Robust Statistics

Basic estimation problem, but we'll go in a new direction:

Given samples from a 1-d Gaussian $\mathcal{N}(\mu, \sigma^2)$, can we estimate its parameters?

Of course! Use:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad , \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

These are examples of <u>maximum likelihood</u> paradigm (Ronald Fisher 1912-1922)

(1) <u>consistent</u>: converges to true parameters as $N \to \infty$ under tame conditions

(2) <u>asymptotically normal</u>: has smallest variance among all unbiased estimators

<u>Main Question</u>: But what if the samples are only approximately Gaussian?

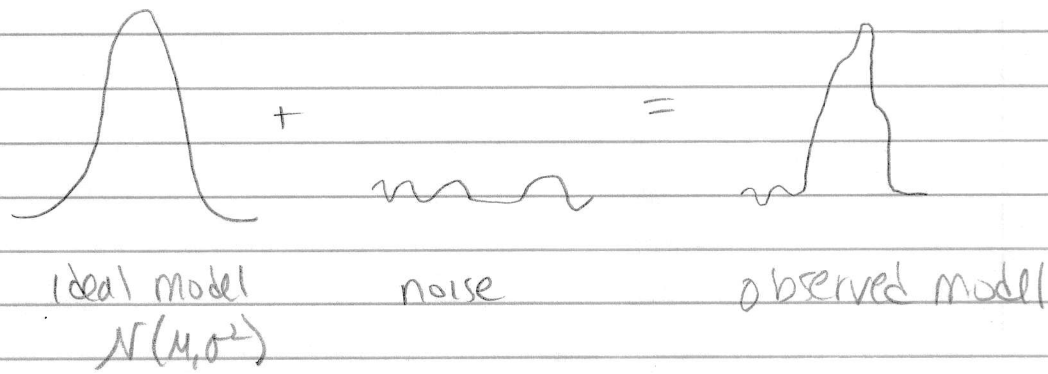
<u>definition</u>: In the <u>strong contamination model</u>:

(1) m samples are drawn iid from $P \in \mathcal{D}$

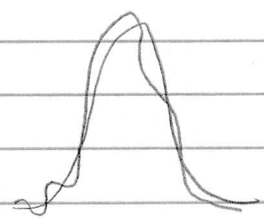
known class of distributions

② adversary is allowed to <u>arbitrarily</u> corrupt an ε-fraction of samples

Pictorially:



we can think of the area between the curves

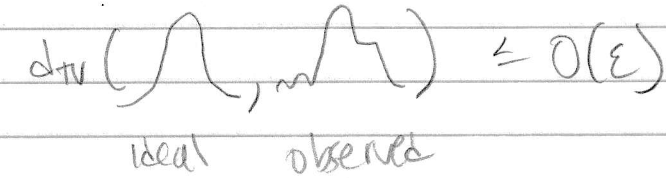


as representing the samples the adversary has added/deleted

<u>definition</u>: the <u>total variation distance</u> between r.v.s. with pdfs $f(x)$ and $g(x)$ is

$$d_{TV}(f,g) = \frac{1}{2}\int |f(x)-g(x)|\,dx$$

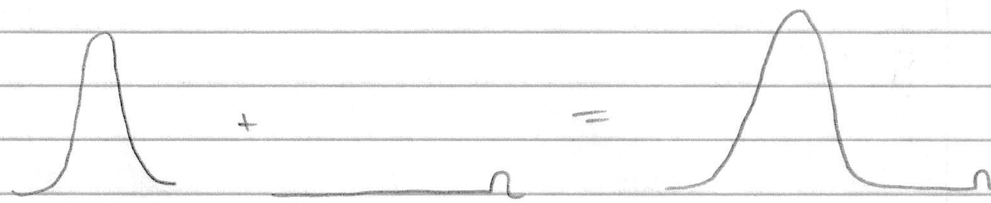
In our model, we have



Can we estimate the true Gaussian in $O(\varepsilon)$ in TV?

<u>Observation</u>: The empirical mean/variance (MLE) do not work!

e.g.



but as the bump $\to \infty$, $\hat{\mu}$ and $\hat{\sigma}$ diverge

So what should we do? Consider

$$\hat{\mu} = \text{median}(\{x_i\}_i)$$

$$MAD = \text{median}(\{|x_i - \hat{\mu}|\}_i)$$

$$\hat{\sigma} = \frac{MAD}{\Phi^{-1}(3/4)}$$

Φ ← c.d.f of a standard Gaussian

<u>Proposition [folklore]</u> Given ε-corrupted samples from a 1-d Gaussian $\mathcal{N}(\mu, \sigma^2)$ we have

$$d_{TV}(\mathcal{N}(\hat{\mu}, \hat{\sigma}^2), \mathcal{N}(\mu, \sigma^2)) \leq O(\varepsilon)$$

provided $m \geq \frac{C \ln 1/\delta}{\varepsilon^2}$ (δ ← failure prob.)

In the nomenclature of TCS

"properly agnostically learning a 1-d Gaussian"

output something from the class

it's not actually Gaussian, but want to do well if it's close

<u>Main Question:</u> what about in high-dimensions?

Given ε-corrupted samples from a d-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$, can we efficiently estimate s.t.

$$d_{TV}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu, \Sigma)) \leq \tilde{O}(\varepsilon)?$$

Special cases:

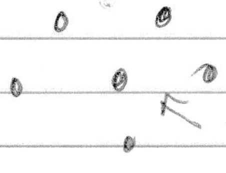
1. unknown mean: $\mathcal{N}(\mu, I)$
2. unknown covariance: $\mathcal{N}(0, \Sigma)$

What's known in robust statistics?

<u>def:</u> The <u>Tukey depth</u> of a point X w.r.t. a dataset $X_1, \ldots X_m$ is

$$\min_{\text{1-d proj}} \min(\# \text{ points to left / right of } X)$$

e.g.



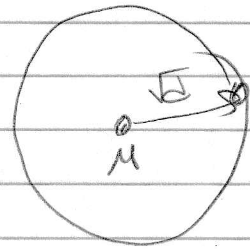
Tukey depth = 2

Fact: Given ε-corrupted samples from $\mathcal{N}(\mu, I)$ the Tukey median (Tukey deepest point over all space) satisfies $d_{TV}(\mathcal{N}(\hat{\mu}, I), \mathcal{N}(\mu, I)) \leq O(\varepsilon)$

Unfortunately:

Lemma: the Tukey Median is NP-hard to compute.

Alternatively we could take coordinatewise median, but that would only get $TV \leq \varepsilon\sqrt{d}$



Because direction of corruption might not be axis aligned

Theorem: [Diakonikolas, Li, Kamath, Kane, Moitra, Stewart] There is a polynomial time/sample complexity algorithm that finds $\hat{\mu}, \hat{\Sigma}$ satisfying

$$d_{TV}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu, \Sigma)) \leq O(\varepsilon \log^{3/2} 1/\varepsilon)$$

in the ε-strong contamination model

[Lai, Rao, Vempala] also gave an algorithm satisfying

$$TV \leq O(\sqrt{\varepsilon} \log d)$$

when the covariance is bounded

General Recipe

① Find an appropriate parameter distance

② Detect when naive estimator has been compromised via method of moments

③ win-win: Find good parameters, or make progress by filtering out corruptions

Unknown Mean

Consider the special case when we get ε-corrupted samples from $\mathcal{N}(\mu, I)$

Definition: The KL-divergence is

$$d_{KL}(f \| g) = \int_{-\infty}^{\infty} f(x) \ln \frac{f(x)}{g(x)} \, dx$$

Fact: For two Gaussians, we have

$$d_{KL}(\mathcal{N}(\mu_1, \Sigma_1) \| \mathcal{N}(\mu_2, \Sigma_2)) =$$

$$\frac{1}{2}\left(\ln \frac{\det(\Sigma_2)}{\det(\Sigma_1)} + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2) - d \right)$$

when $\Sigma_1 = \Sigma_2 = I$ this simplifies to:

$$d_{KL}(\mathcal{N}(\hat{\mu}, I), \mathcal{N}(\mu, I)) = \frac{1}{2} \|\hat{\mu} - \mu\|_2^2$$

Fact [Pinsker's Inequality]

$$d_{TV}(f,g)^2 \le \frac{1}{2} d_{KL}(f,g)$$

Putting it all together, we have

Lemma: If we can estimate

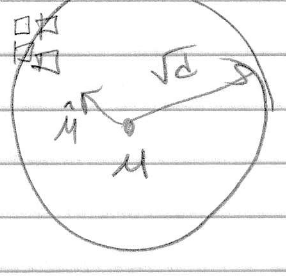
$$\|\hat{\mu} - \mu\|_2 \le \tilde{O}(\varepsilon)$$

then we'd get $d_{TV}(\mathcal{N}(\hat{\mu}, I), \mathcal{N}(\mu, I)) \le \tilde{O}(\varepsilon)$

Thus we have our parameter distance!

Detecting Corruptions

How can the adversary move the empirical mean by $\varepsilon\sqrt{d}$?



But in this case, the projected variance on the direction $\hat{\mu} - \mu \gg 1$

Takeaway: To mess up the first moment, an adversary would have to mess up the second moment too

Key Lemma: If $X_1, \ldots X_m$ are ε-corrupted samples from $\mathcal{N}(\mu, I)$ and

① $m \geq c \dfrac{d \ln 1/\varepsilon}{\varepsilon^2}$

② $\hat{\mu} = \frac{1}{m} \sum X_i, \quad \hat{\Sigma} = \frac{1}{m} \sum (X_i - \hat{\mu})(X_i - \hat{\mu})^T$

then we have w/ probability $\geq 1-\delta$

$$\|\mu - \hat{\mu}\|_2 \geq C' \varepsilon \sqrt{\log 1/\varepsilon} \Rightarrow \|\hat{\Sigma} - I\|_2 \geq C'' \varepsilon \log 1/\varepsilon$$

Thus we can detect when the empirical mean has been corrupted

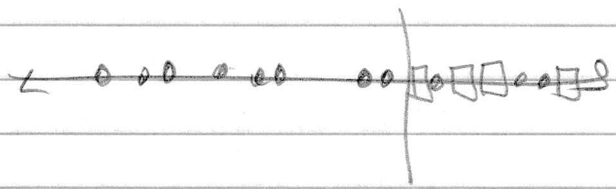
Spectral Filtering

If $\|\hat{\Sigma} - I\|_2 < C'' \varepsilon \log 1/\varepsilon$ then we can just output $\hat{\mu}$ and are guaranteed

$$\|\mu - \hat{\mu}\|_2 \leq C' \varepsilon \sqrt{\log 1/\varepsilon} \Rightarrow$$

$$d_{TV}(\mathcal{N}(\hat{\mu}, I), \mathcal{N}(\mu, I)) \leq O(\varepsilon \sqrt{\log 1/\varepsilon})$$

Otherwise consider $v =$ direction of largest variance



we can compute a threshold T s.t. throwing out the samples above T results in throwing out more corrupted than uncorrupted points

Unknown Covariance

Using the formula for KL-divergence and Pinsker's inequality it can be shown

Fact: For two Gaussians $\mathcal{N}(0,\Sigma)$ and $\mathcal{N}(0,\hat{\Sigma})$

$$d_{TV}(\mathcal{N}(0,\hat{\Sigma}),\mathcal{N}(0,\Sigma)) \le O(\| I - \hat{\Sigma}^{-1/2}\Sigma\hat{\Sigma}^{-1/2}\|_F)$$

Mahalonobis distance

Can we use the fourth moment to detect corruptions in the second moment?

Lemma: Let $X \sim \mathcal{N}(0,\Sigma)$. Then consider

$$M = \mathbb{E}[(X \otimes X)(X \otimes X)^T]$$

restricted to flattenings of symmetric $d \times d$ matrices we have

$$M = 2\Sigma^{\otimes 2} + (\Sigma^\flat)(\Sigma^\flat)^T$$

Now imagine we get ε-corrupted samples $X_1, \ldots X_m$ from $\mathcal{N}(0,\Sigma)$. Then define

$$\hat{\Sigma} = \frac{1}{m} \sum X_i X_i^T \text{ and}$$

$$Y_i = (\hat{\Sigma})^{-1/2} X_i$$

If $\Sigma = \hat{\Sigma}$ then $Y_i \sim N(0, I)$ in which case restricted to subspace of symmetric matrices

$$F = \frac{1}{m} \sum (Y_i \otimes Y_i)(Y_i \otimes Y_i) \sim 2\underset{\substack{\uparrow \\ d \times d}}{I} + (I^b)(I^b)^T$$

If we consider

$$\max_{z \text{ s.t}} z^T F z \sim 2$$

z = flattening of $d \times d$ trace zero, symmetric matrix

We show that if $\|I - \hat{\Sigma}^{-1/2} \Sigma \hat{\Sigma}^{-1/2}\|_F \gg \varepsilon \log 1/\varepsilon$ then we must have

$$\max_{z} z^T (F - 2I) z \gg \varepsilon \log 1/\varepsilon$$

z = flattening...

And similarly if the generalized eigenvalues of F are small, we can output $\hat{\Sigma}$

Otherwise we can find a direction (degree two polynomial) to filter on

Assembling the Algorithm

Given ε-corrupted samples from $\mathcal{N}(\mu, \Sigma)$

① Doubling trick

$$X_i - X_i' \sim_{2\varepsilon} \mathcal{N}(0, 2\Sigma)$$

Use algorithm for unknown covariance

② Agnostic isotropic position

$$\hat{\Sigma}^{-1/2} X_i \sim_{\tilde{O}(\varepsilon)} \mathcal{N}(\hat{\Sigma}^{-1/2}\mu, I)$$

Use algorithm for unknown mean

More Robust Statistics

Many subsequent directions

① Handling more errors ($\varepsilon > 1/2$) with list decoding

② Giving evidence of lower bounds, e.g. Statistical query algorithms can't get $O(\varepsilon)$ error

③ weakening the distributional assumptions, just need bounds on moments

④ exploiting sparsity e.g. $\|\mu\|_0 \leq k$

⑤ more complex generative models

Let's return to GMMs:

Earlier we gave a polynomial time/sample complexity non-robust learner. Implicitly we showed:

definition We say a family of distributions $\mathcal{D}$ is polynomially identifiable if

$\forall P_1, P_2 \in \mathcal{D}$ that have ε-different parameters

$$\Rightarrow d_{TV}(P_1, P_2) \geq \text{poly}(\varepsilon, \tfrac{1}{d})$$

[Kalai, Moitra, Valiant]

Corr: Mixtures of two Gaussians in d-dimensions are polynomially identifiable

Otherwise the algorithm wouldn't work

But to get robust algorithms we need a much stronger notion

definition: We say a family of distributions $\mathcal{D}$ is robustly identifiable if

$\forall P_1, P_2 \in \mathcal{D}$ that have parameter discrepancy ε

$$\Leftrightarrow d_{TV}(P_1, P_2) \geq \text{poly}(\varepsilon)$$

Notice that for unknown mean/covariance

we could take ℓ_2 / Mahalonobis distance in parameters

<u>Theorem</u> [Liu, Moitra] Mixtures of two Gaussians in d-dimensions are robustly identified by a constant number of <u>Hermite moments</u>

The proof will be via generating functions and differential operators

<u>Generating Functions</u>

Consider a Gaussian $G = \mathcal{N}(\mu, I+\Sigma)$ and let

$$M(X) = X^T \mu \quad \text{and} \quad \Sigma(X) = X^T \Sigma X$$

vector of formal variables

<u>Key lemma</u>

$$e^{M(X)y + \frac{1}{2}\Sigma(X)y^2} = \sum_{m=0}^{\infty} \frac{1}{m!} \mathop{\mathbb{E}}_{Z \sim G}\left[H_m(Z, X)\right] y^m$$

where $H_m(Z, X)$ = Hermite moment tensor, i.e.

$$H_m(Z, X=v) = \text{mth Hermite moment of 1-d Gaussian r.v. } \underset{v^T Z}{}$$

Let's make it even simpler to see why

<u>Easier Lemma</u>: Let $G = \mathcal{N}(\mu, 1+\sigma^2)$. Then

$$e^{\mu y + \frac{\sigma^2}{2} y^2} = \sum_{m=0}^{\infty} \frac{1}{m!} \mathop{\mathbb{E}}_{z \sim G}\left[h_m(z)\right] y^m$$

<u>Proof</u>: Expanding the LHS we get

$$1 + \left(\mu y + \frac{\sigma^2}{2} y^2\right) + \frac{\left(\mu y + \frac{\sigma^2}{2} y^2\right)^2}{2} + \frac{\left(\mu y + \frac{\sigma^2}{2} y^2\right)^2}{6}$$

Collecting terms

$$1 + \mu y + \underbrace{\left(\frac{\mu^2 + \sigma^2}{2}\right)} y^2 + \underbrace{\left(\frac{\mu^3 + 3\mu\sigma^2}{6}\right)} y^2 + \dots$$

$$\mathop{\mathbb{E}}_{z \sim G}\left[h_2(z)\right] \stackrel{\Delta}{=} \mathop{\mathbb{E}}_{z \sim G}\left[z^2 - 1\right] = \mu^2 + 1 + \sigma^2 - 1$$

$$\mathop{\mathbb{E}}_{z \sim G}\left[h_3(z)\right] \stackrel{\Delta}{=} \mathop{\mathbb{E}}_{z \sim G}\left[z^3 - 3z\right] = \mu^3 + 3\mu(1+\sigma^2) - 3\mu$$

etc. $\blacksquare$

Now the Key Lemma extends immediately to mixtures

$$\mathcal{M} = w_1 \mathcal{N}(\mu_1, I + \Sigma_1) + \dots + w_k \mathcal{N}(\mu_k, I + \Sigma_k)$$

Then we have:

Key Lemma'

$$\sum_{j=1}^{k} w_j e^{\mu_j(x)y + \frac{1}{2}\Sigma_j(x)y^2} = \sum_{m=0}^{\infty} \frac{1}{m!} \mathbb{E}_{Z\sim\mu}[H_m(Z,x)]\, y^m$$

Back to Identifiability

We WTS that

$$\text{(1)} \quad \sum_{j=1}^{k} w_j e^{\mu_j(x)y + \frac{1}{2}\Sigma_j(x)y^2} = \sum_{j=1}^{k} \hat{w}_j e^{\hat{\mu}_j(x)y + \frac{1}{2}\hat{\Sigma}_j(x)y^2}$$

$$\Updownarrow$$

(2) the parameters match, i.e. the mixtures are the same on a component-by-component basis

Now consider the differential operator

$$D = \partial_y - (\mu + \sigma^2 y)$$

applied to the generating function $e^{\mu y + \frac{1}{2}\sigma^2 y^2}$

Fact: $D(e^{\mu y + \frac{1}{2}\sigma^2 y^2}) = 0$

This holds as a formal identity

Observation D is a polynomial rearrangement on the series expansion

This works in high dimensions too. Let

$$\mathcal{D} \triangleq d_y - (\mu(x) - \Sigma(x) y)$$

Then:

Fact: $\mathcal{D}\left(e^{\mu(x)y + \frac{1}{2}\Sigma(x)y^2}\right) = 0$

which is a complicated, but explicit multivariate polynomial rearrangement

Main Question: what happens when you apply $\mathcal{D}$ to another component?

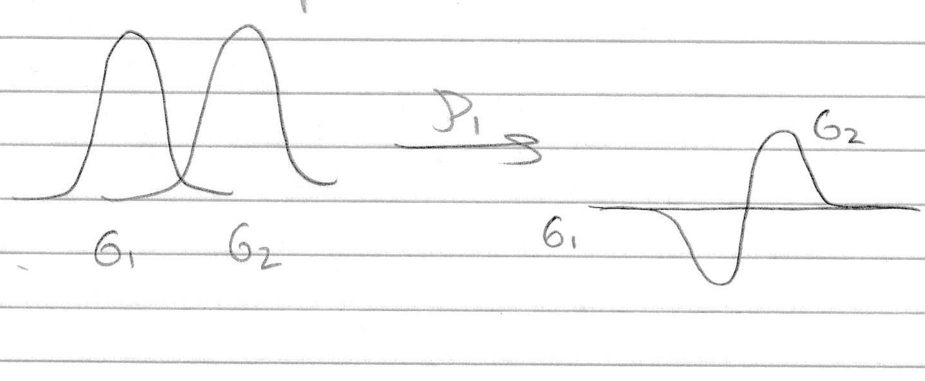
Fact: $\mathcal{D}\left(e^{\hat{\mu}(x)y + \frac{1}{2}\hat{\Sigma}(x)y^2}\right) = P e^{\hat{\mu}(x)y + \frac{1}{2}\hat{\Sigma}(x)y^2}$

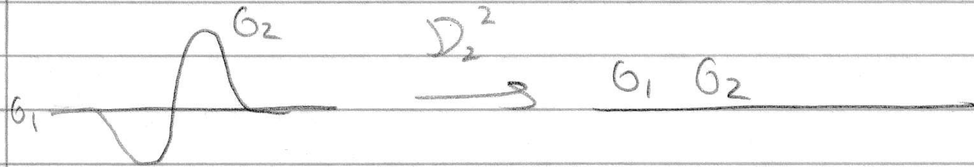
where $P = \left(\hat{\mu}(x) - \mu(x) + y\left(\hat{\Sigma}(x) - \Sigma(x)\right)\right)$

And finally

Fact: $\mathcal{D}\left(P e^{\mu(x)y + \frac{1}{2}\Sigma(x)y^2}\right) = \frac{dP}{dy} e^{\mu(x)y + \frac{1}{2}\Sigma(x)y^2}$

Now we can use differential operators to isolate a component





Thus consider

$$\begin{array}{ccc} \mathcal{M} & D_{k-1}^{2^{2k-2}} \cdots D_1^{2^k} \quad \text{vs.} \quad \hat{D}_k^{2^{k-1}} \cdots \hat{D}_1 & \hat{\mathcal{M}} \\ \downarrow & & \downarrow \\ P(x,y)\, e^{\mu_j(x) y + \frac{1}{2}\Sigma_j(x) y^2} & & 0 \end{array}$$

This implies one of their $f(x)$ degree moments is different!

Can show this argument gets robust identifiability!

Theorem [Liu, Moitra] There is a polynomial time algorithm for robustly learning a GMM with accuracy that depends polynomially on the corruption rate

[Bakshi et al] get related results, but for weaker notion of proper density estimation

[Liu, Moitra] get $\tilde{O}(\varepsilon)$ accuracy for semi-proper density estimation