最適輸送と情報幾何

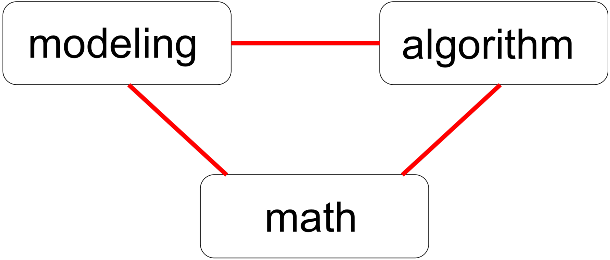
松田 孟留

東京大・情報理工，理研CBS・統計数理

2023年11月1日 @ IBIS2023

自己紹介

- 2017年3月：東京大・情報理工・数理情報・博士
- 2020年6月～：理研CBS（脳センター）統計数理研究ユニットリーダー
- 2022年10月～：東京大・情報理工・数理情報・准教授
- 研究分野：統計学・数理工学・神経科学
  - 統計的モデリング：データのモデリングと解析
  - 計算統計：データ解析のためのアルゴリズムの開発
  - 理論統計：データ解析の基礎数理



からくれなゐにみつくくるとは

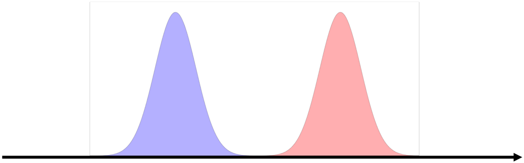
概要

- Wasserstein距離：確率分布間の最適輸送コスト
  - 台集合（確率変数が値をとる空間）の幾何構造を反映
- Kullback–Leiblerダイバージェンス：分布間の見分けやすさ
  - 台集合の変数変換について不変
  - 情報幾何の基礎（cf. Fisher情報量）
- 本講演：Wasserstein距離から誘導される統計モデルの幾何構造と統計的推測との関係について考察
  - Wasserstein距離に関する射影推定量 (Amari and M., 2022)
  - Wasserstein損失のもとでのベイズ予測 (M. and Strawderman, 2021)
  - Wasserstein–Cramer–Rao不等式とロバスト性 (Amari and M., 2023)

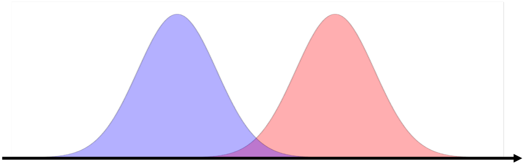
Wasserstein距離と Kullback–Leiblerダイバージェンス

クイズ

- $\mathrm{N}(\mu, \sigma^2)$：平均 μ, 分散 σ^2の正規分布
- $\mathrm{N}(-4, 1)$ v.s. $\mathrm{N}(4, 1)$



- $\mathrm{N}(-4, 4)$ v.s. $\mathrm{N}(4, 4)$



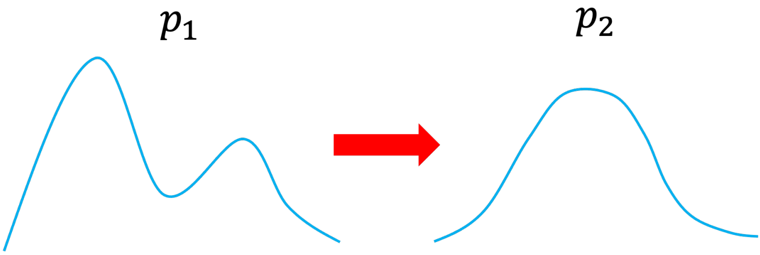
- どちらのペアの方が「近い」??

L^2-Wasserstein距離

- $\mathbb{R}^d$上の確率分布p_1, p_2のL^2-Wasserstein距離

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

  - infはX_1, X_2の周辺分布がp_1, p_2となる(X_1, X_2)の同時分布（カップリング）にわたる下限



1次元の場合 ($d = 1$)

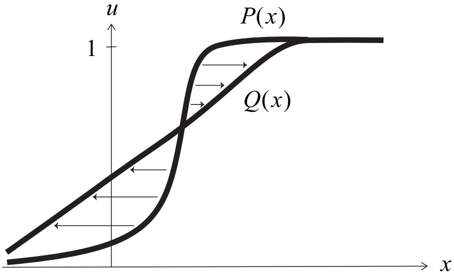
- 1次元ではW_2は累積分布関数P_1, P_2を用いて陽に書ける：

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

$$P_1(x) = \Pr[X_1 \leq x], \quad P_2(x) = \Pr[X_2 \leq x]$$

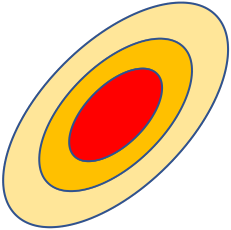
- 最適カップリング＝単調輸送写像

$$x \mapsto P_2^{-1}(P_1(x))$$



楕円対称分布族

- 一般に2次元以上ではW_2は計算困難。
- 計算できる例：楕円対称分布族
  - μ：平均，Σ：共分散，f：形
  - 例：多変量正規分布



$$p(x \mid \mu, \Sigma) = (\det \Sigma)^{-1/2} f(\|\Sigma^{-1/2}(x-\mu)\|)$$

Proposition (Gelbrich, 1990)

$$W_2(p(x \mid \mu_1, \Sigma_1), p(x \mid \mu_2, \Sigma_2)) = \left(\|\mu_1 - \mu_2\|^2 + \mathrm{tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right)\right)^{1/2}$$

- 注意：W_2 は f によらない！

Kullback–LeiblerダイバージェンスとFisher情報量

- 確率分布p_1, p_2のKullback–Leiblerダイバージェンス

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

- 局所的にはFisher情報量

$$D_{\mathrm{KL}}(p(x \mid \theta), p(x \mid \theta + \delta)) = \frac{1}{2} \delta^\top G_{\mathrm{F}}(\theta) \delta + o(\|\delta\|^2)$$

$$G_{\mathrm{F}}(\theta)_{ij} = \mathrm{E}_\theta \left[\frac{\partial}{\partial \theta_i} \log p(x \mid \theta) \frac{\partial}{\partial \theta_j} \log p(x \mid \theta) \right]$$

Cramer–Rao不等式

- データ $X \sim p(x \mid \theta)$ をもとに θ を推定

Cramer–Rao不等式

推定量 $\hat{\theta} = \hat{\theta}(x)$ が不偏 $(\mathrm{E}_\theta[\hat{\theta}] = \theta)$ のとき

$$\mathrm{Var}_\theta(\hat{\theta}) \succeq G_{\mathrm{F}}(\theta)^{-1}$$

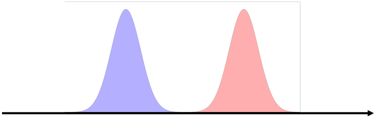
- Fisher情報量＝推定精度の限界＝分布の見分けやすさ
  - 情報が多い $\Leftrightarrow$ 見分けやすい
- 例：$X \sim \mathrm{B}(n, \theta)$ （二項分布：確率θのコイン投げn回）

$$G_{\mathrm{F}}(\theta) = \frac{n}{\theta(1-\theta)}, \quad \hat{\theta} = \frac{x}{n}$$

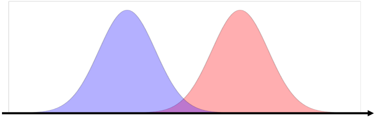
  - $n = 100, \theta = 0.1 \to \hat{\theta} = 0.1 \pm 0.03$
  - $n = 100, \theta = 0.5 \to \hat{\theta} = 0.5 \pm 0.05$
  - $\theta = 0.5$の方が推定しにくい：$G_{\mathrm{F}}(0.1) > G_{\mathrm{F}}(0.5)$

クイズ答

- $\mathrm{N}(-4, 1)$ v.s. $\mathrm{N}(4, 1)$



- $\mathrm{N}(-4, 4)$ v.s. $\mathrm{N}(4, 4)$



- Wasserstein距離だと同じ
  - 各点を右に8ずつ動かすのが最適
  - $W_2(\mathrm{N}(-4, 1), \mathrm{N}(4, 1)) = W_2(\mathrm{N}(-4, 4), \mathrm{N}(4, 4)) = 8$
- Kullback–Leiblerダイバージェンスだと下の方が「近い」
  - 下の方が見分けにくい
  - $D_{\mathrm{KL}}(\mathrm{N}(-4, 1), \mathrm{N}(4, 1)) = 64$, $D_{\mathrm{KL}}(\mathrm{N}(-4, 4), \mathrm{N}(4, 4)) = 16$

台集合の変数変換に関する不変性

- 一対一の変数変換

$$y = g(x) \quad \to \quad \tilde{p}(y) = \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| p(x)$$

- Kullback–Leiblerダイバージェンス：不変
  - 分布の見分けやすさは変数のとり方によらない

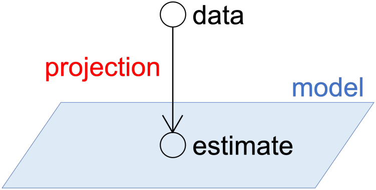
$$D_{\mathrm{KL}}(\tilde{p}, \tilde{q}) = D_{\mathrm{KL}}(p, q)$$

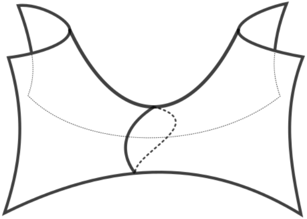
- Wasserstein距離：不変でない
  - 輸送コストは変数のとり方による

$$W_2(\tilde{p}, \tilde{q}) \neq W_2(p, q)$$

統計モデルの幾何

- 統計モデル $\{p(x \mid \theta)\}$ は多様体とみなせる
  - 例：正規分布 $\mathrm{N}(\mu, \sigma^2) \to$ 2次元多様体（曲面）





- Kullback–Leiblerダイバージェンスによって定まる幾何構造（長さ・曲率など）は統計的推測と密接に関連（情報幾何）

- Wasserstein距離だと??

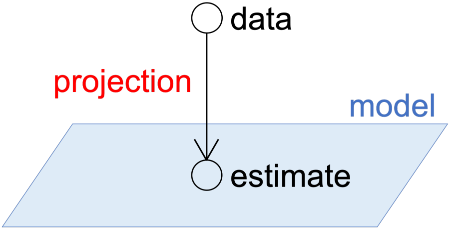
Estimation with Wasserstein distance

- projection w.r.t. Wasserstein distance

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

  - $\hat{p}$: empirical distribution
  - cf. projection w.r.t. Kullback–Leibler divergence = MLE

- Amari and M. (2022): asymptotic distribution of $\hat{\theta}_{\mathrm{W}}$ in 1d location-scale models
  - Fisher efficient in Gaussian case
  - 詳しくは付録スライド参照



Bayesian predictive density under Wasserstein loss

(M. and Strawderman, 2021)

Predictive density problem

$$X \sim p(x \mid \theta), \quad Y \sim p(y \mid \theta)$$

- predict Y based on X by predictive density $\hat{p}(y \mid x)$
- plug-in predictive density with estimate $\hat{\theta}(x)$

$$\hat{p}_{\mathrm{plug-in}}(y \mid x) = p(y \mid \hat{\theta}(x))$$

  - cf. AIC considers plug-in of MLE
- Bayesian predictive density with prior $\pi(\theta)$

$$\hat{p}_{\pi}(y \mid x) = \int p(y \mid \theta)\pi(\theta \mid x)\mathrm{d}\theta$$

- Which predictive density is better ??

Example: Gaussian

$$X \sim \mathrm{N}(\theta, \sigma^2), \quad Y \sim \mathrm{N}(\theta, \tau^2)$$

- plug-in predictive density with MLE

$$\hat{p}_{\mathrm{plug-in}}(y \mid x) = \mathrm{N}(x, \tau^2)$$

- Bayesian predictive density with uniform prior $\pi(\theta) \equiv 1$

$$\hat{p}_{\mathrm{U}}(y \mid x) = \mathrm{N}(x, \sigma^2 + \tau^2)$$

- Bayesian predictive density has larger variance due to the uncertainty of θ

Prediction under Kullback–Leiber loss

- Kullback-Leibler loss

$$D_{\mathrm{KL}}(p(y \mid \theta), \hat{p}(y \mid x)) = \int p(y \mid \theta) \log \frac{p(y \mid \theta)}{\hat{p}(y \mid x)} \mathrm{d}y$$

Proposition (Aitchison, 1975)

Bayesian predictive density minimizes Bayes risk:

$$p_\pi(y \mid x) = \arg\min_{\hat{p}} \int \mathrm{E}_\theta[D_{\mathrm{KL}}(p(y \mid \theta), \hat{p}(y \mid x))] \pi(\theta) \mathrm{d}\theta$$

where

$$\mathrm{E}_\theta[D_{\mathrm{KL}}(p(y \mid \theta), \hat{p}(y \mid x))] = \int D_{\mathrm{KL}}(p(y \mid \theta), \hat{p}(y \mid x)) p(x \mid \theta) \mathrm{d}x$$

Example: Gaussian

$$X \sim \mathrm{N}(\theta, \sigma^2), \quad Y \sim \mathrm{N}(\theta, \tau^2)$$

$$\hat{p}_{\mathrm{plug-in}}(y \mid x) = \mathrm{N}(x, \tau^2)$$

$$\hat{p}_{\mathrm{U}}(y \mid x) = \mathrm{N}(x, \sigma^2 + \tau^2)$$

- Kullback–Leibler risk

$$\mathrm{E}_\theta[D_{\mathrm{KL}}(p(y \mid \theta), \hat{p}_{\mathrm{plug-in}}(y \mid x))] = \frac{\sigma^2}{2\tau^2}$$

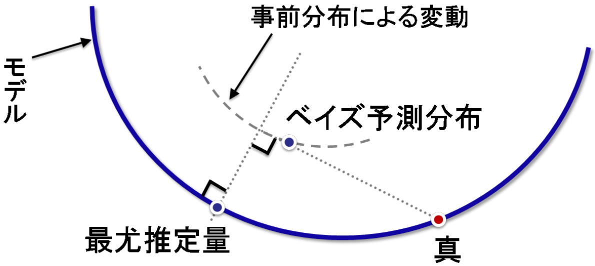
$$\mathrm{E}_\theta[D_{\mathrm{KL}}(p(y \mid \theta), \hat{p}_{\mathrm{U}}(y \mid x))] = \frac{1}{2} \log \left(1 + \frac{\sigma^2}{\tau^2} \right)$$

$\rightarrow$ Bayesian predictive density has smaller risk

Geometry of Bayesian prediction under KL loss

- Komaki (1996): information geometry of Bayesian prediction
- optimal shift from model = m-curvature
- Bayesian predictive density attains optimal shift

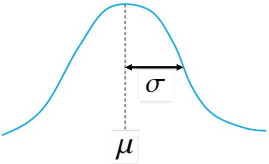
  $\rightarrow$ For curved model, Bayes is better than plug-in



Prediction under Wasserstein loss

- location-scale model

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



Theorem (M. and Strawderman, 2021)

Plug-in predictive density with posterior mean minimizes Bayes risk:

$$p(y \mid \hat{\theta}_\pi(x)) = \arg\min_{\hat{p}} \ \mathrm{E}_\theta[W_2(p(y \mid \theta), \hat{p}(y \mid x))^2]$$

$$\hat{\theta}_\pi(x) = \int \theta \pi(\theta \mid x) \mathrm{d}\theta$$

- no shift $\to$ location-scale model is “flat"
  - Indeed, location-scale model is Euclidean (totally geodesic) in L^2-Wasserstein geometry

Wasserstein–Cramer–Rao inequality and robustness

(Amari and M., 2023)

Li–Zhao framework

- Recently, Li and Zhao (2023) developed Wasserstein counterparts of information geometric concepts

Kullback–Leibler divergence	Wasserstein distance
Fisher score	Wasserstein score
Fisher information matrix	Wasserstein information matrix
covariance	Wasserstein covariance
Cramer–Rao	Wasserstein–Cramer–Rao
Fisher efficiency	Wasserstein efficiency

- We investigate their statistical meaning

Continuity equation

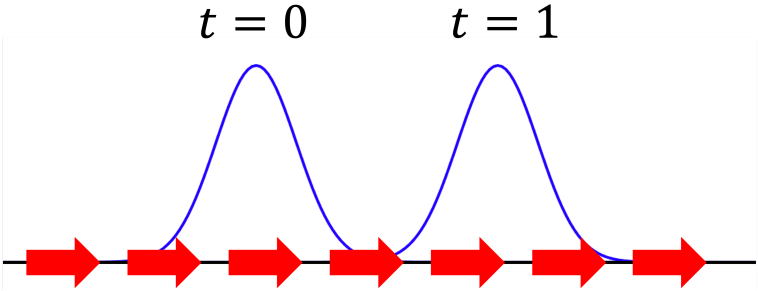
$$\frac{\partial}{\partial t} p(x,t) = -\nabla_x \cdot (p(x,t)\nabla_x \Phi(x))$$

- This PDE describes dynamics of measure transport
- intuition: Many particles are distributed with $p(x,t)$ and they move with velocity $\nabla_x \Phi(x)$

Example: 1d linear potential

$$\frac{\partial}{\partial t}p(x,t) = -\nabla_x \cdot (p(x,t)\nabla_x \Phi(x))$$

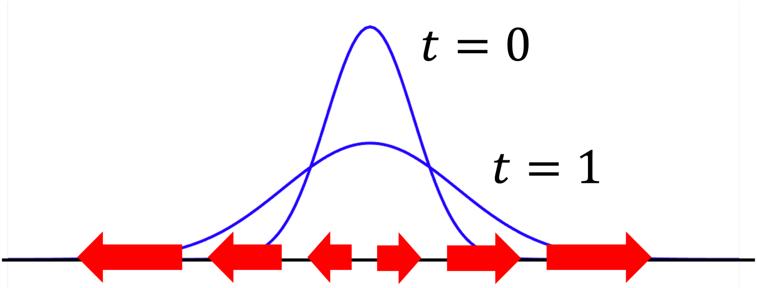
- $\Phi(x) = x \to \nabla_x \Phi(x) \equiv 1$ (const.)
- $p(x,0) = \mathrm{N}(0,1) \to p(x,t) = \mathrm{N}(t,1)$ (shift)



Example: 1d quadratic potential

$$\frac{\partial}{\partial t} p(x,t) = -\nabla_x \cdot (p(x,t) \nabla_x \Phi(x))$$

- $\Phi(x) = x^2 \rightarrow \nabla_x \Phi(x) = 2x$
- $p(x,0) = \mathrm{N}(0,1) \rightarrow p(x,t) = \mathrm{N}(0,t+1)$ (expansion)



Wasserstein score function

Definition (Li and Zhao, 2023)

For $i = 1, \ldots, p$, the Wasserstein score function $\Phi_i^{\mathrm{W}}(x \mid \theta)$ is the solution of

$$-\nabla_x \cdot (p(x \mid \theta) \nabla_x \Phi_i^{\mathrm{W}}(x \mid \theta)) = \frac{\partial}{\partial \theta_i} p(x \mid \theta), \quad \mathrm{E}_\theta[\Phi_i^{\mathrm{W}}(x \mid \theta)] = 0.$$

- For infinitesimal δ, the map $x \mapsto x + \delta \nabla_x \Phi_i^{\mathrm{W}}(x \mid \theta)$ is the optimal transport map from $p(x \mid \theta)$ to $p(x \mid \theta + \delta e_i)$ with transportation cost

$$W_2(p(x \mid \theta), p(x \mid \theta + \delta e_i)) = \left(\int \| \delta \nabla_x \Phi_i^{\mathrm{W}}(x \mid \theta) \|^2 p(x \mid \theta) \mathrm{d}x \right)^{1/2}$$

  - e_i: i-th standard unit vector

Wasserstein information matrix (WIM)

Definition (Li and Zhao, 2023)

The Wasserstein information matrix $G_{\mathrm{W}}(\theta)$ is the $p \times p$ matrix given by

$$G_{\mathrm{W}}(\theta) = \left(\int \frac{\partial}{\partial \theta_i} p(x \mid \theta) \cdot \Phi_j^{\mathrm{W}}(x \mid \theta) \mathrm{d}x \right)_{ij}$$

- cf. Fisher information matrix

$$G_{\mathrm{F}}(\theta) = \left(\int \frac{\partial}{\partial \theta_i} p(x \mid \theta) \cdot \Phi_j^{\mathrm{F}}(x \mid \theta) \mathrm{d}x \right)_{ij}$$

$$\Phi_j^{\mathrm{F}}(x \mid \theta) = \frac{\partial}{\partial \theta_j} \log p(x \mid \theta)$$

- inner product = pairing of tangent vector and cotangent vector
  - information geometry: m-representation and e-representation

Wasserstein information matrix (WIM)

Proposition (Li and Zhao, 2023)

$$G_{\mathrm{W}}(\theta)_{ij} = \mathrm{E}_\theta[(\nabla_x \Phi_i^{\mathrm{W}}(x \mid \theta))^\top (\nabla_x \Phi_j^{\mathrm{W}}(x \mid \theta))]$$

Proposition (Li and Zhao, 2023)

$$W_2(p(x \mid \theta), p(x \mid \theta + \delta))^2 = \delta^\top G_W(\theta)\delta + o(\|\delta\|^2)$$

- WIM = Hessian of Wasserstein distance
  - cf. Fisher information matrix = Hessian of Kullback–Leibler divergence
- WIM appears in Otto calculus and Wasserstein gradient flow

Example: 1d Gaussian

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad \theta = (\mu, \sigma)$$

- Wasserstein distance

$$W_2(p(x \mid \theta_1), p(x \mid \theta_2))^2 = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2$$

- Wasserstein score function

$$\Phi_\mu^{\mathrm{W}}(x \mid \theta) = x - \mu, \quad \Phi_\sigma^{\mathrm{W}}(x \mid \theta) = \frac{(x-\mu)^2 - \sigma^2}{2\sigma}$$

- Wasserstein information matrix

$$G_{\mathrm{W}}(\theta) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

- More generally, 1d location-scale model is Euclidean (totally geodesic) in L^2-Wasserstein geometry

Wasserstein estimator

Definition (Li and Zhao, 2023)

The Wasserstein estimator $\hat{\theta}_{\mathrm{W}}(x)$ is the zero of the Wasserstein score function:

$$\Phi_i^{\mathrm{W}}(x \mid \hat{\theta}_{\mathrm{W}}(x)) = 0, \quad i = 1, \dots, p$$

- cf. MLE = zero of the Fisher score function = projection w.r.t. Kullback–Leibler divergence
- What does it mean??
  - It is different from the projection w.r.t. Wasserstein distance studied in Amari and M. (2022)

Elliptically contoured family

$$p(x \mid \mu, \Sigma) = (\det \Sigma)^{-1/2} f(\|\Sigma^{-1/2}(x-\mu)\|)$$

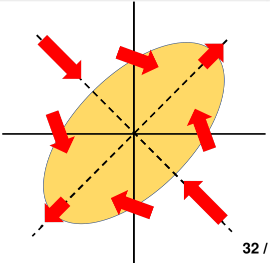
Theorem (Amari and M., 2023)

- Wasserstein score functions are quadratic
- Wasserstein estimator = 2nd-order moment estimator

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$$

e.g. 2d Gaussian $\mathrm{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}\right)$

$$\Phi^{\mathrm{W}}(x \mid \theta) = \frac{1}{4}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^\top \begin{pmatrix} -\theta & 1 \\ 1 & -\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$



Wasserstein covariance & Wasserstein–Cramer–Rao

Definition (Li and Zhao, 2023)

The Wasserstein covariance $\mathrm{Var}_\theta^{\mathrm{W}}[\hat{\theta}]$ of an estimator $\hat{\theta}$ is the $p \times p$ positive semidefinite matrix given by

$$\mathrm{Var}_\theta^{\mathrm{W}}[\hat{\theta}] = (\mathrm{E}_\theta[(\nabla_x \hat{\theta}_i)^\top (\nabla_x \hat{\theta}_j)])_{ij}$$

Theorem (Li and Zhao, 2023)

When $\hat{\theta}$ is unbiased ($\mathrm{E}_\theta[\hat{\theta}] = \theta$),

$$\mathrm{Var}_\theta^{\mathrm{W}}(\hat{\theta}) \succeq G_{\mathrm{W}}(\theta)^{-1}$$

- What does it mean??
  - cf. usual Cramer–Rao = lower bound of mean squared error

Wasserstein covariance and robustness

$$X \sim p(x \mid \theta), \quad Z \sim q(z)$$

- We consider estimation of θ from noisy observation $X + Z$
  - $\mathrm{E}[Z] = 0$, $\mathrm{Var}[Z] = \sigma^2 I$

Theorem (Amari and M., 2023)

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

Wasserstein covariance and robustness

Corollary (Amari and M., 2023)

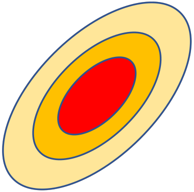
If $\hat{\theta}$ is quadratic,

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

- Thus, Wasserstein covariance quantifies the robustness against additive noise of quadratic estimators.
- e.g. Wasserstein estimator for elliptically contoured family

$$p(x \mid \mu, \Sigma) = (\det \Sigma)^{-1/2} f(\|\Sigma^{-1/2}(x-\mu)\|)$$

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top$$



- “additive noise": not invariant w.r.t. transformation of x
  - noise contamination $\approx$ (random) transportation

まとめと今後の課題

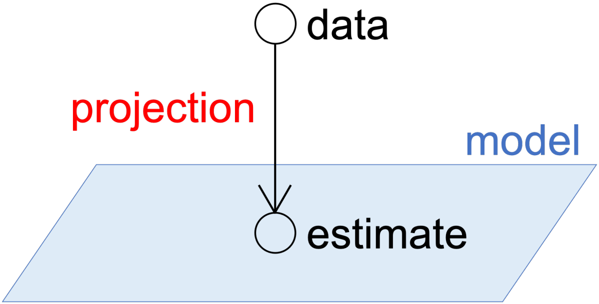
- Wasserstein距離：確率分布間の最適輸送コスト
  - 台集合（確率変数が値をとる空間）の幾何構造を反映
- Kullback–Leiblerダイバージェンス：分布間の見分けやすさ
  - 台集合の変数変換について不変
  - 情報幾何の基礎（cf. Fisher情報量）
- Wasserstein距離から誘導される統計モデルの幾何構造と統計的推測との関係について考察
- 指数型分布族・双対接続のWasserstein版??
  - 例：Cramer–Raoの下限を達成可能 $\Leftrightarrow$ 指数型分布族の期待値パラメータ（m座標）

Wasserstein statistics in one-dimensional location-scale models

(Amari and M., 2022)

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)

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

MLE = KL projection

- Kullback–Leibler divergence

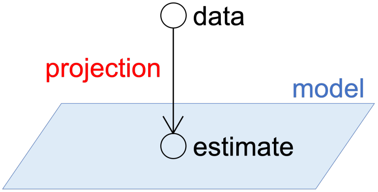
$$D_{\mathrm{KL}}(p_1, p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} \mathrm{d}x$$

- empirical distribution

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

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

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



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)

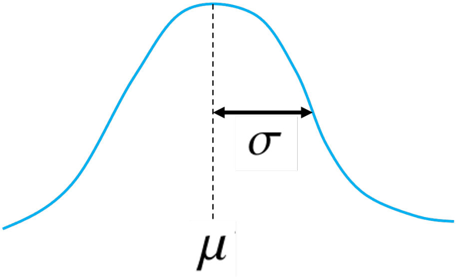
- Here, we focus on one-dimensional location-scale models.

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. $\mathrm{N}(0,1)$)
  $\to p(x \mid \theta)$: 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 $[\mu + \sigma z_{i-1}, \mu + \sigma z_i]$,

$$W_2^2(\hat{p}, p_{\mu,\sigma}) = \sum_{i=1}^{n} \int_{\mu+\sigma z_{i-1}}^{\mu+\sigma z_i} (x - x_{(i)})^2 p(x \mid \mu, \sigma) \mathrm{d}x$$

$$= \left(\mu^2 - \frac{2\mu}{n} \sum_{i=1}^{n} x_{(i)}\right) + \left(\sigma^2 - 2\sigma \sum_{i=1}^{n} k_i x_{(i)}\right) + \frac{1}{n} \sum_{i=}^{n}$$

- 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

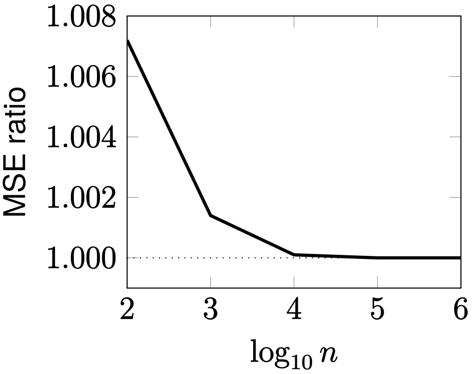
For the Gaussian model ($f(z) = \mathrm{N}(0,1)$), W-estimator is Fisher efficient (attains the Cramer–Rao bound):

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

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

Simulation result (Gaussian model)

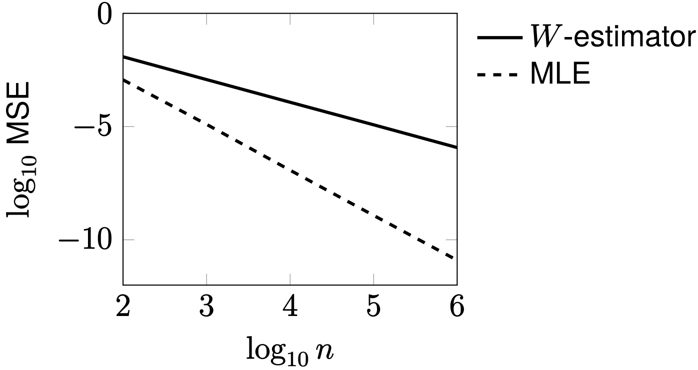
- (MSE of W-estimator) / (MSE of MLE) for Gaussian model
  - mean square error (MSE): $\mathrm{E}[(\hat{\mu}-\mu)^2+(\hat{\sigma}-\sigma)^2]$



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

Simulation result (uniform model)

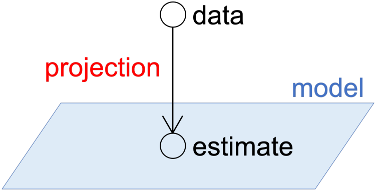
$$f(z) = \begin{cases} \frac{1}{2\sqrt{3}} & (-\sqrt{3} \leq z \leq \sqrt{3}) \\ 0 & (\text{otherwise}) \end{cases}$$



- 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 ??