Robust Mean Estimation

Lecture 10

Outline

Mean estimation, with and without outliers

The good set has small covariance

Small variance pins down the mean

An algorithm: alternating minimization

Outline

Mean estimation, with and without outliers

The good set has small covariance

Small variance pins down the mean

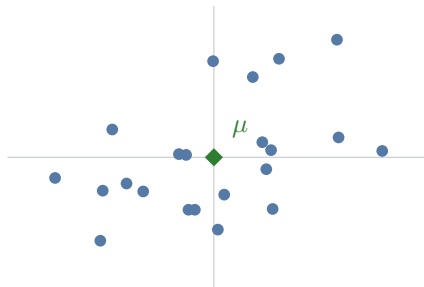
An algorithm: alternating minimization

Estimating a mean

A distribution $\mathcal{D}$ on $\mathbb{R}^d$ with mean μ and covariance

$$\Sigma := \mathbb{E}_{x \sim \mathcal{D}} (x - \mu)(x - \mu)^\top$$

- standing assumption: $\Sigma \preceq I$
- goal: estimate μ from samples



n samples from $\mathcal{D}$

The empirical mean: d/n in squared error

With $\hat{\mu} = \frac{1}{n}\sum_i x^i$ and $z^i := x^i - \mu$: $\mathbb{E}\hat{\mu} = \mu$ and

$$\mathbb{E}\|\hat{\mu} - \mu\|_2^2 = \mathbb{E}\left\|\frac{1}{n}\sum_i z^i\right\|_2^2 \underset{\text{independence}}{=} \frac{1}{n}\mathbb{E}\|z\|_2^2 = \frac{\operatorname{Tr}\Sigma}{n} \leq \frac{d}{n}$$

$$n \gtrsim d/\varepsilon^2 \quad \Longrightarrow \quad \mathbb{E}\|\hat{\mu} - \mu\|_2^2 \leq \varepsilon^2$$

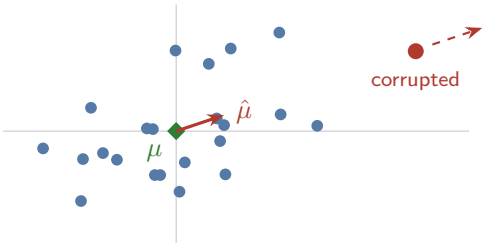
subgaussian $\mathcal{D}$: the same bound holds with high probability;
under a second-moment assumption only, median-of-means still does.

. . . but the empirical mean is not robust

Move a *single* point to distance R:

$$\hat{\mu} \text{ moves by } \frac{R}{n}.$$

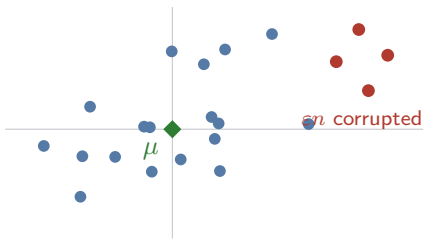
R is unbounded $\Rightarrow$ one corrupted sample destroys the estimate.



The contamination model

1. draw $\tilde{x}_1, \ldots, \tilde{x}_n$ i.i.d. from a subgaussian $\mathcal{D}$
2. an adversary — who knows $\mathcal{D}$, the sample and our algorithm — replaces εn of them arbitrarily
3. we see $x_1, \ldots, x_n$ and must estimate μ

good points: $G := \{i : x_i = \tilde{x}_i\}$,
$|G| \geq (1 - \varepsilon)n$

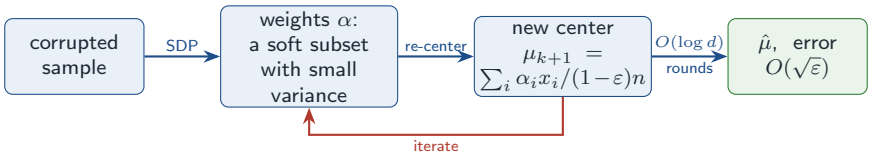


Today's goal

Theorem

A polynomial-time estimator with $\|\hat{\mu} - \mu\| = O(\sqrt{\varepsilon})$ w.h.p., for $n = \tilde{O}(d/\varepsilon^2)$ samples.

the error does not grow with the dimension



Outline

Mean estimation, with and without outliers

The good set has small covariance

Small variance pins down the mean

An algorithm: alternating minimization

What the clean sample gives us

Lemma (A: concentration, stated without proof)

For $n = \tilde{O}(d/\varepsilon^2)$ samples from a subgaussian $\mathcal{D}$ with $\Sigma \preceq I$, with probability $\geq 3/4$

$$\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \mu)(\tilde{x}_i - \mu)^\top \;\preceq\; C\,I.$$

this is the only place where subgaussianity and the sample size are used

. . . and what survives on the good set G

- drop the corrupted indices (each term is PSD): $\frac{1}{n}\sum_{i\in G}(x_i-\mu)(x_i-\mu)^\top \preceq CI$
- renormalize, $|G| \geq (1-\varepsilon)n$: $\frac{1}{|G|}\sum_{i\in G}(x_i-\mu)(x_i-\mu)^\top \preceq C'I$
- recenter at $\mu_G := \frac{1}{|G|}\sum_{i\in G} x_i$: $\Sigma_G = \frac{1}{|G|}\sum_{i\in G}(x_i-\mu_G)(x_i-\mu_G)^\top \preceq C'I$

$\sigma_G^2 := \|\Sigma_G\|_{\mathrm{op}} = O(1)$ the good points have bounded variance in *every* direction

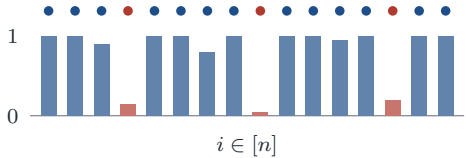
Looking for G: soft subsets

Weights $\alpha_i \in [0,1]$, $\sum_i \alpha_i = (1-\varepsilon)n$, with

$$\frac{1}{(1-\varepsilon)n} \sum_{i \in [n]} \alpha_i (x_i - \mu_\alpha)(x_i - \mu_\alpha)^\top \preceq 2I, \quad (*)$$

$$\mu_\alpha := \frac{1}{(1-\varepsilon)n} \sum_i \alpha_i x_i$$

Feasible: $\alpha = \mathbf{1}_{G'}$ for any $G' \subseteq G$ with $|G'| = (1-\varepsilon)n$.

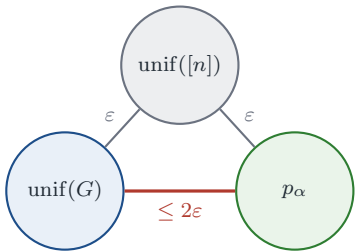


weights α_i; red: corrupted points

Three distributions on the sample

- $\mathrm{unif}([n])$ — all n points
- $\mathrm{unif}(G)$ — the good points
- $p_\alpha(x_i) = \dfrac{\alpha_i}{(1-\varepsilon)n}$ — the soft subset

$$\mathrm{TV}\left(\mathrm{unif}([n]), \mathrm{unif}(G)\right) \le \varepsilon$$
$$\mathrm{TV}\left(\mathrm{unif}([n]), p_\alpha\right) \le \varepsilon$$



$\mathrm{TV}(\mathrm{unif}(G), p_\alpha) \le 2\varepsilon$ — whatever α is, it cannot disagree with G by much.

Outline

Mean estimation, with and without outliers

The good set has small covariance

Small variance pins down the mean

An algorithm: alternating minimization

Key lemma

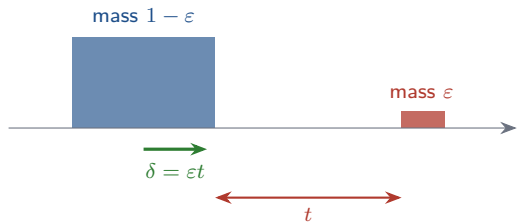
Lemma (TV + bounded variance ⇒ close means)

Let $\mathcal{P}_1, \mathcal{P}_2$ be distributions on $\mathbb{R}^d$ with $\|\Sigma_{\mathcal{P}_1}\|_{\mathrm{op}} \le \sigma_1^2$, $\|\Sigma_{\mathcal{P}_2}\|_{\mathrm{op}} \le \sigma_2^2$ and $\mathrm{TV}(\mathcal{P}_1, \mathcal{P}_2) \le \varepsilon \le \frac{1}{10}$. Then

$$\|\mu_{\mathcal{P}_1} - \mu_{\mathcal{P}_2}\| \;\le\; 2\sqrt{\varepsilon}\,(\sigma_1 + \sigma_2).$$

Why, in one picture: moving the mean by δ needs ε of the mass at distance $t = \delta/\varepsilon$, which costs variance

$$\varepsilon t^2 = \frac{\delta^2}{\varepsilon} \;\le\; \sigma^2.$$



Step 1: one dimension is enough

Fix a unit v and project: $X \mapsto \langle v, X \rangle$.

- variances: $\mathrm{Var}\langle v, X \rangle = v^\top \Sigma_{\mathcal{P}} v \le \sigma^2$
- TV never increases under a map: $\mathrm{TV} \le \varepsilon$ still

So it suffices to prove the lemma for distributions on $\mathbb{R}$.

Step 2: optimal coupling and Cauchy–Schwarz

There is a coupling (X, X') with $X \sim \mathcal{P}_1$, $X' \sim \mathcal{P}_2$ and

$$\Pr[X \neq X'] = \mathrm{TV}(\mathcal{P}_1, \mathcal{P}_2) \leq \varepsilon.$$

$$|\mu_{\mathcal{P}_1} - \mu_{\mathcal{P}_2}| = \big|\mathbb{E}[X - X']\big| \;\leq\; \mathbb{E}\big[|X - X'|\,\mathbf{1}_{X \neq X'}\big] \underset{\text{C.–S.}}{\leq} \|X - X'\|_2 \sqrt{\Pr[X \neq X']}$$

$$|\mu_{\mathcal{P}_1} - \mu_{\mathcal{P}_2}| \;\leq\; \sqrt{\varepsilon}\,\|X - X'\|_2$$

Step 3: close the loop

Write $\Delta := |\mu_{\mathcal{P}_1} - \mu_{\mathcal{P}_2}|$. By the triangle inequality in L_2,

$$\|X - X'\|_2 \;\le\; \underbrace{\|X - \mu_{\mathcal{P}_1}\|_2}_{\le \sigma_1} \;+\; \Delta \;+\; \underbrace{\|X' - \mu_{\mathcal{P}_2}\|_2}_{\le \sigma_2}$$

$$\Delta \;\le\; \sqrt{\varepsilon}\,(\sigma_1 + \sigma_2 + \Delta) \qquad \Longrightarrow \qquad (1 - \sqrt{\varepsilon})\,\Delta \;\le\; \sqrt{\varepsilon}\,(\sigma_1 + \sigma_2)$$

for $\varepsilon \le \frac{1}{10}$: $\Delta \le \dfrac{\sqrt{\varepsilon}}{1 - \sqrt{\varepsilon}}(\sigma_1 + \sigma_2) \le 2\sqrt{\varepsilon}\,(\sigma_1 + \sigma_2)$ $\square$

Consequence: any valid soft subset is already good

Let α be *any* feasible point of $(*)$: $\|\Sigma_{p_\alpha}\|_{\mathrm{op}} \le 2$ and $\mathrm{TV}(p_\alpha, \mathrm{unif}(G)) \le 2\varepsilon$.

$$\|\mu_\alpha - \mu_G\| \;\le\; 2\sqrt{2\varepsilon}\,\big(\sigma_G + \sqrt{2}\big) \;=\; O(\sqrt{\varepsilon})$$

- we never need to *identify* G
- a certificate of small variance on a large soft subset is enough
- remaining problem: $(*)$ is not convex

Outline

Mean estimation, with and without outliers

The good set has small covariance

Small variance pins down the mean

An algorithm: alternating minimization

The obstacle

- ✗ the center μ_α moves with α: the constraint is *quadratic* in α — not an SDP
- ✓ **fix:** freeze the center and alternate

given μ_k $\rightarrow$ optimize α $\rightarrow$ set $\mu_{k+1} := \mu_\alpha$ $\rightarrow$ repeat

One step: a semidefinite program

For a fixed center μ_k:

$$\begin{aligned} \min \quad & \gamma \\ \text{s.t.} \quad & \frac{1}{(1-\varepsilon)n} \sum_{i \in [n]} \alpha_i (x_i - \mu_k)(x_i - \mu_k)^\top \preceq \gamma I \\ & 0 \le \alpha_i \le 1, \qquad \sum_i \alpha_i = (1-\varepsilon)n \end{aligned}$$

$$\text{then} \quad \mu_{k+1} := \frac{1}{(1-\varepsilon)n} \sum_i \alpha_i x_i \quad \text{and iterate.}$$

Why this is an SDP

The matrices $M_i := \dfrac{(x_i - \mu_k)(x_i - \mu_k)^\top}{(1-\varepsilon)n}$ are *fixed* (and PSD).

$$\gamma I - \sum_{i \in [n]} \alpha_i M_i \succeq 0$$

- a linear matrix inequality in the variables (α, γ)
- the remaining constraints $0 \le \alpha_i \le 1$, $\sum_i \alpha_i = (1-\varepsilon)n$ are linear

exactly the generic form from Lecture 9 (pack α, γ and the slacks into one block-diagonal PSD variable)

The optimum γ^* is small

Test the feasible point $\alpha = \mathbf{1}_{G'}$, $G' \subseteq G$, $|G'| = (1-\varepsilon)n$. Exactly:

$$\frac{1}{|G|}\sum_{i\in G}(x_i-\mu_k)(x_i-\mu_k)^\top \;=\; \Sigma_G + (\mu_G-\mu_k)(\mu_G-\mu_k)^\top$$

$$\gamma^* \;\le\; O\big(\sigma_G^2 + \|\mu_k-\mu_G\|^2\big)$$

the only cost of a wrong center is the square of the distance to μ_G

What one iteration buys

The optimal α defines $p_{k+1}(x_i) = \alpha_i/(1-\varepsilon)n$ with

- $\|\Sigma_{p_{k+1}}\|_{\mathrm{op}} \le \gamma^*$ (covariance $\preceq$ second moment about μ_k)
- $\mathrm{TV}\left(p_{k+1}, \mathrm{unif}(G)\right) \le 2\varepsilon$

Key lemma applied to p_{k+1} and $\mathrm{unif}(G)$:

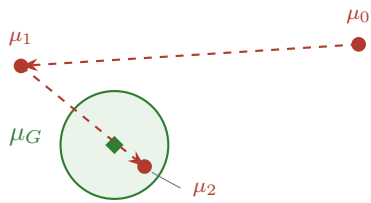
$$\|\mu_{k+1} - \mu_G\| \lesssim \sqrt{\varepsilon}\left(\sigma_G + \|\mu_k - \mu_G\|\right)$$

Contraction

For ε below a constant:

$$\begin{aligned} & \|\mu_k - \mu_G\| \;\geq\; C\sqrt{\varepsilon}\,\sigma_G \\ \Longrightarrow \quad & \|\mu_{k+1} - \mu_G\| \;\leq\; \tfrac{1}{2}\|\mu_k - \mu_G\| \end{aligned}$$

- the distance halves every round
- until it reaches $O(\sqrt{\varepsilon})\,\sigma_G = O(\sqrt{\varepsilon})$

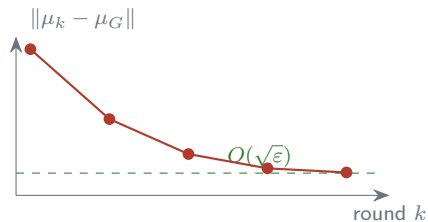


Where to start

Coordinate-wise median μ_0:

- in each coordinate, ε corruptions move the median by $O(\varepsilon)$
- over d coordinates: $\|\mu_0 - \mu_G\| = O(\varepsilon\sqrt{d})$

$$\log \frac{\varepsilon\sqrt{d}}{\sqrt{\varepsilon}\,\sigma_G} = O(\log d) \quad \text{halvings suffice.}$$



The algorithm

1. $\mu_0 :=$ coordinate-wise median of $x_1, \ldots, x_n$
2. repeat $O(\log d)$ times: solve the SDP at μ_k, set $\mu_{k+1} = \frac{1}{(1-\varepsilon)n} \sum_i \alpha_i x_i$
3. output $\hat{\mu} := \mu_K$

Guarantee

$$\|\hat{\mu} - \mu\| \;\le\; \underbrace{\|\hat{\mu} - \mu_G\|}_{O(\sqrt{\varepsilon})} + \underbrace{\|\mu_G - \mu\|}_{O(\sqrt{\varepsilon})} \;=\; O(\sqrt{\varepsilon}) \quad \text{w.p.} \ge 3/4, \ \text{for } n = \tilde{O}(d/\varepsilon^2)$$

polynomial time: $O(\log d)$ semidefinite programs

Remarks and outlook

- $\Theta(\sqrt{\varepsilon})$ is *optimal* when only $\Sigma \preceq I$ is assumed — even with infinitely many samples
- for subgaussian $\mathcal{D}$ the optimal rate is $\Theta\big(\varepsilon\sqrt{\log(1/\varepsilon)}\big)$, reached by refined versions of the same scheme
- *filtering* algorithms replace the SDP by repeated top-eigenvector computations: near-linear time
- the pattern — *certify small variance, then re-center* — extends to robust covariance estimation, regression and clustering