Robust Statistics

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

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

Of course! Use:

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

These are examples of maximum likelihood paradigm (Ronald Fisher 1912-1922)

1. consistent: converges to true parameters as $N \to \infty$ under tame conditions
2. asymptotically normal: has smallest variance among all unbiased estimators

Main question: But what if the samples are only approximately Gaussian?

definition: In the strong contamination model:

1. m samples are drawn i.i.d. from $Pe \mathcal{D}$
The adversary is allowed to arbitrarily corrupt an $\epsilon$-fraction of samples.

Pictorially:

\[
\text{ideal model} + \text{noise} = \text{observed model}
\]

\[\mathcal{N}(4, 0^2)\]

We can think of the area between the curves as representing the samples the adversary has added/deleted.

Definition: The total variation distance 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

\[
d_{TV}(\mathcal{N}, \mathcal{N}) = 0(\epsilon)
\]

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

Observation: The empirical mean/ variance do not work!

e.g.

\[
\begin{align*}
\hat{\mu} & = \text{median } (x_i) \\
\hat{\sigma} & = \frac{\text{MAD}}{\Phi^{-1}(\frac{3}{4})} \\
\text{MAD} & = \text{median } (|x_i - \hat{\mu}|)
\end{align*}
\]

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

So what should we do? Consider

\[
\hat{\mu} = \text{median } (x_i) \\
\text{MAD} = \text{median } (|x_i - \hat{\mu}|) \\
\hat{\sigma} \leq \frac{\text{MAD}}{\Phi^{-1}(\frac{3}{4})}
\]

cdf of a standard Gaussian

Proposition [folklore] Given $\varepsilon$-corrupted samples from a $1-\delta$ Gaussian $N(\mu, \sigma^2)$ we have

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

provided $m \geq C \ln \frac{1}{\varepsilon^2} \\text{failure prob.}$
In the nomenclature of TCS

" properley agnostically learning a 1-d Gaussian"

- output something its not actually
- from the class Gaussian, but want
to do well if its close

Main Question: what about in high-dimensions?

Given ε-corrupted samples from a d-dimensional Gaussian \( N(\mu, \Sigma) \), can we efficiently estimate \( d_{TV}(N(\mu, \Sigma), N(\mu, \Sigma)) \leq \Theta(\varepsilon) \)?

Special cases:

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

What's known in robust statistics?

def: The Tukey depth of a point \( X \) w.r.t. a dataset \( X_1, \ldots, X_m \) is

\[
\min_{1-d proj} \min \left( \# \text{ points left of } X \right) \min \left( \# \text{ points right of } X \right)
\]

e.g. \( \bullet \bullet \bullet \) \( \Rightarrow \) Tukey depth = 2
Fact: Given $\varepsilon$-corrupted samples from $\mathcal{N}(\mu, I)$, the Tukey median (Tukey deepest point over all space) satisfies $d_{TV}(\mathcal{N}(\hat{\mu}, I),\mathcal{N}(\mu, I)) = O(\varepsilon)$. 

Unfortunately:

Lemma: The Tukey median is NP-hard to compute.

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

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 $\hat{\mu}, \hat{\varepsilon}$ satisfying

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

in the $\varepsilon$-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 \( \mathbf{\epsilon} \)-corrupted samples from \( \mathcal{N}(\mu, \mathbf{I}) \)

**Definition:** The KL-divergence is

\[
d_{KL}(f \parallel g) = \int_{-\infty}^{\infty} f(x) \ln \frac{f(x)}{g(x)} \, dx
\]

**Fact:** For two Gaussians, we have

\[
d_{KL}(\mathcal{N}(\mu_1, \Sigma_1) \parallel \mathcal{N}(\mu_2, \Sigma_2)) =
\frac{1}{2} \left( \ln \frac{\det(\Sigma_2)}{\det(\Sigma_1)} + \text{Tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2) - p \right)
\]

When \( \Sigma_1 = \Sigma_2 = \mathbf{I} \) this simplifies to:

\[
d_{KL}(\mathcal{N}(\hat{\mu}, \mathbf{I}), \mathcal{N}(\mu, \mathbf{I})) = \frac{1}{2} \| \hat{\mu} - \mu \|^2_2
\]
Fact: [Pinsker's Inequality]

\[ d_{\text{div}}(f,g)^2 = \frac{1}{2} d_{\text{KL}}(f,g) \]

Putting it all together, we have

**Lemma:** If we can estimate

\[ \| \hat{\mu} - \mu \|_2 = \tilde{O}(\epsilon) \]

then we'd get

\[ d_{\text{div}}(N(\hat{\mu}, I), N(\mu, I)) = \tilde{O}(\epsilon) \]

Thus we have our parameter distance!

**Detecting Corruptions**

How can the adversary move the empirical mean by \( \epsilon \sqrt{d} \)?

\[ \begin{array}{c}
\hat{\mu}
\end{array} \]

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 $\epsilon$-corrupted samples from $N(\mu, I)$ and

1. $m \geq c \frac{d \ln 1/\delta}{\epsilon^2}$

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

then we have with probability $\geq 1 - \delta$

$$||\mu - \hat{\mu}||_2 \leq C' \epsilon \sqrt{\log 1/\delta} \Rightarrow ||\hat{\Sigma} - I||_2 \leq C'' \epsilon \log 1/\delta$$

Thus we can detect when the empirical mean has been corrupted.

Spectral Filtering

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

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

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

Otherwise consider $v = \text{direction of largest variance}$

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 $N(0, \Sigma)$ and $N(0, \Sigma')$

$$\text{div } (N(0, \Sigma), N(0, \Sigma')) = O(\| I - \frac{\Sigma^{-1/2} \Sigma' \Sigma^{-1/2}}{2} \|_F)$$

Mahalanobis distance

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

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

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

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

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

Now imagine we get $\varepsilon$-corrupted samples $X_1, \ldots, X_m$ from $N(0, \Sigma)$. Then define
\[ \Sigma = \frac{1}{m} \sum x_i x_i^T \text{ and} \]
\[ Y_i = (\Sigma)^{-\frac{1}{2}} X_i \]

If \( \Sigma = \Delta \) then \( Y_i \sim N(0, I) \) in which we are restricted to subspace of symmetric matrices.

\[ F = \frac{1}{m} \sum (y_i \otimes y_i) (y_i \otimes y_i) \sim 2I + (I^b)(I^b)^T \]

If we consider

\[ \max \ z^T F z \sim 2 \]

\[ z = \text{flattening of} \]
\[ d \times d \text{ trace zero,} \]
\[ \text{symmetric matrix} \]

We show that if \( \| I - \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} \| = \| \Sigma^{-\frac{1}{2}} \| < \varepsilon \log \frac{1}{\varepsilon} \)
then we must have

\[ \max \ z^T (F - 2I) z \geq \varepsilon \log \frac{1}{\varepsilon} \]

\[ z = \text{flattening...} \]

And similarly if the generalized eigenvalues of \( F \) are small, we can output \( \Delta \)

Otherwise we can find a direction (degree two polynomial) to filter on.
Assembling the Algorithm

Given \( \varepsilon \)-corrupted samples from \( \mathcal{N}(y, \Sigma) \)

1. Doubling trick
   \[ x_i - x_i' \sim \mathcal{N}(0, 2\varepsilon) \]

   Use algorithm for unknown covariance

2. Agnostic isotropic position
   \[ \Sigma^{-1/2} x_i \sim \mathcal{N}(\Sigma^{-1/2} y, I) \]

   Use algorithm for unknown mean

More Robust Statistics

Many subsequent directions

1. Handling more errors \( (\varepsilon > \frac{1}{2}) \) 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. \( \|x\|_0 \leq k \)

5. More complex generative models
Let's return to GMMs:

Earlier we gave a polynomial time/sample complexity, non-robust learner, implicitly we should:

definition we say a family of distributions $D$ is polynomially identifiable if

\[ \forall P_1, P_2 \in D \text{ that have } \epsilon \text{-different parameters} \]

\[ \Rightarrow d_{tv}(P_1, P_2) = \text{poly}(\epsilon, \frac{1}{d}) \]

[Valiant]

Corollary: 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 $D$ is robustly identifiable if

\[ \forall P_1, P_2 \in D \text{ that have parameter discrepancy } \epsilon \]

\[ \Rightarrow d_{tv}(P_1, P_2) = \text{poly}(\epsilon) \]

Notice that for unknown mean/covariance
we could take $l_2$ Mahalonobis distance in parameters

Theorem [Liu, Mitra]. Mixtures of two Gaussians in d-dimensions are robustly identified by a constant number of Hermite moments.

The proof will be via generating functions and differential operators.

Generating Functions.

Consider a Gaussian $G = N(y, I + \Sigma)$ and let

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

1 vector of formal variables

Key Lemma

$$e^{x^T \mu + \frac{1}{2} x^T \Sigma x} = \sum_{m=0}^{\infty} \frac{1}{m!} \mathbb{E} [H_m(\Sigma, x)] y^m$$

where $H_m(\Sigma, x) =$ Hermite moment tensor, i.e.

$$H_m(\Sigma, x = v) = \text{m}^{th} \text{ Hermite moment}$$

of 1-d Gaussian r.v. $v^T \Sigma$
Let's make it even simpler to see why

\textbf{Easier lemma. Let } \mathcal{G} = N(\mu, 1+\sigma^2). \text{ then}

\[e^{\mu y + \frac{\sigma^2}{2} y^2} = \sum_{m=0}^{\infty} \frac{1}{m!} \mathbb{E}[h_m(z)] y^m\]

\textbf{Proof:} Expanding the LHS we get

\[1 + (\mu y + \frac{\sigma^2}{2} y^2) + \frac{(\mu y + \frac{\sigma^2}{2} y^2)^2}{2} + \frac{(\mu y + \frac{\sigma^2}{2} y^2)^3}{6}\]

Collecting terms

\[1 + \mu y + \frac{(\mu^2 + \sigma^2)}{2} y^2 + \frac{(\mu^3 + 3\mu \sigma^2)}{6} y^3 + \ldots\]

\[\mathbb{E}[h_2(z)] = \mathbb{E}[z^2] = \mu^2 + 1 + \sigma^2 - 1\]

\[\mathbb{E}[h_3(z)] = \mathbb{E}[z^3 - 3z] = \mu^3 + 3\mu(1+\sigma^2) - 3\mu\]

\text{etc.}

Now the key lemma extends immediately to mixtures

\[M = w_1 N(\mu_1, I+\Sigma_1) + \ldots + w_k N(\mu_k, I+\Sigma_k)\]

\text{then we have:}
Key Lemma:
\[ \sum_{j=1}^{k} w_j e_{J} \mu_j (x_j + \frac{1}{2} \xi_j (x_j) y^2) = \sum_{m=0}^{\infty} \frac{1}{m!} \left[ H_m (z, x) \right] y^m \]

Back to Identifiability

We wish that
\[ \sum_{j=1}^{k} w_j e_{J} \mu_j (x_j + \frac{1}{2} \xi_j (x_j) y^2) = \sum_{j=1}^{k} w_j e_{J} \]

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

Now consider the differential operator
\[ D = d_y - (\mu + \sigma^2 y) \]

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 of the series expansion.
This works in high dimensions too. Let

\[ D \triangleq dy - (\nu(x) - \Sigma(x)y) \]

Then,

Fact: \( D(e^{\nu(x)y + \frac{1}{2} \Sigma(x)y^2}) = 0 \)

which is a complicated but explicit multivariate polynomial rearrangement

Main Question: What happens when you apply \( D \) to another component?

Fact: \( D(e^{\hat{\nu}(x)y + \frac{1}{2} \hat{\Sigma}(x)y^2}) = P e^{\hat{\nu}(x)y + \frac{1}{2} \hat{\Sigma}(x)y^2} \)

where \( P = (\hat{\nu}(x) - \nu(x) + y(\hat{\Sigma}(x) - \Sigma(x))) \)

And finally,

Fact: \( D(P e^{\nu(x)y + \frac{1}{2} \Sigma(x)y^2}) = \frac{dP}{dy} e^{\nu(x)y + \frac{1}{2} \Sigma(x)y^2} \)

Now we can use differential operators to isolate a component.
Thus consider

\[ M^{2n-2} \quad \text{vs.} \quad M^{2n} \quad \hat{M} \]

\[ \Delta_1 \quad \Delta_2 \quad \Delta_3 \quad \Delta_4 \]

\[ P(x, y) e^{\mu(x)y + \frac{1}{2} \xi_i(x)y^2} \]

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 \( \mathsf{GMM} \) with accuracy that depends polynomially on the corruption rate.

[Rakhsha et al.] get related results, but for weaker notion of density estimation:

- Proper
- Semi-proper

[Liu, Moitra] get \( O(\varepsilon) \) accuracy for density estimation.