*
(5) *χ^2-divergence: $f(t) = (t-1)^2$ and $g(x) = \frac{1}{\sqrt{x(2-x)}}$.*

For the optimization procedure in (47), the following result is very useful.

Theorem 3 (**Poincaré Separation Theorem**)**.** *Let Σ be a real symmetric $d \times d$ matrix and A be a semi-orthogonal $r \times d$ matrix. The eigenvalues of Σ (sorted in the descending order) and the eigenvalues of $A\Sigma A^T$ denoted by $\{\gamma_i, i = 1, \ldots, r\}$ (sorted in the descending order) satisfy*

$$\lambda_{d-(r-i)} \leq \gamma_i \leq \lambda_i, i = 1, \ldots, r.$$

Proposition 9. *Let $\mathbf{X}, \mathbf{Y}$ be two Gaussian PDs with the same covariance matrix: $\mathbf{X} \sim N(\mu_1, \Sigma)$, $\mathbf{Y} \sim N(\mu_2, \Sigma)$. Suppose that matrix Σ is non-singular. Then,*

$$\tau(\mathbf{X}, \mathbf{Y}) = 2\Phi(||\Sigma^{-1/2}(\mu_1 - \mu_2)||/2) - 1. \tag{48}$$

Proof. Here, the set $A := \{\mathbf{x} \in \mathbf{R}^d : p(\mathbf{x}|\mu_1, \Sigma) > p(\mathbf{x}|\mu_2, \Sigma)\}$ is a half-space. Indeed,

$$p(\mathbf{x}|\mu_1, \Sigma) > p(\mathbf{x}|\mu_2, \Sigma) \Leftrightarrow 2\mathbf{x}^T\Sigma^{-1}(\mu_2 - \mu_1) < \mu_2^T\Sigma^{-1}\mu_2 - \mu_1^T\Sigma^{-1}\mu_1. \tag{49}$$

After the change of variables $\mathbf{x} \to \mathbf{x} + \mu_1$, we need to evaluate the expression

$$\begin{aligned} I &:= \tfrac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \int_{\mathbf{R}^d} \mathbf{1}\Big(\mathbf{x}^T\Sigma^{-1}\delta < \tfrac{1}{2}||\Sigma^{-1/2}\delta||^2\Big) \\ &\times \Big(e^{-\mathbf{x}^T\Sigma^{-1}\mathbf{x}/2} - e^{-(\mathbf{x}-\delta)^T\Sigma^{-1}(\mathbf{x}-\delta)/2}\Big)\mathrm{d}\mathbf{x}. \end{aligned} \tag{50}$$

Take an orthogonal $d \times d$ matrix O such that $O\Sigma^{-1/2}\delta = ||\Sigma^{-1/2}\delta||e_1$ and change the variables $\mathbf{x} = \Sigma^{1/2}O^T\mathbf{u}$. Then,

$$\begin{aligned} &\mathbf{x}^T\Sigma^{-1}\delta = ||\Sigma^{-1/2}\delta||u_1, \mathbf{x}^T\Sigma^{-1}\mathbf{x} = \mathbf{u}^T\mathbf{u}, \\ &(\mathbf{x} - \delta)^T\Sigma^{-1}(\mathbf{x} - \delta) = \mathbf{u}^T\mathbf{u} + ||\Sigma^{-1/2}\delta||^2 - 2||\Sigma^{-1/2}\delta||u_1. \end{aligned} \tag{51}$$

Thus,

$$\begin{aligned} I &= \tfrac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^{d-1}} e^{-\mathbf{v}^T\mathbf{v}/2}\mathrm{d}\mathbf{v} \int_{-\infty}^{||\Sigma^{-1/2}\delta||/2} \Big(e^{-u_1^2/2} - e^{-(u_1-||\Sigma^{-1/2}\delta||)^2/2}\Big)\mathrm{d}u_1 \\ &= \Phi(b) - \Phi(-b) \end{aligned} \tag{52}$$

where $b = ||\Sigma^{-1/2}\delta||/2$. □

6. Proofs for the Multi-Dimensional Gaussian Case

Let $X_i \sim \mathrm{N}(\mu_i, \Sigma_i), i = 1, 2$. W.l.o.g., assume that Σ_1, Σ_2 are positively definite, and the general case may be followed from the identity

$$\mathrm{TV}(\mathrm{N}(0, \Sigma_1), \mathrm{N}(0, \Sigma_2)) = \mathrm{TV}(\mathrm{N}(0, \Pi^T\Sigma_1\Pi), \mathrm{N}(0, \Pi^T\Sigma_2\Pi)) \tag{53}$$

where Π is $d \times r$ matrix whose columns form an orthogonal basis for range $(\Sigma_{1,2})$. Denote $u = (\mu_1 + \mu_2)/2, \delta = \mu_1 - \mu_2$ and decompose $\forall w \in \mathbf{R}^d$ as

$$w = u + f_1(w)\delta + f_2(w), f_2(w)^T\delta = 0.$$

Then,

$$\begin{aligned} &\max[\mathrm{TV}(f_1(X_1), f_1(X_2)), \mathrm{TV}(f_2(X_1), f_2(X_2))] \leq \mathrm{TV}(X_1, X_2) \\ &\leq \mathrm{TV}(f_1(X_1), f_1(X_2)) + \mathrm{TV}(f_2(X_1), f_2(X_2)) \end{aligned} \tag{54}$$

All the components are Gaussian and $f_1(X_1) \sim \mathrm{N}\Big(\frac{1}{2}, \frac{\delta^T\Sigma_1\delta}{\delta^T\delta}\Big)$, $f_1(X_2) \sim \mathrm{N}\Big(-\frac{1}{2}, \frac{\delta^T\Sigma_2\delta}{\delta^T\delta}\Big)$, $f_2(X_1) \sim \mathrm{N}(0, \mathbf{P}\Sigma_1\mathbf{P})$, $f_2(X_2) \sim \mathrm{N}(0, \mathbf{P}\Sigma_2\mathbf{P})$, $\mathbf{P} = \mathbf{I}_d - \frac{\delta\delta^T}{\delta^T\delta}$. We claim that

$$\begin{aligned} &\tfrac{1}{200}\min[1, \max[\tfrac{\delta^T(\Sigma_1-\Sigma_2)\delta}{2\delta^T\Sigma_1\delta}, \tfrac{40\sqrt{\delta^T\delta}}{\sqrt{\delta^T\Sigma_1\delta}}]] \\ &\leq \mathrm{TV}(f_1(X_1), f_1(X_2)) \leq \tfrac{3\delta^T(\Sigma_1-\Sigma_2)\delta}{2\delta^T\Sigma_1\delta} + \tfrac{\sqrt{\delta^T\delta}}{2\sqrt{\delta^T\Sigma_1\delta}}, \end{aligned} \tag{55}$$

$$\frac{1}{100}\min[1,\lambda] \le \mathrm{TV}(f_2(X_1), f_2(X_2)) \le \frac{3}{2}\lambda \tag{56}$$

where $\lambda = \left(\sum\limits_{j=1}^{d} \lambda_j\right)^{1/2}$ and λ_i are the eigenvalues of $\Sigma_1^{-1}\Sigma_2 - \mathbf{I}_d$.

Proof of upper bound. It follows from Pinsker's inequality. Let $d = 1$ and $\sigma_2 \ge \sigma_1$. Then, for $x = \frac{\sigma_2^2}{\sigma_1^2}$, we have $x - 1 - \ln x \le (x-1)^2$ and, by Pinsker's inequality,

$$\begin{aligned}&\mathrm{TV}\big(\mathrm{N}(\mu_1,\sigma_1^2),\mathrm{N}(\mu_2,\sigma_2^2)\big) \le \tfrac{1}{2}\sqrt{\tfrac{\sigma_2^2}{\sigma_1^2} - 1 - \ln\tfrac{\sigma_2^2}{\sigma_1^2} + \tfrac{\Delta^2}{\sigma_1^2}}\\ &\le \tfrac{1}{2}\sqrt{\tfrac{\sigma_2^2}{\sigma_1^2} - 1 - \ln\tfrac{\sigma_2^2}{\sigma_1^2}} + \tfrac{1}{2}\sqrt{\tfrac{\Delta^2}{\sigma_1^2}} \le \tfrac{1}{2}\tfrac{|\sigma_2^2-\sigma_1^2|}{\sigma_1^2} + \tfrac{1}{2}\tfrac{\Delta}{\sigma_1}.\end{aligned} \tag{57}$$

For $d > 1$, it is enough to obtain the upper bound in the case $\mu_1 = \mu_2 = 0$. Again, Pinsker's inequality implies: if $\lambda_i > -\frac{2}{3}\ \forall i$,

$$4\mathrm{TV}(\mathrm{N}(0,\Sigma_1),\mathrm{N}(0,\Sigma_2))^2 \le \sum_{i=1}^{d} \lambda_i - \ln(1+\lambda_i) \le \sum_{i=1}^{d} \lambda_i^2 = \lambda^2 \tag{58}$$

□

Sketch of proof for lower bound, cf. [7]. In a 1D case with $X_i \sim \mathrm{N}(\mu_i,\sigma_i^2)$ $(\mu_1 \le \mu_2)$,

$$\begin{aligned}&\mathrm{TV}(\mathrm{N}(\mu_1,\sigma_1^2),\mathrm{N}(\mu_2,\sigma_2^2)) \ge \mathbf{P}(X_2 \ge \mu_2) - \mathbf{P}(X_1 \ge \mu_2) =\\ &\tfrac{1}{2} - \left(\tfrac{1}{2} - \mathbf{P}(X_1 \in (\mu_1,\mu_2))\right) = \mathbf{P}(X_1 \in (\mu_1,\mu_2))\\ &\ge \tfrac{1}{5}\min[1,\tfrac{\Delta}{\sigma_1}]\end{aligned} \tag{59}$$

Next,

$$\mathrm{TV}(\mathrm{N}(\mu_1,\sigma_1^2),\mathrm{N}(\mu_2,\sigma_2^2)) \ge \frac{1}{2}\mathrm{TV}(\mathrm{N}(0,\sigma_1^2),\mathrm{N}(0,\sigma_2^2)) \tag{60}$$

Indeed, assume w.l.o.g. $\mu_1 \le \mu_2, \sigma_1 \le \sigma_2$. Then, $\exists c = c(\sigma_1,\sigma_2)$:

$$\mathrm{TV}(\mathrm{N}(0,\sigma_1^2),\mathrm{N}(0,\sigma_2^2)) = \mathbf{P}(N(0,\sigma_2^2) \notin [-c,c]) - \mathbf{P}(N(0,\sigma_1^2) \notin [-c,c])$$

Hence,

$$\begin{aligned}&\mathrm{TV}(\mathrm{N}(\mu_1,\sigma_1^2),\mathrm{N}(\mu_2,\sigma_2^2)) \ge \mathbf{P}(\mathrm{N}(\mu_2,\sigma_2^2) > c+\mu_1) - \mathbf{P}(\mathrm{N}(\mu_1,\sigma_1^2) > c+\mu_1)\\ &\ge \tfrac{1}{2}\mathrm{TV}(\mathrm{N}(0,\sigma_1^2),\mathrm{N}(0,\sigma_2^2))\end{aligned} \tag{61}$$

Thus, it is enough to study the case $\mu_1 = \mu_2 = 0$. Let $C = \mathrm{diag}(1+\lambda_i)$. Then,

$$\mathrm{TV}(\mathrm{N}(0,\Sigma_1),\mathrm{N}(0,\Sigma_2)) = \mathrm{TV}(\mathrm{N}(0,C^{-1}),\mathrm{N}(0,\mathbf{I}_d))$$

In the case when there exists i: $|\lambda_i| > 0.1$,

$$\begin{aligned}&\mathrm{TV}(\mathrm{N}(0,C^{-1}),\mathrm{N}(0,\mathbf{I}_d)) \ge \mathrm{TV}(\mathrm{N}(0,(1+\lambda_i)^{-1}),\mathrm{N}(0,1)) =\\ &\mathrm{TV}(\mathrm{N}(0,1),\mathrm{N}(0,1+\lambda_i)) \ge \mathbf{P}(\mathrm{N}(0,1) \in [-1,1])\\ &-\mathbf{P}(\mathrm{N}(0,1.1) \in [-1,1]) > 0.68 - 0.66 > 0.01\end{aligned} \tag{62}$$

Finally, in the case when $|\lambda_i| \le 0.1\ \forall i$, the result follows from the lower bound

$$\mathrm{TV}(\mathrm{N}(0,C^{-1}),\mathrm{N}(0,\mathbf{I}_d)) \ge \frac{\lambda}{6} - \frac{\lambda^2}{8} - \frac{1}{2}\left(e^{\lambda^2} - 1\right) \tag{63}$$

The bound (63) $> \frac{\lambda}{100}$ if $\lambda < 0.17$ and > 0.01 if $\lambda \geq 0.17$ and $|\lambda_i| < 0.1 \forall i$. We refer to [7] for the proofs of these facts. □

7. Estimation of Lévy–Prokhorov Distance

Let $\mathbf{P}_i, i = 1, 2$, be probability distributions on a metric space W with metric r. Define the Lévy–Prokhorov distance $\rho^{L-P}(\mathbf{P}_1, \mathbf{P}_2)$ between $\mathbf{P}_1, \mathbf{P}_2$ as the infimum of numbers $\epsilon > 0$ such that, for any closed set $C \subset W$,

$$\mathbf{P}_1(C) - \mathbf{P}_2(C_\epsilon) < \epsilon, \mathbf{P}_2(C) - \mathbf{P}_1(C_\epsilon) < \epsilon \tag{64}$$

where C_ϵ stands for the ϵ-neighborhood of C in metric r. It could be checked that $\rho^{L-P}(\mathbf{P}_1, \mathbf{P}_2) \leq \tau(\mathbf{P}_1, \mathbf{P}_2)$, i.e., the total variance distance. Equivalently,

$$\rho^{L-P}(\mathbf{P}_1, \mathbf{P}_2) = \inf_{\bar{\mathbf{P}} \in \mathcal{P}(\mathbf{P}_1, \mathbf{P}_2)} \inf[\epsilon > 0 : \mathbf{P}(r(X_1, X_2) > \epsilon) < \epsilon] \tag{65}$$

where $\mathcal{P}(\mathbf{P}_1, \mathbf{P}_2)$ is the set of all joint $\bar{\mathbf{P}}$ on $W \times W$ with marginals $\mathbf{P}_i$.

Next, define the Wasserstein distance $W_p^r(\mathbf{P}_1, \mathbf{P}_2)$ between $\mathbf{P}_1, \mathbf{P}_2$ by

$$W_p^r(\mathbf{P}_1, \mathbf{P}_2) = \inf_{\bar{\mathbf{P}} \in \mathcal{P}(\mathbf{P}_1, \mathbf{P}_2)} (\mathbf{E}_{\bar{\mathbf{P}}}[r(X_1, X_2)^p])^{1/p}. \tag{66}$$

In the case of Euclidean space with $r(x_1, x_2) = ||x_1 - x_2||$, the index r is omitted.

Total Variation, Wasserstein and Kolmogorov–Smirnov distances defined above are stronger than weak convergence (i.e., convergence in distribution, which is weak* convergence on the space of probability measures, seen as a dual space). That is, if any of these metrics go to zero as $n \to \infty$, then we have weak convergence. However, the converse is not true. However, weak convergence is metrizable (e.g., by the Lévy–Prokhorov metric).

Theorem 4 (**Dobrushin's bound**).

$$\rho^{L-P}(\mathbf{P}_1, \mathbf{P}_2) \leq [W_1^r(\mathbf{P}_1, \mathbf{P}_2)]^{1/2}. \tag{67}$$

Proof. Suppose that there exists a closed set C for which at least one of the inequalities (64) fails, say $\mathbf{P}_1(C) \geq \epsilon + \mathbf{P}_2(C_\epsilon)$. Then, for any joint $\bar{\mathbf{P}}$ with marginals $\mathbf{P}_1$ and $\mathbf{P}_2$,

$$\begin{aligned}
&\mathbf{E}_{\bar{\mathbf{P}}}[r(X_1, X_2)] \geq \mathbf{E}_{\bar{\mathbf{P}}}[\mathbf{1}(r(X_1, X_2) \geq \epsilon) r(X_1, X_2)] \\
&\geq \epsilon \bar{\mathbf{P}}(r(X_1, X_2) \geq \epsilon) \geq \epsilon \bar{\mathbf{P}}(X_1 \in C, X_2 \in W \setminus C_\epsilon) \\
&\geq \epsilon[\bar{\mathbf{P}}(X_1 \in C) - \bar{\mathbf{P}}(X_1 \in C, X_2 \in C_\epsilon)] \\
&\geq \epsilon[\bar{\mathbf{P}}(X_1 \in C) - \bar{\mathbf{P}}(X_2 \in C_\epsilon)] = \epsilon[\mathbf{P}_1(X_1 \in C) - \mathbf{P}_2(X_2 \in C_\epsilon)] \geq \epsilon^2.
\end{aligned} \tag{68}$$

This leads to (67), as claimed. □

The Lévy–Prokhorov distance is quite tricky to compute, whereas the Wasserstein distance can be found explicitly in a number of cases. Say, in a 1D case $W = \mathbf{R}^1$, we have

Theorem 5. *For $d = 1$,*

$$W_1(\mathbf{P}_1, \mathbf{P}_2) = \int_{\mathbf{R}} |F_1(x) - F_2(x)| \mathrm{d}x. \tag{69}$$

Proof. First, check the upper bound $W_1(\mathbf{P}_1, \mathbf{P}_2) \leq \int_{\mathbf{R}} |F_1(x) - F_2(x)| \mathrm{d}x$. Consider $\xi \sim \mathrm{U}[0, 1]$, $X_i = F_i^{-1}(\xi), i = 1, 2$. Then, in view of the Fubini theorem,

$$\mathbf{E}[|X_1 - X_2|] = \int_0^1 |F_1^{-1}(y) - F_2^{-1}(y)| \mathrm{d}y = \int_{\mathbf{R}} |F_1(x) - F_2(x)| \mathrm{d}x. \tag{70}$$

For the proof of the inverse inequality, see [8]. □

Proposition 10. *For $d = 1$ and $p > 1$,*

$$\begin{aligned} W_p(\mathbf{P}_1, \mathbf{P}_2)^p = p(p-1) \int_{-\infty}^{\infty} \mathrm{d}y \int_y^{\infty} \max[F_2(y) - F_1(x), 0](x-y)^{p-2}\mathrm{d}x \\ +p(p-1) \int_{-\infty}^{\infty} \mathrm{d}x \int_x^{\infty} \max[F_1(x) - F_2(y), 0](y-x)^{p-2}\mathrm{d}y. \end{aligned} \tag{71}$$

Proof. It follows from the identity

$$\begin{aligned} \mathbf{E}[|X-Y|^p] = p(p-1) \int_{-\infty}^{\infty} \mathrm{d}y \int_y^{\infty} [F_2(y) - F(x,y)](x-y)^{p-2}\mathrm{d}x \\ +p(p-1) \int_{-\infty}^{\infty} \mathrm{d}x \int_x^{\infty} [F_1(x) - F(x,y)](y-x)^{p-2}\mathrm{d}y \end{aligned} \tag{72}$$

The minimum is achieved for $\bar{F}(x,y) = \min[F_1(x), F_2(y)]$. For an alternative expression (see [9]):

$$W_p(\mathbf{P}_1, \mathbf{P}_2)^p = \int_0^1 |F_1^{-1}(t) - F_2^{-1}(t)|^p \mathrm{d}t. \tag{73}$$

□

Proposition 11. *Let $(\mathbf{X}, \mathbf{Y}) \in \mathbf{R}^{2d}$ be jointly Gaussian random variables (RVs) with $\mathbf{E}[\mathbf{X}] = \mu^X, \mathbf{E}[\mathbf{Y}] = \mu^Y$. Then, the Frechet-1 distance*

$$\begin{aligned} \rho^{F_1}(\mathbf{X}, \mathbf{Y}) &:= \mathbf{E}\left[\sum_{j=1}^{d} |X_j - Y_j|\right] \\ &= \sum_{j=1}^{d} \left[(\mu_j^X - \mu_j^Y)\left(1 - 2\Phi(-\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j})\right) + 2\hat{\sigma}_j \varphi(-\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j})\right]. \end{aligned} \tag{74}$$

where $\hat{\sigma}_j = \left((\sigma_j^X)^2 + (\sigma_j^Y)^2 - 2\mathrm{Cov}(X_j, Y_j)\right)^{1/2}$, φ and Φ are PDF and CDF of the standard Gaussian RV. Note that, in the case $\mu^X = \mu^Y$, the first term in (74) vanishes, and the second term gives

$$\rho^{F_1}(\mathbf{X}, \mathbf{Y}) = \sqrt{\frac{2}{\pi}} \sum_{j=1}^{d} \hat{\sigma}_j. \tag{75}$$

We also present expressions for the Frechet-3 and Frechet-4 distances

$$\begin{aligned} &\rho^{F_3}(\mathbf{X}, \mathbf{Y}) = \left(\sum_{j=1}^{d} |X_j - Y_j|^3\right)^{1/3} = \Bigg(\sum_{j=1}^{d} (\mu_j^X - \mu_j^Y)^3 \left(1 - 2\Phi(-\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j})\right) \\ &+6(\mu_j^X - \mu_j^Y)^2 \hat{\sigma}_j \varphi(-\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j}) + 3(\hat{\sigma}_j)^2(\mu_j^X - \mu_j^Y)\Big[1 - 2\Phi(-\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j}) - \\ &2\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j}\varphi(-\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j})\Big] + 2(\hat{\sigma}_j)^3 \varphi(-\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j})\left[\left(\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j}\right)^2 + 2\right]\Bigg)^{1/3} \\ &\rho^{F_4}(\mathbf{X}, \mathbf{Y}) = \left(\sum_{j=1}^{d} |X_j - Y_j|^4\right)^{1/4} = \left(\sum_{j=1}^{d} (\mu_j^X - \mu_j^Y)^4 + 6(\mu_j^X - \mu_j^Y)^2(\hat{\sigma}_j)^2 + 3(\hat{\sigma}_j)^4\right)^{1/4}. \end{aligned} \tag{76}$$

All of these expressions are minimized when $\mathrm{Cov}(X_j, Y_j), j = 1, \ldots, d$ are maximal. However, this fact does not lead immediately to the explicit expressions for Wasserstein's metrics. The problem here is that the joint covariance matrix $\Sigma_{\mathbf{X},\mathbf{Y}}$ should be positively definite. Thus, the straightforward choice $\mathrm{Corr}(X_j, Y_j) = 1$ is not always possible; see Theorem 6 below and [10].

[Maurice René Fréchet (1878–1973), a French mathematician, worked in topology, functional analysis, probability theory and statistics. He was the first to introduce the concept of a metric space (1906) and prove the representation theorem in L_2 (1907). However, in both cases, the credit was given to other people: Hausdorff and Riesz. Some sources claim that he discovered the Cramér–Rao inequality before anybody else, but such a claim was impossible to verify since lecture notes of his class appeared to be lost. Fréchet worked in several places in France before moving to Paris in 1928. In 1941, he succeeded Borel at the

Chair of Calculus of Probabilities and Mathematical Physics in Sorbonne. In 1956, he was elected to the French Academy of Sciences, at the age of 78, which was rather unusual. He influenced and mentored a number of young mathematicians, notably Fortet and Loève. He was an enthusiast of Esperanto; some of his papers were published in this language].

8. Wasserstein Distance in the Gaussian Case

In the Gaussian case, it is convenient to use the following extension of Dobrushin's bound for $p = 2$:

$$\rho^{L-P}(\mathbf{P}_1, \mathbf{P}_2) \le [W_p(\mathbf{P}_1, \mathbf{P}_2)]^{p/2}, p \ge 1. \tag{77}$$

Theorem 6. *Let $\mathbf{X}_i \sim N(\mu_i, \Sigma_i^2), i = 1, 2$, be d-dimensional Gaussian RVs. For simplicity, assume that both matrices Σ_1^2 and Σ_2^2 are non-singular (In the general case, the statement holds with Σ_1^{-1} understood as Moore–Penrose inversion). Then, the L_2-Wasserstein distance $W_2(\mathbf{X}_1, \mathbf{X}_2) = W_2(\mathrm{N}(\mu_1, \Sigma_1^2), \mathrm{N}(\mu_2, \Sigma_2^2))$ equals*

$$W_2(\mathbf{X}_1, \mathbf{X}_2) = \left[||\mu_1 - \mu_2||^2 + \mathrm{tr}(\Sigma_1^2) + \mathrm{tr}(\Sigma_2^2) - 2\mathrm{tr}[(\Sigma_1 \Sigma_2^2 \Sigma_1)^{1/2}] \right]^{1/2} \tag{78}$$

where $(\Sigma_1 \Sigma_2^2 \Sigma_1)^{1/2}$ stands for the positively definite matrix square-root. The value (78) is achieved when $\mathbf{X}_2 = \mu_2 + A(\mathbf{X}_1 - \mu_1)$ where $A = \Sigma_1^{-1}(\Sigma_1 \Sigma_2^2 \Sigma_1)^{1/2} \Sigma_1^{-1}$.

Corollary 1. *Let $\mu_1 = \mu_2 = 0$. Then, for $d = 1$, $W_2(X_1, X_2) = |\sigma_1 - \sigma_2|$. For $d = 2$,*

$$W_2(\mathbf{X}_1, \mathbf{X}_2) = \left[\mathrm{tr}(\Sigma_1^2) + \mathrm{tr}(\Sigma_2^2) - 2[\mathrm{tr}(\Sigma_1^2 \Sigma_2^2) + 2\sqrt{\det(\Sigma_1 \Sigma_2)}]^{1/2} \right]^{1/2}. \tag{79}$$

Note that the expression in (79) vanishes when $\Sigma_1^2 = \Sigma_2^2$.

Example 3. *(a) Let $\mathbf{X} \sim N(0, \Sigma_X^2)$, $\mathbf{Y} \sim N(0, \Sigma_Y^2)$ where $\Sigma_X^2 = \sigma_X^2 \mathbf{I}_d$ and $\Sigma_Y^2 = \sigma_Y^2 \mathbf{I}_d$. Then, $W_2(\mathbf{X}, \mathbf{Y}) = \sqrt{d}|\sigma_X - \sigma_Y|$.*
(b) Let $d = 2$, $\mathbf{X} \sim N(0, \Sigma_X^2)$, $\mathbf{Y} \sim N(0, \Sigma_Y^2)$, where $\Sigma_X^2 = \sigma_X^2 \mathbf{I}_2$, $\Sigma_Y^2 = \sigma_Y^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and $\rho \in (-1, 1)$. Then,

$$W_2(\mathbf{X}, \mathbf{Y}) = 2^{1/2} \left(\sigma_X^2 + \sigma_Y^2 - \sigma_X \sigma_Y \left[2 + 2(1 - \rho^2)^{1/2} \right]^{1/2} \right)^{1/2}.$$

(c) Let $d = 2$, $\mathbf{X} \sim N(0, \Sigma_X^2)$, $\mathbf{Y} \sim N(0, \Sigma_Y^2)$, where $\Sigma_X^2 = \sigma_X^2 \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}$, $\Sigma_Y^2 = \sigma_Y^2 \begin{pmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{pmatrix}$ and $\rho_1, \rho_2 \in (-1, 1)$. Then,

$$W_2(\mathbf{X}, \mathbf{Y}) = 2^{1/2} \left(\sigma_X^2 + \sigma_Y^2 - \sigma_X \sigma_Y \left[2 + 2\rho_1 \rho_2 + 2(1 - \rho_1^2)^{1/2} (1 - \rho_2^2)^{1/2} \right]^{1/2} \right)^{1/2}.$$

Note that, in the case $\rho_1 = \rho_2$, $W_2(\mathbf{X}, \mathbf{Y}) = \sqrt{2}|\sigma_X - \sigma_Y|$ as in (a).

Proof. First, reduce to the case $\mu_1 = \mu_2 = 0$ by using the identity $W_2^2(\mathbf{X}_1, \mathbf{X}_2) = ||\mu_1 - \mu_2||^2 + W_2^2(\xi_1, \xi_2)$ with $\xi_i = X_i - \mu_i$. Note that the infimum in (19) is always attained on Gaussian measures as $W_2(\mathbf{X}_1, \mathbf{X}_2)$ is expressed in terms of the covariance matrix $\Sigma^2 = \Sigma_{X,Y}^2$ only (cf. (81) below). Let us write the covariance matrix in the block form

$$\Sigma^2 = \begin{pmatrix} \Sigma_1^2 & K \\ K^T & \Sigma_2^2 \end{pmatrix} = \begin{pmatrix} \Sigma_1 & 0 \\ K^T \Sigma_1^{-1} & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} \Sigma_1 & \Sigma_1^{-1} K \\ 0 & I \end{pmatrix} \tag{80}$$

where the so-called Shur's complement $S = \Sigma_2^2 - K^T \Sigma_1^{-2} K$. The problem is reduced to finding the matrix K in (80) that minimizes the expression

$$\int_{\mathbf{R}^d \times \mathbf{R}^d} ||\mathbf{x} - \mathbf{y}||^2 \mathrm{d}\mathbf{P}_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = \mathrm{tr}(\Sigma_1^2) + \mathrm{tr}(\Sigma_2^2) - 2\mathrm{tr}(K) \tag{81}$$

subject to a constraint that the matrix Σ^2 in (80) is positively definite. The goal is to check that the minimum (81) is achieved when the Shur's complement S in (80) equals 0. Consider the fiber $\sigma^{-1}(S)$, i.e., the set of all matrices K such that $\sigma(K) := \Sigma_Y^2 - K^T(\Sigma_X^2)^{-1}K = S$. It is enough to check that the maximum value of $\mathrm{tr}(K)$ on this fiber equals

$$\max_{F \in \sigma^{-1}(S)} \mathrm{tr}(K) = \mathrm{tr}\Big[(\Sigma_Y(\Sigma_X^2 - S)\Sigma_Y)^{1/2}\Big]. \tag{82}$$

Since the matrix S is positively defined, it is easy to check that the fiber $S = 0$ should be selected. In order to establish (82), represent the positively definite matrix $\Sigma_Y^2 - S$ in the form $\Sigma_Y^2 - S = UD_r^2 U^T$, where the diagonal matrix $D_r^2 = \mathrm{diag}(\lambda_1^2, \ldots, \lambda_r^2, 0, \ldots, 0)$ and $\lambda_i > 0$. Next, $U = (U_r | U_{d-r})$ is the orthogonal matrix of the corresponding eigenvectors. We obtain the following $r \times r$ identity:

$$(\Sigma_X^{-1} K U_r D_r^{-1})^T (\Sigma_X^{-1} K U_r D_r^{-1}) = \mathbf{I}_r. \tag{83}$$

It means that $\Sigma_X^{-1} K U_r D_r^{-1} = O_r$, an 'orthogonal' $d \times r$ matrix, with $O_r^T O_r = \mathbf{I}_r$, and $K = \Sigma_X O_r D_r U_r^T$. The matrix O_r parametrises the fiber $\sigma^{-1}(S)$. As a result, we have an optimization problem

$$\mathrm{tr}(O^T M) \to \max, M = \Sigma_X U_r D_r \tag{84}$$

in a matrix-valued argument O_r, subject to the constraint $O_r^T O_r = \mathbf{I}_r$. A straightforward computation gives the answer $\mathrm{tr}[(M^T M)^{1/2}]$, which is equivalent to (82). Technical details can be found in [11,12]. □

Remark 3. *For general zero means RVs* $\mathbf{X}, \mathbf{Y} \in \mathbf{R}^d$ *with the covariance matrices* $\Sigma_i^2, i = 1, 2$, *the following inequality holds* [13]*:*

$$\mathrm{tr}(\Sigma_1^2) + \mathrm{tr}(\Sigma_2^2) - 2\mathrm{tr}[(\Sigma_1 \Sigma_2^2 \Sigma_1)^{1/2}] \le \mathbf{E}[||\mathbf{X} - \mathbf{Y}||^2] \le \mathrm{tr}(\Sigma_1^2) + \mathrm{tr}(\Sigma_2^2) + 2\mathrm{tr}[(\Sigma_1 \Sigma_2^2 \Sigma_1)^{1/2}]. \tag{85}$$

9. Distance between Distributions of Different Dimensions

For $m \le d$, define a set of matrices with orthonormal rows:

$$O(m, d) = \{V \in \mathbf{R}^{m \times d} : VV^T = \mathbf{I}_m\} \tag{86}$$

and a set of affine maps $\varphi : \mathbf{R}^d \to \mathbf{R}^m$ such that $\varphi_{V,b}(x) = Vx + b$.

Definition 1. *For any measures* $\mu \in M(\mathbf{R}^m)$ *and* $\nu \in M(\mathbf{R}^d)$, *the embeddings of* μ *into* $\mathbf{R}^d$ *are the set of d-dimensional measures* $\Phi^+(\mu, d) := \{\alpha \in \mathrm{M}(\mathbf{R}^n) : \varphi_{V,\beta}(\alpha) = \mu\}$ *for some* $V \in O(m, d), b \in \mathbf{R}^m$, *and the projections of* ν *onto* $\mathbf{R}^m$ *are the set of m-dimensional measures* $\Phi^-(\nu, m) := \{\beta \in \mathrm{M}(\mathbf{R}^m) : \varphi_{V,\beta}(\nu) = \beta\}$ *for some* $V \in O(m, d), b \in \mathbf{R}^m$.

Given a metric κ between measures of the same dimension, define the projection distance $d^-(\mu, \nu) := \inf_{\beta \in \Phi^-(\nu,m)} \kappa(\mu, \beta)$ and the embedding distance $d^+(\mu, \nu) := \inf_{\alpha \in \Phi^+(\mu,d)} \kappa(\alpha, \nu)$. It may be proved [14] that $d^+(\mu, \nu) = d^-(\mu, \nu)$; denote the common value by $\hat{d}(\mu, \nu)$.

Example 4. *Let us compute Wasserstein distance between one-dimensional $X \sim N(\mu_1, \sigma^2)$ and d-dimensional $Y \sim N(\mu_2, \Sigma)$. Denote by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ the eigenvalues of Σ. Then,*

$$\hat{W}_2(X,Y) = \begin{cases} \sigma - \sqrt{\lambda_1} \; if \; \sigma > \sqrt{\lambda_1} \\ 0 \; if \; \sqrt{\lambda_d} \leq \sigma \leq \sqrt{\lambda_1} \\ \sqrt{\lambda_d} - \sigma \; if \; \sigma < \sqrt{\lambda_d}. \end{cases} \tag{87}$$

Indeed, in view of Theorem 6, write

$$(W_2^-(X,Y))^2 = \min_{||\mathbf{x}||_2=1, b\in\mathbf{R}} \Big[||\mu_1 - \mathbf{x}^T\mu_2 - b||_2^2 \\ + \mathrm{tr}(\sigma^2 + \mathbf{x}^T\Sigma\mathbf{x} - 2\sigma\sqrt{\mathbf{x}^T\Sigma\mathbf{x}}) \Big] = \min_{||\mathbf{x}||_2=1} (\sigma - \sqrt{\mathbf{x}^T\Sigma\mathbf{x}})^2, \tag{88}$$

and (87) follows.

Example 5 (Wasserstein-2 distance between Dirac measure on $\mathbf{R}^m$ and a discrete measure on $\mathbf{R}^d$). *Let $\mathbf{y} \in \mathbf{R}^m$ and $\mu_1 \in M(\mathbf{R}^m)$ be the Dirac measure with $\mu_1(\mathbf{y}) = 1$, i.e., all mass centered at $\mathbf{y}$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbf{R}^d$ be distinct points, $p_1, \ldots, p_k \geq 0, p_1 + \ldots + p_k = 0$, and let $\mu_2 \in M(\mathbf{R}^d)$ be the discrete measure of point masses with $\mu_2(\mathbf{x}_i) = p_i$, $i = 1, \ldots, k$. We seek the Wasserstein distance $\hat{W}_2(\mu_1, \mu_2)$ in a closed-form solution. Suppose $m \leq d$; then,*

$$(W_2^-(\mu_1,\mu_2))^2 = \inf_{V\in O(m,d)} \inf_{b\in\mathbf{R}^m} \sum_{i=1}^k p_i ||V\mathbf{x}_i + b - \mathbf{y}||_2^2 \\ = \inf_{V\in O(m,d)} \sum_{i=1}^k p_i ||V\mathbf{x}_i - \sum_{i=1}^k p_i V\mathbf{x}_i||_2^2 = \inf_{V\in O(m,d)} \mathrm{tr}(VCV^T) \tag{89}$$

noting that the second infimum is attained by $b = \mathbf{y} - \sum_{i=1}^k p_i V\mathbf{x}_i$ and defining C in the last infimum to be

$$C := \sum_{i=1}^k p_i \left(\mathbf{x}_i - \sum_{i=1}^k p_i\mathbf{x}_i\right)\left(\mathbf{x}_i - \sum_{i=1}^k p_i\mathbf{x}_i\right)^T \in \mathbf{R}^{d\times d}. \tag{90}$$

Let the eigenvalue decomposition of the symmetric positively semidefinite matrix C be $C = Q\Lambda Q^T$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$. Then,

$$\inf_{V\in O(m,d)} \mathrm{tr}(VCV^T) = \sum_{i=0}^{m-1} \lambda_{d-i} \tag{91}$$

and is attained when $V \in O(m,d)$ has row vectors given by the last m columns of $Q \in O(d)$.

Note that the geodesic distance (7) and (8) between Gaussian PDs (or corresponding covariance matrices) is equivalent to the formula for the Fisher information metric for the multivariate normal model [15]. Indeed, the multivariate normal model is a differentiable manifold, equipped with the Fisher information as a Riemannian metric; this may be used in statistical inference.

Example 6. *Consider i.i.d. random variables $Z_l, \ldots, Z_n$ to be bi-variately normally distributed with diagonal covariance matrices, i.e., we focus on the manifold $M_{diag} = \{\mathrm{N}(\mu, \Lambda) : \mu \in \mathbf{R}^2, \Lambda \text{ diagonal}\}$. In this manifold, consider the submodel $M_{diag}^* = \{\mathrm{N}(\mu, \sigma^2\mathbf{I}) : \mu \in \mathbf{R}^2, \sigma^2 \in \mathbf{R}_+\}$ corresponding to the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$. First, consider the standard statistical estimates $\bar{Z}$ for the mean and s_1, s_2 for the variances. If $\bar{\sigma}^2$ denotes the geodesic estimate of the* *common variance, the squared distance between the initial estimate and the geodesic estimate under the hypothesis H_0 is given by*

$$\frac{n}{2}\left[\left(\ln\frac{\bar{\sigma}^2}{s_1^2}\right)^2+\left(\ln\frac{\bar{\sigma}^2}{s_2^2}\right)^2\right] \tag{92}$$

which is minimized by $\bar{\sigma}^2 = s_1 s_2$. Hence, instead of the arithmetic mean of the initial standard variation estimates, we use as an estimate the geometric mean of these quantities.

Finally, we present the distance between the symmetric positively definite matrices of different dimensions. Let $m \le d$, A is $m \times m$ and $B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$ is $d \times d$; here, B_{11} is a $m \times m$ block. Then, the distance is defined as follows:

$$d_2(A,B) := \left(\sum_{j=1}^{m}\Big(\max[0, \ln\lambda_j(A^{-1}B_{11})]\Big)^2\right)^{1/2}. \tag{93}$$

In order to estimate the distance (93), after the simultaneous diagonalization of matrices A and B, the following classical result is useful:

Theorem 7 (**Cauchy interlacing inequalities**)**.** *Let $B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$ be a $d \times d$ symmetric positively definite matrix with eigenvalues $\lambda_1(B) \le \ldots \le \lambda_d(B)$ and $m \times m$ block B_{11}. Then,*

$$\lambda_j(B) \le \lambda_j(B_{11}) \le \lambda_{j+d-m}(B), j = 1, \ldots, m. \tag{94}$$

10. Context-Sensitive Probability Metrics

The weighted entropy and other weighted probabilistic quantities generated a substantial amount of literature (see [16,17] and the references therein). The purpose was to introduce a disparity between outcomes of the same probability: in the case of a standard entropy, such outcomes contribute the same amount of information/uncertainty, which is appropriate in context-free situations. However, imagine two equally rare medical conditions, occurring with probability $p \ll 1$, one of which carries a major health risk while the other is just a peculiarity. Formally, they provide the same amount of information: $-\log p$, but the value of this information can be very different. The applications of the weighted entropy to the clinical trials are in the process of active development (see [18] and the literature cited therein). In addition, the contribution to the distance (say, from a fixed distribution $\mathbf{Q}$) related to these outcomes, is the same in any conventional sense. The weighted metrics, or weight functions, are supposed to fulfill the task of samples graduation, at least to a certain extent.

Let the weight function or graduation $\varphi > 0$ on the phase space $\mathcal{X}$ be given. Define the total weighted variation (TWV) distance

$$\tau_\varphi(\mathbf{P}_1,\mathbf{P}_2) = \frac{1}{2}\left(\sup_A\Big[\int_A \varphi \mathrm{d}\mathbf{P}_1 - \int_A \varphi \mathrm{d}\mathbf{P}_2\Big] + \sup_A\Big[\int_A \varphi \mathrm{d}\mathbf{P}_2 - \int_A \varphi \mathrm{d}\mathbf{P}_1\Big]\right). \tag{95}$$

Similarly, define the weighted Hellinger distance. Let p_1, p_2 be the densities of $\mathbf{P}_1, \mathbf{P}_2$ w.r.t. to a measure ν. Then,

$$\eta_\varphi(\mathbf{P}_1,\mathbf{P}_2) := \frac{1}{\sqrt{2}}\left(\int \varphi(\sqrt{p_1}-\sqrt{p_2})^2 \mathrm{d}\nu\right)^{1/2}. \tag{96}$$

Lemma 1. *Let p_1, p_2 be the densities of $\mathbf{P}_1, \mathbf{P}_2$ w.r.t. to a measure ν. Then, $\tau_\varphi(\mathbf{P}_1, \mathbf{P}_2)$ is a distance and*

$$\tau_\varphi(\mathbf{P}_1, \mathbf{P}_2) = \frac{1}{2}\int \varphi|p_1 - p_2|\mathrm{d}\nu \tag{97}$$

Proof. The triangular inequality and other properties of the distance follow immediately. Next,

$$\begin{array}{c} \int_{p_1>p_2} \varphi(p_1 - p_2) = \frac{1}{2}(\int \varphi p_1 - \int \varphi p_2) + \frac{1}{2}\int \varphi|p_1 - p_2|\mathrm{d}\nu \\ \int_{p_2>p_1} \varphi(p_2 - p_1) = \frac{1}{2}(\int \varphi p_2 - \int \varphi p_1) + \frac{1}{2}\int \varphi|p_1 - p_2|\mathrm{d}\nu \end{array} \tag{98}$$

Summing up these equalities implies (97). □

Let $\int \varphi p_1 \mathrm{d}\nu \geq \int \varphi p_2 \mathrm{d}\nu$. Then, by the weighted Gibbs inequality [16], $\mathrm{KL}_\varphi(\mathbf{P}_1||\mathbf{P}_2) := \int \varphi p_1 \log \frac{p_1}{p_2} \geq 0$.

Theorem 8 (Weighted Pinsker's inequality).

$$\frac{1}{2}\int \varphi|p_1 - p_2| \leq \sqrt{\mathrm{KL}_\varphi(\mathbf{P}_1||\mathbf{P}_2)/2}\sqrt{\int \varphi p_1}. \tag{99}$$

Proof. Define the function $G(x) = x\log x - x + 1$. The following bound holds, cf. (12):

$$G(x) = x\log x - x + 1 \geq \frac{3}{2}\frac{(x-1)^2}{x+2}, x > 0. \tag{100}$$

Now, by the Cauchy–Schwarz inequality,

$$\begin{array}{l} \left(\int \varphi p_2|\frac{p_1}{p_2} - 1|\right)^2 \leq \int \varphi\frac{(\frac{p_1}{p_2}-1)^2}{\frac{p_1}{p_2}+2}p_2 \int \varphi\left(\frac{p_1}{p_2} + 2\right)p_2 \\ \leq 3\int \varphi\frac{(\frac{p_1}{p_2}-1)^2}{\frac{p_1}{p_2}+2}p_2 \int \varphi p_1 \leq 2\int \varphi G(\frac{p_1}{p_2})p_2 \int \varphi p_1 \leq \mathrm{KL}_\varphi(\mathbf{P}_1||\mathbf{P}_2) \int \varphi p_1. \end{array} \tag{101}$$

□

Theorem 9 (Weighted Le Cam's inequality).

$$\tau_\varphi(\mathbf{P}_1, \mathbf{P}_2) \geq \eta_\varphi(\mathbf{P}_1, \mathbf{P}_2)^2. \tag{102}$$

Proof. In view of inequality

$$\frac{1}{2}|p_1 - p_2| = \frac{1}{2}p_1 + \frac{1}{2}p_2 - \min[p_1, p_2] \geq \frac{1}{2}p_1 + \frac{1}{2}p_2 - \sqrt{p_1 p_2},$$

one obtains

$$\tau_\varphi(\mathbf{P}_1, \mathbf{P}_2) \geq \frac{1}{2}\int \varphi p_1 + \frac{1}{2}\int \varphi p_2 - \int \varphi\sqrt{p_1 p_2} = \eta_\varphi(\mathbf{P}_1, \mathbf{P}_2)^2. \tag{103}$$

□

Next, we relate TWV distance to the sum of sensitive errors of both types in statistical estimation. Let C be the critical domain for the checking the hypothesis H_1:$\mathbf{P}_1$ versus the alternative H_2:$\mathbf{P}_2$. Define by $\alpha_\varphi = \int_C \varphi p_1$ and $\beta_\varphi = \int_{\mathcal{X}\setminus C} \varphi p_2$ the weighted error probabilities of the I and II types.

Lemma 2. *Let $d = d_C$ be the decision rule with the critical domain C. Then,*

$$\inf_d[\alpha_\varphi + \beta_\varphi] = \frac{1}{2}\left[\int \varphi \mathrm{d}\mathbf{P}_1 + \int \varphi \mathrm{d}\mathbf{P}_2\right] - \tau_\varphi(\mathbf{P}_1, \mathbf{P}_2). \tag{104}$$

Proof. Denote $C^* = \{x : p_2(x) > p_1(x)\}$. Then, the result follows from the equality $\forall C$

$$\begin{aligned} &\textstyle\int_C \varphi \mathrm{d}\mathbf{P}_1 + \int_{\mathcal{X}\setminus C} \varphi \mathrm{d}\mathbf{P}_2 = \frac{1}{2}[\int \varphi \mathrm{d}\mathbf{P}_1 + \int \varphi \mathrm{d}\mathbf{P}_2] \\ &\textstyle+ \int \varphi |p_1 - p_2| [\mathbf{1}(x \in C \cap \mathcal{X} \setminus C^*) - \mathbf{1}(x \in C \cap C^*)]. \end{aligned} \tag{105}$$

□

Theorem 10 (Weighted Fano's inequality). *Let $\mathbf{P}_1, \dots, \mathbf{P}_M$, $M \geq 2$ be probability distributions such that $\mathbf{P}_j \ll \mathbf{P}_k$, $\forall j, k$. Then,*

$$\begin{aligned} &\inf_d \max_{1 \leq j \leq M} \int \varphi(x) \mathbf{1}(d(x) \neq j) \mathrm{d}\mathbf{P}_j(x) \geq \tfrac{\log(M)}{\log(M-1)} \tfrac{1}{M} \sum_{j=1}^{M} \int \varphi p_j \\ &- \tfrac{1}{\log(M-1)} \left[\tfrac{1}{M^2} \sum_{j,k=1}^{M} \mathrm{KL}_\varphi(\mathbf{P}_j, \mathbf{P}_k) + \log 2 \tfrac{1}{M} \sum_{j=1}^{M} \int \varphi p_j \right] \end{aligned} \tag{106}$$

where the infimum is taken over all tests with values in $\{1, \dots, M\}$.

Proof. Let $Z \in \{1, \dots, M\}$ be a random variable such that $\mathbf{P}(Z = i) = \frac{1}{M}$ and let $X \sim \mathbf{P}_Z$. Note that $\mathbf{P}_Z$ is a mixture distribution so that, for any measure ν such that $\mathbf{P}_Z \ll \nu$, we have $\frac{\mathrm{d}\mathbf{P}_Z}{\mathrm{d}\nu} = \frac{1}{M} \sum_{k=1}^{M} \frac{\mathrm{d}\mathbf{P}_j}{\mathrm{d}\nu}$ and so

$$\mathbf{P}(Z = j | X = x) = \mathrm{d}\mathbf{P}_j(x) \left(\sum_{k=1}^{M} \mathrm{d}\mathbf{P}_k(x) \right)^{-1}.$$

It implies by Jensen's inequality applied to the convex function $-\log x$

$$\begin{aligned} &\int \varphi(x) \sum_{j=1}^{M} \mathbf{P}(Z = j | X = x) \log \mathbf{P}(Z = j | X = x) \mathrm{d}\mathbf{P}_X(x) \\ &\leq \tfrac{1}{M^2} \sum_{j,k=1}^{M} \int \varphi \log\left(\tfrac{\mathrm{d}\mathbf{P}_j}{\mathrm{d}\mathbf{P}_k} \right) \mathrm{d}P_j - \log(M) \tfrac{1}{M} \sum_{j=1}^{M} \int \varphi p_j \\ &= \sum_{j,k=1}^{M} \mathrm{KL}_\varphi(\mathbf{P}_j, \mathbf{P}_k) - \log(M) \tfrac{1}{M} \sum_{j=1}^{M} \int \varphi p_j. \end{aligned} \tag{107}$$

On the other hand, denote by $q_j = \frac{\mathbf{P}(Z=j|X)}{\mathbf{P}(Z \neq d(X)|X)}$ and $h(x) = x \log x + (1-x) \log(1-x)$. Note that $h(x) \geq -\log 2$ and by Jensen's inequality $\sum_{j \neq d(X)} q_j \log q_j \geq -\log(M-1)$. The following inequality holds:

$$\begin{aligned} &\sum_{j=1}^{M} \mathbf{P}(Z = j | X) \log \mathbf{P}(Z = j | X) \\ &= (1 - \mathbf{P}(Z \neq d(X) | X)) \log(1 - \mathbf{P}(Z \neq d(X) | X)) + \sum_{j \neq d(X)} \mathbf{P}(Z = j | X) \log \mathbf{P}(Z = j | X) \\ &= h(\mathbf{P}(Z = d(X) | X)) + \mathbf{P}(Z \neq d(X) | X) \sum_{j \neq d(X)} q_j \log q_j \\ &\geq -\log 2 - \log(M-1) \mathbf{P}(d(X) \neq Z | X). \end{aligned} \tag{108}$$

Integration of (108) yields

$$\begin{aligned} &\int \varphi(x) \sum_{j=1}^{M} \mathbf{P}(Z = j | X = x) \log \mathbf{P}(Z = j | X = x) \mathrm{d}\mathbf{P}_X(x) \\ &\geq \left(-\log 2 \tfrac{1}{M} \sum_{j=1}^{M} \int \varphi p_j - \log(M-1) \max_{1 \leq j \leq M} \int \varphi(x) \mathbf{1}(d(x) \neq j) \mathrm{d}\mathbf{P}_j \right). \end{aligned} \tag{109}$$

Combining (107) and (109) proves (106). □

11. Conclusions

The contribution of the current paper is summarized in the Table 1 below. The objects 1–8 belong to the treasures of probability theory and statistics, and we present a number of examples and additional facts that are not easy to find in the literature. The objects 9–10, as well as the distances between distributions of different dimensions, appeared quite recently. They are not fully studied and quite rarely used in applied research. Finally, objects 11–12 have been recently introduced by the author and his collaborators. This is the field of the current and future research.

Table 1. The main metrics and divergencies.

Number	Name	Reference	Comment
1	Kullback–Leibler	(2)	Divergence but not a distance
2	Total variation (TV)	(1)	Bounded by Pinsker's inequality
3	Kolmogorov–Smirnov	p. 2	Specific for 1D case
4	Hellinger	(16)	Bounded by Le Cam's inequality
5	Lévy–Prohorov	(1)	Metrization of the weak convergence
6	Fréchet	(8, 80)	Requires the joint distribution
7	Wasserstein	(69)	Marginal distributions only
8	χ^2	p. 5	Divergence but not a distance
9	Jensen–Shannon	(6)	Constructed from Kullback–Leibler
10	Geodesic	(8)	Specific for Gaussian case
11	Weighted TV	(97)	Context sensitive
12	Weighted Hellinger	(98)	Context sensitive

Funding: This research is supported by the grant 23-21-00052 of RSF and the HSE University Basic Research Program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Suhov, Y.; Kelbert, M. *Probability and Statistics by Example: Volume I. Basic Probability and Statistics*; Second Extended Edition; Cambridge University Press: Cambridge, UK, 2014; 457p.
2. Rachev, S.T. *Probability Metrics and the Stability of Stochastic Models*; Wiley: New York, NY, USA, 1991.
3. Zeifman, A.; Korolev, V.; Sipin, A. (Eds.) *Stability Problems for Stochastic Models: Theory and Applications*; MDPI: Basel, Switzerland, 2020.
4. Endres, D.M.; Schindelin, J.E. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860. [CrossRef]
5. Kelbert, M.; Suhov, Y. What scientific folklore knows about the distances between the most popular distributions. *Izv. Sarat. Univ. (N.S.) Ser. Mat. Mekh. Inform.* **2022**, *22*, 233–240. [CrossRef]
6. Dwivedi, A.; Wang, S.; Tajer, A. Discriminant Analysis under f-Divergence Measures. *Entropy* **2022**, **24**, 188. [CrossRef] [PubMed]
7. Devroye, L.; Mehrabian, A.; Reddad, T. The total variation distance between high-dimensional Gaussians. *arXiv* **2020**, arXiv:1810.08693v5.
8. Vallander, S.S. Calculation of the Wasserstein distance between probability distributions on the line. *Theory Probab. Appl.* **1973**, *18*, 784–786. [CrossRef]

9. Rachev, S.T. The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory Probab. Appl.* **1985**, *29*, 647–676. [CrossRef]
10. Gelbrich, M. On a formula for the L_2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Math. Nachrichten* **1990**, *147*, 185–203. [CrossRef]
11. Givens, R.M.; Shortt, R.M. A class of Wasserstein metrics for probability distributions. *Mich. Math J.* **1984**, *31*, 231240. [CrossRef]
12. Olkin, I.; Pwelsheim, F. The distances between two random vectors with given dispersion matrices. *Lin. Algebra Appl.* **1982**, *48*, 267–2263. [CrossRef]
13. Dowson, D.C.; Landau, B.V. The Fréchet distance between multivariate Normal distributions. *J. Multivar. Anal.* **1982**, *12*, 450–456. [CrossRef]
14. Cai, Y.; Lim, L.-H. Distances between probability distributions of different dimensions. *IEEE Trans. Inf. Theory* **2022**, *68*, 4020–4031. [CrossRef]
15. Skovgaard, L.T. A Riemannian geometry of the multivariate normal model. *Scand. J. Stat.* **1984**, *11*, 211–223.
16. Stuhl, I.; Suhov, Y.; Yasaei Sekeh, S.; Kelbert, M. Basic inequalities for weighted entropies. *Aequ. Math.* **2016**, *90*, 817–848.
17. Stuhl, I.; Kelbert, M.; Suhov, Y.; Yasaei Sekeh, S. Weighted Gaussian entropy and determinant inequalities. *Aequ. Math.* **2022**, *96*, 85–114. [CrossRef]
18. Kasianova, K.; Kelbert, M.; Mozgunov, P. Response-adaptive randomization for multi-arm clinical trials using context-dependent information measures. *Comput. Stat. Data Anal.* **2021**, *158*, 107187. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.