Geometric Data Analysis Reading Group

Robust Optimization and Inference on Manifolds

Paper Authors:
Lizhen Lin, Drew Lazar, Bayan Sarpabayeva, David B. Dunson

Paper link:
https://arxiv.org/abs/2006.06843

Presented by *Yikun Zhang*
May 2, 2022

Table of Contents

Background: Mean Estimation and Robust Statistics

Do We Know How to Estimate the Mean?[1]

Problem: Given a random sample $\{X_1, ..., X_n\} \sim P$, consider estimating the population mean $\mu = \mathbb{E}_P(X_i) = \int x \, dP$.

- We want to construct an estimator $\widehat{\mu}_n \equiv \widehat{\mu}_n(X_1, ..., X_n)$.

Most popular estimator: the sample mean $\bar{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \int x \, dP_n$,

where P_n is the empirical distribution.

Consistency: By the (strong) law of large number,

$$\lim_{n \to \infty} \bar{\mu}_n = \mu \quad \text{with probability one.}$$

Drawbacks:

- Require strict assumptions on P for tight confidence bounds.
- Sensitive to outliers.
- ...

[1] The first few slides are modified from the Breiman Lecture of NeurIPS 2021 delivered by Gabor Lugosi (https://nips.cc/virtual/2021/invited-talk/22279).

Mean Estimation: Confidence Bounds

Question

Given a confidence level $\delta \in (0,1)$, what is the smallest $\epsilon \equiv \epsilon(n,\delta)$ such that

$$\|\widehat{\mu}_n - \mu\| \leq \epsilon \quad \textit{with probability at least } 1-\delta \quad ?$$

Consider the sample mean $\bar{\mu}_n$:

- If we know $\sigma^2 = \mathbb{E}_P(X_i - \mu)^2 < \infty$, then by Chebyshev's inequality,

$$|\bar{\mu}_n - \mu| \leq \sigma\sqrt{\frac{1}{n\delta}} \quad \text{with probability at least } 1-\delta. \tag{1}$$

- If P is sub-Gaussian, i.e., $\mathbb{E}_P \exp\left[\lambda(X-\mu)\right] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right)$, then

$$|\bar{\mu}_n - \mu| \leq \sigma\sqrt{\frac{2\log(2/\delta)}{n}} \quad \text{with probability at least } 1-\delta. \tag{2}$$

Mean Estimation: Heavy-Tailed Distribution

Theorem (Theorem 1 in Lugosi and Mendelson 2019a)

Let $n > 5$ be an integer, $\sigma > 0$, and $\delta \in \left(\frac{e^{-n}}{2}, \frac{1}{2}\right)$. Then, for any mean estimator $\widehat{\mu}_n$, there exists a distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ such that

$$\mathbb{P}\left\{|\widehat{\mu}_n - \mu| > \sigma\sqrt{\frac{\log\left(1/(2\delta)\right)}{4n}}\right\} \geq \delta.$$

Without sub-Gaussianity (i.e., P is heavy-tailed), the $\sqrt{1/\delta}$-bound is the best that $\bar{\mu}_n$ can achieve:

- for any $\delta \in (0, 1)$, there is a distribution with variance σ^2 such that

$$\mathbb{P}\left(|\bar{\mu}_n - \mu| > \sigma\sqrt{\frac{C}{n\delta}}\right) > \delta \quad \text{for some constant } C > 0.$$

Question: Is there any estimator $\widehat{\mu}_n$ that can achieve the (sub-gaussian) $\sqrt{\log(1/\delta)}$-bound (2) for all distributions with finite variance?

Mean Estimation: Median-of-Means (MoM)

Median-of-Means (Nemirovskij and Yudin, 1983; Jerrum et al., 1986; Alon et al., 1999): Divide the random sample $\{X_1, ..., X_n\}$ into m groups $B_1, ..., B_m$ with (roughly) equal size $B = \lfloor \frac{n}{m} \rfloor$ and define

$$\widehat{\mu}_{MM} \equiv \texttt{Median}\left(Z_1, ..., Z_m\right), \tag{3}$$

where $Z_i = \frac{1}{|B_i|} \sum_{j \in B_i} X_j$ for $i = 1, ..., m$.

- The MoM estimator is consistent as long as $B \to \infty$ as $n \to \infty$.
- For any $\delta \in (0, 1)$, if $m = \lfloor 8 \log(1/\delta) \rfloor$, then

$$|\widehat{\mu}_{MM} - \mu| \leq \sigma \sqrt{\frac{32 \log(1/\delta)}{n}} \quad \text{with probability at least } 1 - \delta.$$

See Theorem 2 in Lugosi and Mendelson (2019a) and Proposition 1 in Yen-Chi's notes (http://faculty.washington.edu/yenchic/short_note/note_MoM.pdf).

Discussions on the MoM Estimator

• The MoM estimator attains the (sub-gaussian) $\sqrt{\log(1/\delta)}$-bound (2) for all distributions with finite variance, and this bound is *sharp*.

- One undesirable point is that the number of blocks $m = \lfloor 8\log(1/\delta) \rfloor$ depends on the confidence level $\delta \in (0,1)$.
- However, if $\tau = \mathbb{E}_P\left[(X_i - \mu)^3\right] < \infty$ exists, we may take $m = \frac{2\sigma^3}{\tau}\sqrt{n}$ to achieve the sub-Gaussian performance; see Theorem 4 in Lugosi and Mendelson (2019a).
- Other mean estimators that attain the sub-Gaussian bound include
  - Catoni's estimator (Catoni, 2012): the solution to $\sum_{i=1}^{n} \Psi\left(\alpha(X_i - y)\right) = 0$, where $\Psi : \mathbb{R} \to \mathbb{R}$ is an increasing odd function.
  - Trimmed mean (Tukey and McLaughlin, 1963): $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \phi_{\alpha,\beta}(X_i)$ with

$$\phi_{\alpha,\beta}(x) = \begin{cases} \alpha & \text{if } x < \alpha, \\ x & \text{if } \alpha \le x \le \beta, \\ \beta & \text{if } x > \beta. \end{cases}$$

• MoM can be used even if P only has a finite moment $\mathbb{E}_P\left[|X_i - \mu|^{1+\gamma}\right]$ of order $1+\gamma$ with $\gamma \in (0,1)$ (Bubeck et al., 2013; Devroye et al., 2016).

Robustness of the MoM Estimator

More importantly, the MoM estimator is robust to outliers!

- Consider a set $\mathcal{D}_s = \{\mathcal{Y} = \{Y_1, ..., Y_n\} : |\boldsymbol{Y}| = n, |\boldsymbol{X} \cap \boldsymbol{Y}| = n - s\}$. The robustness of $\widehat{\mu}_n(\mathcal{X})$ with $\mathcal{X} = \{X_1, ..., X_n\}$ can be measured by the *breakdown point* as (Huber, 2004):

$$\epsilon^* \left(\widehat{\mu}_n(\mathcal{X})\right) = \max \left\{ \frac{s}{n} : ||\widehat{\mu}_n(\mathcal{Y})|| < \infty \text{ for all } \mathcal{Y} \in \mathcal{X}_s \right\}.$$

- For instance, the sample mean has a breakdown point of 0 while the median has a breakdown point of $1/2$.

- The MoM estimator $\widehat{\mu}_{MM}(\boldsymbol{X})$ has the breakdown point as $\frac{m-1}{2n}$, where m is the number of blocks (Rodriguez and Valdora, 2019).

Multivariate Mean Estimation

Let $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ be an i.i.d. sample in $\mathbb{R}^d$ with $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X}_i)$ and $\Sigma = \mathbb{E}\left[(\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^T\right]$.

- The sample mean $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$ does not have a sub-Gaussian behavior for non-Gaussian and possibly heavy-tailed distributions.

Definition

We say that a mean estimator $\widehat{\boldsymbol{\mu}}_n$ is *sub-Gaussian* if, for $\delta \in (0, 1)$,

$$||\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}|| \leq \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{2\lambda_{\max}\log(1/\delta)}{n}} \quad \text{with probability at least } 1-\delta,$$

where $\lambda_{\max}$ is the maximal eigenvalue of Σ and $||\cdot||$ is the Euclidean norm in $\mathbb{R}^d$.

Question: Can the (multivariate) MoM estimator attain the above sub-Gaussian bound?

Multivariate Median-of-Means

There is no standard notion of a median for multivariate data!

We partition the dataset $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \mathbb{R}^d$ into m groups $U_1, ..., U_m$ and compute the within-group means $\boldsymbol{Z}_i = \frac{1}{|U_i|} \sum_{j \in U_i} \boldsymbol{X}_j$.

- *Coordinate-wise median*: for any $\delta \in (0, 1)$, take $m = \lfloor 8 \log(1/\delta) \rfloor$,

$$||\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}|| \leq \sqrt{\frac{32 \mathrm{Tr}(\Sigma) \log(d/\delta)}{n}} \quad \text{with probability at least } 1 - \delta.$$

- ***Geometric median***: $\widehat{\boldsymbol{\mu}}_n \equiv \arg\min_{\boldsymbol{p} \in \mathbb{R}^d} \frac{1}{m} \sum_{j=1}^{m} \left|\left|\boldsymbol{p} - \boldsymbol{Z}_j\right|\right|$. (It is close to the sub-Gaussian bound.)
- The estimators that truly yield the sub-Gaussian performance are
  1. *Catoni-Giulini estimator* (Catoni and Giulini, 2018):
     $\widehat{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \cdot \min\left(1, \frac{1}{\alpha ||\boldsymbol{X}_i||}\right)$ with tuning parameter $\alpha > 0$.
  2. the *median-of-means tournaments* (Lugosi and Mendelson, 2019b).

Geometric Median on Manifolds

Geometric Median on Manifolds

For a metric space $(\mathcal{M}, \rho)$, the *geometric median* p^* of $p_1, ..., p_m \in \mathcal{M}$ minimizes the sum of distances to the points (Minsker, 2015):

$$p^* = \text{med}(p_1, ..., p_m) = \arg\min_{p \in \mathcal{M}} \frac{1}{m} \sum_{k=1}^{m} \rho(p, p_k). \tag{4}$$

assuming that p^* exists. It is unique (Theorem 1 in Fletcher et al. 2008)
(i) if the sectional curvatures of $\mathcal{M}$ is nonpositive or
(ii) if the section curvatures of $\mathcal{M}$ are bounded by $\Delta > 0$ and $\text{diam}(p_1, ..., p_m) \leq \frac{\pi}{2\sqrt{\Delta}}$.

When $\mathcal{M}$ is a manifold, there are two different ways to define ρ.

1. (*Extrinsic distance*) Given an embedding $J : \mathcal{M} \to \mathbb{R}^d$ into the ambient space $\mathbb{R}^d$,

   $$\rho(p, q) = ||J(p) - J(q)|| \quad \text{with } ||\cdot|| \text{ being the Euclidean norm in } \mathbb{R}^d.$$

2. (*Intrinsic distance*) Take ρ as the geodesic distance arising from a Riemannian structure on $\mathcal{M}$.

Estimating the Geometric Median

To compute $p^* = \arg\min_{p \in \mathcal{M}} \sum_{k=1}^{m} \rho(p, p_k) \equiv h(p)$, we leverage the Ostresh's modification of the Weiszfeld Algorithm (Weiszfeld, 1937; Ostresh Jr, 1978; Fletcher et al., 2008):

1. Compute the (Riemannian) gradient

$$\nabla h(p) = -\sum_{k=1}^{m} \frac{\mathrm{Log}_p(p_k)}{\rho(p, p_k)} \quad \text{when } p \neq p_k.$$

2. Apply the gradient descent iteration

$$p^{(t+1)} \leftarrow \mathrm{Exp}_{p^{(t)}}\left(\eta' \cdot v^{(t)}\right) \text{ with } v^{(t)} = \sum_{k \in I_t} \frac{\mathrm{Log}_{p^{(t)}}(p_k)}{\rho(p^{(t)}, p_k)} \cdot \left(\sum_{k \in I_t} \frac{1}{\rho(p^{(t)}, p_k)}\right)^{-1}$$

where $\eta' \in [0, 2]$ is the step size and $I_t = \left\{k \in \{1, ..., m\} : p_k \neq p^{(t)}\right\}$.

Convergence: $\lim_{t \to \infty} p^{(t)} = p^*$ when $\mathcal{M}$ has a nonnegative sectional curvature.

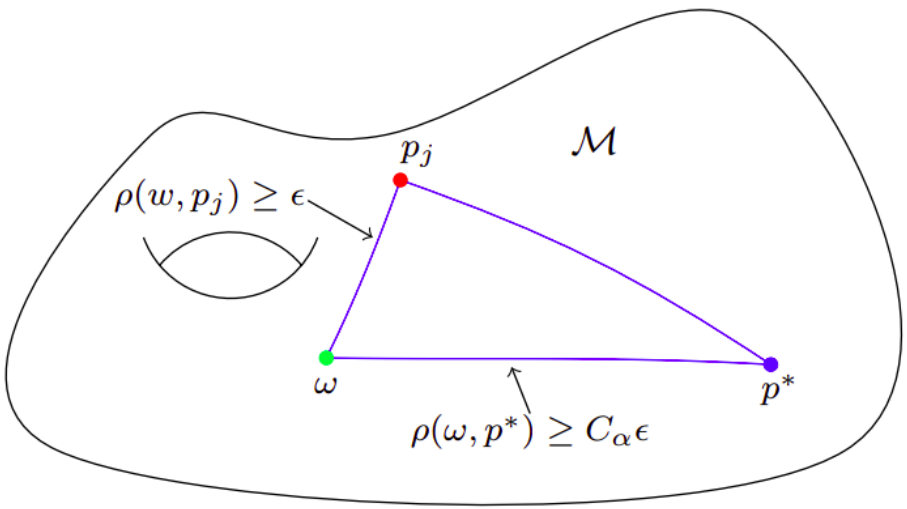
Properties of the Geometric Median

Lemma (Lemma 2.1 in Minsker 2015; Lin et al. 2020)

Let $p_1, ..., p_m \in \mathcal{M}$ and $p^* = med(p_1, ..., p_m)$ as in (4).
(a) Let ρ be the extrinsic distance of an embedding $J : \mathcal{M} \to \widetilde{\mathcal{M}} \subset \mathbb{R}^d$, $w \in \mathcal{M}$, ψ be the angle between $J(w) - J(p^*)$ and the tangent space $T_{J(p^*)}\widetilde{\mathcal{M}}$, and

$$C_\alpha = \frac{1-\alpha}{\sqrt{1-2\alpha}\cos\psi - \alpha\sin\psi} \quad \text{with} \quad \alpha \in \left(0, \cot\psi \tan\frac{\psi}{2}\right).$$

If $\rho(w, p^*) \geq C_\alpha \epsilon$, then there exists an index set $T \subset \{1, ..., m\}$ with $|T| \geq \alpha m$ such that $\rho(p_j, w) \geq \epsilon$ for any $j \in T$.



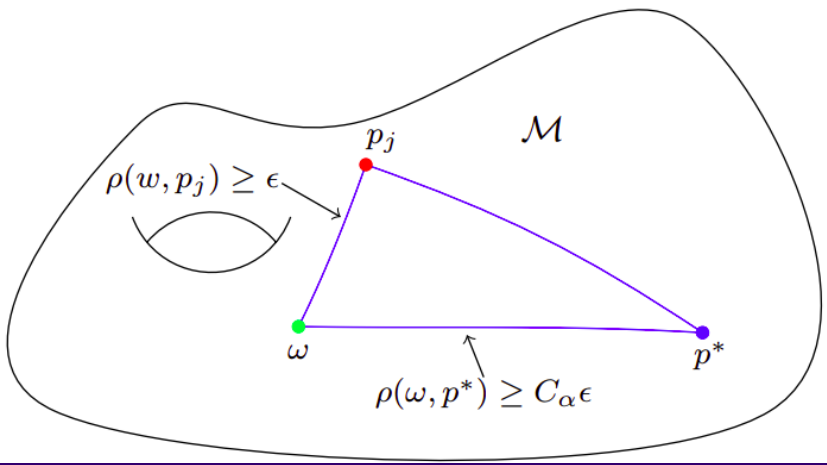
Properties of the Geometric Median (Cont'd)

Lemma (Continued)

Let $p_1, ..., p_m \in \mathcal{M}$ and $p^* = med(p_1, ..., p_m)$ as in (4).
(b) Let ρ be an intrinsic distance on $\mathcal{M}$ with respect to some Riemannian structure, $w \in \mathcal{M}$, the logarithm map Log_{p^*} be K-Lipschitz continuous from $B(w, \epsilon)$ to $T_{p^*}\mathcal{M}$, and

$$C_\alpha = K(1-\alpha)\sqrt{\frac{1}{1-2\alpha}} \quad \text{with} \quad \alpha \in \left(0, \frac{1}{2}\right).$$

If $\rho(w, p^*) \geq C_\alpha \epsilon$, then there exists an index set $T \subset \{1, ..., m\}$ with $|T| \geq \alpha m$ such that $\rho(p_j, w) \geq \epsilon$ for any $j \in T$.



Proof of the Lemma (Sketch)

Proof. Let $L(p) = \sum_{k=1}^{m} \rho(p, p_k)$. Consider the geodesic curve $\gamma(t) = \texttt{Exp}_{p^*}(tv)$ with $v = \texttt{Log}_{p^*} w \in T_{p^*}\mathcal{M}$. Then,

$$dL_{p^*}(v) = \lim_{t \to 0^+} \frac{L(\gamma(t)) - L(\gamma(0))}{t} = \lim_{t \to 0^+} \frac{L(\gamma(t)) - L(p^*)}{t} \geq 0,$$

since $L(p^*)$ minimizes L for all $p \in \mathcal{M}$. By some algebra, one obtains that

$$\frac{dL_{p^*}(v)}{||v||} = -\sum_{j: p_j \neq p^*} \frac{\langle v, v_j \rangle}{||v|| \, ||v_j||} + \sum_{j=1}^{m} \mathbb{1}_{\{p_j = p^*\}},$$

where $v_j = \texttt{Log}_{p^*} p_j$. Assume, by contradiction and without the loss of generality, that

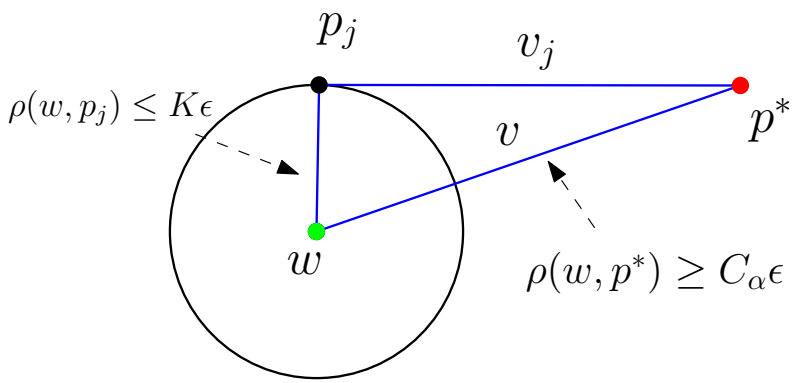
$$\rho(w, p_j) \leq \epsilon \quad \text{for } j = 1, ..., \lfloor (1 - \alpha)m \rfloor + 1,$$

while $\rho(w, p^*) \geq C_\alpha \epsilon$.

By the Lipschitz continuity of Log_{p^*} from $B(w,\epsilon)$ to $T_{p^*}\mathcal{M}$, for $j = 1, ..., \lfloor (1-\alpha)m \rfloor + 1$,

$$||v_j - v|| = \left|\left|\mathrm{Log}_{p^*} p_j - \mathrm{Log}_{p^*} w\right|\right| \leq K \cdot d_g(p_j, w) = K \cdot \rho(p_j, w) \leq K\epsilon.$$

This implies that $\sin(\widehat{v_j, v}) \leq \frac{K}{C_\alpha}$.



Thus, whenever $C_\alpha > K(1-\alpha)\sqrt{\frac{1}{1-2\alpha}}$, we have that

$$\frac{dL_{p^*}(v)}{||v||} = -\sum_{j: p_j \neq p^*} \cos(\widehat{v_j, v}) + \sum_{j=1}^{m} \mathbb{1}_{\{p_j = p^*\}} \leq -(1-\alpha)m\sqrt{1 - \frac{K^2}{C_\alpha^2}} + \alpha m < 0,$$

which is a contradiction. □

There are many Riemannian manifolds with K-Lipschitz continuous logarithm map.

1. *d-dimensional sphere* $S^d = \left\{p \in \mathbb{R}^{d+1} : ||p|| = 1\right\}$: $\text{Log}_p(\cdot)$ on S^d is 2-Lipschitz continuous from $B(p, \pi/2)$ to T_pS^d for all $p \in S^d$.

2. *Planar shape space* $\Sigma_2^k = S^{2k-3}/S^1$: $\text{Log}_p(\cdot)$ on Σ_2^k is 2-Lipschitz continuous from $B(p, \pi/4)$ to $T_p\Sigma_2^k$ for all $p \in S^d$.

3. *Positive definite matrices* $PD(n) \subset \mathbb{R}^{n\times n}$: $\text{Log}_p(\cdot)$ is 1-Lipschitz continuous at any $p \in PD(n)$.

Robust Optimization on Manifolds

Robust Optimization on Manifolds

Let Q be a probability distribution on some space $\mathcal{X}$ and $\mathcal{M}$ be a manifold. Consider estimating the *population parameter*

$$\mu = \arg\min_{p \in \mathcal{M}} L^*(p),$$

where, for some loss function L,

$$L^*(p) = \int_{\mathcal{X}} L(p, x)\, Q(dx).$$

- *Fréchet mean*: $\arg\min_{p \in \mathcal{M}} \int_{\mathcal{M}} \rho^2(p, x)\, Q(dx)$ with Q supported on $\mathcal{M}$.
- *Geometric median*: $\arg\min_{p \in \mathcal{M}} \int_{\mathcal{M}} \rho(p, x)\, Q(dx)$ with Q supported on $\mathcal{M}$.

In practice, given a random sample $\{X_1, ..., X_n\} \sim Q$, the population parameter μ can be estimated by the *empirical risk estimator*

$$\widehat{\mu}_n = \arg\min_{p \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} L(p, X_i).$$

Proposed Geometric Median of Subset Optimizers

Objective: $\mu = \arg\min_{p \in \mathcal{M}} \int_{\mathcal{X}} L(p, x)\, Q(dx)$.

The *geometric median of subset optimizers* is defined as follows.

1. Divide the dataset $\{X_1, ..., X_n\}$ into m subsets $U_1, ..., U_m$ with (roughly) equal size $\lfloor n/m \rfloor$.
2. Compute $\mu_j = \arg\min_{p \in \mathcal{M}} \frac{1}{|U_j|} \sum_{k \in U_j} L(p, X_k)$ for $j = 1, ..., m$.
3. The final estimator is $\widehat{\mu}^* = \arg\min_{p \in \mathcal{M}} \sum_{j=1}^{m} \rho(p, \mu_j)$.

$\widehat{\mu}^*$ inherits the desired robustness properties in estimating the population parameter μ.

Concentration Bound

Theorem (Theorem 3.1 in Lin et al. 2020)

Let $\mu_1, ..., \mu_m$ be some independent estimators of μ and $\mu^* = med(\mu_1, ..., \mu_m)$.

(a) If $\rho(p, q) = ||J(p) - J(q)||$ with $J : \mathcal{M} \to \widetilde{\mathcal{M}} \subset \mathbb{R}^d$, we assume that for any $w \in \mathcal{M}$, the angle between $J(w) - J(\mu^*)$ and the tangent space $T_{J(\mu^*)}\widetilde{\mathcal{M}}$ is no bigger than $\bar{\psi}$. For any $\alpha \in \left(0, \cot\bar{\psi}\tan\frac{\bar{\psi}}{2}\right)$, set $\bar{C}_\alpha = \frac{1-\alpha}{\sqrt{1-2\alpha}\cos\bar{\psi} - \alpha\sin\bar{\psi}}$.

(b) Let ρ be an intrinsic distance on $\mathcal{M}$. Assume that $\mathtt{Log}_{\mu^*}$ is K-Lipschitz continuous from $B(\mu^*, \epsilon)$ to $T_{\mu^*}\mathcal{M}$. For any $\alpha \in \left(0, \frac{1}{2}\right)$, set

$$\bar{C}_\alpha = K(1-\alpha)\sqrt{\frac{1}{1-2\alpha}}.$$

Under (a) or (b), if $\mathbb{P}\left\{\rho(\mu_j, \mu) > \epsilon\right\} \leq \eta$ for $j = 1, ..., m$ with $\eta < \alpha$, then

$$\mathbb{P}\left\{\rho(\mu^*, \mu) > \bar{C}_\alpha\epsilon\right\} \leq \exp\left[-m \cdot \phi(\alpha, \eta)\right],$$

where $\phi(\alpha, \eta) = (1-\alpha)\log\left(\frac{1-\alpha}{1-\eta}\right) + \alpha\log\frac{\alpha}{\eta}$.

Proof of the Theorem (Sketch)

Proof. Note that when $\psi < \bar{\psi}$, we have that $C_\alpha \leq \bar{C}_\alpha$ and $\cot\bar{\psi}\tan\frac{\bar{\psi}}{2} \leq \cot\psi\tan\frac{\psi}{2}$. By the previous lemma,

$$\begin{aligned}
\mathbb{P}\left\{\rho(\mu^*,\mu) > \bar{C}_\alpha\epsilon\right\} &\leq \mathbb{P}\left\{\rho(\mu^*,\mu) > C_\alpha\epsilon\right\} \\
&\leq \mathbb{P}\left(\sum_{j=1}^{m} \mathbb{1}_{\{\rho(\mu_j,\mu)>\epsilon\}} > \alpha m\right) \\
&\leq \exp\left[-m\cdot\phi(\alpha,\eta)\right],
\end{aligned}$$

where we leverage a coupling result (Lemma 23 in Lerasle and Oliveira 2011) and Chernoff's bound to obtain the last inequality. □

Applications of the Concentration Bound

Recall our mean estimation problem via the MoM estimator.

1. Partition the dataset $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \mathbb{R}^d$ into m groups $U_1, ..., U_m$ and compute the within-group means $\boldsymbol{Z}_i = \frac{1}{|U_i|} \sum_{j \in U_i} \boldsymbol{X}_j$.
2. Define the geometric median estimator $\widehat{\boldsymbol{\mu}}_n = \arg\min_{\boldsymbol{p} \in \mathbb{R}^d} \frac{1}{m} \sum_{j=1}^{m} ||\boldsymbol{p} - \boldsymbol{Z}_j||$.

Set $\alpha_* = \frac{7}{18}$ and $\eta_* = 0.1$. For any $\delta \in (0, 1)$, we take

$$m = \left\lfloor \frac{\log(1/\delta)}{\phi(\alpha_*, \eta_*)} \right\rfloor + 1 \leq \lfloor 3.5 \log(1/\delta) \rfloor + 1.$$

Then,

$$||\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}|| \leq 11 \sqrt{\frac{\mathrm{Tr}(\Sigma) \log(1.4/\delta)}{n}} \quad \text{with probability at least } 1 - \delta.$$

See Corollary 4.1 in Minsker (2015).

Caveats in Applying the Proposed Estimator

The geometric median of subset optimizers is $\widehat{\mu}^* = \arg\min_{p \in \mathcal{M}} \sum_{j=1}^{m} \rho(p, \mu_j)$.

- Larger $m \implies$ more robust and tighter concentration bound around the population parameter $\mu = \arg\min_{p \in \mathcal{M}} \int_{\mathcal{X}} L(p, x)\, Q(dx)$.
- However, the within-group sample size $\lfloor n/m \rfloor$ should also be large so that each subset estimator behaves well, i.e.,

$$\mathbb{P}\left\{\rho(\mu_j, \mu) > \epsilon\right\} \leq \eta \quad \text{for } j = 1, ..., m \text{ with a small } \eta.$$

For a given confidence level $\delta \in (0, 1)$, one can determine the number of subsets, m, to achieve a small η.

- However, in practice, η may depend on the unknown parameter; see Example 2 in Lin et al. (2020).

Simulations and Real-World Applications

Simulation Studies on S^d

Problem: Estimate the intrinsic and extrinsic means of the von Mises Fisher distribution in the presence of outliers.

$$\text{vMF}(\mu, \kappa) \text{ on } \mathbb{R}^d \quad \sim \quad f_d(x; \mu, \kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2}\mathcal{I}_{d/2-1}(\kappa)} \cdot \exp(\kappa \mu^T x).$$

Computing sample statistics on $\{p_1, ..., p_n\} \subset S^d$.

• *Intrinsic mean*: $\arg\min_{x \in S^d} \sum_{i=1}^{n} \arccos^2\left(x^T p_i\right)$. By the Lagrangian multiplier, the intrinsic mean on S^d can be obtained by a fixed-point iteration

$$\mu^{(t+1)} \leftarrow \frac{\sum_{i=1}^n \gamma_i(\mu^{(t)}) p_i}{\left|\left|\sum_{i=1}^n \gamma_i(\mu^{(t)}) p_i\right|\right|} \quad \text{for } t = 0, 1, ... \text{ with } \gamma_i(x) = \frac{\arccos(x^T p_i)}{\sqrt{1 - (x^T p_i)^2}}.$$

Notes: Its derivation is similar to our directional mean shift algorithm; see Section 2.2 in Zhang and Chen (2021).

• *Intrinsic median*: $\arg\min_{x \in S^d} \sum_{i=1}^{n} \arccos\left(x^T p_i\right)$ by the modified Weiszfeld's algorithm.

Simulation Studies on S^d (Cont'd)

• *Extrinsic mean*: $\mathcal{P}\left(\frac{1}{n}\sum_{i=1}^{n} J(p_i)\right)$, where $J : \mathcal{M} \to \widetilde{\mathcal{M}} \subset \mathbb{R}^d$ is the embedding map and $\mathcal{P} : \mathbb{R}^d \to \mathcal{M}$ is the projection map. When $\mathcal{M} = S^d$, the extrinsic mean is $\frac{\sum_{i=1}^{n} p_i}{||\sum_{i=1}^{n} p_i||}$, i.e., the spherical mean.

• *Extrinsic median*: $\arg\min_{p \in S^d} \sum_{i=1}^{n} ||x - p_i||$ by the projected gradient descent on S^d (Weiszfeld's algorithm).

Evaluation metric. Repeat the simulation for several times and compute the averages based on the following measures:

- the intrinsic distance $\rho(\mu^*, \mu)$ from the true mean μ to the geometric median of subset means μ^*.
- the average intrinsic distance $\overline{\rho(\mu_i, \mu)} = \frac{1}{m}\sum_{k=1}^{m} \rho(\mu_i, \mu)$ from μ to the subset means $\mu_k, k = 1, ..., m$.

Results on S^2

k	$\overline{\rho(\hat{\mu}, \mu)}$	$\overline{\rho(\mu^*, \mu)}$	$\overline{\overline{\rho(\mu_i, \mu)}}$	$\overline{\rho(\mu^*, \mu)}$	$\overline{\overline{\rho(\mu_i, \mu)}}$
0	0.0597	0.0583	0.0947	0.0514	0.1496
5	0.0647	0.0615	0.1159	0.0531	0.1652
10	0.1194	0.1116	0.1414	0.1018	0.2113
15	0.1819	0.1731	0.1973	0.1631	0.2419
	sample mean (m=1)	m=5		m=15	

k	$\overline{\rho(\mu^*, \mu)}$	$\overline{\overline{\rho(\mu_i, \mu)}}$	$\overline{\rho(\hat{m}, \mu)}$	$\overline{\overline{\rho(\mu_i, \mu)}}$
0	0.0455	0.2118	0.0424	0.2829
5	0.0453	0.2350	0.0447	0.2959
10	0.0776	0.2501	0.0614	0.3259
15	0.1383	0.2954	0.0925	0.3738
	m=30		sample median (m=60)	

Figure 1: Estimating the mean of vMF($\mu, \kappa = 30$) on S^2 with k being the number of outliers and ρ being the intrinsic distance.

Results on S^7

k	$\overline{\rho(\hat{\mu}, \mu)}$	$\overline{\rho(\mu^*, \mu)}$	$\overline{\overline{\rho(\mu_i, \mu)}}$	$\overline{\rho(\mu^*, \mu)}$	$\overline{\overline{\rho(\mu_i, \mu)}}$
0	0.0396	0.0399	0.1186	0.0384	0.2570
10	0.0565	0.0541	0.1258	0.0514	0.2669
20	0.0897	0.0900	0.1462	0.0834	0.2827
40	0.1656	0.1678	0.2082	0.1596	0.3376
	Sample mean (m=1)	m=10		m=50	

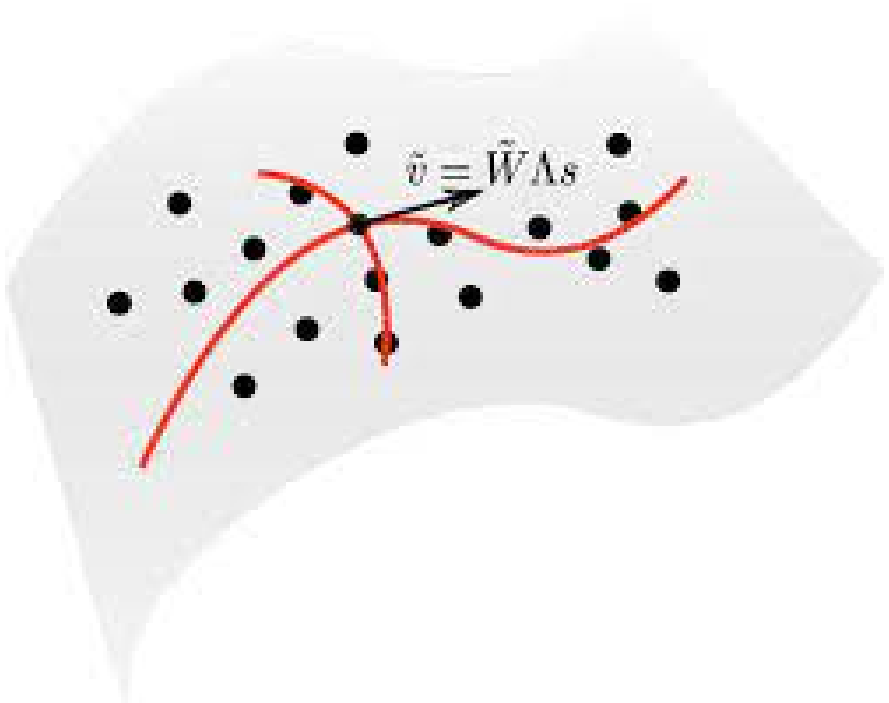
k	$\overline{\rho(\mu^*, \mu)}$	$\overline{\overline{\rho(\mu_i, \mu)}}$	$\overline{\rho(\hat{m}, \mu)}$	$\overline{\overline{\rho(\mu_i, \mu)}}$
0	0.0398	0.3590	0.0387	0.4896
10	0.0469	0.3676	0.0457	0.4978
20	0.0760	0.3896	0.0682	0.5301
40	0.1513	0.5176	0.1305	0.5987
	m=100		sample median (m=200)	

Figure 2: Estimating the mean of vMF($\mu, \kappa = 30$) on S^7 with k being the number of outliers and ρ being the intrinsic distance.

Principal Geodesic Analysis (PGA)

Principal Geodesic Analysis (Fletcher and Joshi, 2007; Lazar and Lin, 2017):

1. Compute the center of the data.
2. Successively find some orthogonal tangent vectors at the center so that their exponentiated space best fits the data according to the intrinsic sum of squared residuals.



Robust Principal Geodesic Analysis (RPGA)

Robust Principal Geodesic Analysis (RPGA):

1. Divide the data $\{X_1, ..., X_n\}$ into m groups $U_1, ..., U_m$, compute the within-group intrinsic mean μ_j, and take $\mu^* = \text{med}(\mu_1, ..., \mu_m)$.

2. Calculate $\boldsymbol{V}_k = \left\{\text{vec}\left(\text{Log}_{\mu^*}(X_j)\right) : j \in U_i\right\}$ and the sample covariance matrix Σ_k of points in $\boldsymbol{V}_k$ for $k = 1, ..., m$.

3. Compute
$$\widehat{\Sigma} = \text{med}\left(\Sigma_1, ..., \Sigma_m\right),$$
where the median is taken with respect to the Frobenius norm $||A||_F = \text{Tr}(A^T A)$.

4. Compute the eigenvectors of $\widehat{\Sigma}$, $\{\omega_1, ..., \omega_6\}$, arranged in order by largest to smallest eigenvalues.

RPGA Results

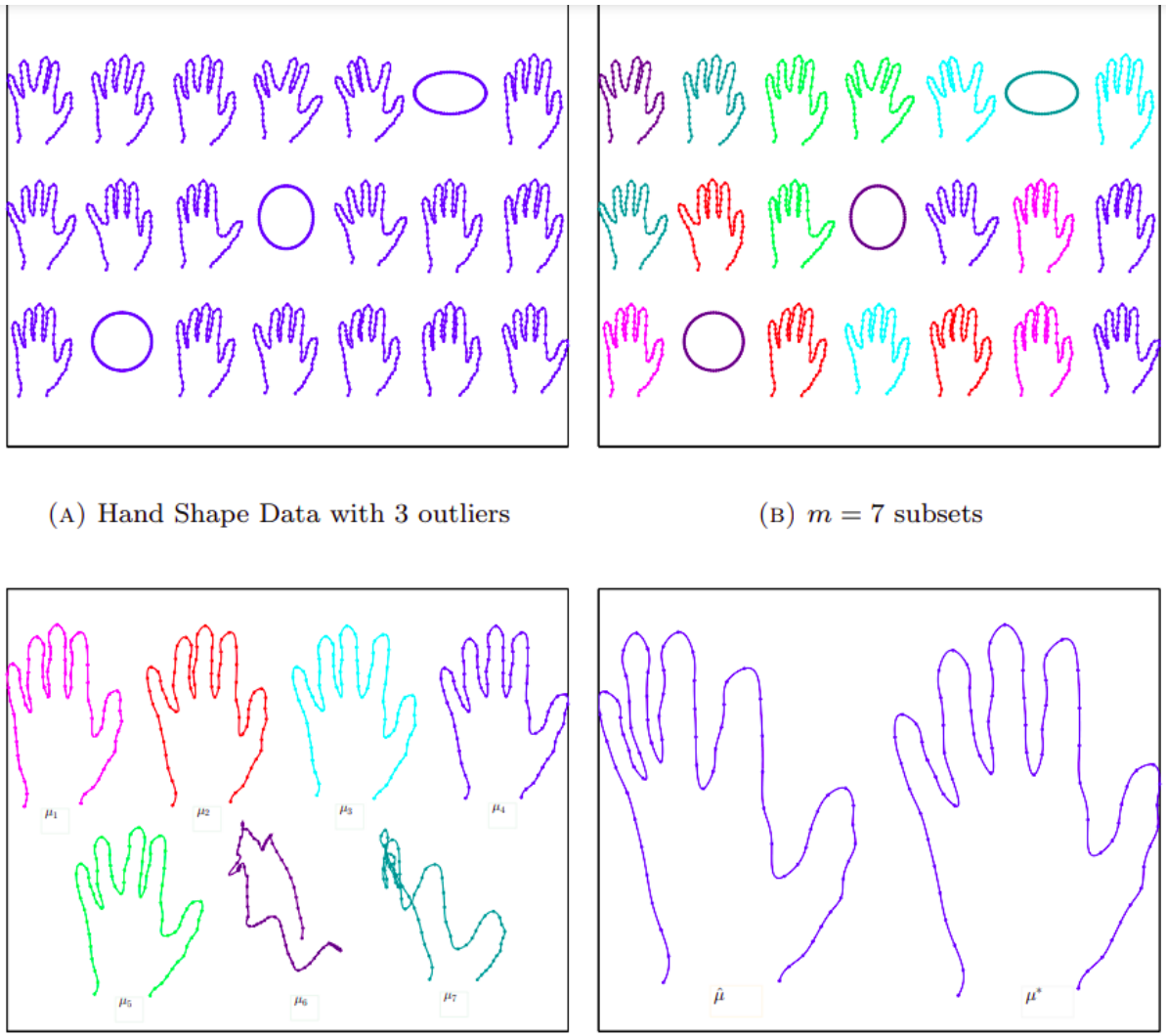
k	PGA	RPGA	RPGA	RPGA
0	0.4206	0.4265	0.4259	0.4320
5	0.4529	0.4465	0.4314	0.4342
10	0.4541	0.4438	0.4508	0.4374
15	0.4540	0.4445	0.4492	0.4442
20	0.4527	0.4473	0.4507	0.4496
m groups		m=5	m=10	m=15

k	PGA	RPGA	RPGA	RPGA
0	0.2629	0.2686	0.2691	0.2751
5	0.2924	0.2870	0.2803	0.2795
10	0.2963	0.2838	0.2925	0.2791
15	0.2994	0.2835	0.2758	0.2850
20	0.3041	0.2841	0.2889	0.2775
m groups		m=5	m=10	m=15

k	PGA	RPGA	RPGA	RPGA
0	0.1472	0.1497	0.1533	0.1608
5	0.1919	0.1801	0.1600	0.1588
10	0.2242	0.2102	0.1940	0.1743
15	0.2208	0.2149	0.2134	0.2079
20	0.2305	0.2259	0.2169	0.2206
m groups		m=5	m=10	m=15

Figure 3: Average mean sum of square residuals to explanatory submanifolds computed with k outliers to data without outliers in $PD(3)$.

Real-World Application: Hand Shape Data



(A) Hand Shape Data with 3 outliers

(B) $m = 7$ subsets

(C) Subset means, μ_i

(D) Sample mean, $\hat{\mu}$ and geometric median, μ^*

Figure 4: Median-of-means on hand shape data.

Thank you!

Reference I

N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

O. Catoni and I. Giulini. Dimension-free pac-bayesian bounds for the estimation of the mean of a random vector. *arXiv preprint arXiv:1802.04308*, 2018.

L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

P. T. Fletcher and S. Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262, 2007.

P. T. Fletcher, S. Venkatasubramanian, and S. Joshi. Robust statistics on riemannian manifolds via the geometric median. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

P. J. Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical computer science*, 43:169–188, 1986.

D. Lazar and L. Lin. Scale and curvature effects in principal geodesic analysis. *Journal of Multivariate Analysis*, 153:64–82, 2017.

M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

L. Lin, D. Lazar, B. Sarpabayeva, and D. B. Dunson. Robust optimization and inference on manifolds. *arXiv preprint arXiv:2006.06843*, 2020.

Reference II

G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019a.

G. Lugosi and S. Mendelson. Sub-gaussian estimators of the mean of a random vector. *The annals of statistics*, 47(2):783–794, 2019b.

S. Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

A. S. Nemirovskij and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983.

L. M. Ostresh Jr. On the convergence of a class of iterative methods for solving the weber location problem. *Operations Research*, 26(4):597–609, 1978.

D. Rodriguez and M. Valdora. The breakdown point of the median of means tournament. *Statistics & Probability Letters*, 153:108–112, 2019.

J. W. Tukey and D. H. McLaughlin. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/winsorization 1. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 331–352, 1963.

E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386, 1937.

Y. Zhang and Y.-C. Chen. The em perspective of directional mean shift algorithm. *arXiv preprint arXiv:2101.10058*, 2021.

Median-of-Means Tournaments

Given an i.i.d. sample $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \mathbb{R}^d$, we partition it into m groups $U_1, ..., U_m$ and compute the within-group means $\boldsymbol{Z}_i = \frac{1}{|U_i|} \sum_{j \in U_i} \boldsymbol{X}_j$.

For each $\boldsymbol{a} \in \mathbb{R}^d$, let

$$T_{\boldsymbol{a}} = \left\{ \boldsymbol{x} \in \mathbb{R}^d : \exists J \subset \{1, ..., m\} \text{ with } |J| \geq m/2 \text{ such that for all } j \in J, ||\boldsymbol{Z}_j - \boldsymbol{x}|| \leq ||\boldsymbol{Z}_j - \boldsymbol{a}|| \right\}$$

and define the "median-of-means tournaments" estimator by

$$\widehat{\boldsymbol{\mu}}_n \in \arg\min_{\boldsymbol{a} \in \mathbb{R}^d} \text{radius}(T_{\boldsymbol{a}}),$$

where $\text{radius}(T_{\boldsymbol{a}}) = \sup_{\boldsymbol{x} \in T_{\boldsymbol{a}}} ||\boldsymbol{x} - \boldsymbol{a}||$.

Theorem (Lugosi and Mendelson 2019b)

Let $\delta \in (0, 1)$ and $k = \lceil 200 \log(2/\delta) \rceil$. If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are i.i.d. random vectors in $\mathbb{R}^d$ with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix Σ, then for all n,

$$||\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}|| \leq \max \left\{ 960 \sqrt{\frac{Tr(\Sigma)}{n}}, 240 \sqrt{\frac{\lambda_{\max} \log(2/\delta)}{n}} \right\}$$

with probability at least $(1 - \delta)$, where $\lambda_{\max}$ is the maximal eigenvalue of Σ.