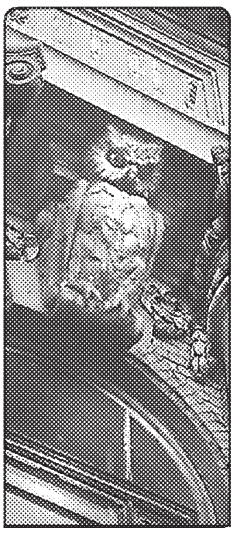

Научный отдел

МАТЕМАТИКА

Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика. 2023. Т. 23, вып. 4. С. 422–434
Izvestiya of Saratov University. Mathematics. Mechanics. Informatics, 2023, vol. 23, iss. 4, pp. 422–434
mmi.sgu.ru
https://doi.org/10.18500/1816-9791-2023-23-4-422-434
EDN: ANLRAB

Article

Wasserstein and weighted metrics for multidimensional Gaussian distributions

M. Y. Kelbert[1]✉, **Y. Suhov**[2]

[1]Higher School of Economics — National Research University, 20 Myasnitskaya St., Moscow 101000, Russia
[2]DPMMS, Penn State University, 201 Old Main, State College, PA 16802, USA
Mark Y. Kelbert, mkelbert@hse.ru, https://orcid.org/0000-0002-3952-2012, AuthorID: 1137288
Yurii Suhov, yms@statslab.cam.ac.uk, AuthorID: 1131362

Abstract. We present a number of low and upper bounds for Lévy – Prokhorov, Wasserstein, Frechét, and Hellinger distances between probability distributions of the same or different dimensions. The weighted (or context-sensitive) total variance and Hellinger distances are introduced. The upper and low bounds for these weighted metrics are proved. The low bounds for the minimum of different errors in sensitive hypothesis testing are proved.

Keywords: Lévy – Prokhorov distance, Wasserstein distance, weighted total variance distance, Dobrushin's inequality, weighted Pinsker's inequality, weighted le Cam's inequality, weighted Fano's inequality

Acknowledgements: This research is supported by the Russian Science Fund (project No. 23-21-00052).

For citation: Kelbert M. Y., Suhov Y. Wasserstein and weighted metrics for multidimensional Gaussian distributions. *Izvestiya of Saratov University. Mathematics. Mechanics. Informatics*, 2023, vol. 23, iss. 4, pp. 422–434. https://doi.org/10.18500/1816-9791-2023-23-4-422-434, EDN: ANLRAB

This is an open access article distributed under the terms of Creative Commons Attribution 4.0 International License (CC-BY 4.0)

© Kelbert M. Y., Suhov Y., 2023

Научная статья
УДК 519.85

Метрика Вассерштейна и взвешенные метрики для многомерных распределений Гаусса

М. Я. Кельберт[1]✉, **Ю. Сухов**[2]

[1]Национальный исследовательский университет «Высшая школа экономики», Россия, 101000, г. Москва, ул. Мясницкая, д. 20
[2]Университет штата Пенсильвания, Соединенные Штаты Америки, Пенсильвания, 16802, г. Стейт-Колледж, кампус Юниверсити-Парк, ул. Олд Мейн, д. 201

Кельберт Марк Яковлевич, кандидат физико-математических наук, профессор-исследователь департамента статистики и анализа данных факультета экономических наук, mkelbert@hse.ru, https://orcid.org/0000-0002-3952-2012, AuthorID: 1137288
Сухов Юрий, кандидат физико-математических наук, профессор математического факультета, yms@statslab.cam.ac.uk, AuthorID: 1131362

Аннотация. Приводится ряд нижних и верхних оценок для расстояний Леви – Прохорова, Вассерштейна, Фреше и Хеллингера между вероятностными распределениями одной и той же или разных размерностей. Вводится взвешенное (или контекстно зависимое) расстояние полной вариации и расстояние Хеллингера. Доказаны верхняя и нижняя оценки для этих взвешенных метрик. Доказаны нижние оценки минимума суммы различных ошибок при проверке чувствительных гипотез.

Ключевые слова: расстояние Леви – Прохорова, расстояние Вассерштейна, взвешенное расстояние полной вариации, неравенство Добрушина, взвешенное неравенство Пинскера, взвешенное неравенство Ле Кама, взвешенное неравенство Фано
Благодарности: Исследование выполнено при поддержке Российского научного фонда (проект № 23-21-00052).
Для цитирования: *Kelbert M. Y., Suhov Y.* Wasserstein and weighted metrics for multidimensional Gaussian distributions [*Кельберт М. Я., Сухов Ю.* Метрика Вассерштейна и взвешенные метрики для многомерных распределений Гаусса] // Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика. 2023. Т. 23, вып. 4. С. 422–434. https://doi.org/10.18500/1816-9791-2023-23-4-422-434, EDN: ANLRAB
Статья опубликована на условиях лицензии Creative Commons Attribution 4.0 International (CC-BY 4.0)

Introduction

In this note, we review basic facts about the metrics for probability measures and provide specific formulae and simplified proofs that could not be easily found in the literature. Alongside the classical results such as the evaluation of the Lévy – Prokhorov distance in terms of the Wasserstein distance presented in Section 1, we discuss some novel approaches. In Section 2, we review a recent development related to the distances between the distributions of different dimensions. Finally, in Section 3, we present the context-sensitive (or weighted) total variance distance and establish a number of new inequalities mimicking some classical results from the information theory. Sections 1 and 2 of the paper are basically a review but contain several improvements. Section 3 is purely original and was never published before.

1. Lévy – Prokhorov and Wasserstein distances

Let $\mathbf{P}_i$, $i = 1, 2$, be probability distributions on a metric space $\mathcal{W}$ 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 $\varepsilon > 0$ such that for any closed set $C \subset \mathcal{W}$,

$$\mathbf{P}_1(C) - \mathbf{P}_2(C_\varepsilon) < \varepsilon, \quad \mathbf{P}_2(C) - \mathbf{P}_1(C_\varepsilon) < \varepsilon \tag{1}$$

where C_ε stands for the ε-neighborhood of C in metric r. It could be easily checked that $\rho^{L-P}(\mathbf{P}_1, \mathbf{P}_2) \leqslant \tau(\mathbf{P}_1, \mathbf{P}_2)$, i. e. the total variance distance. Next, define the Wasserstein distance $W_p^r(\mathbf{P}_1, \mathbf{P}_2)$ between $\mathbf{P}_1, \mathbf{P}_2$ by

$$W_p^r(\mathbf{P}_1, \mathbf{P}_2) = \inf_{\bar{\mathbf{P}}} \left(\mathbf{E}_{\bar{\mathbf{P}}} \left[r(X_1, X_2)^p\right]\right)^{1/p}$$

where the infimum is taken over all joint $\bar{\mathbf{P}}$ on $\mathcal{W} \times \mathcal{W}$ with marginals $\mathbf{P}_i$. In the case of Euclidean space with $r(x_1, x_2) = \|x_1 - x_2\|$, the index r is omitted.

Theorem 1 (Dobrushin's bound)**.**

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

Proof. Suppose that there exists a closed set C for which at least one of the inequalities (1) fails, say $\mathbf{P}_1(C) \geqslant \varepsilon + \mathbf{P}_2(C_\varepsilon)$. Then, for any joint $\bar{\mathbf{P}}$ with marginals $\mathbf{P}_1$ and $\mathbf{P}_2$,

$$\begin{gathered}\mathbf{E}_{\bar{\mathbf{P}}} \left[r(X_1, X_2)\right] \geqslant \mathbf{E}_{\bar{\mathbf{P}}} \left[\mathbf{1}(r(X_1, X_2) \geqslant \varepsilon) r(X_1, X_2)\right] \geqslant \varepsilon \bar{\mathbf{P}}(r(X_1, X_2) \geqslant \varepsilon) \geqslant \\ \geqslant \varepsilon \bar{\mathbf{P}}(X_1 \in C, X_2 \in W \setminus C_\varepsilon) \geqslant \varepsilon \left[\bar{\mathbf{P}}(X_1 \in C) - \bar{\mathbf{P}}(X_1 \in C, X_2 \in C_\varepsilon)\right] \geqslant \\ \geqslant \varepsilon \left[\bar{\mathbf{P}}(X_1 \in C) - \bar{\mathbf{P}}(X_2 \in C_\varepsilon)\right] = \varepsilon \left[\mathbf{P}_1(X_1 \in C) - \mathbf{P}_2(X_2 \in C_\varepsilon)\right] \geqslant \varepsilon^2.\end{gathered}$$

This leads to (2), 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 1D case $\mathcal{W} = \mathbf{R}^1$ we have (cf. [1]).

Theorem 2.

$$W_1(\mathbf{P}_1, \mathbf{P}_2) = \int_{\mathbf{R}} |F_1(x) - F_2(x)|\, dx. \tag{3}$$

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

$$\mathbf{E}[|X_1 - X_2|] = \int_0^1 |F_1^{-1}(y) - F_2^{-1}(y)|\, dy = \int_{\mathbf{R}} |F_1(x) - F_2(x)|\, dx.$$

Let us now prove the inverse inequality. Set $Y = X_2 - X_1 \vee 0, Z = X_1 - X_2 \vee 0$ then $\mathbf{E}[|X_1 - X_2|] = \mathbf{E}[Y] + \mathbf{E}[Z]$. It could be easily checked that

$$\mathbf{E}[Z] = \int_{\mathbf{R}} \bar{\mathbf{P}}(X_1 \leqslant y, X_2 \geqslant y)|\, dy.$$

A similar argument can be done for Y, by swapping X_1 and X_2. This yields

$$\mathbf{E}[|X_1 - X_2|] = \int_{\mathbf{R}} \left[\bar{\mathbf{P}}(X_1 \leqslant y, X_2 \geqslant y) + \bar{\mathbf{P}}(X_2 \leqslant y, X_1 \geqslant y)\right]\, dy =$$

$$= \int_{\mathbf{R}} \left[\mathbf{P}_1(X_1 \leqslant y) + \mathbf{P}_2(X_2 \leqslant y) - 2\bar{\mathbf{P}}(X_1 \leqslant y, X_2 \leqslant y)\right] dy \geqslant$$

$$\geqslant \int_{\mathbf{R}} \left[F_1(x) + F_2(x) - 2\min[F_1(x), F_2(x)]\right] dx = \int_{\mathbf{R}} |F_1(x) - F_2(x)|\, dx. \qquad \square$$

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

$$W_p(\mathbf{P}_1, \mathbf{P}_2)^p = p(p-1) \int_{-\infty}^{\infty} dy \int_{y}^{\infty} \max[F_2(y) - F_1(x), 0](x-y)^{p-2}\, dx +$$

$$+ p(p-1) \int_{-\infty}^{\infty} dx \int_{x}^{\infty} \max[F_1(x) - F_2(y), 0](y-x)^{p-2}\, dy.$$

Proof. Follows from the identity

$$\mathbf{E}[|X - Y|^p] = p(p-1) \int_{-\infty}^{\infty} dy \int_{y}^{\infty} [F_2(y) - F(x,y)](x-y)^{p-2}\, dx +$$

$$+ p(p-1) \int_{-\infty}^{\infty} dx \int_{x}^{\infty} [F_1(x) - F(x,y)](y-x)^{p-2}\, dy.$$

The minimum is achieved for $\bar{F}(x,y) = \min[F_1(x), F_2(y)]$. Alternative expression (see [2]):

$$W_p(\mathbf{P}_1, \mathbf{P}_2)^p = \int_0^1 |F_1^{-1}(t) - F_2^{-1}(t)|^p\, dt. \qquad \square$$

Proposition 2. *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*

$$\rho^{F_1}(\mathbf{X}, \mathbf{Y}) := \mathbf{E}\left[\sum_{j=1}^{d} |X_j - Y_j|\right] =$$

$$= \sum_{j=1}^{d} \left[(\mu_j^X - \mu_j^Y)\left(1 - 2\Phi(-\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j})\right) + 2\hat{\sigma}_j \varphi(-\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j})\right], \qquad (4)$$

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 (4) vanishes, and the second term gives

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

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

$$\rho^{F_3}(\mathbf{X}, \mathbf{Y}) = \left(\sum_{j=1}^{d} \mathbf{E}|X_j - Y_j|^3\right)^{1/3} = \left(\sum_{j=1}^{d} (\mu_j^X - \mu_j^Y)^3 \left(1 - 2\Phi\left(-\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j}\right)\right) +\right.$$

$$+ 6(\mu_j^X - \mu_j^Y)^2 \hat{\sigma}_j \varphi\left(-\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j}\right) + 3(\hat{\sigma}_j)^2 (\mu_j^X - \mu_j^Y)\left[1 - 2\Phi\left(-\frac{(\mu_j^X - \mu_j^Y)}{\hat{\sigma}_j}\right) -\right.$$

$$-2\frac{(\mu_j^X-\mu_j^Y)}{\hat\sigma_j}\varphi\left(-\frac{(\mu_j^X-\mu_j^Y)}{\hat\sigma_j}\right)\Bigg]+2(\hat\sigma_j)^3\varphi\left(-\frac{(\mu_j^X-\mu_j^Y)}{\hat\sigma_j}\right)\left[\left(\frac{(\mu_j^X-\mu_j^Y)}{\hat\sigma_j}\right)^2+2\right]\Bigg)^{1/3},$$

$$\rho^{F_4}(\mathbf{X},\mathbf{Y})=\left(\sum_{j=1}^d\mathbf{E}|X_j-Y_j|^4\right)^{1/4}=\left(\sum_{j=1}^d(\mu_j^X-\mu_j^Y)^4+6(\mu_j^X-\mu_j^Y)^2(\hat\sigma_j)^2+3(\hat\sigma_j)^4\right)^{1/4}.$$

Let $\mu_j^X=\mu_j^Y$. The expressions for $\rho^{F_1}-\rho^{F_4}$ are minimized when $\mathrm{Cov}(X_j,Y_j), j=1,\dots,d$ 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 positive-definite. So, the straightforward choice $\mathrm{Corr}(X_j,Y_j)=1$ is not always possible, see Theorem 3 below.

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 the 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 as 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.

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

$$\rho^{L-P}(\mathbf{P}_1,\mathbf{P}_2)\leqslant[W_p(\mathbf{P}_1,\mathbf{P}_2)]^{p/2},\quad p\geqslant 1.$$

Theorem 3. *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*[1]*. 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*

$$W_2(\mathbf{X}_1,\mathbf{X}_2)=\left[\|\mu_1-\mu_2\|^2+\mathrm{tr}(\Sigma_1^2)+\mathrm{tr}(\Sigma_2^2)-2\mathrm{tr}[(\Sigma_1\Sigma_2^2\Sigma_1)^{1/2}]\right]^{1/2}\tag{5}$$

where $(\Sigma_1\Sigma_2^2\Sigma_1)^{1/2}$ stands for the positive-definite matrix square-root. The value (5) 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. *Let $\mu_1=\mu_2=0$. Then for $d=1$: $W_2(X_1,X_2)=|\sigma_1-\sigma_2|$. For $d=2$*

$$W_2(\mathbf{X}_1,\mathbf{X}_2)=\left[\mathrm{tr}(\Sigma_1^2)+\mathrm{tr}(\Sigma_2^2)-2[\mathrm{tr}(\Sigma_1^2\Sigma_2^2)+2\sqrt{\det(\Sigma_1\Sigma_2)}]^{1/2}\right]^{1/2}.\tag{6}$$

Note that the expression in (6) vanishes when $\Sigma_1^2=\Sigma_2^2$.

Example 1. (a) Let $\mathbf{X}\sim\mathrm{N}(0,\Sigma_X^2)$, $\mathbf{Y}\sim\mathrm{N}(0,\Sigma_Y^2)$ 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 $d=2$, $\mathbf{X}\sim\mathrm{N}(0,\Sigma_X^2)$, $\mathbf{Y}\sim\mathrm{N}(0,\Sigma_Y^2)$ 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

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

[1] In general case the statement holds with Σ_1^{-1} understood as Moore – Penrose inversion.

(c) Let $d = 2$, $\mathbf{X} \sim \mathrm{N}(0, \Sigma_X^2)$, $\mathbf{Y} \sim \mathrm{N}(0, \Sigma_Y^2)$ 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 (5) 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. (8) below). Let us write the covarianve matrix in the block form

$$\Sigma^2 = \begin{pmatrix} \Sigma_1^2 & K \\ K^T & \Sigma_2^2 \end{pmatrix} = \begin{pmatrix} \Sigma_1 & 0 \\ K^T\Sigma_1^{-1} & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} \Sigma_1 & \Sigma_1^{-1}K \\ 0 & I \end{pmatrix} \tag{7}$$

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 (7) that minimizes the expression

$$\int_{\mathbf{R}^d \times \mathbf{R}^d} \|\mathbf{x} - \mathbf{y}\|^2 \, d\mathbf{P}_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = \mathrm{tr}(\Sigma_1^2) + \mathrm{tr}(\Sigma_2^2) - 2\mathrm{tr}(K) \tag{8}$$

subject to constraining that the matrix Σ^2 in (7) is positively definite. The goal is to check that the minimum (5) is achieved when Shur's complement S in (7) equals 0. Consider the fiber $\sigma^{-1}(S)$, i.e. the set of all matrix K such that $\sigma(K) := \Sigma_Y^2 - K^T(\Sigma_X^2)^{-1}K = S$. It is enough to check the maximum value of $\mathrm{tr}(K)$ on this fiber equals

$$\max_{F \in \sigma^{-1}(S)} \mathrm{tr}(K) = \mathrm{tr}\left[(\Sigma_Y(\Sigma_X^2 - S)\Sigma_Y)^{1/2} \right]. \tag{9}$$

Since the matrix S is positively defined, it is easy to check that the fiber $S = 0$ should be selected. In order to establish (9), represent the positively definite matrix $\Sigma_Y^2 - S$ in the form $\Sigma_Y^2 - S = UD_r^2U^T$ 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 eigen vectors. We obtain the following $r \times r$ identity:

$$(\Sigma_X^{-1}KU_rD_r^{-1})^T(\Sigma_X^{-1}KU_rD_r^{-1}) = \mathbf{I}_r.$$

It means that $\Sigma_X^{-1}KU_rD_r^{-1} = O_r$, an 'orthogonal' $d \times r$ matrix, with $O_r^TO_r = \mathbf{I}_r$, 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

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

in a matrix-valued argument O_r, subject to the constraint $O_r^TO_r = \mathbf{I}_r$. A straightforward computation gives the answer $\mathrm{tr}[(M^TM)^{1/2}]$ which is equivalent to (9). The technical details can be found in [3] and [4]. □

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 [5]

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

2. The distances between distributions of different dimensions

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

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

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 $\mu \in \mathrm{M}(\mathbf{R}^m)$ and $\nu \in \mathrm{M}(\mathbf{R}^d)$, the embeddings of μ into $\mathbf{R}^d$ are the set of d-dimensional measures $\Phi^+(\mu,d) := \{\alpha \in \mathrm{M}(\mathbf{R}^d) : \varphi_{V,\beta}(\alpha) = \mu\}$ 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 γ between measures of the same dimension, define the projection distance $\gamma^-(\mu,\nu) := \inf_{\beta\in\Phi^-(\nu,m)} \gamma(\mu,\beta)$ and the embedding distance $\gamma^+(\mu,\nu) := \inf_{\alpha\in\Phi^+(\mu,d)} \gamma(\alpha,\nu)$. It may be proved [6] that $\gamma^+(\mu,\nu) = \gamma^-(\mu,\nu)$, denote the common value by $\hat{\gamma}(\mu,\nu)$.

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

$$\hat{W}_2(X,Y) = \begin{cases} \sigma - \sqrt{\lambda_1}\ if\ \sigma > \sqrt{\lambda_1}, \\ 0\ if\ \sqrt{\lambda_d} \leqslant \sigma \leqslant \sqrt{\lambda_1}, \\ \sqrt{\lambda_d} - \sigma\ if\ \sigma < \sqrt{\lambda_d}. \end{cases} \tag{10}$$

Indeed, in view of Theorem 3, write

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

and (10) follows.

Example 3 (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 $\mu_1 \in \mathrm{M}(\mathbf{R}^m)$ 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, $p_1,\ldots,p_k \geqslant 0$, $p_1 + \ldots + p_k = 0$, and let$\mu_2 \in \mathrm{M}(\mathbf{R}^d)$ 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 $m \leqslant d$, then

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

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

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

Let the eigenvalue decomposition of the symmetric positive semi-definite matrix C be $C = Q\Lambda Q^T$ with $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$, $\lambda_1 \geqslant \dots \geqslant \lambda_d \geqslant 0$. Then

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

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

A closely related question is to find a projection of zero-mean Gaussian models to the space of a low dimension r such distance between the projections of $\mathbf{X}$ and $\mathbf{Y}$ is maximal. We start the discussion with the TV distance. 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 $\mathbf{X} \sim N(\mu_1, \Sigma_1)$ and $\mathbf{Y} \sim N(\mu_2, \Sigma_2)$ such that $\mathrm{TV}(A\mathbf{X}, A\mathbf{Y}) \to \max$. The problem may be reduced to the case $\mu_1 = \mu_2 = 0$, $\Sigma_1 = \mathbf{I}_d$, $\Sigma_2 = \Sigma$, cf. [7]. Based on the results from [7, 8] it is natural to maximize

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

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

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

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

Remark. For zero-mean Gaussian models, this procedure may be repeated mutatis mutandis for any of the so-called f-divergences $D_f(\mathbf{P}\|\mathbf{Q}) := \mathbf{E}_\mathbf{P}\left[f\left(\frac{d\mathbf{Q}}{d\mathbf{P}}\right)\right]$ where f is a convex function such that $f(1) = 0$, cf. [7]. 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 estimations, the following result is utterly useful.

Theorem 4 (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 $\{\gamma_i, i = 1, \dots, r\}$ (sorted in the descending order) satisfy*

$$\lambda_{d-(r-i)} \leqslant \gamma_i \leqslant \lambda_i, \quad i = 1, \dots, r.$$

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

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

The KL-divergence is not symmetric and does not satisfy the triangle inequality. However, it gives rise to the so-called Jensen – Shannon metric [9]

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

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

$$0 \leqslant \mathrm{JS}(\mathbf{P}, \mathbf{Q}) \leqslant \mathrm{TV}(\mathbf{P}, \mathbf{Q}).$$

Jensen – Shannon metric is not easy to compute in terms of covariance matrices in a 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 positive-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 $d(\Sigma_1, \Sigma_2) = \sqrt{\sum\limits_{j=1}^{d} (\ln \lambda_j)^2}$, and a *geodesic* metric between Gaussian PDs $X_1 \sim \mathrm{N}(\mu_1, \Sigma_1)$ and $X_2 \sim \mathrm{N}(\mu_2, \Sigma_2)$:

$$d(X_1, X_2) = \left(\delta^T S^{-1} \delta\right)^{1/2} + \left(\sum_{j=1}^{d} (\ln \lambda_j)^2\right)^{1/2} \tag{11}$$

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

$$d^2(\Sigma_1, \Sigma_2) = \mathrm{tr}\left[(\ln(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}))^2\right]. \tag{12}$$

Remark. It may be proved that the set of symmetric positively-definite matrices $M^+(d, \mathbf{R})$ is a Riemannian manifold, and (12) 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 $M(d, \mathbf{R})$.

Note that the geodesic distance (11) and (12) between Gaussian PDs (or corresponding covariance matrices) is equivalent to the formula for the Fisher information metric for the multivariate normal model [5]. Indeed, the multivariate normal model is a differentiable manifold, equipped with the Fisher information as the Riemannian metric, which may be used in statistical inference.

Example 4. Consider i.i.d. random variables $Z_l, \ldots, Z_n$ being 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_1^2, s_2^2 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

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

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

Finally, we present the distance between the symmetric positively definite matrices of different dimensions. Let $m \leqslant d$, 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 $m \times m$ block. Then the distance is defined as follows

$$d_2(A, B) := \left(\sum_{j=1}^{m} \left(\max[0, \ln \lambda_j(A^{-1} B_{11})] \right)^2 \right)^{1/2}. \tag{13}$$

In order to estimate the distance (13), after the simultaneous diagonalization of matrices A and B, the following classical result is useful.

Theorem 5 (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* $\lambda_1(B) \leqslant \ldots \leqslant \lambda_d(B)$ *and* $m \times m$ *block* B_{11}*. Then*

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

3. Context sensitive probability metrics

Let the weight function or graduation $\varphi > 0$ of the phase space $\mathcal{X}$ is given (cf. [10,11]). Define the total weighted variation (TWV) distance

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

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

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

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

$$\tau_\varphi(\mathbf{P}_1, \mathbf{P}_2) = \frac{1}{2} \int \varphi |p_1 - p_2| \, d\nu. \tag{14}$$

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

$$\int_{p_1 > p_2} \varphi(p_1 - p_2) = \frac{1}{2} \left(\int \varphi p_1 - \int \varphi p_2 \right) + \frac{1}{2} \int \varphi |p_1 - p_2| \, d\nu,$$
$$\int_{p_2 > p_1} \varphi(p_2 - p_1) = \frac{1}{2} \left(\int \varphi p_2 - \int \varphi p_1 \right) + \frac{1}{2} \int \varphi |p_1 - p_2| \, d\nu.$$

Summing these equalities, one gets (14). □

Let $\int \varphi p_1 \, d\nu \geqslant \int \varphi p_2 \, d\nu$. Then, by the weighted Gibbs inequality [10], $\mathrm{KL}_\varphi(\mathbf{P}_1 \| \mathbf{P}_2) \geqslant 0$.

Theorem 6 (Weighted Pinsker's inequality)**.**

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

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

$$G(x) = x\log x - x + 1 \geqslant \frac{3}{2}\frac{(x-1)^2}{x+2}, \quad x > 0. \tag{15}$$

Indeed, since both terms of the inequality (15) coincide at $x = 1$, and their first derivatives coincide at $x = 1$, the following inequality $f''(x) = \frac{1}{x} \geqslant \frac{27}{(x+2)^3}$ proves the result. Now, by the Cauchy – Schwarz inequality

$$\left(\int \varphi p_2 |\frac{p_1}{p_2} - 1|\right)^2 \leqslant \int \varphi \frac{(\frac{p_1}{p_2}-1)^2}{\frac{p_1}{p_2}+2} p_2 \int \varphi \left(\frac{p_1}{p_2}+2\right) p_2 \leqslant$$

$$\leqslant 3\int \varphi \frac{(\frac{p_1}{p_2}-1)^2}{\frac{p_1}{p_2}+2} p_2 \int \varphi p_1 \leqslant \int \varphi g(\frac{p_1}{p_2}) p_2 \int \varphi p_1 \leqslant \mathrm{KL}_\varphi(\mathbf{P}_1||\mathbf{P}_2)\int \varphi p_1. \qquad \square$$

Theorem 7 (Weighted Le Cam's inequality)**.**

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

Proof. In view of inequality

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

one gets

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

Next, we relate TWV distance to the sum of sensitive errors of both types in statistical estimates. Let C be the critical domain for 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*

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

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

$$\int_C \varphi\, d\mathbf{P}_1 + \int_{\mathcal{X}\setminus C} \varphi\, d\mathbf{P}_2 = \frac{1}{2}\left[\int \varphi\, d\mathbf{P}_1 + \int \varphi\, d\mathbf{P}_2\right] +$$

$$+ \int \varphi |p_1 - p_2| \left[\mathbf{1}(x \in C \cap \mathcal{X} \setminus C^*) - \mathbf{1}(x \in C \cap C^*)\right]. \qquad \square$$

Theorem 8 (Weighted Fano's inequality)**.** *Let $\mathbf{P}_1, \ldots, \mathbf{P}_M$, $M \geqslant 2$ be probability distributions such that $\mathbf{P}_j \ll \mathbf{P}_k$, $\forall j, k$. Then*

$$\inf_d \max_{1\leqslant j\leqslant M} \int \varphi(x)\mathbf{1}(d(x) \neq j)\, d\mathbf{P}_j(x) \geqslant \frac{1}{M}\sum_{j=1}^M \int \varphi\, d\mathbf{P}_j -$$

$$-\frac{1}{\log(M-1)}\left[\frac{1}{M^2}\sum_{j,k}\mathrm{KL}_\varphi(\mathbf{P}_j, \mathbf{P}_k) + \log 2\frac{1}{M}\sum_{j=1}^M \int \varphi\, d\mathbf{P}_j\right] \tag{16}$$

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

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 $\frac{d\mathbf{P}_Z}{d\nu} = \frac{1}{M} \sum\limits_{k=1}^{M} \frac{d\mathbf{P}_j}{d\nu}$ and so

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

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

$$\begin{aligned} &\int \varphi(x) \sum_{j=1}^{M} \mathbf{P}(Z = j|X = x) \log \mathbf{P}(Z = j|X = x)\, d\mathbf{P}_X(x) \leqslant \\ &\leqslant \frac{1}{M^2} \sum_{j,k}^{M} \int \varphi \log \left(\frac{d\mathbf{P}_j}{d\mathbf{P}_k} \right) dP_j - \log(M) \frac{1}{M} \sum_{j=1}^{M} \int \varphi p_j = \\ &= \frac{1}{M^2} \sum_{j,k}^{M} \mathrm{KL}_\varphi(\mathbf{P}_j, \mathbf{P}_k) - \log(M) \frac{1}{M} \sum_{j=1}^{M} \int \varphi p_j. \end{aligned} \tag{17}$$

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 $h(x) \geqslant -\log 2$ and by Jensen's inequality $\sum_{j \neq d(X)} q_j \log q_j \geqslant -\log(M-1)$. The following inequality holds

$$\begin{aligned} &\sum_{j=1}^{M} \mathbf{P}(Z = j|X) \log \mathbf{P}(Z = j|X) = \\ &= h(\mathbf{P}(Z \neq d(X)|X)) + \mathbf{P}(Z \neq d(X)|X) \sum_{j \neq d(X)} q_j \log q_j \geqslant \\ &\geqslant -\log 2 - \log(M-1) \mathbf{P}(d(X) \neq Z|X) \log(M-1). \end{aligned} \tag{18}$$

Integration of (18) yields

$$\begin{aligned} &\int \varphi(x) \sum_{j=1}^{M} \mathbf{P}(Z = j|X = x) \log \mathbf{P}(Z = j|X = x)\, d\mathbf{P}_X(x) \geqslant \\ &\geqslant -\log 2 \frac{1}{M} \sum_{j=1}^{M} \int \varphi\, d\mathbf{P}_j - \log(M-1) \max_{1 \leqslant j \leqslant M} \int \varphi(x) \mathbf{1}(d(x) \neq j)\, d\mathbf{P}_j. \end{aligned} \tag{19}$$

Combining (17) and (19) proves (16). □

References

1. Vallander S. S. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 1974, vol. 18, iss. 4, pp. 784–786. https://doi.org/10.1137/1118101
2. Rachev S. T. The Monge – Kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, 1985, vol. 29, iss. 4, pp. 647–676. https://doi.org/10.1137/1129093
3. Givens C. R., Shortt R. M. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 1984, vol. 31, iss. 2, pp. 231–240. https://doi.org/10.1307/mmj/1029003026

4. Olkin I., Pukelsheim F. The distances between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 1982, vol. 48, pp. 257–263. https://doi.org/10.1016/0024-3795(82)90112-4
5. Dowson D. C., Landau B. V. The Fréchet distance between multivariate Normal distributions. *Journal of Multivariate Analysis*, 1982, vol. 12, iss. 3, pp. 450–455. https://doi.org/10.1016/0047-259X(82)90077-X
6. Cai Y., Lim L.-H., Distances between probability distributions of different dimensions. *IEEE Transactions on Information Theory*, 2022, vol. 68, iss. 6, pp. 4020–4031. https://doi.org/10.1109/TIT.2022.3148923
7. Dwivedi A., Wang S., Tajer A. Discriminant analysis under f-divergence measures. *Entropy*, 2022, vol. 24, iss. 2, art. 188, 26 p. https://doi.org/10.3390/e24020188
8. Devroye L., Mehrabian A., Reddad T. The total variation distance between high-dimensional Gaussians. *ArXiv*, 2020, ArXiv:1810.08693v5, pp. 1–12.
9. Endres D. M., Schindelin J. E. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 2003, vol. 49, iss. 7, pp. 1858–1860. https://doi.org/10.1109/TIT.2003.813506
10. Stuhl I., Suhov Y., Yasaei Sekeh S., Kelbert M. Basic inequalities for weighted entropies. *Aequationes Mathematicae*, 2016, vol. 90, iss. 4, pp. 817–848. https://doi.org/10.1007/s00010-015-0396-5
11. Stuhl I., Kelbert M., Suhov Y., Yasaei Sekeh S. Weighted Gaussian entropy and determinant inequalities. *Aequationes Mathematicae*, 2022, vol. 96, iss. 1, pp. 85–114. https://doi.org/10.1007/s00010-021-00861-3

Поступила в редакцию / Received 09.12.2022
Принята к публикации / Accepted 25.12.2022
Опубликована / Published 30.11.2023