Wasserstein Statistics in 1D Location-Scale Model

Shun-ichi Amari
RIKEN Center for Brain Science

Abstract

Wasserstein geometry and information geometry are two important structures introduced in a manifold of probability distributions. The former is defined by using the transportation cost between two distributions, so it reflects the metric structure of the base manifold on which distributions are defined. Information geometry is constructed based on the invariance criterion that the geometry is invariant under reversible transformations of the base space. Both have their own merits for applications. Statistical inference is constructed on information geometry, where the Fisher metric plays a fundamental role, whereas Wasserstein geometry is useful for applications to computer vision and AI. We propose statistical inference based on the Wasserstein geometry in the case that the base space is 1-dimensional. By using the location-scale model, we derive the $W$-estimator explicitly and studies its asymptotic behaviors.

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 structure of the underlying manifold $X$ on which probability distributions are defined. Information geometry, on the hand, studies an invariant structure such that geometry does not change under transformations of $X$ which would change the distance within $X$. So it is independent of the metric of $X$.

Both geometries have their own histories (see e.g., Villani, 2003, 2009; Amari, 2016). Information geometry has been successful for elucidating statistical inference, where the Fisher
information metric plays a fundamental role. It has successfully been applied, not only to
statistics, but also to machine learning, signal processing, systems theory, physics and many
others (Amari, 2016). Wasserstein geometry has been a useful tool for geometry, where the
Ricci flow has played an important role (Villani, 2009; Li and Montúfar, 2018). Recently, it has
a wide scope of applications in computer vision, deep learning and more (e.g., Fronger et al.,
2015; Arjovsky et al., 2017; Montavon et al., 2015; Peyré et al., 2019). There are some trials
to connect the two geometries. Li and Zhao (2019) gave a unified theory connecting them.
See also Wang and Li (2019) and Amari et al. (2018, 2019).

It is natural to consider statistical inference from the Wasserstein geometry point of view and
compare the results with information-geometrical inference based on the likelihood (Li and Zhao,
2019). The present short article studies the statistical inference based on the Wasserstein ge-
ometry from a different point of view of Li and Zhao (2019). Given a number of independent
observations from a probability distribution belonging to a statistical model with a finite num-
ber of parameters, we define the $W$-estimator that minimizes $W$-divergence from the empirical
distribution $\hat{p}(x)$ derived from observed data to the statistical model. In contrast, the infor-
mation geometry estimator is the one that minimizes Kullback-Leibler divergence from the
empirical distribution to the model, and it is the maximum likelihood estimator.

We use 1D base space $X = R^1$, and define the transportation cost to be equal to the square
of the Euclidean distance between two points in $R^1$. We further focus on the location-scale
model to obtain explicit solutions in the asymptotic resume, that is, the number of observations
is sufficiently large. We then give an explicit expression of the $W$-estimator, proving that it is
asymptotically consistent and further calculate its asymptotic variance. Although they are not
Fisher efficient, it minimizes the divergence between the empirical distribution and the model.
We may say that it is $W$-efficient estimator in this sense.

The present $W$-estimator is different from Li and Zhao (2019), based on the Wasserstein
score function. The $W$-efficiency of this estimator is defined. Although this is a fundamental
theory, opening a new paradigm connecting information geometry and $W$ geometry, it does not
minimizes 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. After introduction, we formulate estimating equations for a general parametric statistical model in the 1D-case. We show in section 2 that the optimal estimator uses only a linear function of observations. We then focus on the location-scale model in section 3. We give an explicit form of the $W$-estimator. We analyze the asymptotic properties of the $W$-estimator. We studies the geometry of the location-scale model in section 4, showing that it is Euclidean (Li and Zhao, 2019), although it is a curved submanifold in the function space of $W$-geometry (Takatsu, 2011). We finally give characteristic features of the $W$-estimator, comparing it with the maximum likelihood estimator.

2 $W$-estimator

We first show the optimal transportation cost sending $p(x)$ to $q(x)$, $x \in \mathbb{R}^1$ when the transportation cost from $x$ to $y$, $x, y \in \mathbb{R}^1$, is $(x - y)^2$. Let $P(x)$ and $Q(x)$ be the cumulative distributions of $p$ and $q$, respectively,

\[
P(x) = \int_{-\infty}^{x} p(u)du, \tag{1}
\]
\[
Q(x) = \int_{-\infty}^{x} q(u)du. \tag{2}
\]

Then, it is known that the optimal transportation plan is to send mass of $p(x)$ at $x$ to $x'$, such that

\[
P^{-1}(x) = Q^{-1}(x'),\tag{3}
\]

$P^{-1}$ and $Q^{-1}$ being the inverse functions of $P$ and $Q$. See Fig. 1. The total cost sending $p$ to $q$ is

\[
C(p, q) = \int_{0}^{1} |P^{-1}(z) - Q^{-1}(z)|^2 dz. \tag{4}
\]

We consider a regular statistical model

\[
S = \{p(x, \theta)\}, \tag{5}
\]

parameterized by a vector parameter $\theta$, where $p(x, \theta)$ is a probability density function of
Figure 1: Optimal transportation plan from $p$ to $q$

random variable $x \in \mathbb{R}^1$ with respect to the Lebesgue measure of $\mathbb{R}^1$. Let

$$D = \{x_1, \cdots, x_n\}$$

be $n$ independently observed data subject to $p(x, \theta)$. We rearrange them in the increasing order,

$$x_1 \leq x_2 \leq \cdots \leq x_n.$$ 

Then, $D$ is composed of order statistics. We denote the empirical distribution by

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

where $\delta$ is the delta function.

The optimal transportation plan from $\hat{p}(x)$ to $p(x, \theta)$ is explicitly solved when $x$ is 1-dimensional, $x \in \mathbb{R}^1$. The optimal plan is to transport a mass at $x$ to $x$ defined by

$$\hat{P}^{-1}(x) = P^{-1}(x', \theta),$$

where $\hat{P}(x)$ and $P(x, \theta)$ are the cumulative distributions of $\hat{p}(x)$ and $p(x, \theta)$, respectively,

$$\hat{P}(x) = \int_{-\infty}^{x} \hat{p}(u)du,$$

$$P(x, \theta) = \int_{-\infty}^{x} p(u, \theta)du,$$

and $\hat{P}^{-1}, P^{-1}$ are their inverse functions. The total cost of transporting $\hat{p}(x)$ to $p(x, \theta)$ optimally is given by

$$C(\theta) = \int_0^1 \left| \hat{P}^{-1}(z) - P^{-1}(z, \theta) \right|^2 dz.$$
Let \( z_1, \ldots, z_n \) be the points of equi-probability partition of \( X \) for distribution \( p(x, \theta) \) such that

\[
\int_{z_{i-1}}^{z_i} p(x, \theta) dx = \frac{1}{n},
\]

(13)

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

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

(14)

and

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

(15)

See Fig. 2.

The optimal transportation cost is rewritten as

\[
C(\theta) = \sum_i \int_{z_{i-1}}^{z_i} (x_i - z)^2 p(z) dz
\]

(16)

\[
= \frac{1}{n} \sum x_i^2 - 2 \sum k_i(\theta)x_i + S(\theta),
\]

(17)
where we use (13) and put

$$k_i(\theta) = \int_{z_{i-1}}^{z_i} z p(z, \theta) dz$$

(18)

$$S(\theta) = \sum \int_{z_{i-1}}^{z_i} z^2 p(z, \theta) dz = \int z^2 p(z, \theta) dz.$$  

(19)

By using the mean and variance of \(p(x, \theta)\),

$$\mu(\theta) = \int z p(z, \theta) dz,$$

(20)

$$\sigma^2(\theta) = \int z^2 p(z, \theta) dz - \mu^2.$$  

(21)

We have

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

(22)

We define the \(W\)-estimator \(\hat{\theta}\) by the minimizer of \(C(\theta)\). Differentiating \(C(\theta)\) with respect to \(\theta\) and putting it equal to 0, we have the estimating equation.

**Theorem 1.** The \(W\)-estimator \(\hat{\theta}\) satisfies

$$\sum \frac{\partial}{\partial \theta} k_i(\theta) x_i = \frac{1}{2} \frac{\partial}{\partial \theta} S.$$  

(23)

It is interesting to see that the estimating equation is linear in \(n\) observations \(x_1, \cdots, x_n\) for any statistical model. This is quite different from the maximum likelihood estimator or Bayes estimator.

We give a rough sketch that the estimator is asymptotically consistent, that is, it converges to the true \(\theta_0\) as \(n\) tends to infinity. 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 \(z_i(\theta_0)\), when the true parameter is \(\theta_0\). From (18), we see that

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

(24)

as \(n \to \infty\), so (23) is written as

$$\frac{1}{2n} \frac{\partial}{\partial \theta} \sum z_i(\theta) z_i(\theta_0) = \frac{1}{2} S(\theta).$$  

(25)

We further remark that, as \(n\) tends to infinity,

$$\frac{1}{n} \sum z_i^2 = \int z^2 p(z, \theta) dz = S(\theta).$$  

(26)
Therefore, $\theta = \theta_0$ is the solution of (23) for $x_i = z_i(\theta_0)$, showing the consistency of the estimator.

3 Location-scale model

Let $f(x)$ be a standard probability density function, satisfying

\begin{align}
\int f(x)dx &= 1, \\
\int xf(x)dx &= 0, \\
\int x^2f(x)dx &= 1,
\end{align}

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

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

where $\theta = (\mu, \sigma)$ is the parameters to specify a distribution.

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

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

where $F$ is the cumulative distribution function

$$F(x) = \int_{-\infty}^{x} f(u)du.$$

We use the following transformation of the location and scale,

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

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

The equi-probability partition points $\bar{x}_i$ of $p(x, \theta)$ is given by

$$\bar{x}_i = \sigma z_i + \mu.$$

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

\begin{align}
C(\mu, \sigma) &= \sum \int_{x_{i-1}}^{\bar{x}_i} (x_i - x)^2 p(x, \mu, \sigma)dx \\
&= \mu^2 + \sigma^2 + \sum x_i^2 - 2 \sum x_i \int (\sigma z + \mu) f(z)dz.
\end{align}
By differentiating (36), we obtain

\[
\frac{1}{2} \frac{\partial}{\partial \mu} C = \mu - \frac{1}{n} \sum x_i, \tag{37}
\]

\[
\frac{1}{2} \frac{\partial}{\partial \sigma} C = \sigma - \sum k_i x_i, \tag{38}
\]

where

\[
k_i = \int_{z_{i-1}}^{z_i} zf(z)dz, \tag{39}
\]

which does not depend on \( \mu \) and \( \sigma \) but 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

\[
\hat{\mu} = \frac{1}{n} \sum x_i, \tag{40}
\]

\[
\hat{\sigma} = \sum k_i x_i. \tag{41}
\]

**Remark** The \( W \)-estimator of the mean is the arithmetic average of observed data irrespective of the form of \( f \). The \( W \)-estimator of variance is also a linear function of observed data \( x_1, \cdots, x_n \), but it depends on \( f \), since \( k_i \) depend on \( f \).

The estimator \( \hat{\mu} \) is consistent, asymptotically subject to the Gaussian distribution \( N \left( \mu, \frac{\sigma^2}{n} \right) \).

We next show the asymptotic consistency of \( \hat{\sigma} \) and its asymptotic variance. Since the probability distribution of the order statistics \( x_1, \cdots, x_n \) is explicitly given in literatures of statistics, it is, in principle, possible to calculate the variance, but we need complicated calculations. So we here give a rough estimate based on speculative ideas.

**Theorem 3.** \( \hat{\sigma} \) is asymptotically consistent with asymptotic variance

\[
V(\hat{\sigma}) = \frac{\sigma^2}{n} \int z^4 f(z) dz, \tag{42}
\]

where \( V[\cdot] \) is the variance.

**Sketch of proof.** We evaluate \( k_i \) when \( n \) is large. When \( n \) is large, \( z_{i-1} \) and \( z_i \) are close and

\[
\Delta z_i = z_i - z_{i-1} \tag{43}
\]
is of order $1/n$. More precisely, from
\[
\int_{z_{i-1}}^{z_i} f(z)dz = \frac{1}{n},
\]
we have
\[
\Delta f(z_i) = \frac{1}{n} + O\left(\frac{1}{n^2}\right).
\]
Hence, from (39), we have
\[
k_i = \frac{1}{n}z_i + O\left(\frac{1}{n^2}\right).
\]
Thus, we have an asymptotic relation
\[
\hat{\sigma} = \frac{\sigma}{n} \sum z_i \hat{z}_i + \frac{\mu}{n} \sum \hat{z}_i,
\]
where
\[
\hat{z}_i = \frac{x_i - \mu}{\sigma}.
\]
We further use the following asymptotic relations
\[
\frac{1}{n} \sum z_i^2 \approx \int z^2 p(z)dz = 1,
\]
\[
\frac{1}{n} \sum z_i \approx \int z p(z)dz = 0.
\]
We finally have
\[
\lim_{n \to \infty} \hat{\sigma} = \sigma,
\]
showing that $\hat{\sigma}$ is asymptotically unbiased.

In order to evaluate the asymptotic variance, we use daring speculation. To this end, we divide the $x$-axis into $n$ intervals $I_1 = [-\infty, z_1], I_2 = [z_1, z_2], \ldots, I_n = [z_{n-1}, \infty]$, the probability of each interval being equal to $1/n$. When we select $n$ points from $f(x)$ independently, each observation $\hat{z}_i$ will fall into one interval randomly. One interval may include multiple or no observations. Let $s_i$ be a random variable to show the number of observations that fall in interval $I_i = [z_{i-1}, z_i]$. Then, each random variable $s_i$ is subject to Poisson distribution with mean and variance equal to 1. They are independent except for the total constraint
\[
\sum s_i = n.
\]
The observed order statistic \(\hat{z}_i\) will fall in interval \(I_i = [z_{i-1}, z_i]\) most probably and takes value close to \(z_i\). It may fall in other nearby intervals.

When \(\hat{z}\), one of \(\hat{z}'s\), falls in \(I_i\), its value is written as

\[
\hat{z} = z_i - \varepsilon_i, \quad (53)
\]

where \(\varepsilon_i\)

\[
0 \leq \varepsilon_i \leq z_i - z_{i-1}, \quad (54)
\]

is deviation within \(I_i\). It is a random variable of order \(1/n\).

Let us denote the interval \(i'\) in which \(\hat{z}_i\) falls. Since \(i\) and \(i'\) are close,

\[
|z_i - z_{i'}| = O\left(\frac{1}{n}\right), \quad (55)
\]

with high probability, we can rewrite (47) as

\[
\hat{\sigma} = \frac{\sigma}{n} \sum_i s_i z_i \hat{z}_i + O\left(\frac{1}{n}\right) \quad (56)
\]

by neglecting high-order terms, where summation with respect to \(i\) is replaced by summation with respect to the intervals \(I_{i'}\) with weight \(s_{i'}\). When \(s_i = 0\), interval \(I_i\) includes no observation. When \(s_i > 1\), \(I_i\) includes multiple observations.

We calculate the variance of (41) as

\[
V[\hat{\sigma}] = V\left[\frac{\sigma}{n} \sum_i s_i z_i^2 \right] + O\left(\frac{1}{n^2}\right). \quad (57)
\]

We further note that \(s_i\) are asymptotically independent. Hence, we have

\[
V[\hat{\sigma}] \approx \frac{\sigma^2}{n^2} \sum V[s_i] z_i^4 \quad (58)
\]

\[
\approx \frac{\sigma^2}{n} \int z^4 f(z) dz, \quad (59)
\]

proving the theorem.

It is easy to see from (40) and (41) that \(\hat{\mu}\) and \(\hat{\sigma}\) are asymptotically non-correlated, since \(x_i\)'s are independent.

When \(f\) is Gaussian

\[
f(z) = \frac{1}{\sqrt{2\pi}} \exp \left\{-\frac{1}{2} z^2 \right\}, \quad (60)
\]
the asymptotic variance is
\[ V[\hat{\sigma}] = \frac{3}{n}\sigma^2. \]  
(61)

Hence, it is consistent but not efficient.

When \( f \) is uniform,
\[ f(z) = \begin{cases} \frac{1}{2\sqrt{3}}, & |z| \leq \sqrt{3}, \\ 0, & \text{otherwise}, \end{cases} \]  
(62)
the asymptotic variance is
\[ V[\hat{\sigma}] = \frac{9}{5n}\sigma^2. \]  
(63)

However, the Fisher information divergence to infinity for the uniform distribution and the maximum likelihood estimator \( \hat{\sigma} \) converges to 0 exponentially fast.

In general, the \( W \)-estimator is not sensitive to changes of the waveform \( f \), whereas the maximum likelihood estimator is sensitive.

## 4 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 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 \( \square \).

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
\[ (P + \delta P)^{-1}(z) = P^{-1}(z) - \frac{\delta P\{x(z)\}}{P\{x(z)\}}, \]  
(64)
where
\[ x(z) = P^{-1}(z). \]  
(65)

This equation is derived from
\[ \frac{d}{dz} F^{-1}(z) = \frac{1}{f'\{x(z)\}}. \]  
(66)
which we have from the differentiation of the identity

\[ F^{-1} \{ F(x) \} = x. \]  \hfill (67)

We thus have

\[ C(p, p + \delta p) = \int_{-\infty}^{\infty} \frac{1}{p(x)} \left( \int_{-\infty}^{x} \delta p(y) dy \right)^2 dx \]  \hfill (68)

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 \( M \). We have for the location-scale model \( (69) \),

\[ \delta p(y) = \frac{\partial}{\partial \mu} p(y, \theta) d\mu + \frac{\partial}{\partial \sigma} p(y, \theta) d\sigma. \]  \hfill (69)

The Riemannian metric tensor \( G = (g_{ij}) \) is derived from

\[ C(p, p + \delta p) = \sum g_{ij}(\theta) d\theta_i d\theta_j. \]  \hfill (70)

See also Li and Zhao (2019).

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

\[ g_{ij} = \delta_{ij}. \]  \hfill (71)

**Proof.** We need to calculate \( (68) \). Technical details are given in Appendix. \hfill \( \square \)

It is surprising that \( G = (g_{ij}) \) is the identity matrix for the location-scale model, so that \( S \) 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 \( \mathbb{R}^2 \).

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

\[ C(\hat{p}, \hat{p}_\theta') = C(\hat{p}, \hat{p}_\theta) + C(\hat{p}_\theta, \hat{p}_\theta') \]  \hfill (72)

and the difference of the cost between the two estimators is

\[ C(\hat{p}_\theta, \hat{p}_\theta') = \| \theta' - \hat{\theta}' \|^2. \]  \hfill (73)
Li and Zhao (2019) studies the properties of the W estimator given by the W score function. They give the W-efficiency and W Cramer-Rao inequality. However, their W-estimator does not minimize the transportation cost. It is interesting to study the relation between the two W-estimators.

5 Conclusions

We studied the behaviors of the W-estimator minimizing the transportation cost from the observed empirical distribution to the underlying statistical model on $\mathbb{R}^1$. It is a consistent estimator having a simple form of the estimating equation. We focused on the location-scale model and showed that the estimator can be represented by a simple linear form of observations. Its asymptotic variance was calculated. Although its error variance is worse than the maximum likelihood estimator, it is simple, and further it is the estimator that minimizes the transportation cost from the observed sample to the model.

We need to study further its merits and demerits. We hope to find good applications to computer vision and AI. It is an interesting problem to compare the W-estimator of Li and Zhao (2019) which uses the W score function with the minimum cost W-estimator.

References

Amari, S., Information Geometry and Its Applications. Springer (2016).

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

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

Arjovsky, M., Chintala, S., Bottou, L., Wasserstein GAN. arXiv:1701.07875, (2017).
Appendix: The Riemannian metric of the location scale model

We have
\[
\delta p(x, \theta) = -\frac{1}{\sigma^2} f' \left( \frac{x - \mu}{\sigma} \right) d\mu - \frac{1}{\sigma^3} \left\{ \sigma f \left( \frac{x - \mu}{\sigma} \right) + (x - \mu) f' \left( \frac{x - \mu}{\sigma} \right) \right\} d\sigma. \tag{74}
\]

By integration, we have
\[
\int_{-\infty}^{x} \delta p(y, \theta) dy = -p(x, \theta)d\mu - (x - \mu)p(x, \theta)d\sigma. \tag{75}
\]

Hence, we have
\[
C(\theta, \theta + d\theta) = d\mu^2 + d\sigma^2. \tag{76}
\]