Mean estimation for entangled single-sample distributions

Po-Ling Loh

University of Wisconsin - Madison
Department of Statistics

Workshop on recent themes in resource tradeoffs
IMA

June 18, 2019

Joint work with Ankit Pensia and Varun Jog
Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)
Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)

**Goals:**

1. Develop estimators $T(\cdot)$ that are reliable under deviations from model assumptions
2. Quantify performance with respect to deviations
Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)

Goals:
1. Develop estimators $T(\cdot)$ that are reliable under deviations from model assumptions
2. Quantify performance with respect to deviations

Local stability captured by influence function

$$IF(x; T, F) = \lim_{\epsilon \to 0} \frac{T((1 - \epsilon)F + \epsilon\delta_x) - T(F)}{\epsilon}$$
Brief intro to robust statistics

- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)

**Goals:**

1. Develop estimators $T(\cdot)$ that are reliable under deviations from model assumptions
2. Quantify performance with respect to deviations

- Local stability captured by *influence function*

$$\text{IF}(x; T, F) = \lim_{\epsilon \to 0} \frac{T((1 - \epsilon)F + \epsilon \delta_x) - T(F)}{t}$$

- Global stability captured by *breakdown point*

$$\epsilon^*(T; X_1, \ldots, X_n) = \min \left\{ \frac{m}{n} : \sup_{X^m} \| T(X^m) - T(X) \| = \infty \right\}$$
Adversarial contamination

Instead of drawing i.i.d. data from an $\epsilon$-contaminated mixture, draw i.i.d. data points $\{x_1, \ldots, x_n\}$ and arbitrarily contaminate $\epsilon$-fraction
Adversarial contamination

- Instead of drawing i.i.d. data from an $\epsilon$-contaminated mixture, draw i.i.d. data points $\{x_1, \ldots, x_n\}$ and arbitrarily contaminate $\epsilon$-fraction
- “Adversarial machine learning”: Targeted attacks to neural networks

- Image showing a panda and a gibbon with confidence scores: 57.7% for panda and 99.3% for gibbon.
High-dimensional considerations

- Computational feasibility
High-dimensional considerations

- Computational feasibility
  - Median computation in high dimensions (e.g., Tukey median)
High-dimensional considerations

- Computational feasibility
  - Median computation in high dimensions (e.g., Tukey median)
  - Trimmed means
High-dimensional considerations

- Computational feasibility
  - Median computation in high dimensions (e.g., Tukey median)
  - Trimmed means

- Statistical efficiency
  - Huber loss is minimax optimal in terms of asymptotic variance:

\[
\min_{\{T_n\}} \max_{F \in \mathcal{P}_{\epsilon}(\Phi)} V(\{T_n\}, F)
\]
High-dimensional considerations

- Computational feasibility
  - Median computation in high dimensions (e.g., Tukey median)
  - Trimmed means

- Statistical efficiency
  - Huber loss is minimax optimal in terms of asymptotic variance:
    \[
    \min_{\{T_n\}} \max_{F \in \mathcal{P}_\epsilon(\Phi)} V(\{T_n\}, F)
    \]

- Correct notion of (non-asymptotic) efficiency?
This talk: Non-i.i.d. data

- Independent but non-identically distributed data:
  \[ X_i \sim N(\mu, \sigma^2_i) \]
- Single observation from each distribution
Independent but non-identically distributed data:

\[ X_i \sim N(\mu, \sigma_i^2) \]

Single observation from each distribution

Goal: Estimate \( \mu \) without knowledge of \( \sigma_i's \)

Observations come from a Gaussian mixture, but number of mixing components is equal to \( n \)
Motivation: Crowdsourcing

- Want to determine true value of an item based on different users’ ratings, each with different expertise
Prior work

Prior work

- Rather complicated method (details later) gives high-probability bound of

\[ |\hat{\mu} - \mu| \leq C \sqrt{n \sigma} \log n \]
Prior work

- Rather complicated method (details later) gives high-probability bound of
  \[ |\hat{\mu} - \mu| \leq C \sqrt{n} \sigma \log(n) \]

- **Questions:** Is this the optimal rate? Do data actually need to be Gaussian?
Maximum likelihood

- Can easily compute $\hat{\mu}_{MLE}$, assuming $\{\sigma_i\}_{i=1}^n$ known:

$$L_\mu(X) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma_i^2}\right)$$

$$= C \exp\left(-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma_i^2} - \log(\sigma_i)\right)$$
Maximum likelihood

- Can easily compute $\hat{\mu}_{\text{MLE}}$, assuming $\{\sigma_i\}_{i=1}^n$ known:

$$L_{\mu}(X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} \exp \left( -\frac{(x_i - \mu)^2}{2\sigma_i^2} \right)$$

$$= C \exp \left( -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma_i^2} - \log(\sigma_i) \right)$$

- So MLE solves equation

$$\sum_{i=1}^{n} \frac{x_i - \mu}{\sigma_i^2} = 0,$$

implying $\hat{\mu}_{\text{MLE}} = \left( \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} \right) / \left( \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \right)$
Since $\hat{\mu}_{MLE} \sim N\left(\mu, \frac{1}{\left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right)}\right)$, error is

$$|\hat{\mu}_{MLE} - \mu| = O\left(\frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}}\right)$$
Maximum likelihood

- Since $\hat{\mu}_{\text{MLE}} \sim N \left( \mu, 1 / \left( \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \right) \right)$, error is

  $$|\hat{\mu}_{\text{MLE}} - \mu| = \mathcal{O} \left( \frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}} \right)$$

- Note that MLE may not be consistent: For instance, $\sigma_i = i$ gives

  $$|\hat{\mu}_{\text{MLE}} - \mu| = \Theta(1)$$
Maximum likelihood

- Since $\hat{\mu}_{MLE} \sim N\left(\mu, \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}\right)$, error is

\[ |\hat{\mu}_{MLE} - \mu| = O\left(\frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}}\right) \]

- Note that MLE may not be consistent: For instance, $\sigma_i = i$ gives

\[ |\hat{\mu}_{MLE} - \mu| = \Theta(1) \]

- Turns out the rate is optimal; however, we cannot evaluate $\hat{\mu}_{MLE}$ without knowledge of $\sigma_i$’s
Simplest estimate would be $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$, which is consistent but incurs $\Omega \left( \frac{\sigma(n)}{\sqrt{n}} \right)$ error—not ideal if $\sigma(n)$ is very large.
Mean and median

- Simplest estimate would be $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$, which is consistent but incurs $\Omega\left(\frac{\sigma(n)}{\sqrt{n}}\right)$ error—not ideal if $\sigma(n)$ is very large.

- Median is more robust, and incurs $O\left(\sigma(\sqrt{n \log n})\right)$ error: Show that w.h.p., at most $\frac{n}{2}$ points lie to right and left of $[\mu - \epsilon, \mu + \epsilon]$.

\[
\geq \frac{1}{2} + \frac{C \sqrt{n \log n}}{n}
\]
• More concretely, we have independent Bernoulli random variables \( \{ Y_i \}_{i=1}^n \) such that

\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_i] \geq \frac{1}{2} + \frac{C \sqrt{n} \log n}{n},
\]

and we can apply Hoeffding’s inequality to show that \( \frac{1}{n} \sum_{i=1}^{n} Y_i \geq \frac{1}{2} \), w.h.p.
Mean and median

More concretely, we have independent Bernoulli random variables \( \{ Y_i \}_{i=1}^n \) such that

\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_i] \geq \frac{1}{2} + \frac{C \sqrt{n} \log n}{n},
\]

and we can apply Hoeffding’s inequality to show that \( \frac{1}{n} \sum_{i=1}^{n} Y_i \geq \frac{1}{2} \), w.h.p.

Calculation for mean holds for general sub-Gaussian distributions with common mean; calculation for median holds as long as mean and median agree.
Proposed estimator of Chierichetti et al.

- Idea: Points with smallest variance will be clustered around $\mu$
- "k-shortest gap estimator" identifies $k$ points with smallest spread, returns any point within interval

Also perform initial screening procedure based on computing $(c \sqrt{n \log n})$-median

Overall error is $O(\sqrt{n\sigma(\log n)})$

However, proofs are very Gaussian-specific, and largely incorrect...
Proposed estimator of Chierichetti et al.

- **Idea:** Points with smallest variance will be clustered around $\mu$
- **“$k$-shortest gap estimator”** identifies $k$ points with smallest spread, returns any point within interval $\mu$

![](image)

- **Also perform initial screening procedure based on computing** $(c\sqrt{n \log n})$-median

Overall error is $O(\sqrt{n} \sigma (\log n))$

However, proofs are very Gaussian-specific, and largely incorrect...
Proposed estimator of Chierichetti et al.

- Idea: Points with smallest variance will be clustered around $\mu$
- “$k$-shortest gap estimator” identifies $k$ points with smallest spread, returns any point within interval

$\mu$

- Also perform initial screening procedure based on computing $(c\sqrt{n \log n})$-median
- Overall error is $O(\sqrt{n\sigma(\log n)})$
Proposed estimator of Chierichetti et al.

- Idea: Points with smallest variance will be clustered around $\mu$
- "$k$-shortest gap estimator" identifies $k$ points with smallest spread, returns any point within interval

$$\mu$$

Also perform initial screening procedure based on computing $(c\sqrt{n \log n})$-median
- Overall error is $O(\sqrt{n\sigma(\log n)})$
- However, proofs are very Gaussian-specific, and largely incorrect . . .
Toward more rigorous theory
Shorth estimator

- Shortest gap also known as “shorth estimator” in robust statistics (Andrews et al. ’72)
- Method was meant for robust mode estimation in i.i.d. data, using $k = \frac{n}{2}$
Shortest gap also known as “shorth estimator” in robust statistics (Andrews et al. ’72)

Method was meant for robust mode estimation in i.i.d. data, using $k = \frac{n}{2}$

Kim & Pollard ’90 showed that shorth has error rate $\mathcal{O}\left(\frac{1}{3\sqrt{n}}\right)$ rather than $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$
Shortest gap also known as “shorth estimator” in robust statistics (Andrews et al. ’72)

Method was meant for robust mode estimation in i.i.d. data, using $k = \frac{n}{2}$

Kim & Pollard ’90 showed that shorth has error rate $O\left(\frac{1}{\sqrt{n}}\right)$ rather than $O\left(\frac{1}{3\sqrt{n}}\right)$

For non-i.i.d. data, shorth is superior to mean
Modal interval estimator

- Instead, recast problem as estimation of *mode* rather than mean
- Various proposals from classical statistics for i.i.d. setting
Modal interval estimator

- Instead, recast problem as estimation of mode rather than mean
- Various proposals from classical statistics for i.i.d. setting
- Modal interval estimator (Chernoff '64): Find interval of fixed length containing maximum number of points, return the center

\[ r - \text{modal interval} \]

\[ \mu \]
Define indicator functions $f_{x,r} = 1_{[x-r,x+r]}$ of intervals.
Define indicator functions $f_{x,r} = 1_{[x-r,x+r]}$ of intervals

$$f_{x,r}$$

Define empirical bin counts $R_n(f_{x,r}) = \frac{1}{n} \sum_{i=1}^{n} f_{x,r}(X_i)$, and population-level version $R(f_{x,r}) = \frac{1}{n} \sum_{i=1}^{n} P(|X_i - x| \leq r)$
Define indicator functions $f_{x,r} = 1_{[x-r, x+r]}$ of intervals.

Define empirical bin counts $R_n(f_{x,r}) = \frac{1}{n} \sum_{i=1}^{n} f_{x,r}(X_i)$, and population-level version $R(f_{x,r}) = \frac{1}{n} \sum_{i=1}^{n} P(|X_i - x| \leq r)$.

$r$-modal interval estimator is

$$\hat{\mu}_{M,r} := \arg \max_{x} R_n(f_{x,r}(X_i))$$
Notation and estimators

- $k$-shorth estimator is

$$
\hat{r}_k := \inf_r \sup_x \left\{ R_n(f_{x,r}) \geq \frac{k}{2n} \right\}, \quad \hat{\mu}_{S,k} := \hat{\mu}_{M,\hat{r}_k}
$$

(center of shortest interval containing $\frac{k}{2}$ points)
Notation and estimators

- $k$-shorth estimator is

$$
\hat{r}_k := \inf_r \sup_{x} \left\{ R_n(f_{x,r}) \geq \frac{k}{2n} \right\}, \quad \hat{\mu}_{S,k} := \hat{\mu}_{M,\hat{r}_k}
$$

(center of shortest interval containing $\frac{k}{2}$ points)

- We assume $X_i$'s drawn from symmetric, unimodal distributions with common mean
**Statistical guarantee: Modal interval**

**Theorem**

Suppose \( r \) is chosen such that \( R(f_{\mu}, r) \geq \frac{C \log n}{n} \). Then w.h.p.,

\[
|\hat{\mu}_{M,r} - \mu| \leq \frac{2r}{R(f_{\mu}, r)} \leq \frac{2nr}{C \log n}.
\]
Statistical guarantee: Modal interval

Theorem

Suppose $r$ is chosen such that $R(f_\mu, r) \geq \frac{C \log n}{n}$. Then w.h.p.,

$$|\hat{\mu}_{M,r} - \mu| \leq \frac{2r}{R(f_\mu, r)} \leq \frac{2nr}{C \log n}.$$

- Quality of bound depends on distributions, which affect choice of $r$:
  - For i.i.d. data, can choose $r = \frac{c \log n}{n}$
Statistical guarantee: Modal interval

**Theorem**

Suppose $r$ is chosen such that $R(f_{\mu}, r) \geq \frac{C \log n}{n}$. Then w.h.p.,

$$|\hat{\mu}_{M,r} - \mu| \leq \frac{2r}{R(f_{\mu}, r)} \leq \frac{2nr}{C \log n}.$$

- Quality of bound depends on distributions, which affect choice of $r$:
  - For i.i.d. data, can choose $r = \frac{c \log n}{n}$
  - However, $r = \Theta \left( \sigma_{(\log n)} \right)$ will always work, so we have an upper bound of $O \left( \frac{n\sigma_{(\log n)}}{\log n} \right)$, meaning we only need log $n$ “good” points
Statistical guarantee: Modal interval

**Theorem**

Suppose \( r \) is chosen such that \( R(f_\mu, r) \geq \frac{C \log n}{n} \). Then w.h.p.,

\[
\left| \hat{\mu}_{M,r} - \mu \right| \leq \frac{2r}{R(f_\mu, r)} \leq \frac{2nr}{C \log n}.
\]

- Quality of bound depends on distributions, which affect choice of \( r \):
  - For i.i.d. data, can choose \( r = \frac{c \log n}{n} \)
  - However, \( r = \Theta \left( \sigma_{\log n} \right) \) will always work, so we have an upper bound of \( O \left( \frac{n\sigma_{\log n}}{\log n} \right) \), meaning we only need \( \log n \) “good” points
- Caveat: Need to choose \( r \) adaptively from data
Theorem

Suppose \( k \geq C \log n \). Then w.h.p.,

\[
|\hat{\mu}_{S,k} - \mu| \leq \frac{4nr_{2k}}{k}.
\]

- Here, \( r_{2k} = \inf \{ r : R(f_\mu, r) \geq \frac{k}{n} \} \) is population-level shorth.
- Does not require adaptive choice of shorth parameter.
Theorem

Suppose $k \geq C \log n$. Then w.h.p.,

$$|\hat{\mu}_{S,k} - \mu| \leq \frac{4nr_{2k}}{k}.$$ 

- Here, $r_{2k} = \inf \{ r : R(f_\mu, r) \geq \frac{k}{n} \}$ is population-level shorth.
- Does not require adaptive choice of shorth parameter.
- For $k = C \log n$, bound is of same order as modal interval estimator.
Hybrid estimator

- However, if we compare with estimator of Chierichetti et al. ’14 (with incorrect proof), we want $O(\sqrt{n}\sigma(\log n))$
- Can achieve this with two-step screening procedure based on $k'$-median

![Diagram showing three horizontal lines with dots, indicating the shortest gap and the median.]

$\mu$

$3$ – shortest gap

$5$ – median
Hybrid estimator

1. Compute $(\sqrt{n \log n})$-median
2. Compute $r$-modal interval / $k$-shorth estimator
3. Output projection of estimator (2) onto set defined by (1)

Theorem

Suppose all $X_i$'s are drawn independently from symmetric, unimodal distributions with common mean $\mu$. Then w.h.p., the output of the hybrid estimator satisfies

$$|\hat{\mu} - \mu| = O\left(\sqrt{n} \sigma \left(\log n\right)\right).$$
Hybrid estimator

1. Compute \((\sqrt{n}\log n)\)-median
2. Compute \(r\)-modal interval / \(k\)-shorth estimator
3. Output projection of estimator (2) onto set defined by (1)

Theorem

Suppose all \(X_i\)'s are drawn independently from symmetric, unimodal distributions with common mean \(\mu\). Then w.h.p., the output of the hybrid estimator satisfies

\[
|\hat{\mu} - \mu| = \mathcal{O}\left(\sqrt{n}\sigma (\log n)\right).
\]
Proof ideas

- Key concentration inequality: For any fixed $t \in (0, 1]$ and $r > 0$,

$$
\mathbb{P} \left( \sup_x \left| R_n(f_{x,r}) - R(f_{x,r}) \right| \geq tR(f_{\mu,r}) \right) \leq 2 \exp \left( -cnt^2 R(f_{\mu,r}) \right)
$$
Proof ideas

- Key concentration inequality: For any fixed \( t \in (0, 1] \) and \( r > 0 \),

\[
P \left( \sup_x |R_n(f_{x,r}) - R(f_{x,r})| \geq tR(f_{\mu,r}) \right) \leq 2 \exp \left( -cnt^2 R(f_{\mu,r}) \right)
\]

- Technical challenge: Data are independent but not i.i.d.
Proof ideas

- For hybrid estimator, we have two cases: Let \( r' = \sqrt{n} r \log n \)
  
  1. \( R(f_\mu, r') \geq \frac{C \log n}{\sqrt{n}} \implies \) enough low-variance points, so median screening gives good estimator

  \[
  \begin{array}{c}
  \mu - r' \\
  \mu \\
  \mu + r'
  \end{array}
  \]

  \[
  \begin{array}{c}
  \mu - r' \\
  \mu \\
  \mu + r'
  \end{array}
  \]

  2. \( R(f_\mu, r') < \frac{C \log n}{\sqrt{n}} \implies \) relatively fast decay, so modal interval estimator does well; projecting onto median only does better

  \[
  \begin{array}{c}
  \mu - r' \\
  \mu \\
  \mu + r'
  \end{array}
  \]
Still a noticeable gap between estimation error $O\left(\frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}} \right)$ of MLE and error rate $O(\sqrt{n}\sigma_{\log n})$ of hybrid estimator.
Still a noticeable gap between estimation error $O \left( \frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}} \right)$ of MLE and error rate $O(\sqrt{n} \sigma_{(\log n)})$ of hybrid estimator.

Can show minimax optimality over certain classes of distributions, e.g., $\Omega(\log n)$ points from $N(\mu, 1)$ distribution and remaining points from $N(\mu, n^\alpha)$ distribution.

Lower bounds based on KL divergence show hybrid estimator is within $\log n$ factor of optimum.

Optimal rates for other problem scalings remains an open question.
Still a noticeable gap between estimation error $O \left( \frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}} \right)$ of MLE and error rate $O(\sqrt{n}\sigma(\log n))$ of hybrid estimator.

Can show minimax optimality over certain classes of distributions, e.g., $\Omega(\log n)$ points from $N(\mu, 1)$ distribution and remaining points from $N(\mu, n^\alpha)$ distribution.

Lower bounds based on KL divergence show hybrid estimator is within $\log n$ factor of optimum.
Still a noticeable gap between estimation error $\mathcal{O}\left(\frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}} \right)$ of MLE and error rate $\mathcal{O}(\sqrt{n} \sigma_{(\log n)})$ of hybrid estimator.

Can show minimax optimality over certain classes of distributions, e.g., $\Omega(\log n)$ points from $N(\mu, 1)$ distribution and remaining points from $N(\mu, n^\alpha)$ distribution.

Lower bounds based on KL divergence show hybrid estimator is within $\log n$ factor of optimum.

Optimal rates for other problem scalings remains an open question.
Question: How to extend ideas to dimension $d > 1$?
Multidimensional estimator

**Question:** How to extend ideas to dimension $d > 1$?

Need to figure out right notion of symmetry, right shape of shorth/modal interval functions
**Question:** How to extend ideas to dimension $d > 1$?

Need to figure out right notion of symmetry, right shape of shorth/modal interval functions.

We assume all mixing components are radially symmetric, unimodal, with common mean (to be relaxed later).
Define $f_{x,r} = 1_{B_x(r)}$, ball of radius $r$ around $x$. Then the $r$-modal interval estimator is $\hat{\mu}_{M,r} = \arg \max_x R_n(f_{x,r})$. Theorem: Suppose $R_n(f_{\mu,r}) \geq Cd \log n/ n$. Then w.h.p., $\|\hat{\mu}_{M,r} - \mu\|_2 \leq 4r(2R_n(f_{\mu,r}))^{1/d}$. For minimal choice of $r$, error is $\tilde{O}(n^{1/d} \sqrt{d \log n})$. 
Define \( f_{x,r} = 1_{B_x(r)} \), ball of radius \( r \) around \( x \)

Then \( r \)-modal interval estimator is \( \hat{\mu}_{M,r} = \arg \max_x R_n(f_{x,r}) \)
Define $f_{x,r} = 1_{B_x(r)}$, ball of radius $r$ around $x$.

Then $r$-modal interval estimator is $\hat{\mu}_{M,r} = \arg \max_x R_n(f_{x,r})$.

**Theorem**

Suppose $R(f_{\mu,r}) \geq \frac{Cd \log n}{n}$. Then w.h.p.,

$$\|\hat{\mu}_{M,r} - \mu\|_2 \leq 4r \left(\frac{2}{R(f_{\mu,r})}\right)^{1/d}.$$ 

For minimal choice of $r$, error is $\tilde{O}(n^{1/d} \sqrt{d} \sigma(d \log n))$. 
Similarly, define $k$-shorth as center of smallest $\ell_2$-ball containing at least $\frac{k}{2}$ points.

**Theorem**

Suppose $k \geq Cd \log n$. Then w.h.p.,

$$
\|\hat{\mu}_{S,k} - \mu\|_2 \leq 4r_{2k} \left(\frac{4n}{k}\right)^{1/d}.
$$
Similarly, define $k$-shorth as center of smallest $\ell_2$-ball containing at least $\frac{k}{2}$ points.

**Theorem**

Suppose $k \geq Cd \log n$. Then w.h.p.,

$$\|\hat{\mu}_{S,k} - \mu\|_2 \leq 4r_{2k} \left(\frac{4n}{k}\right)^{1/d}.$$ 

As in $d = 1$ case, shorth and modal interval estimators have errors of the same order.
Actually, computation of modal interval and shorth estimators is difficult in higher dimensions: $\Omega(n^d)$ complexity
Actually, computation of modal interval and shorth estimators is difficult in higher dimensions: $\Omega(n^d)$ complexity.

Use idea from Abraham, Biau & Cadre ’04 on mode estimation for i.i.d. data: Only consider balls centered at data points.
Computational considerations

- Actually, computation of modal interval and shorth estimators is difficult in higher dimensions: $\Omega(n^d)$ complexity
- Use idea from Abraham, Biau & Cadre '04 on mode estimation for i.i.d. data: Only consider balls centered at data points

Computational complexity reduced to $O(n^2)$
Technical challenges: Need to show that w.h.p., some data point $X_i$ lies close to $\mu$ and shorth ball centered at $X_i$ has radius close to $r_k$
Computational considerations

- Technical challenges: Need to show that w.h.p., some data point $X_i$ lies close to $\mu$ and shorth ball centered at $X_i$ has radius close to $r_k$
- Need to derive more refined concentration inequality for deviation between $R_n$ and $R$, obtained using peeling technique:

\[
P \left( |R_n(f_x, r) - R(f_x, r)| \leq 2tR(f_x, r), \quad \forall x : \|x\|_2 \leq \bar{r} \right) \geq 1 - C \exp \left( -cnt^2 R(f_{\bar{r}}, r) \right)
\]
Technical challenges: Need to show that w.h.p., some data point $X_i$ lies close to $\mu$ and shorth ball centered at $X_i$ has radius close to $r_k$.

Need to derive more refined concentration inequality for deviation between $R_n$ and $R$, obtained using peeling technique:

$$\Pr \left( |R_n(f_x,r) - R(f_x,r)| \leq 2tR(f_x,r), \quad \forall x : \|x\|_2 \leq \bar{r} \right) \geq 1 - C \exp \left( -cnt^2 R(f_{\bar{r}},r) \right)$$

Error of final estimator only increases by factor of 2.
In fact, we can decrease error rate to $\tilde{O}(\sqrt{n^{1/d}} \sqrt{d} \sigma_{d \log n})$ using median screening.
In fact, we can decrease error rate to $\tilde{O}(\sqrt{n^{1/d}} \sqrt{d} \sigma(d \log n))$ using median screening.

But what median to use in $d$ dimensions?
In fact, we can decrease error rate to $\tilde{O}(\sqrt{n^{1/d}} \sqrt{d} \sigma(d \log n))$ using median screening.

But what median to use in $d$ dimensions?

Turns out coordinate-wise median works: Create a cuboid and project modal interval/shorth estimator on cuboid.
Relaxing symmetry

- Significant limitation of theory is symmetry assumption
- Needed because we want $R(f_{x,r})$ to be maximized at $x = \mu$
Relaxing symmetry

- Significant limitation of theory is symmetry assumption
- Needed because we want $R(f_x,r)$ to be maximized at $x = \mu$
- Assumption can be relaxed to central symmetry + log-concavity in $d$ dimensions, but only guarantees $\widetilde{O}(\sqrt{n})$ error decay

Intuition: For isotropic distributions, points with large variance can be detected in any of $d$ dimensions $\Rightarrow$ better estimation rates
Relaxing symmetry

- Significant limitation of theory is symmetry assumption
- Needed because we want \( R(f_{x,r}) \) to be maximized at \( x = \mu \)
- Assumption can be relaxed to central symmetry + log-concavity in \( d \) dimensions, but only guarantees \( \tilde{O}(\sqrt{n}) \) error decay

**Intuition:** For isotropic distributions, points with large variance can be detected in any of \( d \) dimensions \( \implies \) better estimation rates
Relaxing symmetry

- How many radially symmetric distributions do we need?

Can obtain $\tilde{O}(\sqrt{n^{1/d}})$ rates with only a fraction of $\frac{1}{n^{1/d}}$ radially symmetric distributions (all other distributions centrally symmetric).
Relaxing symmetry

- In one dimension, actually do not need symmetry—unimodality is good enough!

- Reason: $\arg \max_x R(f_x,r)$ cannot drift too far from $\mu$
Contributions

- Estimation of common mean in independent, non-i.i.d. case
- Solidified theory for Gaussian case, extended to general (symmetric) distributions
- Revitalized shorth estimator from robust statistics
Future work

- Lower bounds for more general classes of distributions
- Relaxing symmetry even further (estimation of common mode)
- More than 1 observation from each mixture component
- Partial knowledge of $\{\sigma_i\}_{i=1}^n$

Thank you!