Wasserstein statistics in one-dimensional location-scale models

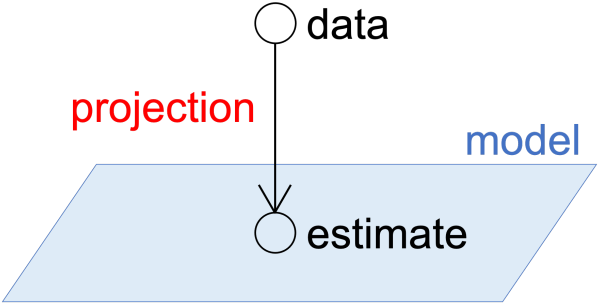
Shun-ichi Amari, Takeru Matsuda

RIKEN Center for Brain Science

GSI 2021

Abstract

- Many estimators can be interpreted as projection w.r.t. some divergence.
  - e.g. maximum likelihood estimator (MLE) = projection w.r.t. Kullback–Leibler divergence



- Here, we focus on projection w.r.t. Wasserstein distance (W-estimator) and study its property for one-dimensional location-scale models.

Problem setting

$$X_1, \ldots, X_n \sim p(x \mid \theta)$$

- task: estimate θ by $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$
- e.g. maximum likelihood estimate (MLE)

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i \mid \theta)$$

MLE = KL projection

- Kullback–Leibler divergence

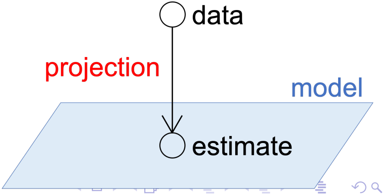
$$D_{\mathrm{KL}}(p_1, p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} \mathrm{d}x$$

- empirical distribution

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$

- MLE = KL projection ("m-projection" in information geometry)

$$\hat{\theta}_{\mathrm{MLE}} = \arg\min_{\theta} D_{\mathrm{KL}}(\hat{p}, p_\theta)$$

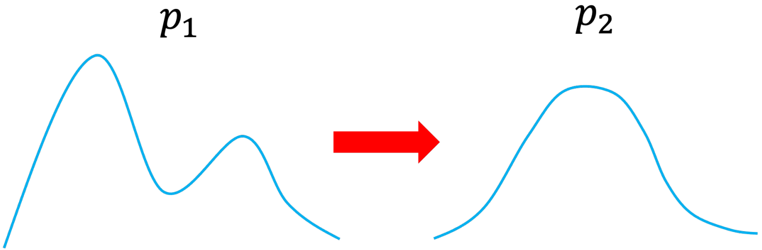


Wasserstein distance

- L^2 Wasserstein distance (= optimal transportation cost) between p_1 and p_2 on $\mathbb{R}^d$

$$W_2(p_1, p_2) = \inf_{X_1, X_2} \mathrm{E}(\|X_1 - X_2\|^2)^{1/2}$$

  - infimum over all joint distributions of (X_1, X_2) with $X_1 \sim p_1$ and $X_2 \sim p_2$ marginally (coupling)



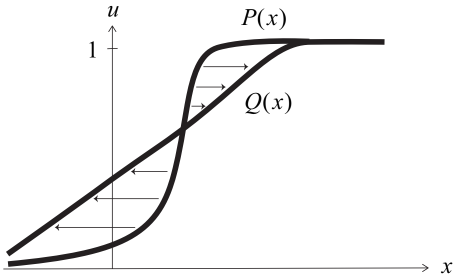
Wasserstein distance in one dimension

- When $d = 1$, W_2 is explicitly given by the cdf P_1 and P_2:

$$W_2(p_1, p_2) = \left(\int_0^1 (P_1^{-1}(u) - P_2^{-1}(u))^2 \mathrm{d}u \right)^{1/2}$$

- optimal coupling = monotone map

$$X_2 = P_2^{-1}(P_1(X_1))$$



W-estimator

- W-estimator = projection w.r.t. Wasserstein distance

$$\hat{\theta}_{\mathrm{W}} = \arg\min_{\theta} W_2(\hat{p}, p_\theta)$$

Kullback–Leibler	MLE
Wasserstein	W-estimator

- Statistical property of W-estimator has been only partially investigated.
  - cf. Bassetti et al. (2006), Montavon et al. (2015), Bernton et al. (2019)

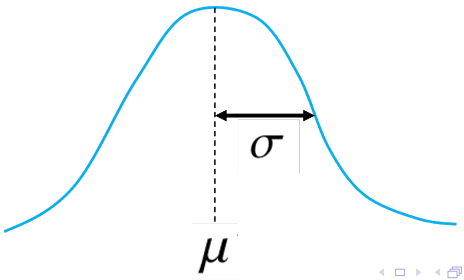
- Here, we focus on one-dimensional location-scale models.

One-dim. location-scale model

Definition

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

- $f(z)$: pdf with mean 0 and variance 1 (e.g. $\mathrm{N}(0,1)$)
  $\rightarrow p(x \mid \theta)$: mean μ, variance σ^2



W-estimator for one-dim. location-scale model

Theorem

$$\hat{\mu}_{\mathrm{W}} = \frac{1}{n}\sum_{i=1}^{n} x_{(i)}, \quad \hat{\sigma}_{\mathrm{W}} = \sum_{i=1}^{n} k_i x_{(i)},$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are order statistics of $x_1, \ldots, x_n$ and

$$k_i = \int_{z_{i-1}}^{z_i} z f(z) dz, \quad z_i = F^{-1}\left(\frac{i}{n}\right).$$

- $\hat{\mu}_{\mathrm{W}}$: arithmetic mean
- $\hat{\sigma}_{\mathrm{W}}$: linear combination of order statistics (L-statistics)

Proof

- Since the optimal coupling of $\hat{p}(x)$ and $p(x \mid \mu, \sigma)$ transports $x_{(i)}$ to $[\mu + \sigma z_{i-1}, \mu + \sigma z_i]$,

$$\begin{aligned} W_2^2(\hat{p}, p_{\mu,\sigma}) &= \sum_{i=1}^{n} \int_{\mu+\sigma z_{i-1}}^{\mu+\sigma z_i} (x - x_{(i)})^2 p(x \mid \mu, \sigma) \mathrm{d}x \\ &= \left(\mu^2 - \frac{2\mu}{n} \sum_{i=1}^{n} x_{(i)} \right) + \left(\sigma^2 - 2\sigma \sum_{i=1}^{n} k_i x_{(i)} \right) + \frac{1}{n} \sum_{i=1}^{n} x_{(i)}^2. \end{aligned}$$

- It is convex and minimized at

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_{(i)}, \quad \sigma = \sum_{i=1}^{n} k_i x_{(i)}.$$

Asymptotic distribution of W-estimator

Theorem

W-estimator is $\sqrt{n}$-consistent and

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

where

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

- proof: functional delta method (Donsker's theorem & L-statistics theory; van der Vaart, 1998)

Gaussian case

Corollary

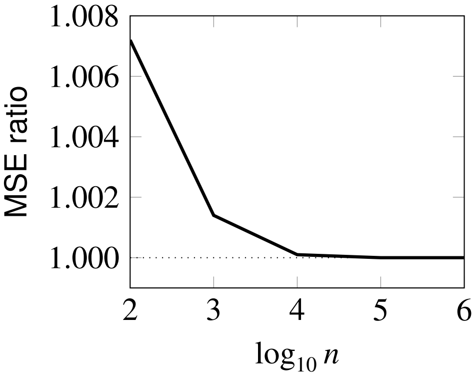
For the Gaussian model ($f(z) = \mathrm{N}(0,1)$), W-estimator is Fisher efficient (attains the Cramer–Rao bound):

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

- proof: $m_4 = 3$, $m_3 = 0$
- For general model, W-estimator is not Fisher efficient
  - MLE is Fisher efficient

Simulation result (Gaussian model)

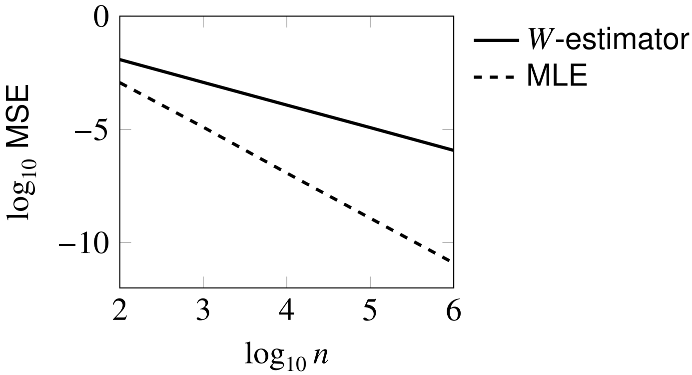
- (MSE of W-estimator) / (MSE of MLE) for Gaussian model
  - mean square error (MSE): $\mathrm{E}[(\hat{\mu} - \mu)^2 + (\hat{\sigma} - \sigma)^2]$



- The ratio converges to one as $n \to \infty$, which indicates that W-estimator is Fisher efficient

Simulation result (uniform model)

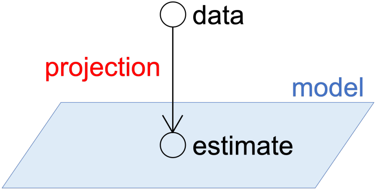
$$f(z) = \begin{cases} \frac{1}{2\sqrt{3}} & (-\sqrt{3} \le z \le \sqrt{3}) \\ 0 & (\text{otherwise}) \end{cases}$$



- W-estimator: $O(n^{-1/2})$, MLE: faster than $O(n^{-1/2})$

Summary

- W-estimator: projection w.r.t. Wasserstein distance



Kullback–Leibler	MLE
Wasserstein	W-estimator

- We derived the asymptotic distribution of W-estimator for one-dimensional location-scale models
  - Fisher efficient in Gaussian case

- future problem: advantage over MLE ?? other models ??