Partial Specification and Agnostic Clustering.

Last time:

- Partial specification for linear regression
  - Non-robust method assuming Gaussianity
  - Robust method based on robust standard errors

This time.

- Clustering

Setting.

Distributions $p_1, \dots, p_k$

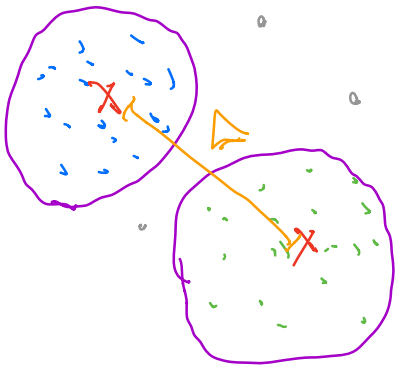
Points $X_1, \dots, X_n$ sampled from $p = \alpha_1 p_1 + \alpha_2 p_2 + \dots + \alpha_k p_k$.

A priori bound: $\alpha_{min} = \min_j \alpha_j$.

Two goals.

- Parameter recovery: estimate parameter e.g. $\mu_j = \mathbb{E}_{p_j}[x]$ of each p_j
- Cluster recovery: figure out which cluster p_j each point x_i came from.

$k = 2$:



Sometimes have separation condition Δ

Settings.

- Parametric: each p_j has known parametric form
  e.g. $N(\mu_j, \Sigma_j)$
- Agnostic: no parametric form, only assume e.g. sub-Gaussianity, bounded moments, etc.
- Robust: also allow ε-fraction of outliers,
  i.e. $\alpha_1 + \dots + \alpha_k = 1 - \varepsilon$

Warm-up: Parametric setting

$$p_j = N(\mu_j, \Sigma_j)$$

$$p = \sum_j \alpha_j p_j = \sum_j \alpha_j N(\mu_j, \Sigma_j)$$

Key questions

- Identifiability: are (μ_j, Σ_j) unique?
  - If not, can't recover (μ_j, Σ_j) even when $n = \infty$.
- Statistical rate: assuming identifiable, how many samples do we need? → $1/\sqrt{n}$ rate w/ MLE

Complicated but "standard"

Only up to permutation

$\alpha_1 \leftrightarrow \alpha_2$

$\mu_1 \leftrightarrow \mu_2$

$\Sigma_1 \leftrightarrow \Sigma_2$

Proposition. As long as (μ_j, Σ_j) are all distinct, then $(\alpha_j, \mu_j, \Sigma_j)_{j=1}^K$ identifiable up to permutation.

Pf. Suppose not, then

$$\sum_j \alpha_j \, N(\mu_j, \Sigma_j) + \sum_j (-\alpha_j') N(\mu_j', \Sigma_j') = 0$$

$\Rightarrow$ linear dependence among PDFs

1D.

$$\sum_{j=1}^m c_j \exp\left(-\frac{(x-\mu_j)^2}{\sigma_j^2}\right) \Big/ \sqrt{2\pi\sigma_j^2} = 0 \quad \forall x$$

$\Rightarrow$ Take MGF

$$\sum_{j=1}^m c_j \exp\left(\frac{1}{2}\sigma_j^2 \lambda^2 + \mu_j \lambda\right) = 0 \quad \forall \lambda.$$

As $\lambda \to \infty$, only largest σ_j's matters.

call it σ_{max},

$\rightarrow$ Divide through by σ_{max}

$\Rightarrow$

$$\sum_{j=1}^m c_j \exp(\mu_j \lambda) \to 0 \quad \text{as } \lambda \to \infty$$

$$\text{as } \lambda \to -\infty$$

$\mu_j > 0$ and $\mu_j < 0$ $\Rightarrow$ contradiction.

$\Rightarrow$ No 1D linear relations.

n-D. Take random projection of n-D Gaussians

Linear relation in n-D $\Rightarrow$ relation in 1-D

$\Rightarrow \Leftarrow$.

$$\sum_{j=1}^{m} c_j N(\mu_j, \Sigma_j) = 0$$

Random projection onto v

$$\Rightarrow \sum_{j=1}^{m} c_j N(v^T\mu_j, v^T\Sigma_j v) = 0$$

Impossible unless parameters are not actually distinct.

probability zero

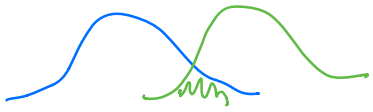
$$v^T\mu_j = v^T\mu_{j'}$$

$$v^T\Sigma_j v = v^T\Sigma_{j'} v \quad \text{for } j \neq j'$$

Cluster recovery: need separation

hard to cluster in overlap

- Overlap b/t $N(\mu, \Sigma)$ and $N(\mu', \Sigma')$?
- Simplify: $\Sigma = \Sigma' = \sigma^2 \cdot I$



okay if $K \cdot \Phi\left(\frac{\|\mu - \mu'\|_2}{\sigma}\right) \leq \varepsilon$

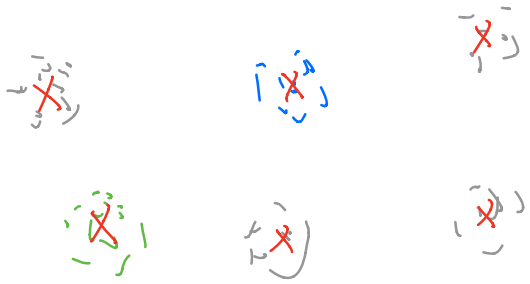
$$\Delta = \min_{j} \|\mu_j - \mu_{j'}\|_2$$

$\Rightarrow$ want $K \cdot \Phi\left(\frac{\Delta}{\sigma}\right) \leq \varepsilon$

$$\Delta \gg \sigma \sqrt{\log(K/\varepsilon)}.$$

Partial specification.

- Just assume sub-Gaussian, resilience
- Take $K = 1$, $\varepsilon = 1 - \alpha$ — 10% good points
  $\alpha = 0.1$ — 90% outliers
- List-decoding. Output $m = O\left(\frac{1}{\alpha}\right)$ "candidates" $\hat{\mu}_\ell$ s.t. $\|\mu - \hat{\mu}_\ell\|_2$ is small for some ℓ.



Can actually handle $k > 1$ clusters automatically via list-decoding

---

Assumption. Set S of an "good" points w/ mean μ, which is $(\rho, \frac{\alpha}{4})$-resilient.

Proposition. Given assumption, can output $m \le \frac{2}{\alpha}$ candidates $\hat{\mu}_1, \dots, \hat{\mu}_m$ s.t.

$$\|\hat{\mu}_j - \mu\|_2 \le \frac{8\rho}{\alpha} \quad \text{for some } j.$$

Interpretation. sub-Gaussian $\Rightarrow (\sigma\alpha\sqrt{\log(1/\alpha)}, \alpha)$-resilient

$$\rho = O(\sigma\alpha\sqrt{\log(1/\alpha)})$$

$$\frac{8\rho}{\alpha} = O(\sigma\sqrt{\log(1/\alpha)}).$$

k clusters each w/ $\alpha = \frac{1}{k}$

$\Rightarrow O(\sigma\sqrt{\log k})$

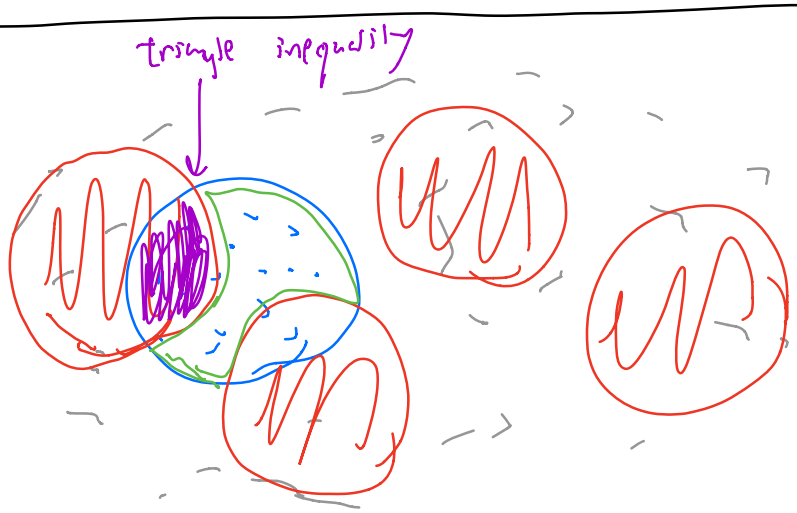
$\Rightarrow$ good cluster recovery if $\Delta > 2\sqrt{\log(k)}$.

bounded covariance

$\Rightarrow (\sigma\sqrt{\alpha}, \alpha)$ -resilient

$\delta_f = O(\sigma/\sqrt{\alpha})$. $\Rightarrow$ $\sigma\sqrt{K}$

↑ Threshold at which MLE can be done "easily" efficiently



Idea.

- Cover set of all points w/ resilient subsets

How to pick red sets: maximal collection of sets S_j s.t.:

- $|S_j| \geq \frac{\alpha}{2} n$
- S_j is $\left(\frac{4}{\alpha}\rho, 1-\frac{\alpha}{2}\right)$-resilient $\nearrow \left(\frac{\frac{\alpha}{2}\cdot\frac{4}{\alpha}\rho}{1-\frac{\alpha}{2}}, \frac{\alpha}{2}\right) \approx (2\rho, \frac{\alpha}{2})$-resilient
- $S_j \cap S_{j'} = \emptyset \quad \forall j \neq j'$.

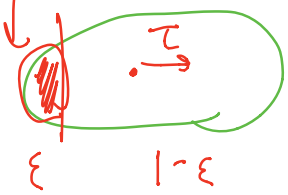
$\downarrow$

$\hat{\mu}_j$ = mean of S_j

Output $\hat{\mu}_1, \ldots, \hat{\mu}_m$.

Lemma P is (ρ, ε)-resilient if and only if it is $\left(\frac{1-\varepsilon}{\varepsilon}\rho, 1-\varepsilon\right)$-resilient.

$\tau' \varepsilon + (1-\varepsilon)\tau = 0 \Rightarrow \tau' = -\frac{(1-\varepsilon)}{\varepsilon}\tau$



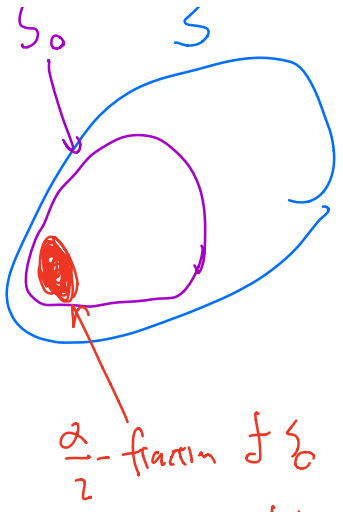
Claim: Some S_j intersects $S \Rightarrow \hat{\mu}_j \approx \mu$.

$S_0 = S \setminus [S_1 \cup \cdots \cup S_m]$.

(1) $|S_0| \geq \frac{\alpha}{2} n$. Then S_0 can be added to S_j.

S is $(\rho, \frac{\alpha}{4})$-resilient

$\Rightarrow \left(\frac{4}{\alpha}\rho, 1-\frac{\alpha}{4}\right)$-resilient (by Lemma)



$\Rightarrow$ $\frac{\alpha}{4}$-fraction of S can't be ∞ far away.

$\Rightarrow$ Any subset of $\frac{\alpha}{4}|S|$ points in S has mean at most $\frac{4}{\alpha}\rho$ away,

$\Rightarrow$ Any subset of $\frac{\alpha}{2}|S_0|$ points in S_0 is a subset of $\frac{\alpha}{4}|S|$ points in $S_0 \subseteq S$.

(because $|S_0| \geq \frac{1}{2}|S|$).

$\Rightarrow$ S_0 is $\left(\frac{4}{\alpha}\rho, 1-\frac{\alpha}{2}\right)$-resilient.

$\Rightarrow$ S_0 could have been added, $\Rightarrow\Leftarrow$.

(2) $|S_0| < \frac{\alpha}{2}n$.

$|S| = \alpha n$

S_i: $S \cap S_1$, $S \cap S_2$, $\dots$, $S \cap S_m$, S_0 ($< \frac{\alpha}{2}n$)

$> \frac{\alpha}{2}n$

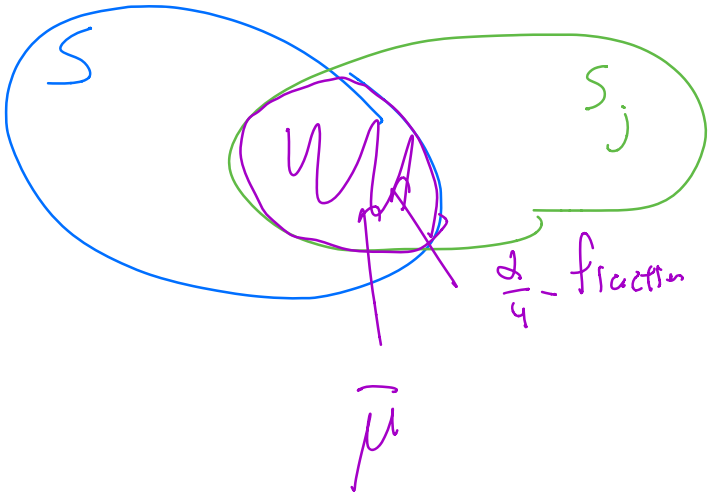
$$\sum_j |S \cap S_j| \geq \frac{\alpha}{2}n$$

$$\sum_j |S_j| \le n \quad \text{(by disjointness)}$$

$$\Rightarrow |S \cap S_j| \ge \frac{\alpha}{2}|S_j| \quad \text{for some } j.$$

$$\ge \frac{\alpha}{4}|S| \quad \leftarrow \text{since } |S_j| \ge \frac{\alpha}{2}n$$

$$= \frac{1}{2}|S|.$$



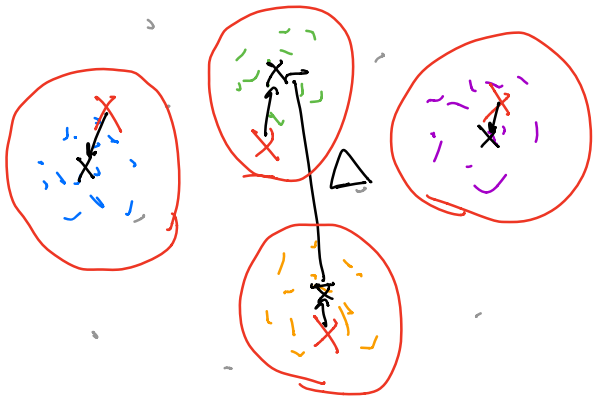
$$\|\bar{\mu} - \mu(S)\|_2 \le \frac{4\rho}{\alpha} \quad \text{(by resilience w/ } \varepsilon = 1 - \tfrac{\alpha}{4}\text{)}$$

$$\|\bar{\mu} - \mu(S_j)\|_2 \le \frac{4\rho}{\alpha}$$

$$\Rightarrow \|\mu(S) - \mu(S_j)\|_2 \le \frac{8\rho}{\alpha}.$$

Some $\hat{\mu}_j$ with $\frac{8\rho}{\alpha}$ of μ, as claimed. ∎

Better result assuming well-separated clusters.



Proposition. Suppose $x_1, \ldots, x_n$ can be partitioned into sets $C_1, \ldots, C_k$ of size $\alpha_1 n, \ldots, \alpha_k n$, plus ξn outliers (so $\alpha_1 + \cdots + \alpha_k = 1 - \xi$).

Further suppose:

- $2\xi \leq \alpha_{\min} = \min_{j=1}^{k} \alpha_j$.
- The means are well-separated: $\Delta > \frac{c\rho}{\xi}$, where $\Delta = \min_{j \neq j'} \|\mu_j - \mu_{j'}\|_2$.
- Each cluster is $(\rho, 2\xi/\alpha)$-resilient. ((ρ, ξ)-resilient)

Then we can output clusters $\hat{C}_j$ such that:

① $|C_j \Delta \hat{C}_j| \le O(\frac{\varepsilon}{\alpha})|C_j|$ (cluster)

② $\|\mu(\hat{C}_j) - \mu(C_j)\|_2 \le 2\rho$. (parameter)

$\rho = \sigma\sqrt{\log(K)}$