Information Geometry of Wasserstein Statistics on Shapes and Affine Deformations

Shun-ichi Amari
Teikyo University, Advanced Comprehensive Research Organization,
RIKEN Center for Brain Science

Takeru Matsuda
The University of Tokyo,
RIKEN Center for Brain Science

Abstract

Information geometry and Wasserstein geometry are two main structures introduced in a manifold of probability distributions, and they capture its different characteristics. We study characteristics of Wasserstein geometry in the framework of [29] for the affine deformation statistical model, which is a multi-dimensional generalization of the location-scale model. We compare merits and demerits of estimators based on information geometry and Wasserstein geometry. The shape of a probability distribution and its affine deformation are separated in the Wasserstein geometry, showing its robustness against the waveform perturbation in exchange for the loss in Fisher efficiency. We show that the Wasserstein estimator is the moment estimator in the case of the elliptically symmetric affine deformation model. It coincides with the information-geometrical estimator (maximum-likelihood estimator) when and only when the waveform is Gaussian. The role of the Wasserstein efficiency is elucidated in terms of robustness against waveform change.

1 Introduction

We study statistics based on probability distribution patterns $p(\boldsymbol{x})$ over $\boldsymbol{x} \in X = \boldsymbol{R}^d$, by using both information geometry [see 1, 8, etc] and Wasserstein geometry [see 47, 40, 42, among many others]. Here, $p(\boldsymbol{x})$ is a probability distribution on $X = \boldsymbol{R}^d$. When $d = 2$, $p(\boldsymbol{x})$ is regarded as a visual pattern on $\boldsymbol{R}^2$.

There are lots of applications of Wasserstein geometry to statistics [see, e.g., 4, 11, 51, 9, 32, 21, 28, 16, and others], machine learning [see, e.g., 7, 18, 49, 40, 35, among many others] and statistical physics [22]. We also recommend the following review paper and book, [38, 39], which include lots of references. However, applications to statistical inference look still premature.

The affine deformation statistical model $p(\boldsymbol{x}, \boldsymbol{\theta})$ is defined as

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = |\Lambda| f\left(\Lambda(\boldsymbol{x} - \boldsymbol{\mu})\right), \tag{1}$$

where $f(\boldsymbol{z})$ is a standard shape distribution satisfying

$$\int f(\boldsymbol{z})d\boldsymbol{z} = 1, \tag{2}$$

$$\int \boldsymbol{z} f(\boldsymbol{z})d\boldsymbol{z} = 0, \tag{3}$$

$$\int \boldsymbol{z}\boldsymbol{z}^\top f(\boldsymbol{z})d\boldsymbol{z} = I, \tag{4}$$

where I is the identity matrix. We also refer to the standard shape f as a "waveform" in the following. The deformation parameter consists of $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Lambda) \in \Theta$ such that $\boldsymbol{\mu}$ is a vector specifying translation of the location and Λ is a non-singular matrix representing scale changes and rotations of $\boldsymbol{x}$. Given a standard f, we have a statistical model parameterized by $\boldsymbol{\theta}$: $\mathcal{M}_f = \{p(\boldsymbol{x}, \boldsymbol{\theta})\}$. Geometrically, it forms a finite-dimensional statistical manifold, where $\boldsymbol{\theta}$ plays the role of a coordinate system. The deformation model is a generalization of the location-scale model. Note that this model is often called the location-scatter model in several fields such as statistics and signal processing [45, 36].

Let $T_{\boldsymbol{\theta}}$ denote the affine deformation from $\boldsymbol{x}$ to $\boldsymbol{z}$ given by

$$\boldsymbol{z} = T_{\boldsymbol{\theta}}\boldsymbol{x} = \Lambda(\boldsymbol{x} - \boldsymbol{\mu}).$$

This may be regarded as deformation of shape f to $\tilde{T}_{\boldsymbol{\theta}} f$,

$$(\tilde{T}_{\boldsymbol{\theta}} f)(\boldsymbol{x}) = f\left(T_{\boldsymbol{\theta}}\boldsymbol{x}\right), \tag{5}$$

where $\tilde{T}_{\boldsymbol{\theta}}$ is an operator to change a standard waveform f to another waveform $\tilde{f} = \tilde{T}_{\boldsymbol{\theta}} f$ defined by (5).

Let $\mathcal{F} = \{p(\boldsymbol{x})\}$ be the space of all smooth positive probability density functions that have finite second moments. Let $\mathcal{F}_S = \{f(\boldsymbol{z})\}$ be its subspace consisting of all the standard distributions $f(\boldsymbol{z})$ satisfying (2), (3) and (4). Then, any $q(\boldsymbol{x}) \in \mathcal{F}$ is written in the form

$$q(\boldsymbol{x}) = |\Lambda| f\left(\Lambda(\boldsymbol{x} - \boldsymbol{\mu})\right)$$

for $f \in \mathcal{F}_S$ and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Lambda) \in \Theta$. Note that $\boldsymbol{\theta}$ is not necessarily unique due to possible symmetries in f. Hence, $\mathcal{F} = \mathcal{F}_S \times \Theta / \sim$, where $\sim$ is the equivalence relation of equality in distribution. See Figure 1.

Geometry of a manifold of probability distributions has so far been studied by information geometry and Wasserstein geometry. The two geometries capture different aspects of a manifold of probability distributions. We use a divergence measure to explain this. Let $D_F[p(\boldsymbol{x}), q(\boldsymbol{x})]$ and $D_W[p(\boldsymbol{x}), q(\boldsymbol{x})]$ be two divergence measures between distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, where subscripts F and W represent Fisher-based information geometry and Wasserstein geometry, respectively. Information geometry uses an invariant divergence D_F, typically the Kullback–Leibler divergence. Wasserstein divergence D_W is defined by the cost of transporting masses distributed in form $p(\boldsymbol{x})$ to another $q(\boldsymbol{x})$. Roughly speaking, D_F measures the vertical differences of $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, for example, represented by their log-ratio $\log(p(\boldsymbol{x})/q(\boldsymbol{x}))$, whereas D_W measures the horizontal differences of $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ which corresponds to the transportation cost from $p(\boldsymbol{x})$ to $q(\boldsymbol{x})$. See Figure 2.

Information geometry is constructed based on the invariance principle of Chentsov [15] such that $D_F[p(\boldsymbol{x}), q(\boldsymbol{x})]$ is invariant under invertible transformations of the coordinates $\boldsymbol{x}$ of the

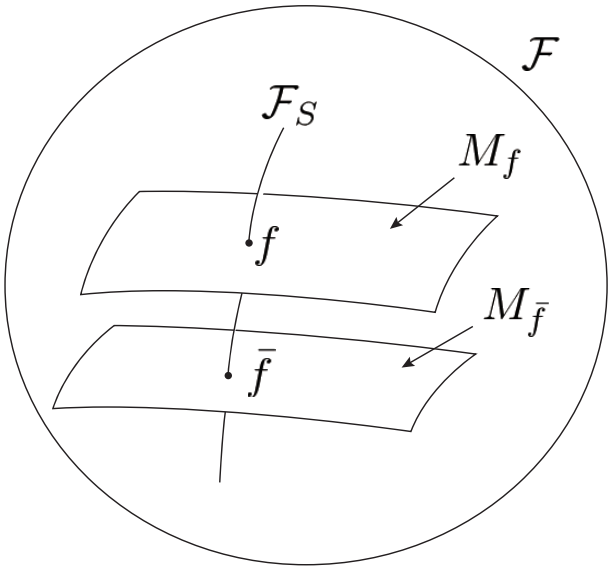

Figure 1: Decomposition of $\mathcal{F}$

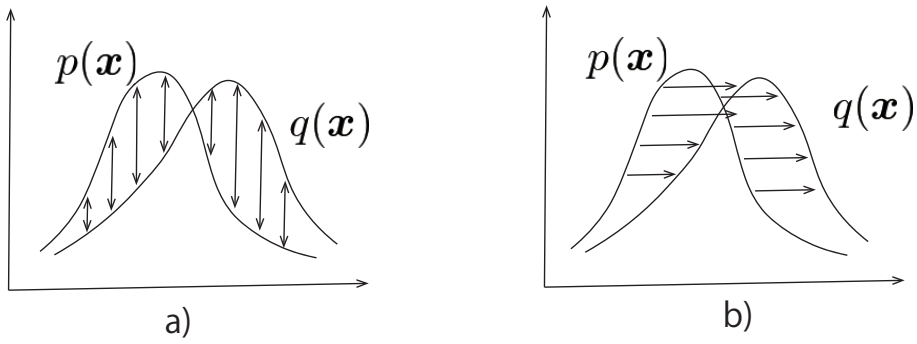

Figure 2: (a) F-divergence. (b) W-divergence.

sample space X. This implies that the divergence does not depend on the coordinate system of X. We then have a unique Riemannian metric, which is Fisher–Rao metric, and also a dual pair of affine connections [5]. This is useful not only for analyzing the performances of statistical inference but also for vision analysis, machine learning, statistical physics, and many others [see 1].

Wasserstein geometry has an old origin, proposed by G. Monge in 1781 as a problem of transporting mass distributed in the form $p(\boldsymbol{x})$ to another $q(\boldsymbol{x})$ such that the total transportation cost is minimized. It depends on the transportation cost $c(\boldsymbol{x}, \boldsymbol{y})$ between two locations $\boldsymbol{x}, \boldsymbol{y} \in X$. The cost is usually a function of the Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. We use the square of the distance as a cost function, which gives L^2-Wasserstein geometry. This Wasserstein geometry directly depends on the Euclidean distance of $X = \boldsymbol{R}^d$. Therefore, it is useful for problems that intrinsically depend on the metric structure of X, such as the transportation problem, non-equilibrium statistical physics, pattern analysis, machine learning and many others.

It is natural to search for the relation between the two geometries. There are a number of such trials, including Amari et al. [2, 3], Khan and Zhang [24], Rankin and Wong [41], Ito [22], Wong and Yang [50], Chizat et al. [17], Kondratyev et al. [26], Liero et al. [30] and others. (See Khan and Zhang [23] for a survey.) Among them, Li and Zhao [29] gave a unified framework for the two geometries. The present article is based on their framework and focuses on the affine deformation model, for which the standard waveform f and the deformation parameter $\boldsymbol{\theta}$ are separated.

Recently, Li and Zhao [29] introduced the Wasserstein score function in parallel to the Fisher score function, defining two estimators $\hat{\boldsymbol{\theta}}_F$ and $\hat{\boldsymbol{\theta}}_W$ thereby. The former $\hat{\boldsymbol{\theta}}_F$ is the maximum likelihood estimator that maximizes the log likelihood. This is the one that minimizes an invariant divergence from the empirical distribution $\hat{p}(\boldsymbol{x})$ to a parametric model, where the empirical distribution is given based on n independent observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ as

$$\hat{p}(\boldsymbol{x}) = \frac{1}{n} \sum_{t=1}^{n} \delta\left(\boldsymbol{x} - \boldsymbol{x}_t\right),$$

where $\delta(\boldsymbol{x})$ is the delta function. The latter Wasserstein estimator $\hat{\boldsymbol{\theta}}_W$ is defined as the zero of the Wasserstein score. It is asymptotically equivalent to the minimizer of the W-divergence between the empirical distribution and model. Also, Li and Zhao [29] further defined the F-efficiency and W-efficiency of an estimator $\hat{\boldsymbol{\theta}}$ given a statistical model $\mathcal{M} = \{p(\boldsymbol{x}, \boldsymbol{\theta})\}$, proving the Cramér-Rao type inequalities.

The present paper is organized as follows. In section 2, we introduce two divergences between distributions, one based on the invariance principle and the other based on the transportation cost. The divergences give two Riemannian structures in the space $\mathcal{F}$ of probability distributions $p(\boldsymbol{x})$ over $X = \boldsymbol{R}^d$. A regular statistical model $\mathcal{M} = \{p(\boldsymbol{x}, \boldsymbol{\theta})\}$ parameterized by $\boldsymbol{\theta}$ is a finite-dimensional submanifold embedded in $\mathcal{F}$. In section 3, we define the F- and W-score functions following [29]. The Riemannian structure of the tangent space of probability distributions is pulled-back to the model submanifold, giving both the Riemannian metrics and score functions. We define the F- and W-estimators $\hat{\boldsymbol{\theta}}_F$ and $\hat{\boldsymbol{\theta}}_W$ by using the F- and W-score functions, respectively. Section 4 defines the affine deformation statistical model. Section 5 studies the elliptically symmetric affine deformation model $\mathcal{M}_f$, where f is a spherically symmetric standard form. For this model, we show that the W-score functions are quadratic functions of $\boldsymbol{x}$. Hence, it is proved that $\hat{\boldsymbol{\theta}}_W$ is a moment estimator. We also show

that $\mathcal{M}_f$ and $\mathcal{F}_S$ are orthogonal in the W-geometry, implying the separation of the waveform and deformation. In Section 6, we elucidate the role of W-efficiency from the point of view of robustness to a change in the waveform f due to observation noise. In Section 7, we prove that the Gaussian shape is a unique model in which F-estimator and W-estimator coincide. Section 8 briefly summarizes the paper and mentions future work.

2 Riemannian structures in the space of probability densities

We consider the space $\mathcal{F} = \{p(\boldsymbol{x})\}$ of all smooth positive probability density functions on $\boldsymbol{R}^d$ that have finite second moments. Later, we may relax the conditions of positivity and smoothness, when we discuss a parametric model, in particular the deformation model[1]. We define a divergence function $D[p(\boldsymbol{x}), q(\boldsymbol{x})]$, which represents the degree of difference between $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$. The square of the L^2 distance between $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ plays this role, but a divergence does not necessarily need to be symmetric with respect to $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$. A divergence function satisfies the following conditions:

1. $D[p(\boldsymbol{x}), q(\boldsymbol{x})] \geq 0$ and the equality holds if and only if $p(\boldsymbol{x}) = q(\boldsymbol{x})$.
2. Let $\delta p(\boldsymbol{x})$ be an infinitesimally small deviation of $p(\boldsymbol{x})$. Then, $D[p(\boldsymbol{x}), p(\boldsymbol{x}) + \delta p(\boldsymbol{x})]$ is approximated by a positive quadratic functional of $\delta p(\boldsymbol{x})$.

A divergence is said to be invariant if

$$D[p(\boldsymbol{x}), q(\boldsymbol{x})] = D[\tilde{p}(\boldsymbol{y}), \tilde{q}(\boldsymbol{y})]$$

holds for every smooth reversible transformation $\boldsymbol{k}$ of the coordinates from $\boldsymbol{x} \in \boldsymbol{R}^d$ to $\boldsymbol{y} = \boldsymbol{k}(\boldsymbol{x})$, where

$$\tilde{p}(\boldsymbol{y}) = \left| \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}} \right| p(\boldsymbol{x}).$$

A typical invariant divergence is the α-divergence ($\alpha \neq \pm 1$) defined by

$$D_\alpha[p(\boldsymbol{x}), q(\boldsymbol{x})] \;=\; \frac{4}{1-\alpha^2} \left(1 - \int p(\boldsymbol{x})^{(1+\alpha)/2} q(\boldsymbol{x})^{(1-\alpha)/2} \mathrm{d}\boldsymbol{x} \right)$$

for $\alpha \neq \pm 1$. For $\alpha = 1$, we define $D_1[p, q]$ by the Kullback–Leibler divergence

$$D_1[p(\boldsymbol{x}), q(\boldsymbol{x})] = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}.$$

For $\alpha = -1$, we define $D_{-1}[p(\boldsymbol{x}), q(\boldsymbol{x})] = D_1[q(\boldsymbol{x}), p(\boldsymbol{x})]$. The case $\alpha = 0$ is equivalent to the Hellinger divergence

$$H^2[p(\boldsymbol{x}), q(\boldsymbol{x})] = \frac{1}{2} \int \left(\sqrt{p(\boldsymbol{x})} - \sqrt{q(\boldsymbol{x})} \right)^2 \mathrm{d}\boldsymbol{x}.$$

A characterization of the α-divergence is given in [1]. The α-divergence gives information-geometric structure to $\mathcal{F}$.

[1]For example, the probability density of the uniform distribution inside a ellipse takes zero outside of the ellipse and thus non-smooth on the boundary.

Another divergence is the Wasserstein divergence. Let us transport masses piled in the form $p(\boldsymbol{x})$ to another $q(\boldsymbol{x})$. To this end, we need to move some mass at $\boldsymbol{x}$ to another position $\boldsymbol{y}$. Let $\pi(\boldsymbol{x}, \boldsymbol{y})$ be a coupling, showing the probability of mass at $\boldsymbol{x}$ to be transported to $\boldsymbol{y}$. We call π a transportation plan when it satisfies the following terminal conditions

$$\int \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} = p(\boldsymbol{x}), \tag{6}$$

$$\int \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} = q(\boldsymbol{y}). \tag{7}$$

Let $c(\boldsymbol{x}, \boldsymbol{y})$ be the cost of transporting a unit of mass from $\boldsymbol{x}$ to $\boldsymbol{y}$. Then, the Wasserstein divergence $D_W[p(\boldsymbol{x}), q(\boldsymbol{x})]$ is the minimum transporting cost from $p(\boldsymbol{x})$ to $q(\boldsymbol{x})$. By using stochastic plan $\pi(\boldsymbol{x}, \boldsymbol{y})$, the Wasserstein divergence between $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ is given by

$$D_W\left[p(\boldsymbol{x}), q(\boldsymbol{x})\right] = \inf_{\pi} \int c(\boldsymbol{x}, \boldsymbol{y}) \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y},$$

where infimum is taken over all stochastic plans π satisfying (6) and (7). When the cost is the square of the Euclidean distance

$$c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2,$$

we call D_W the L^2-Wasserstein divergence. We focus on this divergence in the following. Note that the L^2-Wasserstein divergence is the square of the L^2-Wasserstein distance. From Brenier's theorem, the optimal transport is actually induced by a transport map. In other words, for each point $\boldsymbol{x}$, $\pi(\boldsymbol{x}, \cdot)$ is supported at a single point.

The dynamic formulation of the optimal transport problem proposed by [14] and developed further by [10, 33] is useful. Let $\rho(\boldsymbol{x}, t)$ be a family of probability distributions parameterized by t. It represents the time course $\rho(\boldsymbol{x}, t)$ of transporting $p(\boldsymbol{x})$ to $q(\boldsymbol{x})$, satisfying

$$\rho(\boldsymbol{x}, 0) = p(\boldsymbol{x}), \quad \rho(\boldsymbol{x}, 1) = q(\boldsymbol{x}).$$

We introduce potential $\Phi(\boldsymbol{x}, t)$ such that its gradient $\nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x}, t)$ represents the velocity

$$\boldsymbol{v}(\boldsymbol{x}, t) = \nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x}, t)$$

of mass flow at $\boldsymbol{x}$ and t in the dynamic plan. Then, $\Phi(\boldsymbol{x}, t)$ satisfies the following continuity equation

$$\partial_t \rho(\boldsymbol{x}, t) + \nabla_{\boldsymbol{x}} \cdot \{\rho(\boldsymbol{x}, t) \nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x}, t)\} = 0. \tag{8}$$

The Wasserstein divergence is written in the dynamic formulation as

$$D_W\left[p(\boldsymbol{x}), q(\boldsymbol{x})\right] = \inf_{\Phi} \int_0^1 \int \|\nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x}, t)\|^2 \rho(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x} \mathrm{d}t.$$

We introduce a Riemannian structure to $\mathcal{F}$ by the Taylor expansion of $D[p, p + \delta p]$. The Riemannian metric g gives the squared magnitude ds^2 of an infinitesimal deviation $\delta p(\boldsymbol{x})$ in the tangent space of $\mathcal{F}$, for example, by

$$ds^2 = \langle \delta p(\boldsymbol{x}), g(\delta p(\boldsymbol{x})) \rangle,$$

where

$$\langle a(\boldsymbol{x}), b(\boldsymbol{x}) \rangle = \int a(\boldsymbol{x}) b(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

In the case of the invariant divergence,

$$g_F(\delta p(\boldsymbol{x})) = \frac{\delta p(\boldsymbol{x})}{p(\boldsymbol{x})}.$$

so that

$$ds^2 = \int \frac{\delta p(\boldsymbol{x})^2}{p(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}.$$

In the case of the L^2-Wasserstein divergence, consider the change of density from $\rho(\boldsymbol{x}, 0) = p(\boldsymbol{x})$ at $t = 0$ to $\rho(\boldsymbol{x}, \mathrm{d}t) = p(\boldsymbol{x}) + \delta p(\boldsymbol{x})$ at $t = \mathrm{d}t$ for an infinitesimal $\mathrm{d}t$. By using the potential $\Phi(\boldsymbol{x})$ of this infinitesimal transport,

$$\delta p(\boldsymbol{x}) = -\Delta_p \Phi(\boldsymbol{x}) \mathrm{d}t + o(\mathrm{d}t),$$

where Δ_p is the operator defined by

$$\Delta_p \Phi(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \cdot (p(\boldsymbol{x}) \nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x})) .$$

Then, the L^2-Wasserstein divergence is

$$\begin{aligned} D_W[p(\boldsymbol{x}), p(\boldsymbol{x}) + \delta p(\boldsymbol{x})] &= \int \|\nabla_{\boldsymbol{x}} \Phi(\boldsymbol{x}) \mathrm{d}t\|^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + o(\mathrm{d}t^2) \\ &= \langle \Phi(\boldsymbol{x}) \mathrm{d}t, -\Delta_p \Phi(\boldsymbol{x}) \mathrm{d}t \rangle + o(\mathrm{d}t^2) \\ &= \langle \Phi(\boldsymbol{x}) \mathrm{d}t, \delta p(\boldsymbol{x}) \rangle + o(\mathrm{d}t^2), \end{aligned}$$

where we used

$$\int (\nabla_{\boldsymbol{x}} a(\boldsymbol{x}))^\top (\nabla_{\boldsymbol{x}} b(\boldsymbol{x})) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = -\int a(\boldsymbol{x}) \Delta_p b(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Thus,

$$g_W(\delta p(\boldsymbol{x})) = \Phi(\boldsymbol{x}) \mathrm{d}t.$$

Note that $\Phi(\boldsymbol{x})$ is unique up to an additive constant under regularity conditions (see Theorem 13.8 of [48] and Section 8.1 of [47]). This is Otto's Riemannian metric [37].

We focused on the space $\mathcal{F}$ of smooth and positive densities under L^2-Wasserstein geometry. It is indeed an infinite-dimensional Riemannian manifold [31]. Since the space $\mathcal{F}_S$ is a co-dimension $d + d(d+1)/2$ subspace of $\mathcal{F}$ that specifies the value of linear functionals (first and second moments), it is also an infinite-dimensional Riemannian manifold. However, if we consider the larger space of general probability distributions, then it is not even a Banach manifold under L^2-Wasserstein geometry, because it includes singular (atomic) distributions for which the tangent space is more restricted than that of $\mathcal{F}$.

Note that information geometry and Wasserstein geometry induce different topologies on the space of probability distributions. In information geometry, due to the asymmetry of the divergence, the topology may be different depending on whether we consider open balls with respect to the first or second arguments of divergence functions. Also, the Kullback–Leiber divergence may be infinite even if the distributions have finite second moments and strictly positive densities. The Hellinger distance is symmetric and may be better behaved, although the distributions have to belong to some Orlicz spaces for its finiteness [25]. On the other hand, Wasserstein geometry does not require the absolute continuity of the distributions, because the Wasserstein distance metrizes the weak convergence. Therefore, the identity map between smooth and positive probability densities in terms of the Hellinger (or Kullback–Leibler divergence etc.) and Wasserstein topologies is not a diffeomorphism. Therefore, there are multiple infinite-dimensional manifold structures possible on $\mathcal{F}$.

3 Score functions and estimators

We consider a regular statistical model $\mathcal{M} = \{p(\boldsymbol{x}, \boldsymbol{\theta})\}$ parameterized by m-dimensional vector $\boldsymbol{\theta}$. The tangent space of $\mathcal{M}$ at $\boldsymbol{\theta}$ is spanned by

$$\partial_i p(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\partial}{\partial \theta^i} p(\boldsymbol{x}, \boldsymbol{\theta})$$

for $i = 1, \cdots, m$ so that a tangent vector $\delta p(\boldsymbol{x})$ is given by

$$\delta p(\boldsymbol{x}) = \partial_i p(\boldsymbol{x}, \boldsymbol{\theta}) d\theta^i. \tag{9}$$

Hereafter, the summation convention is used, that is, all indices appearing twice, once as upper and the other as lower indices, e.g. i's in (9), are summed up.

Let us define $S_i(\boldsymbol{x}, \boldsymbol{\theta})$ from the basis functions $\partial_i p(\boldsymbol{x}, \boldsymbol{\theta})$ of the tangent space of $\mathcal{M}$ for $i = 1, \cdots, m$ by applying the Riemannian metric operator g:

$$S_i(\boldsymbol{x}, \boldsymbol{\theta}) = g(\partial_i p(\boldsymbol{x}, \boldsymbol{\theta})).$$

We call them score functions following the tradition of statistics. In the case of invariant Fisher geometry, the Fisher score function $S_i^F(\boldsymbol{x}, \boldsymbol{\theta})$ is

$$S_i^F(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\partial_i p(\boldsymbol{x}, \boldsymbol{\theta})}{p(\boldsymbol{x}, \boldsymbol{\theta})} = \partial_i l(\boldsymbol{x}, \boldsymbol{\theta}), \quad l(\boldsymbol{x}, \boldsymbol{\theta}) = \log p(\boldsymbol{x}, \boldsymbol{\theta}),$$

which is the derivative of log-likelihood. In Wasserstein geometry, the Wasserstein score (W-score) function $S_i^W(\boldsymbol{x}, \boldsymbol{\theta})$ [29] is defined as the solution of

$$\Delta_p S_i^W(\boldsymbol{x}, \boldsymbol{\theta}) = -\partial_i p(\boldsymbol{x}, \boldsymbol{\theta}), \quad \mathrm{E}_\theta[S_i^W(\boldsymbol{x}, \boldsymbol{\theta})] = 0,$$

where the latter condition is imposed to eliminate the indefiniteness due to the integral constant. By using the identity

$$\int a(\boldsymbol{x}) \Delta_p b(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = -\int (\nabla_{\boldsymbol{x}} a(\boldsymbol{x}))^\top (\nabla_{\boldsymbol{x}} b(\boldsymbol{x})) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{10}$$

we see that $S_i^W(\boldsymbol{x}, \boldsymbol{\theta})$ satisfies the Poisson equation:

$$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}, \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{x}} S_i^W(\boldsymbol{x}, \boldsymbol{\theta}) + \Delta_{\boldsymbol{x}} S_i^W(\boldsymbol{x}, \boldsymbol{\theta}) + \frac{\partial}{\partial \theta_i} \log p(\boldsymbol{x}, \boldsymbol{\theta}) = 0. \tag{11}$$

For infinitesimal δ, the map $\boldsymbol{x} \mapsto \boldsymbol{x} + \delta \nabla_{\boldsymbol{x}} S_i^W(\boldsymbol{x}, \boldsymbol{\theta})$ is the optimal transport map from $p(\boldsymbol{x}, \boldsymbol{\theta})$ to $p(\boldsymbol{x}, \boldsymbol{\theta} + \delta \boldsymbol{e}_i)$ with transportation cost

$$D_W(p(\boldsymbol{x}, \boldsymbol{\theta}), p(\boldsymbol{x}, \boldsymbol{\theta} + \delta \boldsymbol{e}_i)) = \int \|\delta \nabla_{\boldsymbol{x}} S_i^W(\boldsymbol{x}, \boldsymbol{\theta})\|^2 p(\boldsymbol{x}, \boldsymbol{\theta}) \mathrm{d}x + o(\delta^2) \tag{12}$$

as $\delta \to 0$, where $\boldsymbol{e}_i$ is the i-th standard unit vector. See Proposition 8.4.6 of [6] for more rigorous statement. In both Fisher and Wasserstein cases, the score function satisfies

$$\mathrm{E}_\theta[S_i(\boldsymbol{x}, \boldsymbol{\theta})] = 0. \tag{13}$$

The Riemannian metric tensor $g_{ij}(\boldsymbol{\theta})$ is pulled-back from g in $\mathcal{F}$ to $\mathcal{M}$, and is derived in terms of the score functions as

$$g_{ij}(\boldsymbol{\theta}) = \langle S_i, g^{-1} S_j \rangle,$$

where

$$\langle a(\boldsymbol{x}), b(\boldsymbol{x}) \rangle = \int a(\boldsymbol{x}) b(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

In the Fisherian case,

$$g^F_{ij}(\boldsymbol{\theta}) = \mathrm{E}\left[\partial_i l(\boldsymbol{x}, \boldsymbol{\theta}) \partial_j l(\boldsymbol{x}, \boldsymbol{\theta})\right] = \int p(\boldsymbol{x}, \boldsymbol{\theta}) \partial_i l(\boldsymbol{x}, \boldsymbol{\theta}) \partial_j l(\boldsymbol{x}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x}.$$

In the Wasserstein case,

$$\begin{aligned} g^W_{ij}(\boldsymbol{\theta}) &= -\int S^W_i(\boldsymbol{x}, \boldsymbol{\theta}) \Delta_p S^W_j(\boldsymbol{x}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} \\ &= \int p(\boldsymbol{x}, \boldsymbol{\theta}) \nabla_{\boldsymbol{x}} S^W_i(\boldsymbol{x}, \boldsymbol{\theta})^\top \nabla_{\boldsymbol{x}} S^W_j(\boldsymbol{x}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} \\ &= \mathrm{E}[\nabla_{\boldsymbol{x}} S^W_i(\boldsymbol{x}, \boldsymbol{\theta})^\top \nabla_{\boldsymbol{x}} S^W_j(\boldsymbol{x}, \boldsymbol{\theta})], \end{aligned} \tag{14}$$

where identity (10) is used.

The score functions $S_i(\boldsymbol{x}, \boldsymbol{\theta})$ give a set of estimating functions from (13), which are used to obtain an estimator $\hat{\boldsymbol{\theta}}$. Suppose that we have n independent observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ from $p(\boldsymbol{x}, \boldsymbol{\theta})$. Let $\hat{p}_{\mathrm{emp}}(\boldsymbol{x})$ be the empirical distribution given by

$$\hat{p}_{\mathrm{emp}}(\boldsymbol{x}) = \frac{1}{n} \sum_{t=1}^{n} \delta\left(\boldsymbol{x} - \boldsymbol{x}_t\right).$$

Then, replacing expectation E in (13) by the expectation with respect to the empirical distribution, we have estimating equations

$$\mathrm{E}_{\mathrm{emp}}[S_i(\boldsymbol{x}, \hat{\boldsymbol{\theta}})] = \frac{1}{n} \sum_{t=1}^{n} S_i(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}) = 0, \quad i = 1, \cdots, m. \tag{15}$$

It is known that the solution $\hat{\boldsymbol{\theta}}$ gives a consistent estimator for large n. Roughly speaking, $\hat{\boldsymbol{\theta}}$ is the projection of $\hat{p}_{\mathrm{emp}}(\boldsymbol{x})$ to the model $\mathcal{M}$ with respect to the metric g (see Figure 3). It is the solution of

$$\langle \hat{p}_{\mathrm{emp}}(\boldsymbol{x}), S_i(\boldsymbol{x}, \boldsymbol{\theta}) \rangle = 0,$$

giving a consistent estimator $\hat{\boldsymbol{\theta}}$. A consistent estimator is Fisher efficient when the projection is orthogonal with respect to the Fisher–Rao metric [5].

For the invariant Fisherian case, the estimator $\hat{\boldsymbol{\theta}}_F$ defined by (15) is the maximum likelihood estimator:

$$\frac{1}{n} \sum_{t=1}^{n} S^F_i(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_F) = 0.$$

Cramér-Rao theorem gives a matrix inequality for any unbiased estimator $\hat{\boldsymbol{\theta}}$,

$$\mathrm{Cov}[\hat{\boldsymbol{\theta}}] \succeq \frac{1}{n} g^{-1}_F(\boldsymbol{\theta}),$$

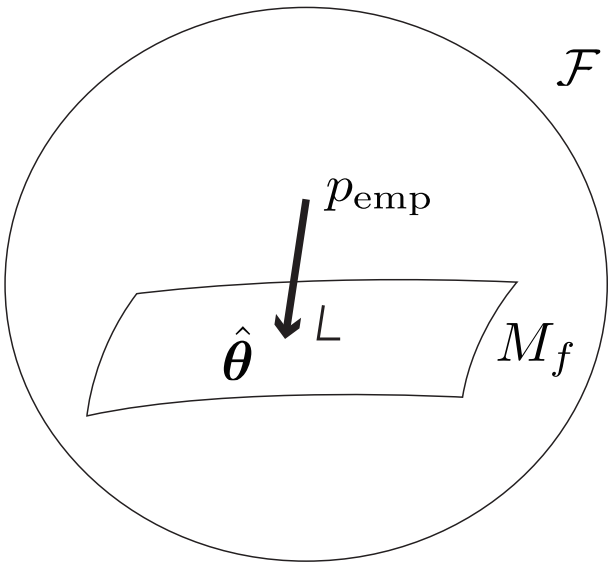

Figure 3: Projection of $\hat{p}_{\mathrm{emp}}$ to $\mathcal{M}$.

where $\mathrm{Cov}[\cdot]$ is the covariance matrix and $\succeq$ denotes the matrix order defined by the positive semidefiniteness. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}_F$ satisfies

$$\mathrm{Cov}[\hat{\boldsymbol{\theta}}_F] \approx \frac{1}{n} g_F^{-1}(\boldsymbol{\theta})$$

asymptotically. Hence, it minimizes the error covariance matrix and the minimized error covariance is given asymptotically by the inverse of the Fisher metric tensor g_F divided by n. Such a property is called the Fisher efficiency.

In the following, we study the characteristics of the estimator $\hat{\boldsymbol{\theta}}_W$ defined by (15) with the Wasserstein score:

$$\frac{1}{n}\sum_{t=1}^{n} S_i^W(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_W) = 0.$$

We call $\hat{\boldsymbol{\theta}}_W$ the Wasserstein estimator (W-estimator) following [29]. In the case of the one-dimensional location-scale model, the Wasserstein estimator is asymptotically equivalent to the estimator obtained by minimizing the Wasserstein divergence (transportation cost) from the empirical distribution $\hat{p}_{\mathrm{emp}}(\boldsymbol{x})$ to model $\mathcal{M}$:

$$\hat{\boldsymbol{\theta}}_{Wp} = \arg\min_{\boldsymbol{\theta}} D_W\left[\hat{p}_{\mathrm{emp}}(\boldsymbol{x}), p(\boldsymbol{x}, \boldsymbol{\theta})\right]. \tag{16}$$

See the end of Section 5. The properties of $\hat{\boldsymbol{\theta}}_{Wp}$ were studied in detail by [4] in the case of the one-dimensional location-scale model. Note that $\hat{\boldsymbol{\theta}}_W \neq \hat{\boldsymbol{\theta}}_{Wp}$ in general, contrary to the Fisher case.

4 Affine deformation model

Now, we focus on the affine deformation model. Let $f(\boldsymbol{z})$ be a standard probability density function satisfying (2), (3), and (4). To define $\mathcal{M}_f$, we use affine deformation of $\boldsymbol{x}$ to $\boldsymbol{z}$ by

$$\boldsymbol{z} = \Lambda(\boldsymbol{x} - \boldsymbol{\mu}),$$

where $\boldsymbol{\mu}$ is a vector representing shift of location and Λ is a non-singular matrix. Hence, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Lambda)$ is m-dimensional where $m \leq d^2 + d$ due to the possible symmetries in f. The model $\mathcal{M}_f$ is defined from

$$p(\boldsymbol{x}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{x} = f(\boldsymbol{z})d\boldsymbol{z},$$

that is

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = |\Lambda| f(\Lambda(\boldsymbol{x} - \boldsymbol{\mu})),$$

satisfying

$$\int \boldsymbol{x} p(\boldsymbol{x}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{x} = \boldsymbol{\mu}, \quad \int \boldsymbol{x}\boldsymbol{x}^\top p(\boldsymbol{x}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{x} = \Lambda^{-2} + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{17}$$

This is a generalization of the location-scale model, which is simply obtained by putting $\Lambda = (1/\sigma)I$, with σ being the scale factor. It should be noted that Λ is decomposed as $\Lambda = UDO$, where U and O are orthogonal matrices and D is a positive diagonal matrix (singular value decomposition). In the following, we denote the log probability of standard shape f by

$$l(\boldsymbol{z}) = \log f(\boldsymbol{z}).$$

As we discussed in Introduction, the set of all standard shape functions $\mathcal{F}_S = \{f\}$ does not form a manifold but has an interesting topological structure due to the rotational invariance for some f. For each standard shape function $f \in \mathcal{F}_S$, an affine deformation model $\mathcal{M}_f$ parameterized by $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Lambda)$ is attached. Thus, $\mathcal{F}$ is decomposed into the direct product of $\mathcal{F}_S$ and $\mathcal{M}_f$,

$$\mathcal{F} = \mathcal{F}_S \times \mathcal{M}_f.$$

For any f, $\mathcal{M}_f$ has cone structure parameterized by $(\boldsymbol{\mu}, D, U, O)$, where $\Lambda = UDO$ and D is a diagonal matrix with diagonal elements $d_i > 0$. Thus, D can be identified with a vector in the open positive quadrant $\boldsymbol{R}^d_+$ of $\boldsymbol{R}^d$, which has the cone structure. Since $\boldsymbol{\mu} \in \boldsymbol{R}^d$, and $U, O \in \mathcal{O}(d)$, we have the decomposition

$$\mathcal{M}_f = \boldsymbol{R}^d \times \boldsymbol{R}^d_+ \times \mathcal{O}(d) \times \mathcal{O}(d).$$

See [44] for the cone structure of $\mathcal{F}$. When f is Gaussian, its structure is studied in detail by [43].

When $p(\boldsymbol{x})$ belongs to $\mathcal{M}_f$, the waveform of $p(\boldsymbol{x})$ is said to be equivalent to that of f. The space $\mathcal{M}_f$ consists of the distributions of all equivalent waveforms. All ellipsoidal shapes are equivalent to a spherical shape. A family of special parallel-piped shapes are equivalent to a cubic form (see Figure 4). Therefore, our model is useful for separating the effect of the shape from location and affine deformation.

We may consider subclasses of the transformation model. One simple example is the location model, in which Λ is fixed to the identity matrix I. A stronger theorem is known in such a simple model [20]. In our context, it can be expressed as follows.

Proposition 1. Wasserstein geometry gives an orthogonal decomposition of the shape and locations,

$$D_W\left[f_1\left(\boldsymbol{x} - \boldsymbol{\mu}_1\right), f_2\left(\boldsymbol{x} - \boldsymbol{\mu}_2\right)\right] = D_W\left[f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\right] + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2.$$

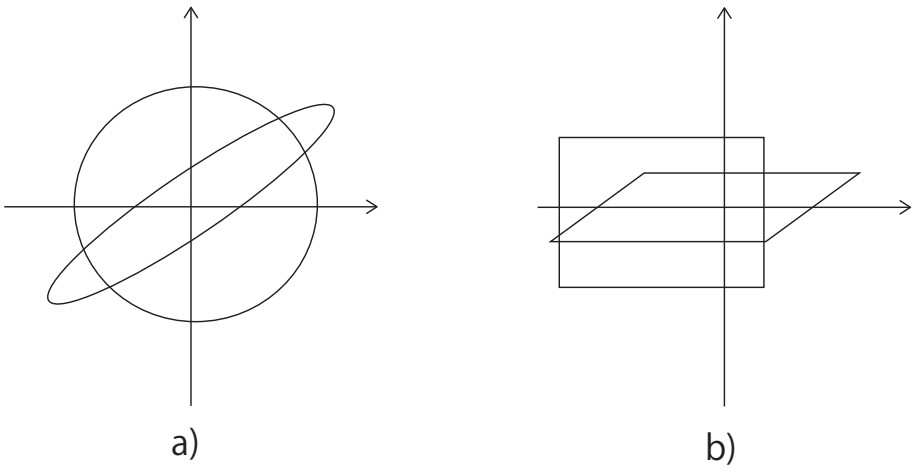

Figure 4: Equivalent shapes.

5 Elliptically symmetric deformation model

Here, we focus on deformation models that are elliptically symmetric:

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = |\Lambda| g(\|\Lambda(\boldsymbol{x} - \boldsymbol{\mu})\|), \tag{18}$$

where $f(\boldsymbol{z}) = g(\|\boldsymbol{z}\|)$ satisfies the standard density conditions (2), (3), and (4). Since $\|\Lambda(\boldsymbol{x} - \boldsymbol{\mu})\| = \|\Lambda^\top(\boldsymbol{x} - \boldsymbol{\mu})\|$, we restrict the parameter Λ to be symmetric in this section. Namely, the dimension of $\mathcal{M}_f$ is $d + d(d+1)/2$.

First, we consider the F-estimator $\hat{\boldsymbol{\theta}}_F$ (maximum likelihood estimator). The log-likelihood is given by

$$\log p(\boldsymbol{x}, \boldsymbol{\theta}) = \log |\Lambda| + \log g(\|\Lambda(\boldsymbol{x} - \boldsymbol{\mu})\|).$$

When there are n independent observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$, summation is taken over them so that we have the likelihood equations

$$\sum_{j=1}^{n} \partial_{\boldsymbol{\theta}} \log p\left(\boldsymbol{x}_j, \boldsymbol{\theta}\right) = 0.$$

The solution $\hat{\boldsymbol{\theta}}_F$ strongly depends of the shape g.

Contrary to this, the W-estimator $\hat{\boldsymbol{\theta}}_W$ does not depend on the shape g as follows. We write the i-th standard unit vector by $\boldsymbol{e}_i \in \mathbb{R}^d$ so that $(\boldsymbol{e}_i)_j = \delta_{ij}$.

Lemma 1.

$$\nabla_{\boldsymbol{x}} \left(\frac{1}{2} \boldsymbol{x}^\top A \boldsymbol{x} + b^\top \boldsymbol{x} + c \right) = \frac{A + A^\top}{2} \boldsymbol{x} + b, \quad \Delta_{\boldsymbol{x}} \left(\frac{1}{2} \boldsymbol{x}^\top A \boldsymbol{x} + b^\top \boldsymbol{x} + c \right) = \mathrm{tr}(A).$$

Proof. Straightforward calculation. Note that $(A + A^\top)/2 \neq A$ when A is not symmetric. □

Lemma 2. *Let $A, B \in \mathbb{R}^{d \times d}$ be symmetric matrices. If A is positive definite, then the Sylvester equation $AX + XA = B$ has the unique solution X, which satisfies $X^\top = X$ and $\mathrm{tr}(X) = \mathrm{tr}(A^{-1}B)/2$.*

Proof. From the positive semidefiniteness of A, the spectra of A and $-A$ are disjoint. Thus, from Theorem VII.2.1 of [12], the Sylvester equation $AX + XA = B$ has a unique solution.

Let X be the solution of the Sylvester equation. From $A^\top = A$, we have $AX^\top + X^\top A = (AX + XA)^\top = B^\top = B$, which means that $X^\top$ is also a solution of the Sylvester equation. Since the solution is unique, it implies $X^\top = X$. Also, from the positive semidefiniteness of A and $AX + XA = B$, we have $X + A^{-1}XA = A^{-1}B$. Taking the trace and using $\mathrm{tr}(A^{-1}XA) = \mathrm{tr}(X)$, we obtain $\mathrm{tr}X = \mathrm{tr}(A^{-1}B)/2$. □

Theorem 1. For the elliptically symmetric deformation model (18), the Wasserstein score functions are quadratic. Specifically, the Wasserstein score function for μ_i is

$$S^W_{\mu_i}(\boldsymbol{x}, \boldsymbol{\theta}) = x_i - \mu_i,$$

and the Wasserstein score function for Λ_{ij} is

$$S^W_{\Lambda_{ij}}(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x} + b^\top \boldsymbol{x} - \mathrm{E}_{\boldsymbol{\theta}}\left[\frac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x} + b^\top \boldsymbol{x}\right],$$

where A is the unique solution of the Sylvester equation $\Lambda^2 A + A\Lambda^2 = -\Lambda \boldsymbol{e}_i \boldsymbol{e}_j^\top - \boldsymbol{e}_i \boldsymbol{e}_j^\top \Lambda$ and $b = -A\mu$.

Proof. We show that the above S^W's satisfy the Poisson equation (11) directly.

First, we consider the mean parameter μ_i. From (18),

$$\frac{\partial}{\partial \mu_i} \log p(\boldsymbol{x}, \boldsymbol{\theta}) = -\frac{\partial}{\partial x_i} \log p(\boldsymbol{x}, \boldsymbol{\theta}) = -\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}, \boldsymbol{\theta})^\top \boldsymbol{e}_i.$$

Also, from Lemma 1,

$$\nabla_{\boldsymbol{x}}(x_i - \mu_i) = \boldsymbol{e}_i, \quad \Delta_{\boldsymbol{x}}(x_i - \mu_i) = 0.$$

Therefore,

$$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}, \boldsymbol{\theta})^\top \nabla_{\boldsymbol{x}}(x_i - \mu_i) + \Delta_{\boldsymbol{x}}(x_i - \mu_i) + \frac{\partial}{\partial \mu_i} \log p(\boldsymbol{x}, \boldsymbol{\theta}) = 0.$$

Thus, the Wasserstein score function for the mean parameter μ_i is

$$S^W_{\mu_i}(\boldsymbol{x}, \boldsymbol{\theta}) = x_i - \mu_i.$$

Next, we consider the deformation parameter Λ_{ij}. Since

$$\begin{aligned}
\frac{\partial}{\partial \Lambda_{ij}} \|\Lambda(\boldsymbol{x} - \boldsymbol{\mu})\| &= \frac{1}{2}\|\Lambda(\boldsymbol{x} - \boldsymbol{\mu})\|^{-1} \frac{\partial}{\partial \Lambda_{ij}} (\boldsymbol{x} - \boldsymbol{\mu})^\top \Lambda^2 (\boldsymbol{x} - \boldsymbol{\mu}) \\
&= \frac{1}{2}\|\Lambda(\boldsymbol{x} - \boldsymbol{\mu})\|^{-1} (\boldsymbol{x} - \boldsymbol{\mu})^\top \frac{\partial \Lambda^2}{\partial \Lambda_{ij}} (\boldsymbol{x} - \boldsymbol{\mu}) \\
&= \frac{1}{2}\|\Lambda(\boldsymbol{x} - \boldsymbol{\mu})\|^{-1} (\boldsymbol{x} - \boldsymbol{\mu})^\top (\Lambda \boldsymbol{e}_i \boldsymbol{e}_j^\top + \boldsymbol{e}_i \boldsymbol{e}_j^\top \Lambda)(\boldsymbol{x} - \boldsymbol{\mu}),
\end{aligned}$$

we have

$$\begin{aligned}\frac{\partial}{\partial\Lambda_{ij}}\log p(\boldsymbol{x},\theta) &= \frac{\partial}{\partial\Lambda_{ij}}\log\det\Lambda + \frac{\partial}{\partial\Lambda_{ij}}\log g(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)\\ &= (\Lambda^{-1})_{ij} + \frac{g'(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}{g(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}\frac{\partial}{\partial\Lambda_{ij}}\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|\\ &= (\Lambda^{-1})_{ij} + \frac{g'(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}{2\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|g(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}(\boldsymbol{x}-\boldsymbol{\mu})^\top(\Lambda\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_i\boldsymbol{e}_j^\top\Lambda)(\boldsymbol{x}-\boldsymbol{\mu})\\ &= -\frac{g'(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}{\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|g(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}\left(-\frac{1}{2}\boldsymbol{x}^\top L\boldsymbol{x} + \boldsymbol{x}^\top L\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^\top L\boldsymbol{\mu}\right) + (\Lambda^{-1})_{ij},\end{aligned}$$

where $L = \Lambda\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_i\boldsymbol{e}_j^\top\Lambda$. Let

$$S(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x} + b^\top\boldsymbol{x} - \mathrm{E}_\theta\left[\frac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x} + b^\top\boldsymbol{x}\right],$$

where A is the unique solution of the Sylvester equation $\Lambda^2 A + A\Lambda^2 = -L$ and $b = -A\boldsymbol{\mu}$. From Lemma 1 and Lemma 2,

$$\Delta_{\boldsymbol{x}}S(\boldsymbol{x}) = \mathrm{tr}A = -\frac{1}{2}\mathrm{tr}(\Lambda^{-2}L) = -\frac{1}{2}\mathrm{tr}(\Lambda^{-2}(\Lambda\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_i\boldsymbol{e}_i^\top\Lambda)) = -(\Lambda^{-1})_{ij}.$$

Also,

$$\begin{aligned}&\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x},\boldsymbol{\theta})^\top\nabla_{\boldsymbol{x}}S(\boldsymbol{x})\\ =&\frac{g'(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}{g(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}\nabla_{\boldsymbol{x}}(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)^\top(A\boldsymbol{x}+b)\\ =&\frac{g'(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}{\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|g(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}(\Lambda^2(\boldsymbol{x}-\boldsymbol{\mu}))^\top(A\boldsymbol{x}+b)\\ =&\frac{g'(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}{\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|g(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}\left(\frac{1}{2}\boldsymbol{x}^\top(\Lambda^2A + A\Lambda^2)\boldsymbol{x} + \boldsymbol{x}^\top(\Lambda^2b - A\Lambda^2\boldsymbol{\mu}) - \boldsymbol{\mu}^\top\Lambda^2b\right)\\ =&\frac{g'(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}{\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|g(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|)}\left(-\frac{1}{2}\boldsymbol{x}^\top L\boldsymbol{x} + \boldsymbol{x}^\top L\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^\top L\boldsymbol{\mu}\right),\end{aligned}$$

where we used

$$\begin{aligned}\nabla_{\boldsymbol{x}}(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|) &= \frac{1}{2\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|}\nabla_{\boldsymbol{x}}(\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|^2)\\ &= \frac{1}{2\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|}\nabla_{\boldsymbol{x}}((\boldsymbol{x}-\boldsymbol{\mu})^\top\Lambda^2(\boldsymbol{x}-\boldsymbol{\mu}))\\ &= \frac{1}{\|\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\|}\Lambda^2(\boldsymbol{x}-\boldsymbol{\mu}).\end{aligned}$$

Therefore,

$$\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x},\boldsymbol{\theta})^\top\nabla_{\boldsymbol{x}}S(\boldsymbol{x}) + \Delta_{\boldsymbol{x}}S(\boldsymbol{x}) + \frac{\partial}{\partial\Lambda_{ij}}\log p(\boldsymbol{x},\boldsymbol{\theta}) = 0,$$

which means that $S(\boldsymbol{x})$ is the Wasserstein score function for Λ_{ij}.

□

The optimal transport map from $p(\boldsymbol{x}, \boldsymbol{\theta})$ to $p(\boldsymbol{x}, \widetilde{\boldsymbol{\theta}})$ is given by the affine map $\boldsymbol{x} \mapsto \Lambda(\Lambda^{-1}\widetilde{\Lambda}^{-2}\Lambda^{-1})^{1/2}\Lambda(\boldsymbol{x} - \boldsymbol{\mu}) + \widetilde{\boldsymbol{\mu}}$ [19]. It provides another proof of Theorem 1.

Now, we consider the Wasserstein estimator $\hat{\boldsymbol{\theta}}_W$ defined as the zero of the Wasserstein score function [29].

Corollary 1. Suppose that we have n independent observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from the elliptically symmetric deformation model (18) where $n \geq d$. Then, the Wasserstein estimator $\hat{\boldsymbol{\theta}}_W = (\hat{\boldsymbol{\mu}}_W, \hat{\Lambda}_W)$ is the second-order moment estimator given by

$$\hat{\boldsymbol{\mu}}_W = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{x}_t, \quad \hat{\Lambda}_W = \left(\frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_W)(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_W)^\top\right)^{-1/2},$$

irrespective of the waveform $f(\boldsymbol{z}) = g(\|\boldsymbol{z}\|)$.

Proof. From Theorem 1, the Wasserstein estimator is the solution of

$$\frac{1}{n}\sum_{t=1}^{n} S_{\mu_i}^W(\boldsymbol{x}_t, \boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}((\boldsymbol{x}_t)_i - \mu_i) = 0 \tag{19}$$

for $i = 1, \ldots, d$ and

$$\frac{1}{n}\sum_{t=1}^{n} S_{\Lambda_{ij}}^W(\boldsymbol{x}_t, \boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{2}\boldsymbol{x}_t^\top A\boldsymbol{x}_t + b^\top \boldsymbol{x}_t\right) - \mathrm{E}_{\boldsymbol{\theta}}\left[\frac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x} + b^\top \boldsymbol{x}\right] = 0 \tag{20}$$

for $i, j = 1, \ldots, d$, where A is the unique solution of the Sylvester equation $\Lambda^2 A + A\Lambda^2 = -\Lambda \boldsymbol{e}_i \boldsymbol{e}_j^\top - \boldsymbol{e}_i \boldsymbol{e}_j^\top \Lambda$ and $b = -A\boldsymbol{\mu}$. (Note that the Wasserstein score function is a function from $\mathbb{R}^d$ to $\mathbb{R}$ and does not depend on n.)

From (19), the Wasserstein estimator of μ is

$$\hat{\boldsymbol{\mu}}_W = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{x}_t.$$

Also, since (20) implies that the second-order empirical moments match the second-order population moments and $\mathrm{Cov}[\boldsymbol{x}] = \Lambda^{-2}$ from (17), the Wasserstein estimator of Λ is

$$\hat{\Lambda}_W = \left(\frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_W)(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_W)^\top\right)^{-1/2}.$$

□

Note that [19] showed that the L^2-Wasserstein divergence for the elliptically symmetric deformation model (18) does not depend on the waveform, and is given by using the Bures–Wasserstein distance between positive definite matrices [13]:

$$D_W(p(\boldsymbol{x}, \boldsymbol{\theta}_1), p(\boldsymbol{x}, \boldsymbol{\theta}_2)) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 + \mathrm{tr}(\Lambda_1^{-2} + \Lambda_2^{-2} - 2(\Lambda_1^{-1}\Lambda_2^{-2}\Lambda_1^{-1})^{1/2}).$$

It is an interesting future problem to derive the Wasserstein score function and Wasserstein estimator for general affine deformation models.

Regarding the geometric structure of the elliptically symmetric deformation model (18), we obtain the following. See Figure 1.

Theorem 2. When f is spherically symmetric, the model $\mathcal{M}_f$ is orthogonal to $\mathcal{F}_S$ at the origin $\boldsymbol{\mu} = 0, \Lambda = I$ of $\mathcal{M}_f$ with respect to the Wasserstein metric.

Proof. Let $\delta p(\boldsymbol{x})$ be a tangent vector of $\mathcal{F}_S$ at the origin. Since all $p(\boldsymbol{x})$ in $\mathcal{F}_S$ satisfy the standard conditions (2), (3), and (4), $\delta p(\boldsymbol{x})$ is orthogonal to any quadratic function of $\boldsymbol{x}$. Since the W-score functions of M_S are quadratic functions from Theorem 1, it implies that $\delta p(\boldsymbol{x})$ is orthogonal to the Waserstein functions of $\boldsymbol{x}$, which form the basis of the tangent space of $\mathcal{M}_f$, with respect to the Wasserstein metric. □

Here, we discuss the relation between the current Wasserstein estimator $\hat{\boldsymbol{\theta}}_W$ and the estimator $\hat{\boldsymbol{\theta}}_{Wp}$ in (16) defined as the projection of the empirical distribution with respect to the Wasserstein distance. For the one-dimensional location-scale model, [4] studied the estimator $\hat{\boldsymbol{\theta}}_{Wp}$ in (16) by using the order statistics $x_{(i)}$. This estimator minimizes the Wasserstein distance between the empirical distribution and the model. Here, we show that this estimator $\hat{\boldsymbol{\theta}}_{Wp} = (\hat{\boldsymbol{\mu}}_{Wp}, \hat{\sigma}_{Wp})$ is asymptotically equivalent to the Wasserstein estimator $\hat{\boldsymbol{\theta}}_W$, which coincides with the second-order moment estimator from Theorem 1. We assume $\mu = 0$ without loss of generality. The estimator (16) of the location is

$$\hat{\mu}_{Wp} = \frac{1}{n}\sum_{i=1}^{n} x_{(i)} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

which is the empirical mean and coincides with the moment estimator. Also, the estimator (16) of the scale is

$$\hat{\sigma}_{Wp} = \sum_{i=1}^{n} k_i x_{(i)},$$

where

$$k_i = \int_{z_{i-1}}^{z_i} z f(z) dz.$$

Here, z_i is the i-th equipartition point of $f(z)$ defined by

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

where F is the cumulative distribution function of $f(z)$. From $\mu = 0$, we have $x_{(i)} \approx \sigma z_i$ asymptotically. Hence,

$$k_i \approx \frac{1}{n} z_i \approx \frac{1}{n}\frac{x_{(i)}}{\sigma},$$

which leads to

$$\hat{\sigma}_{Wp} = \sum_{i=1}^{n} k_i x_{(i)} \approx \frac{1}{n\sigma}\sum_{i=1}^{n} x_{(i)}^2 = \frac{1}{n\sigma}\sum_{i=1}^{n} x_i^2.$$

Since $\hat{\sigma}_{Wp} \approx \sigma$ asymptotically,

$$\hat{\sigma}_{Wp}^2 \approx \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

This shows that $\hat{\boldsymbol{\theta}}_{Wp} = (\hat{\boldsymbol{\mu}}_{Wp}, \hat{\sigma}_{Wp})$ asymptotically coincides with the second-order moment estimator $\hat{\boldsymbol{\theta}}_W = (\hat{\boldsymbol{\mu}}_W, \hat{\sigma}_W)$.

6 Wasserstein covariance and robustness

Following [29], we define the Wasserstein covariance (W-covariance) matrix $\mathrm{Var}_\theta^{\mathrm{W}}[\hat{\boldsymbol{\theta}}]$ of an estimator $\hat{\boldsymbol{\theta}}$ by the positive semidefinite matrix given by

$$\mathrm{Var}_\theta^{\mathrm{W}}[\hat{\boldsymbol{\theta}}] = (\mathrm{E}_\theta[(\nabla_{\boldsymbol{x}}\hat{\boldsymbol{\theta}}_a)^\top(\nabla_{\boldsymbol{x}}\hat{\boldsymbol{\theta}}_b)])_{ab}. \tag{21}$$

[29] showed the Wasserstein–Cramer–Rao inequality

$$\mathrm{Var}_\theta^{\mathrm{W}}(\hat{\boldsymbol{\theta}}) \succeq \left(\frac{\partial}{\partial\theta}\mathrm{E}_\theta[\hat{\boldsymbol{\theta}}]\right)^\top G_{\mathrm{W}}(\theta)^{-1}\left(\frac{\partial}{\partial\theta}\mathrm{E}_\theta[\hat{\boldsymbol{\theta}}]\right), \tag{22}$$

where

$$\frac{\partial}{\partial\theta}\mathrm{E}_\theta[\hat{\boldsymbol{\theta}}] := \left(\frac{\partial}{\partial\theta_j}\mathrm{E}_\theta[\hat{\boldsymbol{\theta}}_i]\right)_{ij}.$$

A consistent estimator $\hat{\boldsymbol{\theta}}$ is said to be Wasserstein efficient (W-efficient) if its Wasserstein covariance asymptotically satisfies (22) with equality. We give a proof of the Wasserstein–Cramer–Rao inequality based on the Cauchy–Schwarz inequality in the Appendix.

We show that the Wasserstein covariance of an estimator can be viewed as a measure of robustness against additive noise. Suppose that $X \sim p(\boldsymbol{x},\boldsymbol{\theta})$ and we estimate $\boldsymbol{\theta}$ from the noisy observation $\widetilde{X} = X + Z$, where Z is independent from X, $\mathrm{E}[Z] = 0$ and $\mathrm{Var}[Z] = \sigma^2 I$ with σ^2 sufficiently small. The probability density of $\widetilde{X}$ is given by $\widetilde{p}(\boldsymbol{x},\boldsymbol{\theta}) = p(\boldsymbol{x},\boldsymbol{\theta}) * q(\boldsymbol{x})$, where $*$ is the convolution and q is the probability density of the noise Z. Namely, the noise changes the waveform from p to $\widetilde{p}$. Generally, the estimator degrades when the noise is added. Here, we quantify the robustness of an estimator against noise based on how much its variance increases due to noise. Namely, we focus on $\mathrm{Var}_\theta[\hat{\boldsymbol{\theta}}(X+Z)] - \mathrm{Var}_\theta[\hat{\boldsymbol{\theta}}(X)]$. If this quantity is small, it implies that the estimator is not much affected by noise contamination, which can be viewed as its robustness. This quantity is closely related to the Wasserstein covariance as follows.

Theorem 3. *The Wasserstein covariance satisfies*

$$\begin{aligned}\mathrm{Var}_\theta^{\mathrm{W}}[\hat{\boldsymbol{\theta}}]_{ab} = \lim_{\sigma^2\to 0}\frac{\mathrm{Var}_\theta[\hat{\boldsymbol{\theta}}(X+Z)]_{ab} - \mathrm{Var}_\theta[\hat{\boldsymbol{\theta}}(X)]_{ab}}{\sigma^2} \\ - \frac{1}{2}\left(\mathrm{Cov}_\theta[\hat{\boldsymbol{\theta}}_a(X), \Delta\hat{\boldsymbol{\theta}}_b(X)] + \mathrm{Cov}_\theta[\hat{\boldsymbol{\theta}}_b(X), \Delta\hat{\boldsymbol{\theta}}_a(X)]\right),\end{aligned}$$

where Δ is the Laplacian. In particular, if $\Delta\hat{\boldsymbol{\theta}}_a(X)$ is constant for every a (e.g. $\hat{\boldsymbol{\theta}}$ is quadratic in $\boldsymbol{x}$), then

$$\mathrm{Var}_\theta^{\mathrm{W}}[\hat{\boldsymbol{\theta}}] = \lim_{\sigma^2\to 0}\frac{\mathrm{Var}_\theta[\hat{\boldsymbol{\theta}}(X+Z)] - \mathrm{Var}_\theta[\hat{\boldsymbol{\theta}}(X)]}{\sigma^2}.$$

Proof. By Taylor expansion, for sufficiently small $\boldsymbol{z}$,

$$\hat{\boldsymbol{\theta}}_a(\boldsymbol{x}+\boldsymbol{z}) \approx \hat{\boldsymbol{\theta}}_a(\boldsymbol{x}) + \sum_i \frac{\partial\hat{\boldsymbol{\theta}}_a}{\partial x_i}(\boldsymbol{x})z_i + \frac{1}{2}\sum_{i,j}\frac{\partial^2\hat{\boldsymbol{\theta}}_a}{\partial x_i\partial x_j}(\boldsymbol{x})z_iz_j.$$

From $\mathrm{E}[Z] = 0$, $\mathrm{Var}[Z] = \sigma^2 I$ and the independence of X and Z,

$$\begin{aligned}\mathrm{E}_\theta[\hat{\boldsymbol{\theta}}_a(X+Z)] &= \mathrm{E}_\theta[\hat{\boldsymbol{\theta}}_a(X)] + \sum_i \mathrm{E}_\theta\left[\frac{\partial \hat{\boldsymbol{\theta}}_a}{\partial x_i}(X)\right]\mathrm{E}[z_i] + \frac{1}{2}\sum_{i,j}\mathrm{E}_\theta\left[\frac{\partial^2 \hat{\boldsymbol{\theta}}_a}{\partial x_i \partial x_j}(X)\right]\mathrm{E}[z_i z_j] \\ &= \mathrm{E}_\theta[\hat{\boldsymbol{\theta}}_a(X)] + \frac{1}{2}\mathrm{E}_\theta[\Delta\hat{\boldsymbol{\theta}}_a(X)]\sigma^2.\end{aligned}$$

Also,

$$\begin{aligned}\mathrm{E}_\theta[\hat{\boldsymbol{\theta}}_a(X+Z)\hat{\boldsymbol{\theta}}_b(X+Z)] &= \mathrm{E}_\theta[\hat{\boldsymbol{\theta}}_a(X)\hat{\boldsymbol{\theta}}_b(X)] + \frac{1}{2}\mathrm{E}_\theta[\hat{\boldsymbol{\theta}}_a(X)\Delta\hat{\boldsymbol{\theta}}_b(X) + \hat{\boldsymbol{\theta}}_b(X)\Delta\hat{\boldsymbol{\theta}}_a(X)]\sigma^2 \\ &\quad + \mathrm{E}_\theta[(\nabla\hat{\boldsymbol{\theta}}_a(X))^\top(\nabla\hat{\boldsymbol{\theta}}_b(X))]\sigma^2 + o(\sigma^2) \\ &= \mathrm{E}_\theta[\hat{\boldsymbol{\theta}}_a(X)\hat{\boldsymbol{\theta}}_b(X)] + \frac{1}{2}\mathrm{E}_\theta[\hat{\boldsymbol{\theta}}_a(X)\Delta\hat{\boldsymbol{\theta}}_b(X) + \hat{\boldsymbol{\theta}}_b(X)\Delta\hat{\boldsymbol{\theta}}_a(X)]\sigma^2 \\ &\quad + \mathrm{Var}_\theta[\hat{\boldsymbol{\theta}}]_{ab}\sigma^2 + o(\sigma^2).\end{aligned}$$

Then,

$$\begin{aligned}&\mathrm{Var}_\theta[\hat{\boldsymbol{\theta}}(X+Z)]_{ab} \\ =&\mathrm{E}_\theta[\hat{\boldsymbol{\theta}}_a(X+Z)\hat{\boldsymbol{\theta}}_b(X+Z)] - \mathrm{E}_\theta[\hat{\boldsymbol{\theta}}_a(X+Z)]\mathrm{E}_\theta[\hat{\boldsymbol{\theta}}_b(X+Z)] \\ =&\mathrm{Var}_\theta[\hat{\boldsymbol{\theta}}(X)]_{ab} + \mathrm{Var}_\theta^{\mathrm{W}}[\hat{\boldsymbol{\theta}}]_{ab}\sigma^2 \\ &\quad + \frac{1}{2}\left(\mathrm{Cov}_\theta[\hat{\boldsymbol{\theta}}_a(X), \Delta\hat{\boldsymbol{\theta}}_b(X)] + \mathrm{Cov}_\theta[\hat{\boldsymbol{\theta}}_b(X), \Delta\hat{\boldsymbol{\theta}}_a(X)]\right)\sigma^2 + o(\sigma^2),\end{aligned}$$

where the covariance term vanishes when $\hat{\boldsymbol{\theta}}$ is quadratic in $\boldsymbol{x}$. □

For the elliptically symmetric deformation model (18), from Theorem 3 and Corollary 1, the Wasserstein covariance quantifies the robustness of the Wasserstein estimator and the Wasserstein–Cramer–Rao inequality gives its limit. It is an interesting future problem to investigate when the Wasserstein estimator attains the Wasserstein efficiency. Note that the Fisher efficiency (in finite samples), which is defined by the usual Cramer–Rao inequality, is attained by the maximum likelihood estimator if and only if the estimand is the expectation parameter of an exponential family.

7 Contribution of waveform to F- and W-efficiencies

We study how the waveform f contributes to the F-efficiency and W-efficiency of estimators. We first show the following theorem.

Theorem 4. When and only when f is Gaussian, F-estimator $\hat{\boldsymbol{\theta}}_F$ and W-estimator $\hat{\boldsymbol{\theta}}_W$ are identical and F-efficient.

Proof. For the standard Gaussian f,

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{d/2}}\exp\left\{-\frac{\|\boldsymbol{z}\|^2}{2}\right\},$$

the F-score functions are

$$S_F(\boldsymbol{x}, \boldsymbol{\theta}) = \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{x}, \boldsymbol{\theta}) = -\boldsymbol{z} \cdot \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\theta}}.$$

Hence, the F-score functions are quadratic with respect to $\boldsymbol{x}$. So they are equivalent to the W-score functions. On the contrary, when the score functions are quadratic with respect to $\boldsymbol{x}$, the waveform f is Gaussian. □

When f is not Gaussian, the F-efficiency of $\hat{\boldsymbol{\theta}}_W$ degrades. When f is close to Gaussian, their cumulants of order larger than two are small. We use the Gram-Charlie expansion of f to study how cumulants of the waveform f contribute to the F-efficiency of $\hat{\boldsymbol{\theta}}_W$.

We study how waveform f contributes to the amounts of Fisher information g_F and Wasserstein information g_W, when f is close to the Gaussian distribution. Let

$$\phi(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{\|\boldsymbol{x}\|^2}{2}\right\}$$

be the standard Gaussian. We use the Gram-Charlie expansion [34]:

$$f(\boldsymbol{x}) = \phi(\boldsymbol{x}) \left\{1 + \frac{\kappa_3}{3!} \circ h_3(\boldsymbol{x}) + \frac{\kappa_4}{4!} \circ h_4(\boldsymbol{x})\right\},$$

where κ_i is the ith order cumulant tensor of f, $h_i(\boldsymbol{x})$ is the ith order tensorial Hermite polynomial, and $\circ$ denotes the tensorial inner product such as

$$\kappa_3 \circ h_3 = \sum \kappa_{3,ijk} h_3^{ijk}.$$

The logarithm of $p(\boldsymbol{x}, \boldsymbol{\theta})$ is expanded as

$$l(\boldsymbol{x}, \boldsymbol{\theta}) = -\log|\Lambda| + \left\{-\frac{1}{2}|\boldsymbol{z}|^2 + \frac{\kappa_3}{3!} \circ h_3(\boldsymbol{z}) + \frac{\kappa_4}{4!} \circ h_4(\boldsymbol{z}) + \dots\right\},$$

where $\boldsymbol{z} = \Lambda(\boldsymbol{x} - \boldsymbol{\mu})$.

The Fisher information g_F is given by

$$g_F = -\mathrm{E}\left[\partial_{\boldsymbol{\theta}} \partial_{\boldsymbol{\theta}} l(\boldsymbol{x}, \boldsymbol{\theta})\right].$$

We have

$$\partial_{\boldsymbol{\theta}} l = \frac{\partial l}{\partial \boldsymbol{z}} \cdot \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\theta}}, \quad \partial_{\boldsymbol{\theta}} \partial_{\boldsymbol{\theta}} l = \frac{\partial^2 l}{\partial \boldsymbol{z} \partial \boldsymbol{z}} \cdot \left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\theta}}\right) + \frac{\partial l}{\partial \boldsymbol{z}} \cdot \frac{\partial^2 \boldsymbol{z}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}},$$

where $\cdot$ denotes the inner product with respect to the indices of $\boldsymbol{z}$. From the derivative of Hermite polynomials, we have

$$\partial_{\boldsymbol{z}} l = \Lambda^{-1} \left\{-\boldsymbol{z} + \frac{\kappa_3}{2} \circ h_2(\boldsymbol{z}) + \frac{\kappa_4}{6} \circ h_3(\boldsymbol{z})\right\},$$
$$\partial_{\boldsymbol{z}} \partial_{\boldsymbol{z}} l = \Lambda^{-1} \left\{-I + \kappa_3 \circ h_1(\boldsymbol{z}) + \frac{\kappa_4}{2} \circ h_2(\boldsymbol{z})\right\},$$

where $\kappa_i \circ h_j(\boldsymbol{z})$ are tensorial polynomials of $\boldsymbol{z}$. On the other hand, $\partial \boldsymbol{z} / \partial \boldsymbol{\theta}$ are given by

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\mu}} = \Lambda, \quad \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\Lambda}} = \boldsymbol{x} - \boldsymbol{\mu}.$$

In order to avoid complicated tensorial calculations, we study only the case of $d = 1$, that is the location-scale model. After simple calculations, we obtain

$$\begin{aligned}
\frac{\partial^2 l}{\partial\mu\partial\mu} &= \lambda^2 \left\{ -1 + \kappa_3 z + \frac{\kappa_4}{2}\left(z^2 - 1\right) \right\}, \\
\frac{\partial^2 l}{\partial\lambda\partial\mu} &= \lambda^2 \left\{ -2z + \frac{\kappa_3}{2}\left(3z^2 - 1\right) + \frac{\kappa_4}{3}\left(2z^3 - 3z\right) \right\}, \\
\frac{\partial^2 l}{\partial\lambda\partial\lambda} &= \lambda^2 \left\{ \left(-3z^2 + 1\right) + \frac{\kappa_3}{2}\left(4z^3 - 3z\right) + \frac{\kappa_4}{6}\left(4z^4 - 6z^2\right) \right\}.
\end{aligned}$$

Here, z is subject to $f(z)$, so we have

$$g_F(\mu, \lambda) = \lambda^2 \begin{pmatrix} 1 & -\kappa_3 \\ -\kappa_3 & 2 - \kappa_4 \end{pmatrix}.$$

This shows how $g_F(\mu, \lambda)$ deviates from the Gaussian case depending on κ_3 and κ_4.

It is also interesting to consider the case when f has high-frequency wavy structure. Since F-score functions are derivatives of the log probability, high-frequency components are sensitive to them, contributing to the F-metric. However, by adding small noises to $\boldsymbol{x}$, those components are smoothed out. Hence the W-metric is insensitive to the high-frequency components. We observe that, when f includes a high frequency component such as

$$f(x) = \phi(\boldsymbol{x}) \left\{ 1 + \varepsilon \sin \tau x \right\},$$

where ε is small and τ is the frequency of small deviation, $\partial^2 l / \partial z^2$ has a component proportional to $\varepsilon\tau^2$. Hence, the increment due to the high-frequency component is proportional to τ^2, implying that the increment of the Fisher information is proportional to τ^2. Hence, high frequency ripples of waveform f increases g_F.

The W-estimator $\hat{\boldsymbol{\theta}}_W$ is not F-efficient except for the Gaussian case. The loss of F-efficiency depends on the waveform f. We again use the Gram-Charlie expansion and see the effect of κ_3 and κ_4, assuming they are small. We focus on the location-scale model with $d = 1$ for convenience. For n observations $\boldsymbol{x} = (x_1, \cdots, x_n)$, define the empirical moments by

$$m_r(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i^r, \quad r = 1, 2, \cdots$$

Note that $\hat{\boldsymbol{\theta}}_W$ uses only m_1 and m_2, discarding higher-order moments. Let

$$C(s_1, s_2) = \{ \boldsymbol{x} \,|m_1(\boldsymbol{x}) = s_1, m_2(\boldsymbol{x}) = s_2 \}$$

be the set of $\boldsymbol{x}$ having moments s_1, s_2. We calculate the Fisher information included in $\hat{\boldsymbol{\theta}}_W$. The Fisher information in $\boldsymbol{x}$ is the covariance of the F-score $\nabla l(\boldsymbol{x}, \boldsymbol{\theta})$ and decomposed into the sum of within-class covariance and between-class covariance:

$$\mathrm{Cov}[\nabla l] = \mathrm{E}_{s_1,s_2}[\mathrm{Cov}[\nabla l | s_1, s_2]] + \mathrm{Cov}[\mathrm{E}[\nabla l | s_1, s_2]],$$

where $\mathrm{E}[\cdot|s_1, s_2]$ and $\mathrm{Cov}[\cdot|s_1, s_2]$ denotes the conditional expectation and conditional covariance conditioned on s_1, s_2. Since $\hat{\boldsymbol{\theta}}_W$ does not use higher-order moment information, the

Fisher information included in $\hat{\boldsymbol{\theta}}_W$ corresponds to the between-class covariance. Thus, the loss of information due to the transformation from $\boldsymbol{x}$ to $\hat{\theta}_W(\boldsymbol{x})$ is

$$\Delta g_{F,W} = \mathrm{E}_{s_1,s_2}\left[\mathrm{Cov}[\nabla l|s_1,s_2]\right].$$

The conditional expectation of ∇l is

$$\begin{aligned}
E\left[\partial_\mu l\,|s_1,s_2\right] &= -s_1 + \frac{\kappa_3}{2}\left(s_2-1\right) + \frac{\kappa_4}{6}\left(\mathrm{E}\left[m_3(\boldsymbol{x})\,|s_1,s_2\right]-3s_1\right),\\
E\left[\partial_\lambda l\,|s_1,s_2\right] &= -s_2 + \frac{\kappa_3}{2}\left\{\left(\mathrm{E}\left[m_3(\boldsymbol{x})\,|s_1,s_2\right]\right)-s_1\right\} + \frac{\kappa_4}{6}\left\{\mathrm{E}\left[m_4(\boldsymbol{x})\,|s_1,s_2\right]-3s_2\right\}.
\end{aligned}$$

Hence, the conditional covariance of ∇l is

$$\begin{aligned}
\mathrm{Cov}\left[\partial_\mu l\,|s_1,s_2\right] &= \frac{\kappa_4^2}{36}\mathrm{Cov}\left[m_3(\boldsymbol{x})\,|s_1,s_2\right],\\
\mathrm{Cov}\left[\partial_\lambda l\,|s_1,s_2\right] &= \frac{\kappa_3^2}{4}\mathrm{Cov}\left[m_3(\boldsymbol{x})\,|s_1,s_2\right] + \frac{\kappa_4^2}{36}\mathrm{Cov}\left[m_4(\boldsymbol{x})\,|s_1,s_2\right],\\
\mathrm{Cov}\left[\partial_\mu l,\partial_\lambda l\,|s_1,s_2\right] &= \frac{\kappa_3\kappa_4}{12}\mathrm{E}\left[m_3(\boldsymbol{x})|s_1,s_2\right]\mathrm{E}\left[m_4(\boldsymbol{x})|s_1,s_2\right].
\end{aligned}$$

It should be noted that $m_3(\boldsymbol{x})$ and $m_4(\boldsymbol{x})$ are asymptotically independent of $m_1(\boldsymbol{x})$ and $m_2(\boldsymbol{x})$, because $(m_1(\boldsymbol{x}),m_2(\boldsymbol{x}),m_3(\boldsymbol{x}),m_4(\boldsymbol{x}))$ are asymptotically independent and jointly Gaussian. We thus have asymptotically

$$\begin{aligned}
\mathrm{E}\left[m_3(\boldsymbol{x})\,|s_1,s_2\right] &= \mathrm{E}\left[x^3\right] = \frac{\kappa_3}{\lambda^3},\\
\mathrm{E}\left[m_4(\boldsymbol{x})\,|s_1,s_2\right] &= \mathrm{E}\left[x^4\right] = \frac{3\left(1+\kappa_4\right)}{\lambda^4} + \frac{6}{\lambda^2}\mu^2 + \mu^4,
\end{aligned}$$

and so on. Summing up all these results, we obtain $\Delta g_{F,W}$ in terms of κ_3 and κ_4. We leave more detail for future work.

8 Conclusion

Statistical inference has so far been studied mostly based on information geometry from the Fisherian point of view, with remarkable success. It is based on the likelihood principle, and the invariant divergence has played a fundamental role. However, Wasserstein divergence gives another viewpoint, which is based on the geometric structure of the sample space X. There are many applications of the Wasserstein geometry not only to the transportation problem but to vision analysis, signal analysis and AI in which the geometry of X is sensible.

We studied the Wasserstein statistics using the framework of [29], proving that the Wasserstein covariance quantifies robustness against the convolutional waveform deformation due to observation noise. We further studied W-statistics of the affine deformation model. We showed F-efficiency and W-efficiency of estimators $\hat{\boldsymbol{\theta}}_F$ and $\hat{\boldsymbol{\theta}}_W$. We elucidated how the waveform f contributes to the efficiencies. The Gaussian distribution gives the only waveform in which the F-estimator and W-estimator coincide.

Other than the elliptically symmetric deformation model, it is difficult in general to derive the Wasserstein score, which corresponds to the infinitesimal optimal transport. It is an interesting future problem to explore other statistical models for which the Wasserstein score

is obtained in closed form. Also, it would be useful to develop approximation techniques for the Wasserstein score.

The present paper is only a first step to construct general Wasserstein statistics. In future work, we need to use more general statistical models. We also need to extend our approach to statistical theories of hypothesis testing, pattern classification, clustering and many other statistical problems based on the Wasserstein geometry.

Acknowledgements

We thank the referees for constructive comments. We thank Asuka Takatsu and Tomonari Sei for helpful comments. We thank Emi Namioka for drawing the figures. Takeru Matsuda was supported by JSPS KAKENHI Grant Numbers 19K20220, 21H05205, 22K17865 and JST Moonshot Grant Number JPMJMS2024.

References

[1] Amari, S. (2016). *Information Geometry and Its Applications*. Springer.

[2] 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.

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

[4] Amari, S. & Matsuda, T. (2022). Wasserstein statistics in one-dimensional location scale models *Annals of the Institute of Statistical Mathematics*, **74**, 33–47.

[5] Amari, S. & Nagaoka, H. (2016). *Methods of Information Geometry*. American Mathematical Society.

[6] Ambrosio, L., Gigli, N., & Savare, G. (2008). *Gradient Flows In Metric Spaces and in the Space of Probability Measures*. Springer.

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

[8] Ay, N. and Jost, J. and Vân Lê, H. & Schwachhöfer, L. (2017). *Information Geometry*. Springer.

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

[10] Benamou, J. D. & Brenier, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, **84**, 375–393.

[11] 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.

[12] Bhatia, R. (1997). *Matrix Analysis*. Springer.

[13] Bhatia, R., Jain, T. and Lim, Y. (2019). On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, **37**, 165–191.

[14] Brenier, Y. (1999). Minimal geodesics on groups of volume-preserving maps and generalized solutions of the Euler equations. *Comm. Pure Appl. Math.*, **52**, 411–452.

[15] Chentsov, N. (1982). *Statistical decision rules and optimal inference*. American Mathematical Society.

[16] Chen, Y., Lin, Z. & Müller, H. G. (2021). Wasserstein regression. *Journal of the American Statistical Association*, accepted.

[17] Chizat, L., Peyre, G., Schmitzer, B. & Vialard, F-X. (2018). An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, **18**, 1–44.

[18] 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).

[19] Gelbrich, M. (1990). On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematics Nachrichten* **147**, 185–203.

[20] Givens, C. R. & Shortt, R. M. (1984). A class of Wasserstein metrics for probability distributions. *Michigan Mathematics Journal* **31**, 231–240.

[21] Imaizumi, M., Ota, H. & Hamaguchi, T. (2022). Hypothesis test and confidence analysis with Wasserstein distance on general dimension. *Neural Computation* **34**. 1448–1487.

[22] Ito, S. (2023). Geometric thermodynamics for the Fokker–Planck equation: stochastic thermodynamic links between information geometry and optimal transport. *Information Geometry*.

[23] Khan, G. & Zhang, J. (2022). When optimal transport meats information geometry. *Information Geometry*, **5**, 47–78.

[24] Khan, G. & Zhang, J. (2020). The Kahler geometry of certain optimal transport problems. *Pure and Applied Analysis* **2**, 397–426.

[25] Khesin, B., Lenells, J., Misiolek, G. & Preston, S. C. (2013). Geometry of diffeomorphism groups, complete integrability and geometric statistics. *Geometric and Functional Analysis* **23**, 334–366.

[26] Kondratyev, S., Monsaingeon, L. & Vorotnikov, D. (2016). A new optimal transport distance on the space of finite Radon measures. *Advances in Differential Equations* **21**, 1117–1164.

[27] Kurose, T., Yoshizawa, S. & Amari, S. (2022). Optimal transportation plans with escort entropy regularization. *Information Geometry*, **5**, 79–95.

[28] Li, W. & Montúfar, G. (2020). Ricci curvature for parametric statistics via optimal transport. *Information Geometry*, **3**, 89-117.

[29] Li, W. & Zhao, J. (2023). Wasserstein information matrix. *Information Geometry*, **6**, 203–255.

[30] Liero, M., Mielke, A. & Savare, G. (2018). Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, **211**, 969–1117.

[31] Lott, J. (2008). Some geometric calculations on Wasserstein space. *Communications in Mathematical Physics*, **2**, 423–437.

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

[33] McCann, R. J. (1997). A convexity principle for interacting gases. *Advances in Mathematics*, **128**, 153–179.

[34] McCullagh, P. (2018). *Tensor Methods in Statistics*. Dover.

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

[36] Ollila, E. & Tyler, D. (2014). Regularized M-estimators of scatter matrix. *IEEE Transactions on Signal Processing*, **62**, 6059–6070.

[37] Otto, F. (2001). The geometry of dissipative evolution equations: The porous medium equation. *Commun. Partial Differ. Equations*, **26**, 101–174.

[38] Panaretos, V. M. & Zemel, Y. (2019). Statistical aspects of Wasserstein distances. arXiv:1806.05500.

[39] Panaretos, V. M. & Zemel, Y. (2022). *An Invitation to Statistics in Wasserstein Space*. Springer.

[40] Peyré, G. & Cuturi, M. (2019). Computational optimal transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, **11**, 355–607.

[41] Rankin, C. & Wong, TK, L. (2023). Bregman-Wasserstein divergence: geometry and applications. arXiv:2302.05833.

[42] Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*. Springer.

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

[44] Takatsu, A. & Yokota, T. (2012). Cone structure of L^2-Wasserstein spaces. *Journal of Topology and Analysis*, **4**, 237–253.

[45] Tyler, D. (1987). A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, **15**, 234–251.

[46] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[47] Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

[48] Villani, C. (2009). *Optimal Transport: Old and New*. Springer.

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

[50] Wong, T. K. L., & Yang, J. (2022). Pseudo-Riemannian geometry encodes information geometry in optimal transport. *Information Geometry*, **5**, 131–159.

[51] Yatracos, Y. G. (2022). Limitations of the Wasserstein MDE for univariate data. *Statistics and Computing volume*, **32**, 32–95.

A Proof of Wasserstein–Cramer–Rao inequality

For random vectors A and B,

$$0 \leq \mathrm{E}\|tA + B\|^2 = \mathrm{E}[\|A\|^2]t^2 + 2\mathrm{E}[A^\top B]t + \mathrm{E}[\|B\|^2]$$

for every t. Thus, by considering the discriminant of the quadratic equation,

$$\mathrm{E}[A^\top B]^2 \leq \mathrm{E}[\|A\|^2]\mathrm{E}[\|B\|^2]. \tag{23}$$

Substituting

$$A = \sum_i a_i \nabla_{\boldsymbol{x}} \hat{\theta}_i, \quad B = \sum_j b_j \nabla_{\boldsymbol{x}} S_j^W$$

into (23) yields

$$\left(\sum_{i,j} a_i b_j \mathrm{E}_\theta[(\nabla_{\boldsymbol{x}} \hat{\theta}_i)^\top (\nabla_{\boldsymbol{x}} S_j^W)] \right)^2 \leq \mathrm{E}_\theta \left[\left\| \sum_i a_i \nabla_{\boldsymbol{x}} \hat{\theta}_i \right\|^2 \right] \mathrm{E}_\theta \left[\left\| \sum_j b_j \nabla_{\boldsymbol{x}} S_j^W \right\|^2 \right]. \tag{24}$$

From the property (11) of the Wasserstein score function,

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} \mathrm{E}_\theta[\hat{\theta}_i] &= \int \hat{\theta}_i(x) \frac{\partial}{\partial \theta_j} p(x,\theta) \mathrm{d}x \\
&= -\int \hat{\theta}_i(x) \nabla_{\boldsymbol{x}} \cdot (p(x,\theta) \nabla_{\boldsymbol{x}} S_j^W(x,\theta)) \mathrm{d}x \\
&= -\int (\nabla_{\boldsymbol{x}} \cdot (\hat{\theta}_i(x) p(x,\theta) \nabla_{\boldsymbol{x}} S_j^W(x,\theta)) - (\nabla_{\boldsymbol{x}} \hat{\theta}_i(x))^\top (\nabla_{\boldsymbol{x}} S_j^W(x,\theta)) p(x,\theta)) \mathrm{d}x \\
&= \mathrm{E}_\theta[(\nabla_{\boldsymbol{x}} \hat{\theta}_i)^\top (\nabla_{\boldsymbol{x}} S_j^W)],
\end{aligned}$$

which yields

$$\sum_{i,j} a_i b_j \mathrm{E}_\theta[(\nabla_{\boldsymbol{x}} \hat{\theta}_i)^\top (\nabla_{\boldsymbol{x}} S_j^W)] = \sum_{i,j} a_i b_j \frac{\partial}{\partial \theta_j} \mathrm{E}_\theta[\hat{\theta}_i] = a^\top \left(\frac{\partial}{\partial \theta} \mathrm{E}_\theta[\hat{\theta}] \right) b. \tag{25}$$

From the definition of the Wasserstein covariance (21),

$$\begin{aligned}
\mathrm{E}_\theta\left[\left\|\sum_i a_i \nabla_{\boldsymbol{x}} \hat{\theta}_i\right\|^2\right] &= \mathrm{E}_\theta\left[\sum_{i,j} a_i a_j (\nabla_{\boldsymbol{x}} \hat{\theta}_i)^\top (\nabla_{\boldsymbol{x}} \hat{\theta}_j)\right] \\
&= \sum_{i,j} a_i a_j \mathrm{Var}_\theta^{\mathrm{W}}(\hat{\theta})_{ij} \\
&= a^\top \mathrm{Var}_\theta^{\mathrm{W}}(\hat{\theta}) a. \qquad (26)
\end{aligned}$$

From the definition of the Wasserstein information matrix (14),

$$\begin{aligned}
\mathrm{E}_\theta\left[\left\|\sum_j b_j \nabla_{\boldsymbol{x}} S_j^W\right\|^2\right] &= \mathrm{E}_\theta\left[\sum_{i,j} b_i b_j (\nabla_{\boldsymbol{x}} S_i^W)^\top (\nabla_{\boldsymbol{x}} S_j^W)\right] \\
&= \sum_{i,j} b_i b_j G_{\mathrm{W}}(\theta)_{ij} \\
&= b^\top G_{\mathrm{W}}(\theta) b. \qquad (27)
\end{aligned}$$

Substituting (25), (26) and (27) into (24), we obtain

$$\left(a^\top \left(\frac{\partial}{\partial \theta} \mathrm{E}_\theta[\hat{\theta}]\right) b\right)^2 \leq a^\top \mathrm{Var}_\theta^{\mathrm{W}}(\hat{\theta}) a \cdot b^\top G_{\mathrm{W}}(\theta) b.$$

By putting

$$b = G_{\mathrm{W}}(\theta)^{-1} \left(\frac{\partial}{\partial \theta} \mathrm{E}_\theta[\hat{\theta}]\right) a,$$

it leads to

$$a^\top \left(\frac{\partial}{\partial \theta} \mathrm{E}_\theta[\hat{\theta}]\right) G_{\mathrm{W}}(\theta)^{-1} \left(\frac{\partial}{\partial \theta} \mathrm{E}_\theta[\hat{\theta}]\right) a \leq a^\top \mathrm{Var}_\theta^{\mathrm{W}}(\hat{\theta}) a,$$

which is equal to the Wasserstein Cramer–Rao inequality (22).