SoS Tutorial, Part II

(loading loading loading – advance slide)

SoS and Robust Statistics, Part 2

Sam Hopkins (Cornell University $\rightarrow$ UC Berkeley)

Agenda:

1. Wrap up from yesterday
2. Mixture models and clustering
3. Which distributions does SoS robustly estimate and cluster?

Recap of Part 1

SoS is really great (Thor's hammer, Sauron's ring, etc.)

Offers a method to design *efficient algorithms* based on *simple identifiability proofs*.

The "usual" identifiability argument for **mean estimation** with ε-corrupted samples under **bounded covariance assumptions** with $O(\sqrt{\varepsilon})$ **error** is the right kind of simple.
Uses only Cauchy-Schwarz, triangle inequalities, hence has an SoS certificate

The holy trinity: simple identifiability proofs, sum-of-squares polynomials, efficient algorithms

Yesterday we saw:

If $X = \{x_1, \ldots, x_m\}$ are ε-corrupted from a distribution $\mathcal{D}$ with mean μ, variance 1, whp there are degree $O(1)$ polynomials $s_i(w, X', \mu', g), q_j(w, X', \mu', g)$ such that

$$O(\varepsilon) - (\mu - \mu')^2 = \sum s_i^2 + \sum q_j p_j$$

where $p_1 = 0, \ldots, p_m = 0$ enforce

$w_i^2 = w_i$ and $\sum w_i = (1-\varepsilon)m$

$w_i(X_i - X_i')$

$\mu' = \frac{1}{(1-\varepsilon)m} \sum w_i X_i'$

$\sum w_i (X_i - \mu')^2 + g^2 = 1$

This is an SoS identifiability proof

Not going to go through the basic identifiability proof yet again, but just a taste:

To *SoS-ify* the Cauchy-Schwarz step, use that

$$\left(\sum_{i\leq n} y_i^2\right)\left(\sum_{i\leq n} x_i^2\right) - \left(\sum_{i\leq n} x_i y_i\right)^2 = \sum_{i,j}(x_i y_j - x_j y_i)^2$$

SoS identifiability proof + meta-theorem $\rightarrow$ efficient algorithm

If $p_1(\widehat{\Theta}, W) = 0, \ldots, p_m(\widehat{\Theta}, W) = 0$ imply $\|\Theta - \widehat{\Theta}\|^2 \leq \delta$ and this has an SoS proof of degree t, then there is an $(mn)^{O(t)}$ time algorithm to output Θ' with $\|\Theta' - \Theta\|^2 \leq \delta$.

$$\delta - \|\Theta - \widehat{\Theta}\|^2 = \sum s_i^2 + \sum q_j p_j$$

Proof of meta-theorem:

Suppose linear operator $\tilde{\mathbb{E}} : \mathbb{R}[\widehat{\Theta}, W]_{\leq t} \rightarrow \mathbb{R}$ such that

1. $\tilde{\mathbb{E}} 1 = 1$
2. $\tilde{\mathbb{E}} p^2 \geq 0$ for all p such that $\deg p^2 \leq t$
3. $\tilde{\mathbb{E}} p_i q = 0$ for all q, p_i such that $\deg p_i q \leq t$

Then $\tilde{\mathbb{E}}\|\Theta - \widehat{\Theta}\|^2 \leq \delta$, expands to $\|\Theta\|^2 + \tilde{\mathbb{E}}\|\widehat{\Theta}\|^2 - 2\langle \Theta, \tilde{\mathbb{E}}\widehat{\Theta}\rangle \leq \delta$.

Since $\tilde{\mathbb{E}}\|\widehat{\Theta}\|^2 \geq \|\tilde{\mathbb{E}}\widehat{\Theta}\|^2$, we find $\|\tilde{\mathbb{E}}\widehat{\Theta} - \Theta\|^2 \leq \delta$.

Have some $p_1(y), \ldots, p_m(y)$

Suppose linear operator $\tilde{\mathbb{E}} : \mathbb{R}[y]_{\leq t} \to \mathbb{R}$ such that

1. $\tilde{\mathbb{E}}1 = 1$
2. $\tilde{\mathbb{E}}p^2 \geq 0$ for all p such that $\deg p^2 \leq t$
3. $\tilde{\mathbb{E}}p_i q = 0$ for all q, p_i such that $\deg p_i q \leq t$

Set of such $\tilde{\mathbb{E}}$ is feasible set of following SDP:

Variables: "$\tilde{\mathbb{E}}y^\alpha$" for every multi-index α with $|\alpha| \leq t$ (assume t even)

They define an operator: $\tilde{\mathbb{E}}p(y) = \tilde{\mathbb{E}}\sum p_\alpha y^\alpha = \sum p_\alpha \tilde{\mathbb{E}}y^\alpha$

Constraints (1) and (3): $\tilde{\mathbb{E}}1 = \tilde{\mathbb{E}}y^\emptyset = 1$ is a linear constraint. So is $\tilde{\mathbb{E}}p_i(y) \cdot y^\alpha = 0$.

Constraint (2): $\tilde{\mathbb{E}}p^2 \geq 0$ is equivalent to $p^\top M p \geq 0$ where $M_{\alpha,\beta} = \tilde{\mathbb{E}}y^\alpha y^\beta$.

Resulting SDP has "intended solution" $(y^{\otimes t/2})(y^{\otimes t/2})^\top$
(compare with $yy^\top$ from basic SDP)

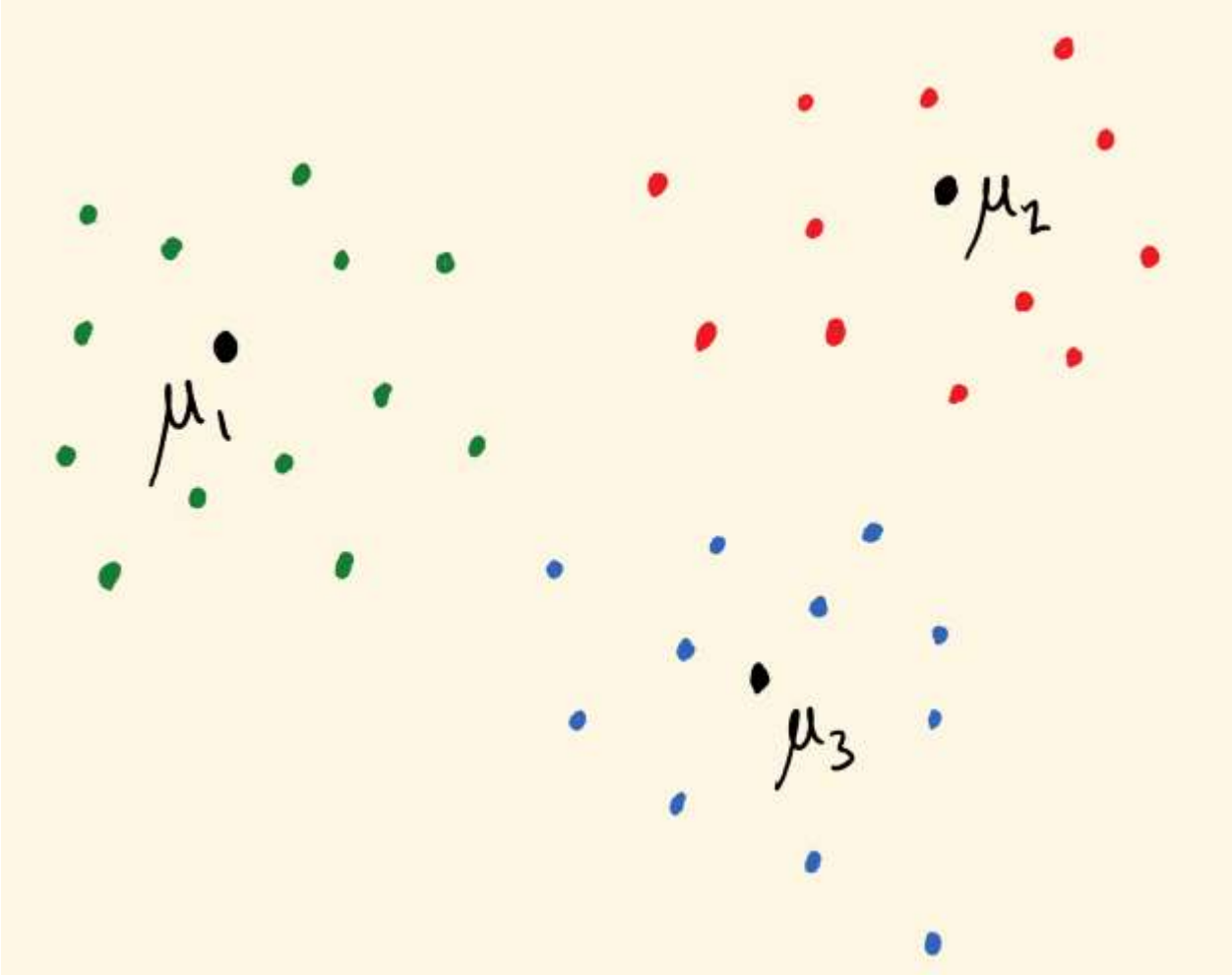
Final comments:

Run through the whole construction for robust mean estimation and will get an SDP with "intended solution"

$$w^{\otimes t} \otimes \mu^{\otimes t} \otimes X^{\otimes t} \otimes g^{\otimes t}$$

where w is $0/1$ indicator of a set of $(1-\varepsilon)m$ samples with mean μ, bounded covariance, and X, g are auxiliary variables.

Mixture Models



Mixture Models

Input: Samples $X_1, \ldots, X_n \in \mathbb{R}^d$ from mixture of $\mathcal{D}_1, \ldots, \mathcal{D}_k$ with means $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$

Goal: cluster $X_1, \ldots, X_n$ and/or estimate $\mu_1, \ldots, \mu_k$

1890s: Pearson *invents method of moments* to learn mixture of $k = 2$ Gaussians in $d = 1$ dimension

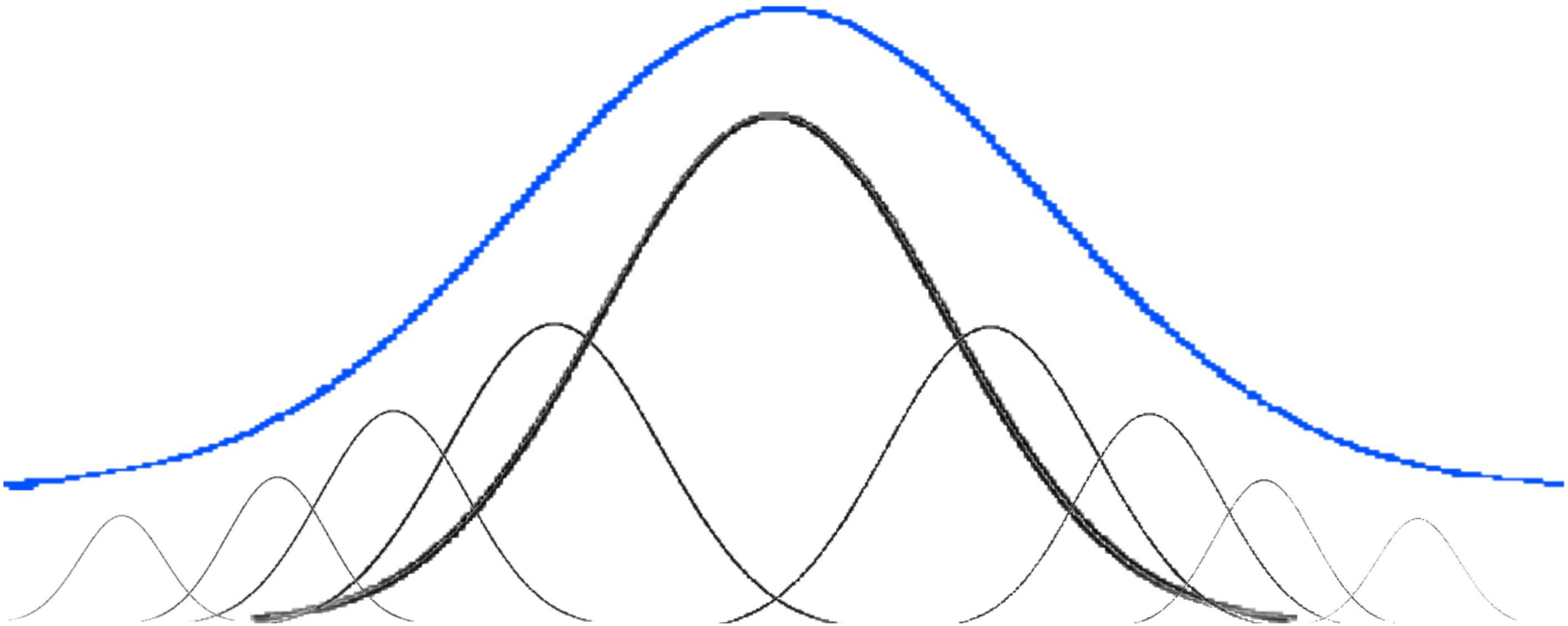
Now: ubiquitous generative model of inhomogeneous data – data from multiple populations

Today, inhomogeneous data is high-dimensional and can have many underlying components

Aim to use $\mathrm{poly}(d, k)$ samples and time

Information-Theoretic Barrier

Mixture of k Gaussians in $d = 1$ dimension can be $2^{-\Omega(k)}$-close to standard Gaussian [Moitra-Valiant]

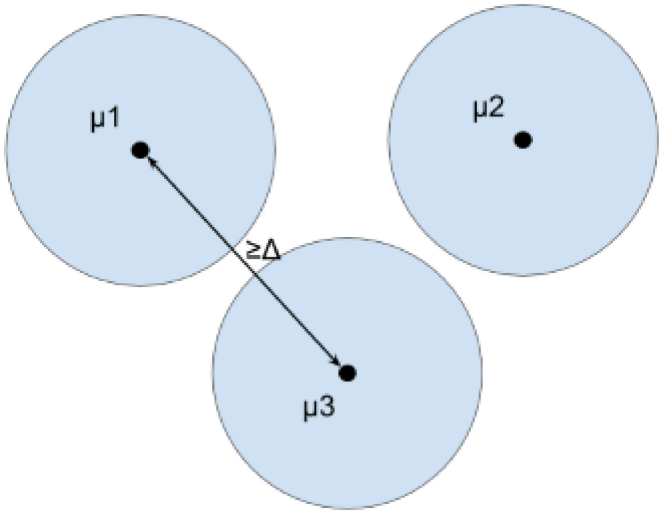


Separation Assumption

Input: Samples $X_1, \ldots, X_n \in \mathbb{R}^d$ from mixture of $\mathcal{D}_1, \ldots, \mathcal{D}_k$ with means $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$

Scaling: Assume covariances $\Sigma_1, \ldots, \Sigma_k \preceq I$

Δ-Separation assumption: $\|\mu_i - \mu_j\| \geq \Delta$ for some $\Delta > 0$.



For which $\Delta > 0$ and which $\mathcal{D}_1, \ldots, \mathcal{D}_k$ can $\mu_1, \ldots, \mu_k$ be estimated in $\mathrm{poly}(d, k)$ time, samples?

Mixture Models – Non-SoS and SoS Results

For now: $\mathcal{D}_i$ Gaussian, for simplicity, covariances $\Sigma \preceq I$, uniform mixture.

Δ	**Algorithm**	**Property of Gaussians**	**Reference**
$10\sqrt{d}$	greedy	distance to mean	[folklore]
$0.01\sqrt{d}$	spectral	bdd covariance	[D99]
$d^{1/4}$	EM (captured by greedy)	pairwise distances	[DS01]
$\min(d,k)^{1/4}$	PCA+EM/greedy	pairwise distances	[VW02]
k^{ε}	**sum of squares**	bdd $1/\varepsilon$ moments	[HL18,KSS18,DKS18]

lower bound: if $\Delta \leq o(\sqrt{\log k})$, need $\gg \mathrm{poly}(d,k)$ samples [RV17]

Theorem 1: If $\Delta = k^{\varepsilon}$, can recover μ_i's and cluster up to $1/\mathrm{poly}(k)$ error in time, samples $d^{O(1)}k^{O(1/\varepsilon)}$.

Theorem 2: If $\Delta = C\sqrt{\log k}$, can recover μ_i's and cluster up to $1/\mathrm{poly}(k)$ error in time, samples $d^{O(1)}k^{O(\log k)}$, for a universal constant C.

Proofs to Algorithms

Recall from yesterday:

Simple identifiability proof $\rightarrow$ SoS identifiability proof $\rightarrow$ SDP-based algorithm

Whiteboard time!

Certifiable Moment Boundedness

For which distributions $\mathcal{D}$ can SoS robustly estimate the mean?
For which Δ-separated $\mathcal{D}_1, \dots \mathcal{D}_k$ can SoS cluster and learn means?

Various names in literature: *certifiable subgaussianity, explicit boundedness*

In identifiability proofs, needed $\mathbb{E}_{X \sim \mathcal{D}} \langle X - \mu, u \rangle^t \leq t^{t/2} \|u\|^t$ for all $u \in \mathbb{R}^d$.

(Implies no event $\mathcal{E}$ with probability ε influences the mean by more than $\varepsilon^{1-1/t}$)

To SoS-ify the identifiability proof, will need

$$C^t t^{t/2} \cdot \|u\|^t - \mathbb{E}_{X \sim \mathcal{D}} \langle X - \mu, u \rangle^t = \sum s_i^2$$

True for t-wise products (next slide) and rotations thereof

Also true for Poincare distributions (an isoperimetry property) $\rightarrow$ strongly log-concave distributions [KSS18]

Certifiable moment bounds for product distributions

Let X on $\mathbb{R}^d$ be t-wise independent, assume $\mathbb{E}\, X = 0$ and $\mathbb{E}\, X_i^t \leq B$.

Assume coordinates X_i are symmetric about 0 (otherwise replace with $X - X'$ for independent draw X')

Then $\mathbb{E}\, X^\alpha = 0$ for any odd α with $|\alpha| \leq t$. E.g $\mathbb{E}\, X_1^2 X_{10}^5 = 0$

$\mathbb{E}\langle X, u\rangle^t = \sum_{|\alpha|=t} u^\alpha\, \mathbb{E}\, X^\alpha = \sum_{|\alpha|=t, \alpha \text{ even}} u^\alpha\, \mathbb{E}\, X^\alpha$

Let $c_\alpha = B - \mathbb{E}\, X^\alpha \geq 0$.

Then $B \cdot \|u\|^t - \mathbb{E}\langle X, u\rangle^t = \sum_{|\alpha|=t, \alpha \text{ even}} c_\alpha u^\alpha$ is an SoS.

Which distributions have certifiably bounded moments?

Known: Poincare (with dimension-independent constant), hence strongly log-concave

The frontier: log-concave (implied by KLS conjecture via Poincare?)

Moonshot: subgaussian?? (probably too broad)

Open problem: prove a hardness result for some subgaussian distribution

Wrapping up

If you remember only one thing: **simple identifiability proofs $\rightarrow$ computationally efficient algorithms.**

SoS offers provable guarantees for broadest known classes of distributions for clustering, robust moment estimation.

And robust regression, robust sparse recovery, . . .

Proofs to algorithms recipe also works for dictionary learning, matrix/tensor completion, tensor principal component analysis, and more

Thanks!!