Wasserstein statistics in one-dimensional location scale models

Shun-ichi Amari[1] · Takeru Matsuda[1]

Abstract

Wasserstein geometry and information geometry are two important structures to be introduced in a manifold of probability distributions. Wasserstein geometry is defined by using the transportation cost between two distributions, so it reflects the metric of the base manifold on which the distributions are defined. Information geometry is defined to be invariant under reversible transformations of the base space. Both have their own merits for applications. In this study, we analyze statistical inference based on the Wasserstein geometry in the case that the base space is one-dimensional. By using the location-scale model, we further derive the *W*-estimator that explicitly minimizes the transportation cost from the empirical distribution to a statistical model and study its asymptotic behaviors. We show that the *W*-estimator is consistent and explicitly give its asymptotic distribution by using the functional delta method. The *W*-estimator is Fisher efficient in the Gaussian case.

Keywords Information geometry · Location-scale model · Optimal transport · Wasserstein distance

1 Introduction

Wasserstein geometry defines a divergence between two probability distributions $p(x)$ and $q(x)$, $x \in X$ by using the cost of transportation from p to q. Hence, it reflects the metric of the underlying manifold X on which the probability distributions are defined. Information geometry, on the hand, studies an invariant structures wherein the geometry does not change under transformations of X which may change the distance within X. So information geometry is constructed independently of the metric of X.

Both geometries have their own histories (see, e.g., Villani 2003, 2009; Amari 2016). Information geometry has been successful in elucidating statistical inference,

✉ Shun-ichi Amari
  amari@brain.riken.jp

[1] RIKEN Center for Brain Science, 2-1 Hirosawa Wako City, Saitama 351-0198, Japan

where the Fisher information metric plays a fundamental role. It has successfully been applied to, not only statistics, but also machine learning, signal processing, systems theory, physics and many other fields (Amari 2016). Wasserstein geometry has been a useful tool in geometry, where the Ricci flow has played an important role (Villani 2009; Li and Montúfar 2020). Recently, it has found a widened scope of applications in computer vision, deep learning, etc. (e.g., Fronger et al. 2015; Arjovsky et al. 2017; Montavon et al. 2015; Peyré and Cuturi 2019). There have been attempts to connect the two geometries (see Amari et al. 2018, 2019; Wang and Li 2020 for examples), and Li and Zhao (2019) has proposed a unified theory connecting them.

It is natural to consider statistical inference from the Wasserstein geometry point of view (Li and Zhao 2019) and compare its results with information-geometrical inference based on the likelihood. The present article studies the statistical inference based on the Wasserstein geometry from a point of view different from that of Li and Zhao (2019). Given a number of independent observations from a probability distribution belonging to a statistical model with a finite number of parameters, we define the W-estimator that minimizes the transportation cost from the empirical distribution $\hat{p}(x)$ derived from observed data to the statistical model. This is the approach taken in many studies (see, e.g., Bernton et al. 2019; Bassetti et al. 2006). In contrast, the information geometry estimator is the one that minimizes the Kullback–Leibler divergence from the empirical distribution to the model, and it is the maximum likelihood estimator. Note that Matsuda and Strawderman (2021) investigated predictive density estimation under the Wasserstein loss.

We use a one-dimensional (1D) base space $X = \boldsymbol{R}^1$ and define the transportation cost equal to the square of the Euclidean distance between two points in $\boldsymbol{R}^1$. We give an equation for the W-estimator $\hat{\boldsymbol{\theta}}$ for a statistical model $S = \{p(x, \boldsymbol{\theta})\}$, where $p(x, \boldsymbol{\theta})$ is the probability density of x parametrized by a vector parameter $\boldsymbol{\theta}$. We then focus on the location-scale model to obtain explicit solutions of the W-estimator. We analyze its behavior, proving that it is consistent and furthermore derives its asymptotic distribution. The W-estimator is not Fisher efficient except for the Gaussian case, but it minimizes the W-divergence, which is the transportation cost between the empirical distribution and the model. We may say that it is W-efficient in this sense.

The present W-estimator is different from the estimator of Li and Zhao (2019), which is based on the Wasserstein score function. While their fundamental theory is a new paradigm connecting information geometry and Wasserstein geometry, their estimator does not minimize the W-divergence from the empirical one to the model. It is an interesting problem to compare these two frameworks of Wasserstein statistics.

The present paper is organized as follows. In Sect. 2, we introduce the W-estimator for a general parametric statistical model in the 1D-case. We show that the W-estimator uses only a linear function of the observations. In Sect. 3, we then focus on the location-scale model. We give an explicit form of the W-estimator. In Sect. 4, we analyze the asymptotic behavior of the W-estimator, proving that it is Fisher efficient in the Gaussian case. We study the geometry of the location-scale model in Sect. 5, showing that it is Euclidean (Li and Zhao 2019), although it is a curved submanifold in the function space of W-geometry (Takatsu 2011). Finally, we prove that the maximum likelihood

estimator asymptotically minimizes the transportation cost from the true distribution to the estimated one.

2 *W*-estimator

First, we show the optimal transportation cost of sending $p(x)$ to $q(x)$, $x \in \mathbf{R}^1$ when the transportation cost from x to y is $(x - y)^2$, where $x, y \in \mathbf{R}^1$. Let $P(x)$ and $Q(x)$ be the cumulative distribution functions of p and q, respectively, defined by

$$P(x) = \int_{-\infty}^{x} p(y)\mathrm{d}y, \quad Q(x) = \int_{-\infty}^{x} q(y)\mathrm{d}y.$$

Then, it is known (Santambrogio 2015; Peyré and Cuturi 2019) that the optimal transportation plan is to send mass of $p(x)$ at x to x' in a way that satisfies

$$P(x) = Q(x').$$

See Fig. 1. Thus, the total cost sending p to q is

$$C(p, q) = \int_{0}^{1} \left| P^{-1}(u) - Q^{-1}(u) \right|^2 \mathrm{d}u, \tag{1}$$

where P^{-1} and Q^{-1} are the inverse functions of P and Q, respectively.

We consider a regular statistical model

$$S = \{p(x, \boldsymbol{\theta})\},$$

parametrized by a vector parameter $\boldsymbol{\theta}$, where $p(x, \boldsymbol{\theta})$ is a probability density function of a random variable $x \in \mathbf{R}^1$ with respect to the Lebesgue measure of $\mathbf{R}^1$. Let

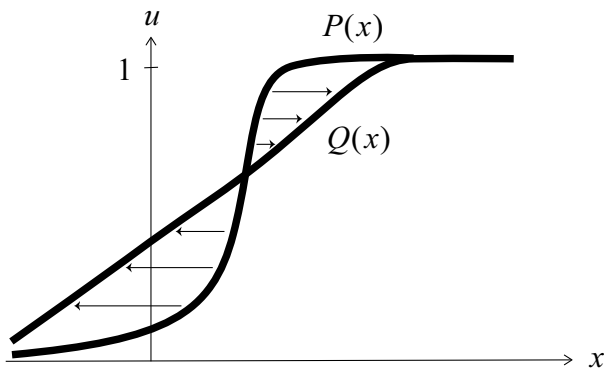$$D = \left\{ x_1, \ldots, x_n \right\}$$

Fig. 1 Optimal transportation plan from p to q

be n independent samples from $p(x, \theta)$. We denote the empirical distribution by

$$\hat{p}(x) = \frac{1}{n} \sum_i \delta(x - x_i),$$

where δ is the Dirac delta function. We rearrange $x_1, \cdots, x_n$ in the increasing order,

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)},$$

which are order statistics.

The optimal transportation plan from $\hat{p}(x)$ to $p(x, \theta)$ is explicitly solved when x is one-dimensional, $x \in \mathbf{R}^1$. The optimal plan is to transport mass at x to those points x' satisfying

$$\hat{P}(x_-) \leq P(x', \theta) \leq \hat{P}(x),$$

where $\hat{P}(x)$ and $P(x, \theta)$ are the (right-continuous) cumulative distribution functions of $\hat{p}(x)$ and $p(x, \theta)$, respectively:

$$\hat{P}(x) = \int_{-\infty}^{x} \hat{p}(y)\mathrm{d}y, \quad P(x, \theta) = \int_{-\infty}^{x} p(y, \theta)\mathrm{d}y,$$

and $\hat{P}(x_-) = \lim_{y \to x-0} \hat{P}(y)$. The total cost C of optimally transporting $\hat{p}(x)$ to $p(x, \theta)$ is given by

$$C(\theta) = \int_0^1 \left| \hat{P}^{-1}(u) - P^{-1}(u, \theta) \right|^2 \mathrm{d}u,$$

where $\hat{P}^{-1}$ and P^{-1} are inverse functions of $\hat{P}$ and P, respectively. Note that

$$\hat{P}^{-1}(u) = \inf\{y \mid P(y) \geq u\}.$$

Let $z_0(\theta), z_1(\theta), \ldots, z_n(\theta)$ be the points of the equi-probability partition of the distribution $p(x, \theta)$ such that

$$\int_{z_{i-1}(\theta)}^{z_i(\theta)} p(x, \theta)\mathrm{d}x = \frac{1}{n}, \tag{2}$$

where $z_0(\theta) = -\infty$ and $z_n(\theta) = \infty$. In terms of the cumulative distribution, $z_i(\theta)$ can be written as

$$P(z_i(\theta), \theta) = \frac{i}{n}$$

and

$$z_i(\theta) = P^{-1}\left(\frac{i}{n}, \theta\right).$$

See Fig. 2.

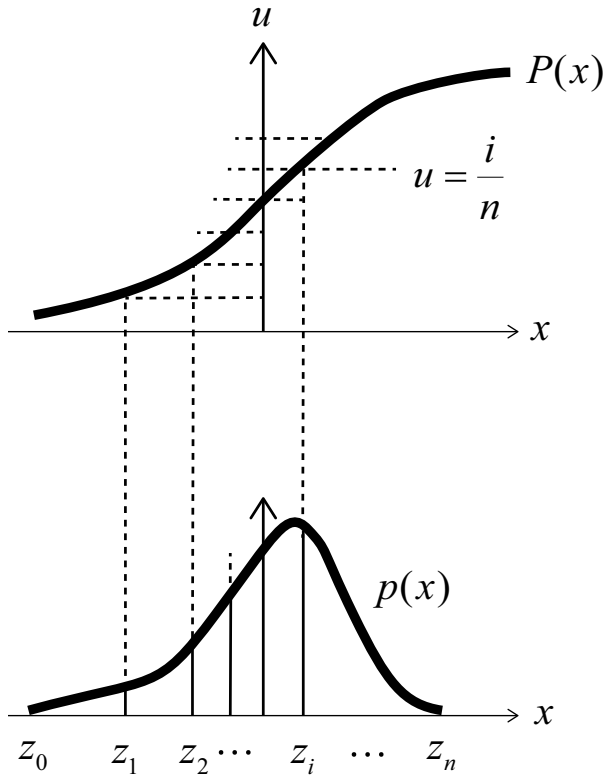The optimal transportation cost is rewritten as

Fig. 2 Equi-partition points $z_0, z_1, \ldots, z_n$ of $p(x)$

$$C(\boldsymbol{\theta}) = \sum_i \int_{z_{i-1}(\boldsymbol{\theta})}^{z_i(\boldsymbol{\theta})} (x_{(i)} - y)^2 p(y, \boldsymbol{\theta}) \mathrm{d}y = \frac{1}{n} \sum_i x_{(i)}^2 - 2 \sum_i k_i(\boldsymbol{\theta}) x_{(i)} + S(\boldsymbol{\theta}),$$

where we have used (2) and put

$$\begin{aligned} k_i(\boldsymbol{\theta}) &= \int_{z_{i-1}(\boldsymbol{\theta})}^{z_i(\boldsymbol{\theta})} y p(y, \boldsymbol{\theta}) \mathrm{d}y, \\ S(\boldsymbol{\theta}) &= \sum_i \int_{z_{i-1}(\boldsymbol{\theta})}^{z_i(\boldsymbol{\theta})} y^2 p(y, \boldsymbol{\theta}) \mathrm{d}y = \int_{-\infty}^{\infty} y^2 p(y, \boldsymbol{\theta}) \mathrm{d}y. \end{aligned} \tag{3}$$

By using the mean and variance of $p(x, \boldsymbol{\theta})$,

$$\begin{aligned} \mu(\boldsymbol{\theta}) &= \int_{-\infty}^{\infty} y p(y, \boldsymbol{\theta}) \mathrm{d}y, \\ \sigma^2(\boldsymbol{\theta}) &= \int_{-\infty}^{\infty} y^2 p(y, \boldsymbol{\theta}) \mathrm{d}y - \mu(\boldsymbol{\theta})^2, \end{aligned}$$

we have

$$S(\boldsymbol{\theta}) = \mu(\boldsymbol{\theta})^2 + \sigma^2(\boldsymbol{\theta}).$$

The *W*-estimator $\hat{\boldsymbol{\theta}}$ is the minimizer of $C(\boldsymbol{\theta})$. Differentiating $C(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and putting it equal to 0, we obtain the estimating equation as follows.

Theorem 1 *The W-estimator $\hat{\boldsymbol{\theta}}$ satisfies*

$$\frac{\partial}{\partial \boldsymbol{\theta}} \sum_i k_i(\boldsymbol{\theta}) x_{(i)} = \frac{1}{2} \frac{\partial}{\partial \boldsymbol{\theta}} S(\boldsymbol{\theta}). \tag{4}$$

It is interesting to see that the estimating equation is linear in *n* observations $x_{(1)}, \ldots, x_{(n)}$ for any statistical model. This is quite different from the maximum likelihood estimator or Bayes estimator.

Here, we will give a rough sketch showing that the *W*-estimator is consistent; that is, it converges to the true $\boldsymbol{\theta}_0$ as *n* tends to infinity (see Bassetti et al. 2006). More detailed discussions are given for the location-scale model in the next section. As *n* tends to infinity, the order statistic $x_{(i)}$ converges to the *i*th partition point $z_i(\boldsymbol{\theta}_0)$, when the true parameter is $\boldsymbol{\theta}_0$. From (3), we see that

$$k_i(\boldsymbol{\theta}) \approx \frac{1}{n} z_i(\boldsymbol{\theta})$$

as $n \to \infty$, so we have

$$\sum_i k_i(\boldsymbol{\theta}) x_{(i)} \approx \frac{1}{n} \sum_i z_i(\boldsymbol{\theta}) z_i(\boldsymbol{\theta}_0).$$

Moreover, as *n* tends to infinity,

$$S(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} z^2 p(z, \boldsymbol{\theta}) \mathrm{d}z \approx \frac{1}{n} \sum_i z_i^2(\boldsymbol{\theta}).$$

Therefore, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is the solution of (4), showing the consistency of the estimator.

Remark Bassetti et al. (2006) investigated existence, measurability and consistency of the *W*-estimator for general models and Bernton et al. (2019) extended this result to mis-specified models. Montavon et al. (2015) studied *W*-estimators for Boltzmann machines. In this study, we focus on the one-dimensional models, for which Theorem 1 gives a closed-form solution of the *W*-estimator.

3 *W*-estimator in location-scale model

Now, we focus on location-scale models. Let *f(z)* be a standard probability density function, satisfying

$$\int_{-\infty}^{\infty} f(z)\mathrm{d}z = 1,$$

$$\int_{-\infty}^{\infty} zf(z)\mathrm{d}z = 0,$$

$$\int_{-\infty}^{\infty} z^2 f(z)\mathrm{d}z = 1,$$

that is, its mean is 0 and the variance is 1. The location-scale model $p(x, \boldsymbol{\theta})$ is written as

$$p(x, \boldsymbol{\theta}) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right), \tag{5}$$

where $\boldsymbol{\theta} = (\mu, \sigma)$ is a parameter for specifying the distribution.

We define the equi-probability partition points z_i for the standard $f(z)$ as

$$z_i = F^{-1}\left(\frac{i}{n}\right),$$

where F is the cumulative distribution function

$$F(z) = \int_{-\infty}^{z} f(x)\mathrm{d}x.$$

We use the following transformation of the location and scale,

$$z = \frac{x - \mu}{\sigma},$$

$$x = \sigma z + \mu.$$

The equi-probability partition points $y_i = y_i(\boldsymbol{\theta})$ of $p(x, \boldsymbol{\theta})$ are given by

$$y_i = \sigma z_i + \mu.$$

The cost of the optimal transport from the empirical distribution $\hat{p}(x)$ to $p(x, \boldsymbol{\theta})$ is then written as

$$
\begin{aligned}
C(\mu, \sigma) &= \sum_i \int_{y_{i-1}}^{y_i} \left(x_{(i)} - x\right)^2 p(x, \mu, \sigma)\mathrm{d}x \\
&= \mu^2 + \sigma^2 + \frac{1}{n}\sum_i x_{(i)}^2 - 2\sum_i x_{(i)} \int_{z_{i-1}}^{z_i} (\sigma z + \mu) f(z)\mathrm{d}z.
\end{aligned} \tag{6}
$$

By differentiating (6), we obtain

$$\frac{1}{2}\frac{\partial}{\partial \mu} C = \mu - \frac{1}{n}\sum_i x_{(i)},$$

$$\frac{1}{2}\frac{\partial}{\partial \sigma} C = \sigma - \sum_i k_i x_{(i)},$$

where

$$k_i = \int_{z_{i-1}}^{z_i} z f(z) \mathrm{d}z, \tag{7}$$

which does not depend on μ or σ and depends only on the shape of f. By putting the derivatives equal to 0, we obtain the following theorem.

Theorem 2 *The W-estimator of a location-scale model is given by*

$$\hat{\mu} = \frac{1}{n} \sum_i x_{(i)}, \tag{8}$$

$$\hat{\sigma} = \sum_i k_i x_{(i)}. \tag{9}$$

Remark The *W*-estimator of the location parameter μ is the arithmetic mean of the observed data irrespective of the form of *f*. The *W*-estimator of the scale parameter σ is also a linear function of the observed data $x_{(1)}, \dots, x_{(n)}$, but it depends on f through k_i.

4 Asymptotic distribution of *W*-estimator

Here, we derive the asymptotic distribution of the *W*-estimator in location-scale models. Our derivation is based on the fact that the *W*-estimator has the form of L-statistics (van der Vaart 1998), which is a linear combination of order statistics.

Theorem 3 *The asymptotic distribution of the W-estimator $(\hat{\mu}, \hat{\sigma})$ in (8) (9) is*

$$\sqrt{n} \begin{pmatrix} \hat{\mu} - \mu \\ \hat{\sigma} - \sigma \end{pmatrix} \Rightarrow N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \frac{1}{2} m_3 \sigma^2 \\ \frac{1}{2} m_3 \sigma^2 & \frac{1}{4}(m_4 - 1)\sigma^2 \end{pmatrix} \right), \tag{10}$$

where

$$m_4 = \int_{-\infty}^{\infty} z^4 f(z) \mathrm{d}z, \quad m_3 = \int_{-\infty}^{\infty} z^3 f(z) \mathrm{d}z,$$

are the fourth and third moments of f(z), respectively.

Proof Without loss of generality, we focus on the case $\mu = 0$ and $\sigma = 1$. Let

$$\phi(\widetilde{F}) = \left(\int_0^1 \widetilde{F}^{-1}(u) \mathrm{d}u, \int_0^1 F^{-1}(u) \widetilde{F}^{-1}(u) \mathrm{d}u \right),$$

where F is the distribution function of f. Note that $\phi(F) = (0, 1)$.

Then, the *W*-estimator in (8) (9) is expressed as

$$(\hat{\mu}, \hat{\sigma}) = \phi(F_n),$$

where F_n is the empirical distribution of $x_1, \ldots, x_n$, because

$$k_i = \int_{(i-1)/n}^{i/n} F^{-1}(u)\mathrm{d}u.$$

To derive the asymptotic distribution of $\phi(F_n)$, we use the functional delta method (van der Vaart 1998). From Donsker's theorem (Theorem 19.3 of van der Vaart 1998),

$$\sqrt{n}(F_n - F) \Rightarrow \mathbb{G}_F = \mathbb{G} \circ F,$$

where $\mathbb{G}$ is the standard Brownian bridge. Namely, $\mathbb{G}_F$ is the mean zero Gaussian process on $(-\infty, \infty)$ with covariance given by

$$\mathrm{E}[\mathbb{G}_F(x)\mathbb{G}_F(y)] = F(x) \wedge F(y) - F(x)F(y),$$

where $s \wedge t = \min(s, t)$. Let $u = F(x)$ and $x_t = (F + tH)^{-1}(u)$ for sufficiently small t. Then, from $x_0 = x$,

$$u = F(x_t) + tH(x_t) = F(x) + f(x)(x_t - x) + tH(x) + O(t^2),$$

which yields

$$x_t = x - t\frac{H(x)}{f(x)} + O(t^2).$$

Thus, by putting $u = F(x)$,

$$\int_0^1 (F + tH)^{-1}(u)\mathrm{d}u = \int_0^1 \left(F^{-1}(u) - t\frac{H(F^{-1}(u))}{f(F^{-1}(u))} \right)\mathrm{d}u + O(t^2)$$
$$= \int_{-\infty}^{\infty} xf(x)\mathrm{d}x - t\int_{-\infty}^{\infty} H(x)\mathrm{d}x + O(t^2).$$

Similarly,

$$\int_0^1 F^{-1}(u)(F + tH)^{-1}(u)\mathrm{d}u = \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x - t\int_{-\infty}^{\infty} xH(x)\mathrm{d}x + O(t^2).$$

Therefore, ϕ is Hadamard differentiable with derivative given by

$$\phi_F'(H) = \lim_{t \to 0} \frac{\phi(F + tH) - \phi(F)}{t} = \left(-\int_{-\infty}^{\infty} H(x)\mathrm{d}x, -\int_{-\infty}^{\infty} xH(x)\mathrm{d}x \right).$$

Thus, from Theorem 20.8 of van der Vaart (1998),

$$\sqrt{n}(\phi(F_n) - \phi(F)) \Rightarrow \phi_F'(\mathbb{G}_F) \sim \mathrm{N}(0, \Sigma),$$

where

$$\Sigma_{11} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (F(x) \wedge F(y) - F(x)F(y)) \mathrm{d}x \mathrm{d}y,$$

$$\Sigma_{12} = \Sigma_{21} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(F(x) \wedge F(y) - F(x)F(y)) \mathrm{d}x \mathrm{d}y,$$

$$\Sigma_{22} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy(F(x) \wedge F(y) - F(x)F(y)) \mathrm{d}x \mathrm{d}y.$$

By using

$$\int_{-\infty}^{y} F(x)\mathrm{d}x = \left[(x-y)F(x)\right]_{x=-\infty}^{x=y} - \int_{-\infty}^{y} (x-y)f(x)\mathrm{d}x$$

$$= -\int_{-\infty}^{y} (x-y)f(x)\mathrm{d}x,$$

$$\int_{x}^{\infty} (y-x)(1-F(y))\mathrm{d}y = \left[\frac{(y-x)^2}{2}(1-F(y))\right]_{y=x}^{y=\infty} - \int_{x}^{\infty} \frac{(y-x)^2}{2}(-f(y))\mathrm{d}y$$

$$= \int_{x}^{\infty} \frac{(y-x)^2}{2}f(y)\mathrm{d}y,$$

and the symmetry of the integrand of Σ_{11}, we have

$$\Sigma_{11} = 2 \int_{-\infty}^{\infty} \int_{-\infty}^{y} F(x)(1-F(y))\mathrm{d}x \mathrm{d}y$$

$$= 2 \int_{-\infty}^{\infty} (1-F(y)) \int_{-\infty}^{y} F(x)\mathrm{d}x \mathrm{d}y$$

$$= -2 \int_{-\infty}^{\infty} (1-F(y)) \int_{-\infty}^{y} (x-y)f(x)\mathrm{d}x \mathrm{d}y$$

$$= 2 \int_{-\infty}^{\infty} f(x) \int_{x}^{\infty} (y-x)(1-F(y))\mathrm{d}y \mathrm{d}x$$

$$= 2 \int_{-\infty}^{\infty} f(x) \int_{x}^{\infty} \frac{(y-x)^2}{2}f(y)\mathrm{d}y \mathrm{d}x$$

$$= \int_{-\infty}^{\infty} \int_{x}^{\infty} (x-y)^2 f(x)f(y)\mathrm{d}y \mathrm{d}x$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-y)^2 f(x)f(y)\mathrm{d}y \mathrm{d}x.$$

Therefore, letting X and Y be independent samples from $f(z)$,

$$\Sigma_{11} = \frac{1}{2}E[(X-Y)^2] = m_2.$$

A similar calculation yields

$$\Sigma_{12} = \Sigma_{21} = E\left[\frac{1}{3}X^3 - \frac{1}{2}X^2Y + \frac{1}{6}Y^3\right] = \frac{1}{2}m_3,$$

$$\Sigma_{22} = E\left[\frac{(X^2 - Y^2)^2}{8}\right] = \frac{1}{4}(m_4 - 1).$$

Hence, we obtain (10). □

In particular, the *W*-estimator is Fisher efficient for the Gaussian model, but it is not efficient for other models.

Corollary 1 *For the Gaussian model, the asymptotic distribution of the W-estimator* $(\hat{\mu}, \hat{\sigma})$ *is*

$$\sqrt{n}\begin{pmatrix} \hat{\mu} - \mu \\ \hat{\sigma} - \sigma \end{pmatrix} \to N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{pmatrix}\right),$$

which attains the Cramer–Rao bound.

Proof For the Gaussian model, we have $m_4 = 3$ and $m_3 = 0$. □

Figure 3 plots the ratio of the mean square error $E[(\hat{\mu} - \mu)^2 + (\hat{\sigma} - \sigma)^2]$ of the *W*-estimator to that of the MLE for the Gaussian model with respect to *n*. The ratio converges to one as *n* goes to infinity, which shows that the *W*-estimator has statistical efficiency.

Figure 4 compares the mean square error of the *W*-estimator and MLE for the uniform model

$$f(z) = \begin{cases} \frac{1}{2\sqrt{3}} & (-\sqrt{3} \le z \le \sqrt{3}) \\ 0 & (\text{otherwise}) \end{cases}.$$

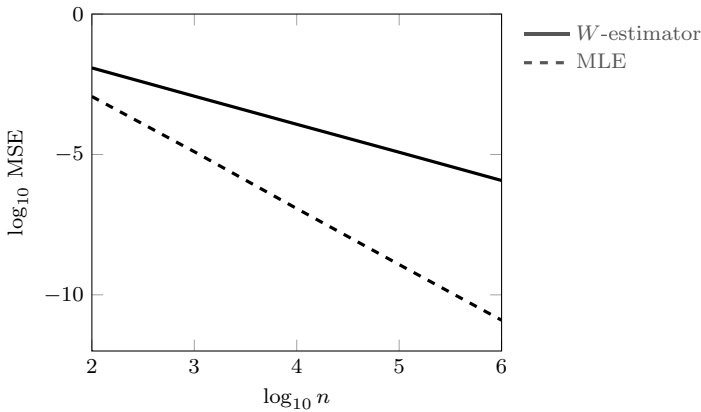**Fig. 3** Ratio of mean square error of *W*-estimator to that of MLE for the Gaussian model
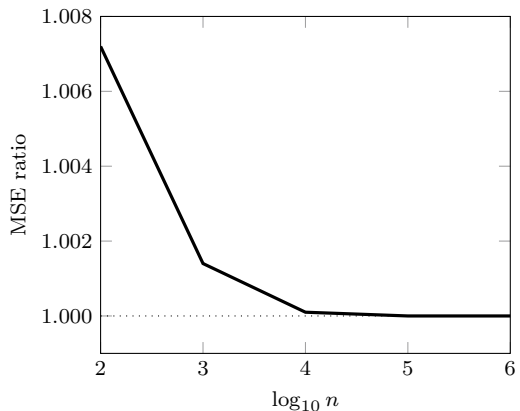
Fig. 4 Mean square error of W-estimator and MLE for the uniform model

In this case, the convergence rate of MLE is faster than $n^{-1/2}$, whereas the W-estimator is only $\sqrt{n}$-consistent.

5 Riemannian structure of *W*-divergence

Consider the manifold $M = \{p(x)\}$ of probability distributions which are absolutely continuous with respect to the Lebesgue measure and have finite second moments. It is known that M has a Riemannian structure due to the Wasserstein distance or the cost function. For two distributions $p(x)$ and $q(x)$, their optimal transportation cost, that is, the divergence between them, is given by (1).

We calculate the optimal transportation cost between two nearby distributions $p(x)$ and $p(x) + \delta p(x)$, where $\delta p(x)$ is infinitesimally small. We have

$$(P + \delta P)^{-1}(u) = P^{-1}(u) - \frac{\delta P\{x(u)\}}{P'\{x(u)\}},$$

where

$$x(u) = P^{-1}(u).$$

This equation is derived from

$$\frac{\mathrm{d}}{\mathrm{d}u} F^{-1}(u) = \frac{1}{f\{x(u)\}},$$

which comes from the differentiation of the identity

$$F^{-1}\{F(x)\} = x.$$

We thus have

$$C(p, p + \delta p) = \int_{-\infty}^{\infty} \frac{1}{p(x)} \left(\int_{-\infty}^{x} \delta p(y) \mathrm{d}y \right)^2 \mathrm{d}x \qquad (11)$$

which is a quadratic form of $\delta p(x)$. This gives a Riemannian metric to M.

The location-scale model S is a finite-dimensional submanifold embedded in M. For the location-scale model (5), we have

$$\delta p(y) = \frac{\partial}{\partial \mu} p(y, \boldsymbol{\theta}) d\mu + \frac{\partial}{\partial \sigma} p(y, \boldsymbol{\theta}) d\sigma.$$

The Riemannian metric tensor $G^W = \left(g_{ij}^W \right)$ is derived from

$$C(p, p + \delta p) = \sum g_{ij}^W(\boldsymbol{\theta}) d\theta_i d\theta_j.$$

See also Li and Zhao (2019).

Theorem 4 *The location-scale model is a Euclidean space, irrespective of f,*

$$g_{ij}^W = \delta_{ij}.$$

Proof We need to calculate (11). We have

$$\delta p(x, \boldsymbol{\theta}) = -\frac{1}{\sigma^2} f'\left(\frac{x-\mu}{\sigma}\right) d\mu - \frac{1}{\sigma^3} \left\{ \sigma f\left(\frac{x-\mu}{\sigma}\right) + (x-\mu) f'\left(\frac{x-\mu}{\sigma}\right) \right\} d\sigma.$$

Integration gives

$$\int_{-\infty}^{x} \delta p(y, \boldsymbol{\theta}) \mathrm{d}y = -p(x, \boldsymbol{\theta}) d\mu - \frac{1}{\sigma}(x-\mu) p(x, \boldsymbol{\theta}) d\sigma.$$

Hence, we have

$$C(\boldsymbol{\theta}, \boldsymbol{\theta} + d\boldsymbol{\theta}) = d\mu^2 + d\sigma^2.$$

$\square$

It is surprising that $G = \left(g_{ij} \right)$ is the identity matrix for the location-scale model, so S is a Euclidean space. See also Li and Zhao (2019). It is flat by itself, but S is a curved submanifold in M (Takatsu 2011), like a cylinder embedded in $\boldsymbol{R}^3$.

When n is large, the cost decreases on the order of $1/n$. The W-estimator is the projection of $\hat{p}(x)$ to S in the tangent space of M. Let $\hat{\boldsymbol{\theta}}'$ be another consistent estimator. Accordingly, we have the Pythagorean relation

$$C(\hat{p}, p_{\hat{\theta}'}) = C(\hat{p}, p_{\hat{\theta}}) + C(p_{\hat{\theta}}, p_{\hat{\theta}'}),$$

and the difference of the cost between the two estimators is

$$C(p_{\hat{\theta}}, p_{\hat{\theta}'}) = \frac{1}{n} \left| \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}' \right|^2.$$

Li and Zhao (2019) studied the properties of a *W*-estimator given by the *W* score function. They gave the *W*-efficiency and *W*-Cramer–Rao inequality. However, their *W*-estimator does not minimize the transportation cost.

6 Maximum likelihood estimator and *W*-divergence

It is an interesting problem to study the estimator that minimizes the transportation cost from the true distribution to the estimated one. Let $\hat{\theta}$ be a consistent estimator and let $e = \hat{\theta} - \theta_0$ be the estimation error vector, where θ_0 is the true parameter. We want to study the minimizer of $C(p_{\theta_0}, p_{\hat{\theta}})$. Since the W-metric g is the identity matrix for the location scale model, for the covariance $V = E[(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^\top]$ of the estimation error, we have

$$C = \mathrm{tr}V.$$

Therefore, the covariance is minimized when the expectations of the sum of the squares of the location error and scale error are at a minimum in the location scale case. Furthermore, we have a more general result.

Theorem 5 *The transportation cost is asymptotically minimized by the maximum likelihood estimator for a general statistical model.*

Proof The error covariance V satisfies the Cramer–Rao inequality

$$V \geq \frac{1}{n} G_F^{-1}$$

in the sense of the matrix positive-definiteness, where G_F is the Fisher information matrix. The minimum is attained asymptotically by the MLE. On the other hand, when $A \succeq B$ for two positive-definite matrices A and B,

$$\mathrm{tr}(G^W A) \geq \mathrm{tr}(G^W B).$$

Since the transportation cost is asymptotically written as

$$C = \mathrm{tr}(G^W V) \geq \frac{1}{n} \mathrm{tr}(G^W G_F^{-1}),$$

it is minimized for the maximum likelihood estimator that asymptotically attains $V = G_F^{-1}/n$.   □

It would be interesting to analyze the transportation cost of the *W*-estimator in general.

7 Discussion

There are three estimators, the MLE, *W*-score estimator and *W*-estimator. They have their own optimal properties and related behaviors. The MLE minimizes the KL divergence from the empirical distribution to the estimated distribution in the model. It minimizes the KL divergence and the *W*-divergence (transportation cost) from the true distribution to the estimated model at the same time. The *W*-estimator minimizes the transportation cost from the empirical distribution to the estimated distribution. However, it does not necessarily minimize the cost from the true distribution to the estimated one. The *W*-score estimator minimizes the integrated W-score function which is not the transportation cost. Further studies should be conducted on the merits and demerits of these estimators and their applicability to various problems.

References

Amari, S. (2016). *Information geometry and its applications*. New York: Springer.

Amari, S., Karakida, R., Oizumi, M. (2018). Information geometry connecting Wasserstein distance and Kullback–Leibler divergence via the entropy-relaxed transportation problem. *Information Geometry, 1,* 13–37.

Amari, S., Karakida, R., Oizumi, M., Cuturi, M. (2019). Information geometry for regularized optimal transport and barycenters of patterns. *Neural Computation, 31,* 827–848.

Arjovsky, M., Chintala, S., Bottou, L. (2017). Wasserstein GAN. arXiv:1701.07875.

Bassetti, F., Bodini, A., Regazzini, E. (2006). On minimum Kantorovich distance estimators. *Statistics & Probability Letters, 76,* 1298–1302.

Bernton, E., Jacob, P. E., Gerber, M., Robert, C. P. (2019). On parameter estimation with the Wasserstein distance. *Information and Inference*: *A Journal of the IMA*, *8,* 657–676.

Fronger, C., Zhang, C., Mobahi, H., Araya-Polo, M., Poggio, T. (2015). Learning with a Wasserstein loss. Advances in Neural Information Processing Systems 28 (NIPS 2015).

Li, W., Montúfar, G. (2020). Ricci curvature for parametric statistics via optimal transport. *Information Geometry, 3,* 89–117.

Li, W., Zhao, J. (2019). Wasserstein information matrix. arXiv:1910.11248.

Matsuda, T., Strawderman, W. E. (2021). Predictive density estimation under the Wasserstein loss. *Journal of Statistical Planning and Inference, 210,* 53–63.

Montavon, G., Müller, K. R., Cuturi, M. (2015). Wasserstein training for Boltzmann machine. Advances in Neural Information Processing Systems 29 (NIPS 2016).

Peyré, G., Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning, 11,* 355–607.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*. New York: Springer.

Takatsu, A. (2011). Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics, 48,* 1005–1026.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.

Villani, C. (2003). *Topics in optimal transportation*. New York: American Mathematical Society.

Villani, C. (2009). *Optimal transport*: *Old and new*. New York: Springer.

Wang, Y., Li, W. (2020). Information Newton's flow: Second-order optimization method in probability space. arXiv:2001.04341.